OPTIMALITY IN NUMERICAL DIFFERENTIATION

A R Davies

1. INTRODUCTION

Consider the problem of estimating the mth derivative of a data function $g(x)$, given only N sampled values

$$y_n = g(x_n) + \varepsilon_n, \quad n = 0,\ldots, N-1, \tag{1}$$

where ε_n are uncorrelated random errors with mean zero and common variance σ^2 (possibly unknown). For simplicity consider equally-spaced sampling points $x_n = n/N$ on the interval $[0,1)$. Let m be a strictly positive integer, and denote the mth derivative by $f(x) = g^{(m)}(x)$, which is to be estimated on the interval $0 \leqslant x \leqslant 1$.

If K denotes an integral operator such that $Kf = g$, then a stabilized derivative can be constructed using pth-order Tikhonov regularization:

$$\min_{f \in F_p} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} [(Kf)x_n) - y_n]^2 + \lambda \|f\|_p^2 \right\}, \tag{2}$$

where F_p is a suitably chosen Hilbert space with norm $\|\cdot\|_p$ parametrized by the order of regularization $p > 0$, and the constant $\lambda \geqslant 0$ is the regularization parameter. Let $f_{N;\underset{\sim}{\alpha}}$ denote the minimizer of (2), where $\underset{\sim}{\alpha}$ is the parameter pair $\underset{\sim}{\alpha} = (p,\lambda)$.

In theory we may define an *absolutely optimal* parameter set $\underset{\sim}{\alpha}$ as that which minimizes (with respect to $\underset{\sim}{\alpha}$) the error

$$\| f_{N;\underset{\sim}{\alpha}} - f \|_F \tag{3}$$

where $\|\cdot\|_F$ denotes the strongest norm consistent with the smoothness of the exact derivative f. In the data space there will exist a norm $\|\cdot\|_G$ such that (3) and

$$\| Kf_{N;\underset{\sim}{\alpha}} - g \|_G \tag{4}$$

are equal; however, in the absolute sense the G-norm is necessarily stronger than the F-norm since K is a smoothing operator. Thus, if we seek an $\underset{\sim}{\alpha}$ which is optimal with respect to the L^2-error in the space of the derivative, it is *not* sufficient to minimize the L^2-error in the data space. This is one reason why absolute optimality is difficult to achieve in practice. Another reason is that the smoothness of f is usually unknown.

It is not our intention in this report to address the quest for an absolutely optimal parameter set $\underset{\sim}{\alpha}$; indeed, the author knows of no practical method which can achieve this. Instead we discuss the weaker concept of D-optimality and the even weaker concept of S-optimality, which we define below.

First consider the *predictive mean-square signal error*

$$S(\underset{\sim}{\alpha}) = \frac{1}{N} \sum_{n=0}^{N-1} [(Kf_{N;\underset{\sim}{\alpha}})(x_n) - g(x_n)]^2. \tag{5}$$

The minimizer of $S(\underset{\sim}{\alpha})$ is estimated quite closely by several practical statistical methods, at least when p, the order of regularization, is fixed. For the present let p be fixed. We emphasize this by writing λ where previously we have used $\underset{\sim}{\alpha}$. Let λ_0 denote the minimizer of $S(\lambda)$ with respect to λ. Following [1] we say that a value of λ is

(i) *strongly S-optimal* if $S(\lambda)/S(\lambda_0) = 1 + O(N^{-a})$, $a \geq a_0 > 0$, as $N \to \infty$; (6a)

(ii) *S-optimal* if $S(\lambda)/S(\lambda_0) = 1 + o(1)$ as $N \to \infty$, where $o(1) \to 0$ as $N \to \infty$; (6b)

(iii) *weakly S-optimal* if $S(\lambda)/S(\lambda_0) = O(1)$ as $N \to \infty$; (6c)

(iv) *S-suboptimal* if $S(\lambda)/S(\lambda_0) \to \infty$ as $N \to \infty$. (6d)

Since $f_{N;\lambda}$ is determined from the data $\underset{\sim}{y}_N$, $S(\lambda)$ is a random variable. Its analysis as such is difficult, and it is easier to study $ES(\lambda)$ instead, where E denotes expectation with respect to the error distribution. With this in mind let $\tilde{\lambda}_0$ denote the minimizer of $ES(\lambda)$. We maintain the definitions (6) if

$S(\lambda)$ is replaced by $ES(\lambda)$ and λ_0 by $\underset{\sim}{\lambda}_0$.

The *predictive mean-square derivative error* is defined by

$$D(\underset{\sim}{\alpha}) = \frac{1}{N}\sum_{n=0}^{n-1}\left[f_{N;\underset{\sim}{\alpha}}(x_n) - f(x_n)\right]^2. \tag{7}$$

When p is fixed we can define a D-optimal value of λ in the same way as (6), simply replacing S by D. Unfortunately, for the same value of p, a value of λ which is S-optimal is not D-optimal, nor is a D-optimal value of λ absolutely optimal unless the natural smoothness of f is no greater than L^2.

Certain practical methods for determining λ are known to yield values which vary from S-suboptimal to strongly S-optimal. (This depends on the method and the regularity of the data; see §2 below). For example, when the noise variance σ^2 is known, the unbiased risk method always yields strongly S-optimal values of λ; when σ^2 is unknown, the unbiased risk method cannot be used, but the method of cross-validation can yield S-optimal (sometimes strongly S-optimal) values of λ. It is therefore natural to pose the question: *can practical methods be devised for determining values of λ which are D-optimal (preferably strongly D-optimal)?* This report presents some background information which may be of value in answering this question.

2. S-OPTIMALITY.

Data regularity Let $\nu = m + p$, and $\beta > 0$ be a constant such that

$$\sum_{q=0}^{\infty} \omega_q^{2\beta\nu}|\hat{g}_q|^2 < \infty, \quad \omega_q = 2\pi q, \tag{8}$$

where $\hat{g}_q$ are the Fourier coefficients of the underlying data function g, i.e.

$$\hat{g}_q = \hat{g}(\omega_q), \text{ where } \hat{g}(\omega) = \int_0^1 g(x)\exp(-i\omega x)dx. \tag{9}$$

Clearly, $r = \beta\nu$ is a measure of the smoothness of g. If there is a natural

limit on this smoothness, i.e. $r \leq r_0$, say, then as p is increased the value of β in (8) is reduced so that $\beta\nu \leq r_0$. On the other hand, when p is fixed, β is a measure of the smoothness of g.

As has been shown by Lukas [2], the asymptotic properties of $\tilde{\lambda}_0$ and $ES(\tilde{\lambda}_0)$ depend quite crucially on whether $0 < \beta < 2$, or whether $\beta \geq 2$. Let us consider the case of Fourier differentiation where F_p in (2) is the space of trigonometric polynomials of degree at most $\frac{1}{2}N$, and $\|f\|_p = \|f^{(p)}\|_{L^2}$. The associated operator K is defined in [1]. We state the following without proof:

Theorem 1. Under the regularity assumption (8), the value of λ which minimizes $ES(\lambda)$ may be written

(i) when $\beta \geq 2$:

$$\tilde{\lambda}_0 = \left[\frac{\kappa_\nu}{4\nu} \cdot \frac{\sigma^2}{N} \cdot \frac{1}{|||g|||_{2\nu}}\right]^{2\nu/(4\nu+1)} [1 + o(1)] \text{ as } N \to \infty, \tag{10}$$

where κ_ν is the constant

$$\kappa_\nu = \frac{1}{\pi} \int_0^\infty \frac{dt}{(1+t^{2\nu})^2}, \tag{11}$$

$$|||g|||_r \equiv \frac{1}{\pi} \int_0^\infty \omega^{2r} |\hat{g}(\omega)|^2 d\omega, \quad r > 0, \tag{12}$$

and $o(1) \to 0$ as $N \to \infty$. The minimizing value of $ES(\lambda)$ as $N \to \infty$ is

$$ES(\tilde{\lambda}_0) = (4\nu + 1)\left[\frac{\kappa_\nu}{4\nu} \cdot \frac{\sigma^2}{N}\right]^{4\nu/(4\nu+1)} |||g|||_{2\nu}^{1/(4\nu+1)} [1 + o(1)]. \tag{13}$$

(ii) When $0 < \beta < 2$, we have instead

$$\tilde{\lambda}_0 \geq \left[\frac{\kappa_\nu}{\beta\nu} \cdot \frac{\sigma^2}{N} \cdot \frac{1}{|||g|||_{\beta\nu}}\right]^{2\nu/(2\beta\nu+1)} [1 + o(1)] \text{ as } N \to \infty, \tag{14}$$

and

$$ES(\underset{\sim}{\lambda}_0) \leq (2\beta\nu+1)\left[\frac{\kappa_\nu}{2\beta\nu}\cdot\frac{\sigma^2}{N}\right]^{2\beta\nu/(2\beta\nu+1)} |||g|||_{\beta\nu}^{1/(2\beta\nu+1)}[1+o(1)]. \tag{15}$$

Corollary. If $\beta \geq 3$, the $o(1)$ terms in (10) and (13) become $O\left(n^{-2\nu/(4\nu+1)}\right)$. (For a proof see [1].)

From (13) and (15) we see that $ES(\underset{\sim}{\lambda}_0)$ is a function of p. As $N \to \infty$ we can seek to optimize p by maximizing the exponents of (σ^2/N) in (13) and (15). This demands that

$$p > \frac{1}{2}(\beta\nu - 2m) \tag{16}$$

and in particular that p must be sufficiently large to bring β into the range $0 < \beta < 2$. Of course this is an asymptotic result, and in practice it need not necessarily be consistent with the minimization of $S(\underset{\sim}{\alpha})$ when N is finite.

Practical methods. In practice $S(\underset{\sim}{\alpha})$ cannot be computed directly without knowing the exact data function and so alternative means of estimating its minimizer (with respect to $\underset{\sim}{\alpha}$) is needed. Practical methods fall into two classes:

Class I. Those which require a knowledge of the noise variance σ^2.

Class II. Those which do not.

Most methods can be implemented in terms of an *influence matrix* $A(\underset{\sim}{\alpha})$. If $\underset{\sim}{g}_{N;\underset{\sim}{\alpha}}$ denotes the N-vector whose elements are the values of the function $(Kf_{N;\underset{\sim}{\alpha}})(x)$ at $x = x_n$, then the influence matrix is such that

$$\underset{\sim}{g}_{N;\underset{\sim}{\alpha}} = A(\underset{\sim}{\alpha})\underset{\sim}{y}_N. \tag{17}$$

For Fourier differentiation, $A(\underset{\sim}{\alpha})$ is the circulant matrix whose eigenvalues constitute the discrete regularization filter imposed by the regularization process [1]. Class I methods invariably involve the mean-square error or discrepancy:

$$T(\underset{\sim}{\alpha};\underset{\sim}{y}_N) = \frac{1}{N}\sum_{n=0}^{N-1}\left[(Kf_{N;\underset{\sim}{\alpha}})(x_n) - y_n\right]^2 = \frac{1}{N}\left\|(I - A(\underset{\sim}{\alpha}))\underset{\sim}{y}_N\right\|^2, \tag{18}$$

or modifications thereof. It may be shown that

$$ET(\underset{\sim}{\alpha};\underset{\sim}{y}_N) = ES(\underset{\sim}{\alpha}) + \sigma^2 - 0\left\{\left[N\lambda^{1/(2\nu)}\right]^{-1}\right\} \tag{19}$$

provided $N\lambda^{1/(2\nu)} \to \infty$ as $N \to \infty$. Notable among Class I methods are the discrepancy principles of Arcangeli [3], Morozov [4], and their generalizations [5,6]; the Bayesian method of Turchin [7,8]; the unbiased risk method [9]; and the Bayesian method of Turchin-Klein [10]. Class II methods involve the optimization of certain functions in lieu of $S(\underset{\sim}{\alpha})$. Among these are cross-calidation [11] and maximum likelihood [12]. We summarize in the adjoining table the practical implementation and classification (with respect to S-optimality) of the main methods in each class.

We see from the table that only unbiased risk minimization, cross-validation, and maximum-likelihood enable the optimization of p as well as λ. Of the Class I methods tabulated, unbiased risk minimization is clearly the most powerful. This is because of the identity

$$ER(\underset{\sim}{\alpha}) = ES(\underset{\sim}{\alpha}) \tag{20}$$

which immediately yields strong S-optimality in the expectational sense with $a = \infty$. Of the Class II methods tabulated, cross-validation is the most powerful.

3. D-OPTIMALITY.

Let p be fixed, and let $\underset{\sim}{\lambda}_0^{(m)}$ denote the minimizer of $ED(\lambda)$. To obtain similar results for $\underset{\sim}{\lambda}_0^{(m)}$ and $ED(\underset{\sim}{\lambda}_0^{(m)})$ to those given for $\underset{\sim}{\lambda}_0$ and $ES(\underset{\sim}{\lambda}_0)$ in Theorem 1, when the data regularity condition (8) is replaced by

$$\sum_{q=0}^{\infty} \omega_q^{2(\beta\nu+m)} |\hat{g}_q|^2 < \infty, \tag{21}$$

	Method	Practical Implementation	Classification	
			$0 < \beta < 2$	$\beta \geq 2$
CLASS I	Morozov discrepancy [4]	Fix p. Solve (T1) for λ: $\frac{1}{N}\|(I - A(\lambda))\underline{y}_N\|^2 = \sigma^2$ (T1)	Weakly S-optimal	Weakly S-optimal
	Turchin [7,8]	Fix p. Solve (T2) for λ: $\frac{\sigma^2}{N}\mathrm{Tr}A(\lambda) + \frac{1}{N}\|(I - A(\lambda))\underline{y}_N\|^2 = \sigma^2$ (T2)	Weakly S-optimal	Weakly S-optimal
	Unbiased risk minimization [9]	Minimize $R(\underline{\alpha})$ w.r.t. $\underline{\alpha} = (p,\lambda)$: $R(\underline{\alpha}) \equiv \frac{1}{N}\|(I - A(\underline{\alpha}))\underline{y}_N\|^2 + \frac{2\sigma^2}{N}\mathrm{Tr}A(\underline{\alpha}) + \sigma^2$	Strongly S-optimal ($a = \infty$)	Strongly S-optimal ($a = \infty$)
CLASS II	Cross-validation [10]	Minimize (T3) w.r.t. $\underline{\alpha}$: $\frac{\frac{1}{N}\|(I - A(\underline{\alpha})\underline{y}_N\|^2}{[\frac{1}{N}\mathrm{Tr}(I - A(\underline{\alpha}))]^2}$ (T3)	S-optimal	S-optimal (Strongly if $\beta \geq 3$)
	Maximum likelihood [12]	Minimize (T4) w.r.t. $\underline{\alpha}$: $\frac{\underline{y}^T(I - A(\underline{\alpha}))\underline{y}}{[\det{}^+(I - A(\underline{\alpha}))]^{1/(N-1)}}$ (T4)	Weakly S-optimal	S-suboptimal

($\det^+$ denotes omission of zero eigenvalues).

it can be shown that

$$\left.\begin{aligned} \tilde{\lambda}_0^{(m)} &= O\left(\left[\frac{\sigma^2}{N}\right]^{2\nu/(4\nu+2m+1)}\right) \\ ED\left(\tilde{\lambda}_0^{(m)}\right) &= O\left(\left[\frac{\sigma^2}{N}\right]^{4\nu/(4\nu+2m+1)}\right) \end{aligned}\right\} \beta \geq 2, \tag{22}$$

and

$$ED\left(\tilde{\lambda}_0^{(m)}\right) \leq O\left(\left[\frac{\sigma^2}{N}\right]^{2\beta\nu/(2\beta\nu+2m+1)}\right), \quad 0 < \beta < 2. \tag{23}$$

From these rates we may deduce that an S-optimal value of λ is D-suboptimal.

To achieve *weak D-optimality* is not difficult in principle. For example if σ^2 is known then for a fixed p satisfying (21) we can choose

$$\lambda \propto \left(\frac{\sigma^2}{N}\right)^{\mu} \tag{24}$$

and

$$\mu = \begin{cases} 2\nu/(4\nu+2m+1), & \beta \geq 2 \\ 2\nu/(2\beta\nu+2m+1), & 0 < \beta < 2. \end{cases}$$

This guarantees the correct convergence rates in (22)-(23), although the choice of constant of proportionality in (24) will greatly affect the quality of the solution.

Weak D-optimality can sometimes be achieved by cross-validation also. Let p,p' and $\beta \geq 2$ be such that

$$p' = (2m+1)p + 2m^2 \tag{25}$$

and (21) is satisfied with ν replaced by $\nu' = m+p'$. Clearly (8) is also satisfied. Using cross-validation with order of regularization p determines an S-optimal value of λ_{CV} with the associated expectational property

$$\tilde{\lambda}_{CV} = O\left(\left[\frac{\sigma^2}{N}\right]^{2\nu/(4\nu+1)}\right) \text{ as } N \to \infty.$$

The choice of p' in (25), however, is such that

$$\frac{2\nu}{4\nu+1} = \frac{2\nu'}{4\nu'+2m+1}$$

and so λ_{CV} is also weakly D-optimal for the order of regularization p' higher than p.

To achieve stronger levels of D-optimality for a given p appears to be a greater challenge in practice. This challenge arises from the fundamental nature of ill-posed inverse problems and is illustrated by the following observations.

In the degenerate case m = 0, which is the case of data smoothing, there is clearly no difference between S- and D-optimality. Thus *cross-validation applied to a smoothing problem achieves exactly the right level of optimality* in the present context. (We are not discussing absolute optimality, which is not achieved.) What then if the problem of differentiating inexact data is first converted to a smoothing problem, and then an S-optimal method for choosing a regularization parameter for the smoothing problem is used? Does this do better than S-optimality for the original problem?

Consider first the direct approach of pth-order regularization applied to mth-order Fourier differentiation, i.e.

$$\min_{f \in T_N} \left\{ \frac{1}{N} \sum_{n=0}^{\infty} [(Kf)(x_n) - y_n]^2 + \lambda \| f^{(p)} \|_{L^2}^2 \right\} \tag{26}$$

where T_N denotes the space of trigonometric polynomials on [0,1] of degree at most $\frac{1}{2}N$. The minimizer $f_{N;\lambda}$ of (26) has discrete Fourier coefficients

$$\frac{(i\omega_q)^m \hat{y}_{N,q}}{1 + \lambda \omega_q^{2\nu}}, \quad -\frac{1}{2}N \leq q \leq \frac{1}{2}N, \quad \nu = m+p, \tag{27}$$

where $\{\hat{y}_{N,q}\}$ denotes the discrete fourier transform (DFT) of the data $\{y_n\}$. The factor $(i\omega_q)^m$ in (27) appears as the result of mth-order differentiation whereas the factor $(1 + \lambda\omega_q^{2\nu})^{-1}$ is the regularization filter acting on the qth coefficient.

Now consider the *a-posteriori* smoothing approach where we first compute inexact derivative data $\{d_n\}$ which we subsequently smooth using pth-order regularization. The DFT of the derivative data will always have the form

$$\hat{d}_{N,q} = (i\omega_q)^m \tau_{N,q} \hat{y}_{N,q} \tag{28}$$

where $\tau_{N,q}$ is an attenuation factor determined by whatever method we use to generate the derivative data $\{d_n\}$. For example if $\{d_n\}$ is obtained from $\{y_n\}$ by mth-order central differencing, we have

$$\tau_{N,q} = \left[\frac{\sin\left(\frac{\omega_q}{N}\right)}{\left(\frac{\omega_q}{N}\right)}\right]^m .$$

Alternatively if $\{d_n\}$ is obtained by direct regularization then $\tau_{N,q}$ is the associated regularization filter (cf.(27)).

The variance-covariance matrix of the Fourier data $\{\hat{d}_{N,q}\}$ is

$$\frac{\sigma^2}{N}\hat{V},$$

where $\hat{V} = \text{diag}\left(\omega_q^{2m}\tau_{N,q}^2\right)$. The associated matrix V for the data $\{d_n\}$ is given by

$$V = \psi\hat{V}\psi^*$$

where ψ is the DFT matrix. Both $\hat{V}$ and V have rank < N. In terms of the generalized inverse $V^\dagger$ of V, the a posteriori smoothing problem may be stated:

$$\min_{\phi \in T_N} \left\{\frac{1}{N}(\underset{\sim}{\phi} - \underset{\sim}{d}_N)^T V^\dagger (\underset{\sim}{\phi} - \underset{\sim}{d}_N) + \lambda \|\phi^{(p)}\|_{L^2}^2\right\} \tag{29}$$

where $\underset{\sim}{\phi}$ denotes the N-vector sampling ϕ at x_n. It is easily shown that the minimizer $\phi_{N;\lambda}$ of (29) has discrete Fourier coefficients

$$\frac{(i\omega_q)^m \tau_{N,q} \hat{y}_{N,q}}{1 + \lambda \tau_{N,q}^2 \omega_q^{2\nu}}, \qquad -\frac{1}{2}N \leq q \leq \frac{1}{2}N. \tag{30}$$

For finite N there will be a difference, therefore, in the derivatives $f_{N;\lambda}$ obtained through (27) by the direct approach and $\phi_{N;\lambda}$ obtained through (30) by the a posteriori smoothing approach. As $N \to \infty$, however, it transpires that $\phi_{N;\lambda} \to f_{N;\lambda}$ since $\tau_{N,q} \to 1$ for all q. What then if we choose an S-optimal value of λ for the smoothing problem (29)? It is not difficult to prove the following:

Theorem 2. Let $0 < \tau_N \leq |\tau_{N,q}| \leq 1 < \infty$ for all q. Then the value of λ which minimizes

$$E\left\{\frac{1}{N}\sum_{n=0}^{N-1}\left[\phi_{N;\lambda}(x_n) - f(x_n)\right]^2\right\}$$

under the regularity condition (8), may be written

$$\tilde{\lambda}_0^{(aps)} = \theta_0 \tilde{\lambda}_0 \quad \text{as } N \to \infty,$$

where θ_0 satisfies

$$\tau_N^{4\nu/(4\nu+1)} \leq \theta_0 \leq \tau_N^{-2(6\nu+1)/(4\nu+1)} \quad \text{when } \beta \geq 2,$$

$$\tau_N^{8\nu} \leq \theta_0^{2\beta\nu+1} \leq \tau_N^{-2(6\nu+1)}, \quad \text{when } 0 < \beta < 2.$$

Corollary. $\tilde{\lambda}_0^{(aps)} \to \tilde{\lambda}_0$ as $N \to \infty$, whichever way the derivative data $\{d_n\}$ are derived. This follows immediately from the fact that $\tau_N \to 1$ as $N \to \infty$, and so $\theta_0 \to 1$.

Thus, in the limit $N \to \infty$, S-optimal a posteriori smoothing (whatever method

is used to provide derivative data) is equivalent to S-optimal regularization of the original differentiation problem. The reason for the lack of improvement is clear. The variance-covariance matrix of the differentiated data reflects the instability of differentiation, and the weighting thus introduced into the a posteriori smoothing problem makes it equivalent to the differentiation problem as $N \to \infty$. Of course there is a difference when N is finite, and improvements in the quality of the derivative may be possible in practice (cf., for example, [13]) using a posteriori smoothing.

REFERENCES

1. A R Davies and R S Anderssen. Numer. Math. 48 (1986) 671-697.
2. M A Lukas. Research Report CMA-R15-87, Centre for Mathematical Analysis, The Australian National University.
3. R Arcangeli. C R Acad. Sci. Paris, Ser. A263 (1966) 282-285.
4. V A Morozov. USSR Comput. Math. and Math. Phys. 8 (1968) 63-67.
5. H Engl and A Neubauer. In "Constructive Methods for the Practical Treatment of Integral Equations" (Eds. G Hämmerlin and K H Hoffmann). pp 120-141. Birkhäuser, Berlin (1985).
6. E Schock. Integral Equations and Operator Theory 7 (1984) 894-898.
7. V F Turchin. USSR Comput. Math. and Math. Phys. 7 (1967) 79-96.
8. V F Turchin. USSR Comput. Math. and Math. Phys. 8 (1968) 328-339.
9. P Craven and G Wahba. Numer. Math. 31 (1979) 377-403.
10. G Wahba. SIAM J. Numer. Anal. 14 (1977) 651-667.
11. G. Glein. J. Comput. Appl. Math. 5 (1979) 259-263.
12. R S Anderssen and P Bloomfield. Numer. Math. 22 (1974) 157-182.
13. I Koch and R S Anderssen. In "Computational Techniques and Applications: CTAC-85". Conference proceedings to be published by North Holland.

Department of Mathematics
University College of Wales
Aberystwyth
SY23 3BZ
United Kingdom