INTERNATIONAL COMPUTER SCIENCE INSTITUTE

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

How to benefit from noise

Leszek Plaskota*

TR-95-056

September 1995

Abstract

We compare nonadaptive and adaptive designs for estimating linear functionals in the (minimax) statistical setting. It is known that adaptive designs are no better in the worst case setting for convex and symmetric classes, as well as in the average case setting with Gaussian distributions.

In the statistical setting, the opposite is true. Namely, adaptive designs can be significantly better. Moreover, using adaptive designs one can obtain much better estimators for noisy data than for exact data. These results hold because adaption and noisy data make the Monte Carlo simulation possible.

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

1 Introduction

Many interesting results have been recently obtained for the minimax statistical setting of estimating linear functionals over convex and symmetric classes. One of the most important results is due to Donoho (1994) who proved, under mild assumptions, that linear estimators are within $11,1\ldots\%$ of being optimal among all nonlinear estimators. He also gave formulas for the optimal linear estimators. This was done by establishing a relation between the minimax statistical setting and the problem of *optimal recovery* in the worst case setting. Optimality properties of linear estimators in the worst case setting have been known before, see, e.g., Smolyak (1965), Sukharev (1986), Magaril–Il'yaev and Osipenko (1991). Hence, Donoho's results for statistical estimation correspond to those of Smolyak and others for optimal recovery in the worst case setting. We add that for classes given as balls in Hilbert spaces, the same results can be obtained by using relations between the statistical and the average case settings with Gaussian distributions, see Plaskota (1996).

The relations between optimal estimators in the statistical, worst and average case settings mentioned above hold for estimators using fixed, nonadaptive designs. It is now natural to ask whether similar relations hold for adaptive designs.

Adaptive designs have been studied in the worst and average case settings. The question how much adaption helps is one of the fundamental problems in *information-based complexity*, see, e.g., Traub *et al.* (1988) and Plaskota (1996). In particular, it is known that adaption does *not* help in the worst case setting for convex and symmetric classes, and, under some additional assumptions, in the average case setting for Gaussian distributions, see Section 3. (This also holds for approximating linear operators. On the other hand, for some nonlinear operators adaption is exponentially more powerful in the worst and average case settings.)

How about adaption in the statistical setting? Remarkably, there is not much on this subject in the statistical literature. It is, however, known that adaption does not help asymptotically for nonparametric regression, see Golubev (1992).

This discussion may lead us to the conjecture that for convex and symmetric classes adaption does *not* help in the statistical setting. However, the opposite is true. More precisely, we provide a simple and natural example which reveals the following two important and rather surprising things:

- Adaptive designs can be exponentially better than nonadaptive designs in the statistical setting for convex and symmetric classes.

- In the statistical setting one can sometimes obtain much better estimators using noisy data rather than exact data.

These results hold because adaption in the statistical setting makes the Monte Carlo simulation possible, and for many problems the error of the Monte Carlo (randomized) method is much smaller than the error of any non–randomized (even adaptive) method.

We now comment on the result that noisy data may lead to smaller errors than exact data. For noisy data, the error is defined, in particular, by taking an average with respect to noise. For exact data, the average over noise disappears and we are back in the worst case error setting. Hence, noisy and exact data really correspond to the different error settings and this change makes the result possible.

The problem for which adaption is significantly better in the statistical setting is multivariate integration of Lipschitz functions. However, from the proof it will be clear that similar results hold for other problems for which randomized algorithms are better than non–randomized ones. It would be interesting to verify whether this is the only reason why adaption helps in the statistical setting. In other words: can adaption help in the statistical setting for problems for which randomization does not help? The answer is unknown.

2 Nonadaptive and adaptive designs

Let X be a linear space of real functions defined on a domain D, and let F be a subclass of X. We assume that F is *convex* and *symmetric* (with respect to zero). Suppose that for an (unknown) $f \in F$ we observe data $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$,

$$y_i = f(t_i) + x_i, \qquad 1 \le i \le n, \tag{1}$$

where $t_i \in D$ and $x = [x_1, \ldots, x_n]$ is the *white noise* vector, i.e., $x_i \sim_{iid} \mathcal{N}(0, \sigma^2)$. We stress that we also allow $\sigma = 0$ which corresponds to the exact (non–noisy) data. Our aim is to estimate the value $S(f)$, where S is a linear functional over X. An estimator is of the form $S_n(f, x) = \varphi(y)$, i.e., it only uses the data y.

In the statistical setting, the *error* of S_n is given as

$$R^{\mathrm{st}}(S_n, T_n, \sigma) = \sup_{f \in F} (E_x (S(f) - S_n(f, x))^2)^{1/2},$$

where E_x denotes the expectation over x. Here $T_n = \{t_i\}_{i=1}^n$ is the *design*.

Although this is not the only way of defining the error, this definition has been most often used in the statistical literature, see, e.g., Sacks and Ylvisaker (1978), Speckman (1979), Ibragimov and Hasminski (1984), Nussbaum (1985), Donoho (1994), Donoho *et al.* (1995).

In (1) we assume that the design points T_n are given in advance. One of possible and natural generalizations is to assume that the successive observations are performed for points which are chosen adaptively depending on the results of previous observations. That is, we now have

$$y_i = f(t_i(y_1, \ldots, y_{i-1})) + x_i, \qquad 1 \le i \le n, \tag{2}$$

where $x_i \sim_{iid} \mathcal{N}(0, \sigma^2)$ and $t_i : \mathbb{R}^{i-1} \to D$ are measurable mappings. Such a design will be called *adaptive*.

Remark 1 *Throughout this paper we assume, for simplicity, that the number n of observations in any adaptive design is fixed. Sometimes it is reasonable to consider adaptive designs with varying n depending on y_i's; see also Remark 3.*

Our aim is to compare the power of adaptive and nonadaptive designs. Define

$$R^{\rm st}_{\rm non}(n,\sigma) = \inf\{R(S_n,T_n,\sigma) : S_n \text{ arbitrary}, T_n \text{ nonadaptive }\}$$

as the minimal error that can be achieved for n nonadaptive observations, and

$$R^{\rm st}_{\rm adp}(n,\sigma) = \inf\{R(S_n,T_n,\sigma) : S_n \text{ arbitrary}, T_n \text{ adaptive }\}$$

as the corresponding minimal error for n adaptive observations. Clearly,

$$R^{\rm st}_{\rm adp}(n,\sigma) \le R^{\rm st}_{\rm non}(n,\sigma).$$

3 Adaptive designs in different settings

Adaptive designs have been studied in the worst and average cases settings. The following sample results are typical and important.

3.1 Worst case setting

Suppose that the noise in (1) and (2) is deterministic rather than random, and we know that x is bounded in a norm, i.e., $\|x\| \le \delta$. Define the error of an estimator S_n as the worst case error,

$$R^{\rm wor}(S_n,T_n,\delta) = \sup_{f\in F}\sup_{\|x\|\le\delta} |S(f) - S_n(f,x)|.$$

Then for the respective nth minimal errors we have

$$R^{\rm wor}_{\rm adp}(n,\delta) = R^{\rm wor}_{\rm non}(n,\delta),$$

see, e.g., Bakhvalov (1971), Gal and Micchelli (1980), Traub and Woźniakowski (1980), Traub *et al.* (1983).

3.2 Average case setting

Assume that data are again of the form (1) or (2), but the function f is now the realization of a zero mean Gaussian stochastic process on D. The error of S_n is defined as the expected (average) error over both f and the noise x, i.e.,

$$R^{\rm avg}(S_n,T_n,\sigma) = (E_f E_x (S(f) - S_n(f,x))^2)^{1/2}.$$

Assuming additionally that the functional S is continuous, we have

$$R^{\rm avg}_{\rm adp}(n,\sigma) = R^{\rm avg}_{\rm non}(n,\sigma),$$

see, e.g., Kadane *et al.* (1988), Plaskota (1996).

Thus adaption does not help in both the worst case and average case settings.

Remark 2 *For S being a linear operator, adaption still does not (essentially) help. Namely, we have $R^{\rm wor}_{\rm adp}(n,\delta) \ge \frac{1}{2} R^{\rm wor}_{\rm non}(n,\delta)$ and $R^{\rm avg}_{\rm adp}(n,\sigma) = R^{\rm avg}_{\rm non}(n,\sigma)$.*

Remark 3 *In the average case setting, it is reasonable to consider adaptive designs with varying n. For such designs, adaption usually does not help for linear S, see, e.g., Wasilkowski (1986) and Plaskota (1996). However, counterexamples where the opposite is true are also known, see Plaskota (1993).*

3.3 Statistical setting

As already mentioned in the introduction, adaption does not help in the statistical setting for the nonparametric regression. The result reads as follows, see Golubev (1992).

Let F be the Sobolev class of functions $f : [0,1] \to \mathbb{R}$ of regularity r such that

$$\sum_{k=0}^{r} \int_0^1 (f^{(k)}(u))^2 du \le 1.$$

Suppose that instead of a functional, we want to estimate the function f in the $\mathcal{L}_2$–norm. That is, the error of an estimator $f_n(u,x)$ is now given as

$$R^{\rm st}(S_n, T_n, \sigma) = \sup_{f \in F} (E_x \int_0^1 (f(u) - f_n(u,x))^2 du)^{1/2}.$$

Then

$$R^{\rm st}_{\rm adp}(n,\sigma) \approx R^{\rm st}_{\rm non}(n,\sigma), \qquad \text{as} \quad n \to +\infty.$$

(Here $a_n \approx b_n$ means that $\lim_{n\to\infty} a_n/b_n = 1$.) Moreover, optimal design is given by equidistant points.

4 Adaption may help in the statistical setting

We now present a problem of multivariate integration for which adaption is exponentially better in the statistical setting.

Let $D = [0,1]^d$ with $d \ge 2$. Let F be the class of 1–Lipschitz functions, i.e.,

$$|f(u_1) - f(u_2)| \le \|u_1 - u_2\|_\infty, \quad \forall u_1, u_2 \in D.$$

Obviously, F is convex and symmetric. Suppose we want to estimate the integral of f,

$$S(f) = \int_D f(u)\, du,$$

using data (1) or (2). Then we have the following result. (Below $a_n \asymp b_n$ means that there exist constants $0 < a \le b < +\infty$ such that for all n we have $a \le a_n/b_n \le b$, $\forall n$.)

Theorem 1 *For estimating the integral of a real* 1*–Lipschitz function defined on the* d*–dimensional unit cube we have*

$$R^{\rm st}_{\rm non}(n,\sigma) \asymp n^{-1/d}$$

and

$$R^{\rm st}_{\rm adp}(n,\sigma) \asymp \begin{cases} n^{-1/d} & \text{for } \sigma = 0 \\ n^{-1/2} & \text{for } \sigma > 0 \end{cases},$$

as $n \to +\infty$.

Hence, for nonadaptive designs the minimal error is of order $n^{-1/d}$ which strongly depends on the dimension d. Actually, in this case we have the *curse of dimensionality*, since we have to perform exponentially (in d) many observations to reduce the error to a desired level. Note that the behavior of $R^{\rm st}_{\rm non}(n,\sigma)$ is the same for exact as well as for noisy data.

However, the situation drastically changes if we allow adaptive observations. For exact data the error is still proportional to $n^{-1/d}$, but for noisy data the minimal error of adaptive estimators surprisingly reduces to $n^{-1/2}$ and is independent of d. In particular, the curse of dimensionality vanishes, and for large d it is much better to deal with noisy than exact data.

Why is this possible? The idea is very simple. Assume that we have noisy data, i.e., $\sigma > 0$. If we make two observations at the same point and subtract their results, we will obtain a Gaussian random variable with known distribution. Hence, the statistical setting with non–zero noise provides us with an additional tool which is the random number generator. This tool together with adaption allows us to implement randomized algorithms and, in particular, the classical Monte Carlo method. For multivariate integration the expected error of Monte Carlo is much smaller than that of non–randomized methods, and we are done.

The formal proof of the theorem follows.

5 Proof

The case of exact data, $\sigma = 0$, corresponds to the worst case setting with exact data ($\delta = 0$). Hence, using the well known results from the worst case, see, e.g., Novak (1988), we obtain

$$R^{\rm st}_{\rm non}(n,0) = R^{\rm st}_{\rm adp}(n,0) \asymp n^{-1/d}.$$

Moreover, the equispaced design $T^*_n = \{t^*_i\}$ and the arithmetic mean

$$S^*_n(f,x) = \frac{1}{n}\sum_{i=1}^{n}(f(t^*_i) + x_i)$$

have error proportional to $n^{-1/d}$.

Let $\sigma > 0$. Consider first a nonadaptive design $T_n = \{t_i\}_{i=1}^n$ and a linear estimator

$$S_n(f, x) = \sum_{i=1}^n w_i(f(t_i) + x_i),$$

where w_i's are some reals. Then

$$R^{\rm st}(S_n, T_n, \sigma) = (\, R^{\rm st}(S_n, T_n, 0)^2 + \sigma^2 \sum_{i=1}^n w_i^2 \,)^{1/2} \ge R^{\rm st}_{\rm non}(n, 0).$$

On the other hand, we have

$$R^{\rm st}(S_n^*, T_n^*, \sigma) = (\, R^{\rm st}(S_n^*, T_n^*, 0)^2 + \sigma^2/n \,)^{1/2} \asymp n^{-1/d},$$

as $n \to +\infty$. To complete the proof of the nonadaptive case, it suffices to show that the error of order $n^{-1/d}$ cannot be reduced by using nonlinear estimators.

Indeed, let $c > 0$ be such that $R^{\rm st}_{\rm non}(n, 0) > c\, n^{-1/d}$, $\forall n$. Then we can select $h_n \in F$ satisfying $h_n(t_i) = 0$, $1 \le i \le n$, and $S(h_n) > c\, n^{-1/d}$. It is clear that the error will not increase if the set F is replaced by the interval $[-h_n, h_n]$. For such a 'reduced' problem the data consist of pure noise, $y_i = x_i$, $\forall i$, and such data are known to be useless. Zero is the best estimator among all nonlinear estimators, and the error is at least $S(h_n)$ which is larger than $c\, n^{-1/d}$, as claimed.

We now construct an adaptive design and an estimator with error proportional to $n^{-1/2}$. Assume without loss of generality that $n = k(2d + 1)$. Let

$$\psi(x) = \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^x \exp\{-u^2/(2\sigma^2)\}\, du.$$

The adaptive design $T_n^{**} = \{t_i\}_{i=1}^n$ with $t_i = (t_i^1, \ldots, t_i^d) \in \mathbb{R}^d$, is given as follows. Let $s = 2kd$. We set $t_i = (0, \ldots, 0)$ for $1 \le i \le s$, and

$$t_{s+i}^j = \psi\left(y_{2d(i-1)+2j} - y_{2d(i-1)+2j-1}\right)$$

for $1 \le i \le k$, $1 \le j \le d$. As the estimator we take

$$S_n^{**}(f, x) = \frac{1}{k} \sum_{i=1}^k (f(t_{s+i}) + x_{s+i}).$$

We claim that $R^{\rm st}(S_n^{**}, T_n^{**}, \sigma) \asymp n^{-1/2}$. Indeed, for any $1 \le i \le s$ and $1 \le j \le d$ the difference $x_i^j = y_{2d(i-1)+2j} - y_{2d(i-1)+2j-1}$ has the zero mean normal distribution with variance $2\sigma^2$. Hence, $t_i^j = \psi(x_i^j)$ is distributed uniformly on the unit interval, and the design points $t_{s+1}, \ldots, t_n$ are distributed uniformly on the cube D. Our estimator is then nothing but the classical Monte Carlo method, see, e.g., Novak (1988), applied

to noisy data. We obtain that for any $f \in F$

$$\begin{aligned} E_x(S(f) - S_n^{**}(f,x))^2 &= E_x \left(\int_D f(u)du - \frac{1}{k} \sum_{i=1}^{k} (f(t_{s+i}) + x_{s+i}) \right)^2 \\ &= \frac{1}{k} \left(\int_D f^2(u)du - \left(\int_D f(u)du \right)^2 \right) + \frac{\sigma^2}{k} \\ &= \frac{\alpha(f) + \sigma^2}{k}, \end{aligned}$$

where

$$0 \le \alpha(f) \le \sup \left\{ \int_D f^2(u)du : \ f \in F, \int_D f(u) = 0 \right\} < 1.$$

Since $k = n/(2d+1)$, the error $R^{\rm st}(S_n^{**}, T_n^{**}, \sigma)$ is proportional to $n^{-1/2}$, as claimed.

The lower bound for $R^{\rm st}(n, \sigma)$ is provided by the following argument. Consider the simpler problem of estimating the integral of a 0–Lipschitz (constant) function. This is equivalent to estimating a real parameter from its n noisy observations with variance σ^2. It is well known that for such estimation the minimal error is just $\sigma\, n^{-1/2}$.

The proof is complete.

Remark 4 *The method S_n^{**}, T_n^{**} constructed in the proof uses the 'continuous' version of the Monte Carlo, i.e., the points are selected randomly from the whole unit cube. The same error bounds can be obtained by using a 'discrete' Monte Carlo, where selection is made from a grid of cardinality at least proportional to n^d.*

Remark 5 *We showed that*

$$R^{\rm st}(S_n^{**}, T_n^{**}, \sigma) \le c\, n^{-1/2},$$

where $c = c(\sigma, d) = (\,(\sigma^2+1)(2d+1)\,)^{1/2}$. One can get rid of the dependence on d by generating all *random sample points from only one random number $y_0 = y_1 - y_2$, $y_i = f(0) + x_i$, $i = 1, 2$, i.e., using only 2 instead of $2kd$ 'preliminary' observations. In the latter case the constant c is roughly $(\sigma^2+1)^{1/2}$.*

Moreover, we will still have a similar upper bound if those two observations are made at different points, but sufficiently close to each other. Hence, the main result also holds in the case when repeated observations are not allowed.

Acknowledgments. I would like to thank David Donoho, Erich Novak, Henryk Woźniakowski for interest and helpful comments on this paper, and Michael Nussbaum for directing my attention to Golubev's result.

References

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

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. Statist. Soc.* ser. B **57** 301–369.

Gal, S. and Micchelli, C.A. (1980). Optimal sequential and nonsequential procedures for evaluating a functional. *Appl. Anal.* **10** 105–120.

Golubev, G.K. (1992). On sequential experimental designs for nonparametric estimation of smooth regression functions. *Problems Inform. Transmission* **28** 76–79.

Ibragimov, I.A. and Hasminski, R.Z. (1984). On nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Prob. Appl.* **29** 19–32. (In Russian.)

Kadane, J.B., Wasilkowski, G.W. and Woźniakowski, H. (1988). On adaption with noisy information. *J. Complexity* **4** 257–276.

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

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 in $\mathcal{L}_2$. *Annals of Statistics* **13** 984–997.

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

Plaskota, L. (1996). Noisy Information and Computational Complexity. To appear in Cambridge University Press.

Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. *Annals of Statistics* **6** 122–137.

Smolyak, S.A. (1965). On optimal recovery of functions and functionals of them. PhD thesis, Moscow State Univ. (In Russian.)

Speckman, P. (1979). Minimax estimates of linear functionals in a Hilbert space. Unpublished manuscript.

Sukharev, A.G. (1986). On the existence of optimal affine methods for approximating linear functionals. *J. Complexity* **2** 317–322.

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.

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