METHODS FOR CHOOSING THE REGULARIZATION PARAMETER

Mark A. Lukas

1. INTRODUCTION

Many inverse problems arising in practice can be modelled in the form of an operator equation

$$Kf = g, \tag{1.1}$$

where the function $g : I\!R^d \to I\!R$ is known only as discrete noisy data $y_i = g(x_i) + \epsilon_i$, $i = 1, \ldots, n$. An example is the estimation of an unknown parameter function f (e.g. a diffusivity) in a partial differential equation $L_f u = h$. Here the operator equation (1.1) is $Kf = u(f) = g$, where $u(f)$ is the solution of the forward problem $L_f u = h$ subject to some boundary conditions, but g is known only at discrete points and with error.

Suppose that equation (1.1) has a unique solution f_0. In many cases (1.1) is ill–posed, meaning that with respect to appropriate normed spaces, the inverse of K is not continuous at f_0. To obtain a reasonably good approximate solution, it is usually necessary to stabilize the problem. This is especially true given only discrete noisy data.

A well-known and effective technique for stabilizing the problem is the method of regularization. This replaces the original problem by the minimization over f in a suitable Hilbert space W of

$$n^{-1}\sum_{i=1}^{n}(Kf(x_i)-y_i)^2 \; + \; \lambda J(f), \tag{1.2}$$

where the stabilizing functional $J(f)$ is either the squared norm $J(f) = \| f \|_W^2$ or a squared seminorm. Typically, the space W is a Sobolev space $W^{m,2}$ and $J(f)$ is the L^2 norm of sums of squares of derivatives of f; for example, with $m = 2$ and $d = 2$,

$$J(f) \; = \; \int\int(f_{t_1t_1}^2 + 2f_{t_1t_2}^2 + f_{t_2t_2}^2)dt_1dt_2.$$

Clearly, minimizing (1.2) represents a tradeoff between fidelity to the data through the first sum of squares and smoothness of the approximate solution through $J(f)$. The regularization parameter $\lambda > 0$ is the weighting factor between the two. In general, if λ is too small, then errors in the data will be severely magnified resulting in a very noisy approximate solution. If λ is too large, then the approximate solution, although smooth, will not be consistent with the data. The choice of λ is therefore crucial.

It is possible of course to choose λ by testing a range of values and selecting the one which appears to give a good approximate solution with the correct degree of smoothness. However this is only a subjective choice, and on its own is not very satisfactory. It would be better to use some objective method of choosing λ. For such a method to be reliable, the choice should be based on the given data and its influence on the resulting approximate solution.

If K is linear and the functionals $W \to \mathbb{R}$, $f \to Kf(x_i)$ are bounded, then it is possible to explicitly write down the solution of (1.2), called the regularized solution $f_{n\lambda}$. Because $f_{n\lambda}$ depends linearly on the data y_i, $i = 1, \ldots, n$, the influence of the data can be expressed in terms of the influence matrix A defined by

$$A\boldsymbol{y}_i \; = \; Kf_{n\lambda}(x_i)$$

for any data vector $\boldsymbol{y}$. It is not hard to derive an explicit form for A. A spectral decomposition of A reveals how regularization achieves a filtering of the signal from the noisy data (see section 2).

The situation is more complicated when K is non-linear. Though a unique regularized solution $f_{n\lambda}$ of (1.2) exists under quite general conditions, it must be obtained by an iterative method. Furthermore, the degree of ill–posedness of (1.1) can depend on the solution f_0, and the influence function $\boldsymbol{y} \rightarrow Kf_{n\lambda}$ is non-linear. However, in principle it is possible to approximate this influence function locally by linearizing $Kf_{n\lambda}$, and base the choice of λ on this linearization. For the problem of estimating a diffusivity f, O'Sullivan and Wong [17] have developed such a method (see also section 9.7 in [22]).

In this paper we will consider four important methods for choosing the regularization parameter: the unbiased risk estimate, the discrepancy principle, generalized cross-validation (GCV) and generalized maximum likelihood (GML). We will confine our attention to linear problems involving functions of one variable, but from above it is clear that the results also have a bearing on non-linear problems.

Our aim is to compare the above methods from a theoretical point of view by determining their asymptotic behaviour as the number of points $n \rightarrow \infty$. To provide a basis for comparison, we need to decide on an appropriate loss function. For the most part, we will use the risk $ER(\lambda)$ defined as

$$ER(\lambda) \; = \; En^{-1}\sum_{i=1}^{n}[Kf_{n\lambda}(x_i) - g(x_i)]^2,$$

where E denotes expectation. Some results for other loss functions involving norms on the input error $f_{n\lambda} - f_0$ will also be mentioned. An estimate λ_X performs well with respect to the risk criterion if the inefficiency ratio

$$I = \frac{ER(\lambda_X)}{\min ER(\lambda)}$$

is close to 1. The method will be called asymptotically optimal (ao) if $I \to 1$ as $n \to \infty$, and weakly ao if $I = O(1)$.

In some special cases, results about the asymptotic behaviour of the above methods have been known for some time. If $K = I$ (the identity), then minimizing (1.2) is a well-known technique for data smoothing. In this case, Craven and Wahba [3] show that GCV is asymptotically optimal with respect to the risk. Davies and Anderssen [5] derive a similar result for the problem of periodic numerical differentiation. Wahba [19, 21] and Davies and Anderssen [5, 6] also derive some asymptotic results in certain cases for the discrepancy principle and GML.

In this paper, we survey the results in [12] and [13] about the above methods. These results apply to an arbitrary linear operator K with general smoothness conditions on f_0, and use both the risk and more general loss functions. In loose terms, the results show that GCV is ao in many cases, while the discrepancy principle is mostly only weakly ao, and GML is in most cases asymptotically sub-optimal meaning that $I \to \infty$ as $n \to \infty$. Therefore, based on these asymptotic results, the GCV estimate is to be recommended. This estimate also has the practical advantage of not requiring knowledge of the error variance. If the error variance is known, then the unbiased risk estimate is also recommended.

2. THE REGULARIZED SOLUTION

Assume that in (1.1), $K : L^2(0,1) \to L^2(0,1)$ is linear and 1-1. For simplicity, we use regularization of the form: minimize

$$n^{-1}\sum_{i=1}^{n}(Kf(x_i) - y_i)^2 + \lambda \parallel f \parallel_W^2 \tag{2.1}$$

over a suitable Hilbert space W, for example the Sobolev space $W = W^{m,2}[0,1]$. Note that the main results of this paper also hold for regularization with the seminorm $\|f^{(m)}\|_{L^2}$ in place of the norm $\|f\|_W$ in (2.1).

Assume that for each $x \in [0,1]$, the linear functional $f \to Kf(x)$ is bounded from $W \to I\!R$. Then there exists a representer $\eta_x \in W$ such that for all $f \in W$, $Kf(x) = (f, \eta_x)_W$. Define the kernel q by

$$q(x,t) = (\eta_x, \eta_t)_W = K\eta_x(t). \tag{2.2}$$

Then it is known (see [9, 20]) that (2.1) has a unique solution $f_{n\lambda}$ (the regularized solution) which can be represented as

$$f_{n\lambda} = \boldsymbol{\eta}^T(Q_n + n\lambda I)^{-1}\boldsymbol{y}, \tag{2.3}$$

where $\eta_i = \eta_{x_i}$, $i = 1, \ldots, n$, and $Q_n = [q(x_i, x_j)]$.

If for example K is an integral operator given by

$$Kf(x) = \int_0^1 k(x,t)f(t)\,dt,$$

and $W = W^{m,2}[0,1]$, then η_i and Q_n are known explicitly as (see [20])

$$\eta_i(t) = \int_0^1 r(x,t)k(x_i,s)\,ds \quad \text{and}$$

$$[Q_n]_{ij} = \int_0^1 \int_0^1 k(x_i,t)r(t,s)k(x_j,s)\,dsdt,$$

where $r(t,s)$ is the Green's function for a certain differential operator of order $2m$.

Define $K_n : W \to I\!R^n$ by $(K_n f)_i = Kf(x_i)$. Clearly, from (2.2) and (2.3), the influence matrix A is given by

$$Ay = K_n f_{n\lambda} = Q_n(Q_n + n\lambda I)^{-1}\boldsymbol{y}. \tag{2.4}$$

We now introduce a spectral decomposition which is important for the analysis of the regularized solution (see also [20]). From the definition, the matrix $n^{-1}Q_n$ is symmetric and non-negative definite. Therefore it has eigenvalues $\bar{\lambda}_i$ such that $\bar{\lambda}_1 \geq \bar{\lambda}_2 \geq \cdots \geq \bar{\lambda}_n \geq 0$ and corresponding eigenvectors $\bar{\phi}_i$ such that

$$n^{-1}(\bar{\phi}_i, \bar{\phi}_j) = \delta_{ij},$$

where $(\cdot,\cdot)$ denotes the usual Euclidean inner product.

In terms of this spectral decomposition, it is easy to show from (2.4) that

$$A\boldsymbol{y} = \sum_{i=1}^{n} n^{-1}(\boldsymbol{y}, \bar{\phi}_i)\,[\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)]\,\bar{\phi}_i. \tag{2.5}$$

Since

$$\boldsymbol{y} = \sum_{i=1}^{n} n^{-1}(\boldsymbol{y}, \bar{\phi}_i)\bar{\phi}_i$$

and $\bar{\lambda}_i/(\bar{\lambda}_i + \lambda)$ is a decreasing sequence, (2.5) shows that $K_n f_{n\lambda}$ achieves a low pass filtering of the data. Note that λ determines the effective cut-off frequency of this filter.

The matrix $n^{-1}Q_n$ and its finite spectral decomposition above can be related to an operator Q with infinite spectral decomposition. Let F_n denote the empirical distribution function of the points x_i and assume that F_n converges in the sup norm to a distribution function F with density bounded away from 0 and ∞. Let $L^2(F)$ denote the space $L^2(0,1)$ with inner product

$$(g,h)_{L^2(F)} \;=\; \int_0^1 gh\,dF.$$

Clearly the norms $\|\cdot\|_{L^2}$ and $\|\cdot\|_{L^2(F)}$ are equivalent. Let $K^*: L^2(F) \to W$ be the adjoint of $K: W \to L^2(F)$.

LEMMA 2.1 *If $K: W \to L^2(F)$ is bounded, then $KK^* = Q$, where Q is the integral operator given by*

$$Qh(x) \;=\; \int_0^1 q(x,t)h(t)\,dF.$$

Proof Since $K: W \to L^2(F)$ is bounded, dom$K^* = L^2(F)$. By definition of η_x and from (2.2), for all $h \in L^2(F)$

$$KK^*h(x) \;=\; (\eta_x, K^*h)_W \;=\; (K\eta_x, h)_{L^2(F)} \;=\; Qh(x).$$

We assume throughout that $K: W \to L^2(F)$ is compact with dense range. Then $Q = KK^*: L^2(F) \to L^2(F)$ is compact, 1-1 and positive. Hence Q has an infinite sequence of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ with $\lambda_i \to 0$, and corresponding eigenfunctions ϕ_i that are orthonormal in $L^2(F)$. We also assume that $q(x,t)$ is bounded on $[0,1] \times [0,1]$.

Suppose that the function $t \to q(x,t)g(t)$ is absolutely continuous for each $x \in [0,1]$. Then we can write

$$n^{-1}Q_n\boldsymbol{g}_i \;=\; \int_0^1 q(x_i,t)g(t)\,dF_n, \quad i = 1,\ldots,n, \tag{2.6}$$

where $\boldsymbol{g} = (g(x_1), \ldots, g(x_n))^T$, and

$$\begin{aligned} |Qg(x_i) - n^{-1}Q_n\boldsymbol{g}_i| &= |\int_0^1 q(x_i,t)g(t)\,d(F - F_n)| \\ &= |\int_0^1 (F - F_n)(t)\,d[q(x_i,t)g(t)]|, \end{aligned}$$

which → 0 as $n \to \infty$ since $F_n \to F$ in the sup norm. This means that (2.6) is a consistent approximation of $Qg(x_i)$. Therefore, one can expect (under certain conditions - see [7]) that the eigenvalues and eigenvectors of $n^{-1}Q_n$ will approximate the eigenvalues and eigenfunctions of Q. In the asymptotic analysis behind the results of this paper, estimates of certain functions of $\bar{\lambda}_i$ and $\bar{\phi}_i$ are obtained by comparison with the corresponding functions of λ_i and ϕ_i.

3. OPTIMAL REGULARIZATION PARAMETER

We first introduce a class of norms which will be used to gauge the smoothness of f_0 or $g = Kf_0$. It is known (see [10]) that the set $H_\rho = Q^{\rho/2}(L^2)$ with the inner product

$$(g,h)_{H_\rho} = (Q^{-\rho/2}g, Q^{-\rho/2}h)_{L^2}$$

defines a Hilbert space, which can also be represented as

$$H_\rho = \left\{ g \in L^2 : \|g\|_{H_\rho}^2 = \sum_{i=1}^{\infty}(g,\phi_i)_{L^2(F)}^2/\lambda_i^\rho < \infty \right\}.$$

From this space we define another Hilbert space W_ρ to be the completion of the set

$$\{f \in W : Kf \in H_\rho\}$$

under the inner product

$$(f,v)_{W_\rho} = (Kf, Kv)_{H_\rho}.$$

Note that for any $f \in W_\rho$, $\|f\|_{W_\rho} = \|Kf\|_{H_\rho}$. It can be shown that $W_1 = W$ and $H_0 = L^2(F)$, and for $\rho > \mu$, $W_\rho \subset W_\mu$ with continuous imbedding. Thus, as ρ increases, the W_ρ and H_ρ norms become increasingly strong. In fact, in some cases W_ρ and H_ρ can be identified with fractional Sobolev spaces (see [2, 11]).

The above norms can be used to define the loss functions

$$EL_\rho(\lambda) = E\|f_{n\lambda} - f_0\|^2_{W_\rho}, \tag{3.1}$$

where E denotes expectation. However, to simplify the presentation, we will concentrate on the risk loss function

$$ER(\lambda) = n^{-1}E\|K_n f_{n\lambda} - \boldsymbol{g}\|^2. \tag{3.2}$$

Note that the risk is an approximation of

$$EL_0(\lambda) = E\|Kf_{n\lambda} - g\|^2_{L^2(F)}. \tag{3.3}$$

Asymptotic estimates of $EL_\rho(\lambda)$ and $ER(\lambda)$, and their minimizers λ_ρ and λ_R respectively, are known (see [2, 10, 12, 16]), but here we only state the results for $ER(\lambda)$ and λ_R.

For convenience we will use the following notation. For two positive sequences a_n and b_n, we write $a_n \lesssim b_n$ if $a_n = O(b_n)$, and $a_n \approx b_n$ if there exist c_1, $c_2 > 0$ such that $c_1 b_n \leq a_n \leq c_2 b_n$. As usual $a_n \sim b_n$ means $a_n = b_n(1 + o(1))$. We will also use the function

$$D(\lambda; a, b) = \begin{cases} \lambda^a, & \lambda \leq 1, \\ \lambda^b, & \lambda > 1. \end{cases}$$

In addition to the earlier conditions, we will make use of the following assumptions.

Assumption 1. The errors ϵ_i are uncorrelated random variables with mean $E\epsilon_i = 0$ and variance $E\epsilon_i^2 = \sigma^2$.

Assumption 2. The eigenvalues λ_i of $Q : L^2(F) \rightarrow L^2(F)$ satisfy $\lambda_i \approx i^{-r}$, $r > 1$.

Assumption 3. There exist $s \in (0, 1 - 1/r)$, $\rho_1, \ldots, \rho_J \subseteq [0, s]$ and a sequence $d_n \rightarrow 0$ such that for all $f, v \in W$,

$$\left| \int_0^1 Kf(x)Kv(x)\,dF \; - \; n^{-1}\sum_{i=1}^{n} Kf(x_i)Kv(x_i) \right| \;\le\; d_n \sum_{j=1}^{J} \|f\|_{W_{\rho_j}} \|v\|_{W_{s-\rho_j}}.$$

Assumption 3 can be interpreted as prescribing a degree of accuracy for the quadrature formula.

Using Assumption 1 and (3.2), the risk can be decomposed as

$$ER(\lambda) \;=\; b^2(\lambda) + v(\lambda), \tag{3.4}$$

where $b^2(\lambda)$ is the squared bias

$$b^2(\lambda) \;=\; n^{-1}\|EK_n f_{n\lambda} - g\|^2 \;=\; n^{-1}\|(I - A)g\|^2 \tag{3.5}$$

and $v(\lambda)$ is the variance

$$v(\lambda) \;=\; n^{-1}E\|K_n f_{n\lambda} - EK_n f_{n\lambda}\|^2 \;=\; n^{-1}E\|A\epsilon\|^2 \;=\; \sigma^2 n^{-1}\mathrm{tr}A^2. \tag{3.6}$$

THEOREM 3.1 *(a) Suppose that Assumptions 1, 2 and 3 hold, and let $f_0 \in W_\beta$, $\beta \ge s$. There exists a sequence $\lambda_n \rightarrow 0$ (known in terms of d_n) such that*

$$\min\{1, \lambda^2\}\|g\|^2_{L^2(F)} \;\lesssim\; b^2(\lambda) \;\lesssim\; \min\{1, \lambda^\beta\}\|f_0\|^2_{W_\beta} \quad \text{if} \quad 0 \le \beta < 2,$$

and

$$b^2(\lambda) \;\approx\; \min\{1, \lambda^2\}\|f_0\|^2_{W_2} \quad \text{if} \quad \beta \ge 2,$$

uniformly in $\lambda \in [\lambda_n, \infty)$.

(b) Suppose that Assumptions 2 and 3 hold, and $\lambda_n \to 0$ *as* $n \to \infty$ *such that* $d_n^2 \lambda_n^{-(s+1)} \to 0$. *Then*

$$v(\lambda) \sim \sigma^2 n^{-1} \sum_{i=1}^{\infty} [\lambda_i/(\lambda_i + \lambda)]^2 \approx \sigma^2 n^{-1} D(\lambda; -1/r, -2),$$

uniformly in $\lambda \in [\lambda_n, \infty)$.

For the proof of Theorem 3.1, see Theorems 4.4 and 4.5 in [12].

Combining (3.4) and Theorem 3.1, it is clear that there exists a sequence $\lambda = \lambda(n)$ such that $ER(\lambda) \to 0$ as $n \to \infty$. Let λ_R minimize $ER(\lambda)$, i.e. λ_R is optimal with respect to the risk. We now consider how λ_R behaves asymptotically.

Define S to be the set of functions for which the upper bound on $b^2(\lambda)$ in Theorem 3.1 is achieved in the sense that

$$b^2(\lambda) \approx \min\{1, \lambda^\beta\} \|f_0\|_\beta^2 \quad \text{if} \quad 0 \le \beta < 2.$$

It is shown in [10] that there is such a set.

COROLLARY 3.1 *Suppose that the assumptions of Theorem 3.1 hold and* λ_R *minimizes* $ER(\lambda)$. *Define*

$$\lambda^* = \begin{cases} (\sigma^2 n^{-1})^{r/(\beta r+1)}, & 0 < \beta < 2, \\ (\sigma^2 n^{-1})^{r/(2r+1)}, & \beta \ge 2, \end{cases} \tag{3.7}$$

and assume that $\lambda^* \ge \lambda_n$. *If either* $f_0 \in W_\beta$, $\beta \ge 2$, *or* $f_0 \in S \cap W_\beta$, $0 < \beta < 2$, *then* $\lambda_R \approx \lambda^*$ *and* $ER(\lambda_R) \approx ER(\lambda^*)$ *as* $n \to \infty$.

Define the functions $\mu_1(\lambda)$ and $\mu_2(\lambda)$ by

$$\mu_1(\lambda) = n^{-1}\mathrm{tr}A = n^{-1}\sum_{i=1}^{n} \bar{\lambda}_i/(\bar{\lambda}_i + \lambda) \quad \text{and}$$

$$\mu_2(\lambda) \;=\; n^{-1}\mathrm{tr}A^2 \;=\; n^{-1}\sum_{i=1}^{n}[\bar{\lambda}_i/(\bar{\lambda}_i+\lambda)]^2.$$

These functions play a crucial role in several places. For example, from (3.6), the variance $v(\lambda) = \sigma^2\mu_2(\lambda)$. From Theorems 4.1 and 4.3 in [12], we have the following estimates. If Assumptions 2 and 3 hold and $\lambda_n \to 0$ such that $d_n^2\lambda_n^{-(s+1)} \to 0$, then

$$(3.8) \qquad \mu_1(\lambda) \;\sim\; n^{-1}\sum_{i=1}^{\infty}\lambda_i/(\lambda_i+\lambda) \;\approx\; n^{-1}D(\lambda;-1/r,-1) \qquad \text{and}$$

$$(3.9) \qquad \mu_2(\lambda) \;\sim\; n^{-1}\sum_{i=1}^{\infty}[\lambda_i/(\lambda_i+\lambda)]^2 \;\approx\; n^{-1}D(\lambda;-1/r,-2),$$

uniformly in $\lambda \in [\lambda_n, \infty)$.

Lastly in this section, we consider the question of existence and uniqueness of λ_R. The following result is derived in [8].

THEOREM 3.2 *A minimizer λ_R of $ER(\lambda)$ exists, and if the sequence*

$$s_i \;=\; n^{-2}(g,\bar{\phi}_i)^2/\bar{\lambda}_i^2, \quad i=1,\ldots,n,$$

is non-increasing, then λ_R is unique.

By examining the proof, it is clear that the condition in Theorem 3.2 can be relaxed to some extent without affecting uniqueness. Hence it is plausible that if $n^{-2}(g,\bar{\phi}_i)^2/\bar{\lambda}_i^2$ approximates $(g,\phi_i)^2_{L^2(F)}/\lambda_i^2$ and the latter is a non-increasing sequence, then λ_R will be unique. Therefore, we can expect that for a significant subset of $g \in H_2$, there will be a unique minimizer λ_R.

4. UNBIASED RISK ESTIMATE

Suppose that the error variance σ^2 (or a very good estimate of it) is known. Define the random function $\hat{R}$ by

$$\hat{R}(\lambda) = n^{-1}\|(I-A)\boldsymbol{y}\|^2 + 2\sigma^2 n^{-1}\mathrm{tr}A - \sigma^2$$

and let $\hat{\lambda}_U$ minimize $\hat{R}(\lambda)$. Note that $\hat{R}(\lambda)$ can be computed from the spectral decomposition in section 2.

From Assumption 1

$$En^{-1}\|(I-A)\boldsymbol{y}\|^2 = n^{-1}\|(I-A)\boldsymbol{g}\|^2 + \sigma^2 n^{-1}\mathrm{tr}(I-A)^2, \tag{4.1}$$

and therefore it is clear from (3.4), (3.5) and (3.6) that $E\hat{R}(\lambda) = ER(\lambda)$. That is, $\hat{R}(\lambda)$ is an unbiased estimate of the risk $ER(\lambda)$. Thus the minimizer λ_U of $E\hat{R}(\lambda)$ equals λ_R, so the inefficiency $I = ER(\lambda_U)/ER(\lambda_R)$ is identically 1, i.e. λ_U is optimal for all n. It is feasible therefore that $\hat{\lambda}_U$ will be a good estimate of λ_R.

This estimate was proposed in this general context by Craven and Wahba [3] and Lukas [8]. Numerical experiments carried out in [8] indicate that $\hat{\lambda}_U$ is very reliable, and is to be recommended.

5. DISCREPANCY PRINCIPLE

Like the unbiased risk estimate, the discrepancy principle method also assumes that σ^2 (or a good estimate of it) is known. Define the discrepancy function $D(\lambda)$ by

$$D(\lambda) = n^{-1}\|(I-A)\boldsymbol{y}\|^2,$$

which from (2.5) can be expressed as

$$D(\lambda) \;=\; \lambda^2 \sum_{i=1}^{n} n^{-2}(\boldsymbol{y}, \bar{\phi}_i)^2/(\bar{\lambda}_i + \lambda)^2. \tag{5.1}$$

The discrepancy principle states that for a good regularized solution $f_{n\lambda}$ the discrepancy $D(\lambda)$ should be of the same order as the error variance σ^2. From (5.1) it is clear that $D(\lambda)$ is strictly increasing, with $D(0) = 0$ and $D(\lambda) \to \sum_{i=1}^{n} n^{-2}(\boldsymbol{y}, \bar{\phi}_i)^2$ as $\lambda \to \infty$. Hence, if $\sigma^2 < \sum_{i=1}^{n} n^{-2}(\boldsymbol{y}, \bar{\phi}_i)^2$, there is a unique solution to the equation $D(\lambda) = \sigma^2$. This solution is the discrepancy principle estimate $\hat{\lambda}_D$.

Taking expectation of $D(\lambda)$ as in (4.1) gives

$$ED(\lambda) \;=\; b^2(\lambda) + \sigma^2 n^{-1}\mathrm{tr}(I - A)^2.$$

It is not hard to show (see [13]) that the equation $ED(\lambda) = \sigma^2$ has a unique solution, which we call the "expected" discrepancy principle estimate λ_D.

The discrepancy principle estimate was proposed by Morozov [14, 15] and has been investigated and extended by a number of authors. However, these investigations have largely been done in a deterministic framework. In the present probabilistic framework for the case of data smoothing, Wahba [19] shows that if f_0 is sufficiently smooth, then λ_D is weakly ao in the sense that $ER(\lambda_D)/ER(\lambda_R) = O(1)$ as $n \to \infty$. Davies and Anderssen [6] extend this result to periodic numerical differentiation and certain convolution integral equations.

The next result (proved in [13] as Theorem 3.2) shows that λ_D is weakly ao in general.

THEOREM 5.1 *Suppose that Assumptions 1, 2 and 3 hold, $f_0 \in W_\beta$, $\beta \geq 2$, or*

$f_0 \in S \cap W_\beta$, $s \leq \beta < 2$, and $\lambda_n \to 0$ in such a way that $n^{-1}\lambda_n^{-1/r} \to 0$ and

$$d_n^2 \lambda_n^{-(s+1)} \to 0 \quad \text{if} \quad s \leq \beta < 3 - 1/r,$$

$$d_n^2 \lambda_n^{-(s+1/r+\beta-2)} \to 0 \quad \text{if} \quad \beta \geq 3 - 1/r.$$

Let λ^* be as defined in (3.7) and assume that $\lambda^* \geq \lambda_n$ and $\lambda_D \geq \lambda_n$. Then $\lambda_D \approx \lambda_R \approx \lambda^*$ as $n \to \infty$ and

$$\frac{ER(\lambda_D)}{ER(\lambda_R)} = O(1).$$

The proof of this theorem begins by rewriting the equation $ED(\lambda) = \sigma^2$ as

$$b^2(\lambda) = \sigma^2[2\mu_1(\lambda) - \mu_2(\lambda)],$$

and then makes use of the estimates of $b^2(\lambda)$, $\mu_1(\lambda)$ and $\mu_2(\lambda)$ given in Theorem 3.1, (3.8) and (3.9) respectively.

In [13] it is shown that λ_D is also weakly ao with respect to a range of the loss functions $EL\rho(\lambda)$ so long as $f_0 \notin W_{2+\delta}$, $\delta > 0$, which can be interpreted as meaning that f_0 should not be too smooth relative to W.

In addition to being weakly ao when f_0 is smooth, the following result (proved in [13] as Theorem 3.4) shows that λ_D consistently oversmooths with respect to the risk.

THEOREM 5.2 *Suppose that Assumptions 1, 2 and 3 hold and $q(x,t)$ has the smoothness properties of the Green's function of a self-adjoint linear differential operator of order $2p$. Also suppose that*

$$d_n^2 n^{r(s+1)/(2r+1)} \to 0 \quad \text{and} \quad h_n^p n^{r/(2r+1)} \to 0,$$

where $h_n = \max\{x_{i+1} - x_i\}$, and for all sufficiently large n, the minimizer λ_R of $ER(\lambda)$ is unique . If $f_0 \in W_3$, then for all n sufficiently large, $\lambda_D > \lambda_R$.

With respect to the stronger loss functions $EL_\rho(\lambda)$, $\rho > 0$, however, λ_D does not oversmooth. In fact it will usually undersmooth because, from Theorem 3.3 in [13], if $f_0 \in W_\beta$, $\beta > 2$, then $\lambda_D/\lambda_\rho \to 0$ as $n \to \infty$.

6. GENERALIZED CROSS-VALIDATION

The GCV method has a significant practical advantage over the previous two methods in that it requires no knowledge of σ^2. First, we briefly describe the motivating idea behind the method.

Let $f_{n\lambda}^{(k)}$ be the minimizer over W of

$$n^{-1}\sum_{i\neq k}(Kf(x_i) - y_i)^2 + \lambda\|f\|_W^2,$$

which is the same functional as in (2.1) but with the kth data point excluded. Intuitively, we would expect that for a good choice $\hat{\lambda}$ of the regularization parameter, $Kf_{n\hat{\lambda}}^{(k)}(x_k)$ should be closer to y_k on average than $Kf_{n\lambda}(x_k)$ for other λ. That is, $Kf_{n\hat{\lambda}}^{(k)}(x_k)$ should be a good predictor of y_k. Define $V(\lambda)$ to be the weighted sum of squares of prediction errors

$$V(\lambda) = n^{-1}\sum_{k=1}^{n}(Kf_{n\lambda}^{(k)}(x_k) - y_k)^2 w_k^2, \tag{6.1}$$

where

$$w_k = (1 - a_{kk})/[n^{-1}\mathrm{tr}(I - A)].$$

The weights w_k are required to reflect the fact that even for a good $\hat{\lambda}$, for different k, $Kf_{n\hat{\lambda}}^{(k)}(x_k)$ will have different reliability as a predictor of y_k. For example, if $K = I$ and the points x_i are sparse about x_k, then $f_{n\hat{\lambda}}^{(k)}(x_k) - y_k$ has relatively large variability, so a small weight w_k is required to reduce the importance of the kth prediction error.

The GCV estimate $\hat{\lambda}_V$ is defined as the minimizer of $V(\lambda)$ over $\lambda > 0$. The expression (6.1) is not suitable for computation, but it can be shown (see [22]) that $V(\lambda)$ has the equivalent form

$$V(\lambda) = \frac{n^{-1}\|(I-A)\boldsymbol{y}\|^2}{[n^{-1}\text{tr}(I-A)]^2}.$$

This can be computed using the spectral decomposition in section 2. Also define the "expected" GCV estimate λ_V to be the minimizer of $EV(\lambda)$.

The GCV estimate was proposed and investigated by Wahba [20]. In the case of data smoothing, Craven and Wahba [3] show that λ_V is ao with respect to the risk (a gap in their argument being filled by Utreras [18]). Davies and Anderssen [5] extend this result to periodic numerical differentiation. We now describe some general results obtained in [12] (and foreshadowed in [11]) for the case of an arbitrary linear operator.

The following result (proved in [12] as Theorem 5.1) shows that in general λ_V is ao with respect to the risk.

THEOREM 6.1 *Suppose that Assumptions 1, 2 and 3 hold, $f_0 \in W_\beta$, $\beta \geq s$, and $\lambda_n \to 0$ as $n \to \infty$ as in Theorem 3.1. Let λ^* be as in (3.7) and assume that $\lambda^* \geq \lambda_n$. Let $\lambda_R = \lambda_R(n)$ minimize $ER(\lambda)$ over $\lambda \geq \lambda_n$. Then there exists a sequence $\lambda_V = \lambda_V(n)$ of minimizers of $EV(\lambda)$ over $\lambda \geq \lambda_n$ such that as $n \to \infty$*

$$\frac{ER(\lambda_V)}{ER(\lambda_R)} \to 1.$$

The proof of this result begins in the same way as the corresponding result in [3], i.e. it is shown that

$$(6.2) \qquad |ER(\lambda) + \sigma^2 - EV(\lambda)|/ER(\lambda) \leq h(\lambda) \equiv (2\mu_1 + \mu_1^2/\mu_2)/(1-\mu_1)^2.$$

From (3.8) and (3.9), if $\lambda \to 0$ as $n \to \infty$ such that $d_n^2 \lambda^{-(s+1)} \to 0$ and $n^{-1}\lambda^{-1/r} \to 0$, then

$$h(\lambda) \approx 3n^{-1}\lambda^{-1/r} \to 0.$$

Combining this with (6.2) means intuitively that for a certain range of λ, the graph of $EV(\lambda)$ tracks the graph of $ER(\lambda)$. It is plausible therefore that the minimizers λ_V and λ_R will be close.

Under much the same conditions as in Theorem 6.1, it is shown in [12] that λ_V is also ao with respect to $EL_0(\lambda)$ defined in (3.3). Note that the asymptotic optimality is independent of the error variance σ^2, the eigenvalue decay rate r and the smoothness index β of f_0.

The situation is more complicated for the general class of loss functions $EL_\rho(\lambda)$ defined in (3.1). Basically, if $f_0 \notin W_{2+\delta}$, $\delta > 0$, and either $f_0 \in S \cap W_\beta$, $s \leq \beta < 2$, or $f_0 \in W_\beta$, $\beta = 2$ (meaning that f_0 should not be too smooth relative to W), then λ_V is weakly ao with respect to $EL_\rho(\lambda)$ for any $0 < \rho \leq \beta$. However, if $f_0 \in W_\beta$, $\beta > 2$, then for any $\rho > 0$, λ_V is asymptotically sub-optimal in that $EL_\rho(\lambda_V)/EL_\rho(\lambda_\rho) \to \infty$.

In [12], these results are applied to the case of periodic numerical differentiation. It is shown there that Davies and Anderssen [5] are incorrect in stating that λ_V is always asymptotically sub-optimal with respect to the mean-square derivative error. As indicated above, λ_V will be weakly ao for a certain class of f_0.

7. GENERALIZED MAXIMUM LIKELIHOOD

Like GCV, the GML method also has the practical advantage of not requiring σ^2. The method derives from the following Bayesian model of the regularized solution $f_{n\lambda}$ (see Wahba [22]).

Suppose that ϵ_i is normally distributed and let $f(t)$ be a Gaussian stochastic process $b^{1/2}Z(t)$ with mean 0 and covariance function

$$Eb^{1/2}Z(s)b^{1/2}Z(t) \;=\; br(s,t),$$

where b is a constant and the kernel $r_s(t) = r(s,t)$ satisfies $Kr_s(x) = \eta_x(s)$. It is known that $f_{n\lambda}$ is the conditional expectation

$$f_{n\lambda}(t) \;=\; E\{f(t)\,|\,Kf(x_i) + \epsilon_i = y_i,\ i = 1,\ldots,n\},$$

with $\lambda = \sigma^2/(nb)$. In addition, the random vector $\boldsymbol{Y}$ with components $Y_i = Kf(x_i)+\epsilon_i$ is normally distributed as

$$\boldsymbol{Y} \;\sim\; N(0, bQ_n + \sigma^2 I).$$

Setting $\lambda = \sigma^2/(nb)$, this becomes

$$\boldsymbol{Y} \;\sim\; N(0, b(Q_n + n\lambda I)).$$

Then the GML estimate $\hat{\lambda}_M$ is defined to be the usual maximum likelihood estimate of λ given the random value $\boldsymbol{y}$ of $\boldsymbol{Y}$. It can be shown that $\hat{\lambda}_M$ is the minimizer over $\lambda > 0$ of $M(\lambda)$ defined by

$$M(\lambda) \;=\; \frac{\boldsymbol{y}^T(Q_n + n\lambda I)^{-1}\boldsymbol{y}}{[\det(Q_n + n\lambda I)^{-1}]^{1/n}},$$

which from (2.4) can also be written as

$$M(\lambda) = \frac{\boldsymbol{y}^T(I-A)\boldsymbol{y}}{[\det(I-A)]^{1/n}}.$$

Define the "expected" GML estimate λ_M to be the minimizer of $EM(\lambda)$.

The estimate $\hat{\lambda}_M$ was first proposed by Anderssen and Bloomfield [1] in the context of numerical differentiation (see also Davies [4]). It was extended to an arbitrary linear operator by Wahba [21], where it was also generalized (hence the name GML) to the case where the stabilizing functional $J(f)$ is a semi-norm. In [21] some asymptotic results for λ_M are derived under certain heuristic assumptions. For the case of periodic numerical differentiation, Davies and Anderssen [5] show that if f_0 is smooth, then λ_M is asymptotically sub-optimal with respect to the risk. In [13] this is shown to be true in general with respect to the risk and the other loss functions $EL_\rho(\lambda)$ defined in (3.1). A more precise statement of the result is as follows (see Theorems 4.2 and 4.3 in [13]). It shows that λ_M is asymptotically undersmoothing.

THEOREM 7.1 *Suppose that Assumptions 1, 2 and 3 hold and let $f_0 \in W_\beta$, $\beta > 1$, and $0 \le \rho < 2 - s - 1/r$. If $d_n \to 0$ sufficiently quickly, then there exists a sequence $\lambda_n \to 0$, and a sequence of minimizers λ_R of $ER(\lambda)$, λ_ρ of $EL_\rho(\lambda)$ and λ_M of $EM(\lambda)$ over $\lambda \ge \lambda_n$ such that $\lambda_M/\lambda_R \to 0$, $\lambda_M/\lambda_\rho \to 0$,*

$$\frac{ER(\lambda_M)}{ER(\lambda_R)} \to \infty \quad \text{and} \quad \frac{EL_\rho(\lambda_M)}{EL_\rho(\lambda_\rho)} \to \infty.$$

It is known that the situation is different if f_0 is "rough" relative to W (see [21]). If, for example,

$$\sum_{i=1}^{n}(g,\phi_i)^2/\lambda_i \approx n$$

(which implies that $f_0 \notin W$), then λ_M may in fact perform reasonably well. However, from Theorem 7.1, on the whole λ_M has unfavourable asymptotic properties.

Acknowledgements The work for this paper was completed while the author was on sabbatical from Murdoch University and visiting the Centre for Mathematical Analysis, The Australian National University, and the Johannes- Kepler–Universität, Linz. He is grateful for the support of these institutions and wishes to thank, in particular, Dr. R.S. Anderssen, Professor N.S. Trudinger and Professor H.W. Engl.

REFERENCES

[1] R.S. Anderssen and P. Bloomfield, Numerical differentiation procedures for non-exact data. *Numer. Math.* **22**(1974), 157-182.

[2] D.D. Cox, Approximation of regularization estimators. *Ann. Statist.* **16**(1988), 694-712.

[3] P. Craven and G. Wahba, Smoothing noisy data with spline functions. *Numer. Math.* **31**(1979), 377-403.

[4] A.R. Davies, On the maximum likelihood regularization of Fredholm convolution equations of the first kind. *Treatment of Integral Equations by Numerical Methods* (C.T.H. Baker, G.F. Miller, eds.), 95-105, Academic Press, London, 1982.

[5] A.R. Davies and R.S. Anderssen, Improved estimates of statistical regularization parameters in Fourier differentiation and smoothing. *Numer. Math.* **48**(1986), 671-697.

[6] A.R. Davies and R.S. Anderssen, Optimization in the regularization of ill–posed problems. *J. Austral. Math. Soc. Ser. B* **28**(1986), 114-133.

[7] P. Linz, On the numerical computation of eigenvalues and eigenvectors of symmetric integral equations. *Math. Comp.* **24**(1970), 905-910.

[8] M.A. Lukas, *Regularization of Linear Operator Equations*, Ph.D. thesis, Australian National University, 1981.

[9] M.A. Lukas, Regularization. *The Application and Numerical Solution of Integral Equations* (R.S. Anderssen, F.R. de Hoog, M.A. Lukas, eds.), 151-182, Sijthoff and Noordhoff, 1980.

[10] M.A. Lukas, Convergence rates for regularized solutions. *Math. Comp.* **51**(1988), 107-131.

[11] M.A. Lukas, Assessing regularised solutions. *J. Austral. Math. Soc. Ser. B* **30**(1988), 24-42.

[12] M.A. Lukas, Asymptotic optimality of generalized cross–validation for choosing the regularization parameter. Report No. 43, Centre for Mathematical Analysis, Australian National University, 1990.

[13] M.A. Lukas, Asymptotic behaviour of the discrepancy principle and generalized maximum likelihood for choosing the regularization parameter. Report No. 4, Centre for Mathematics and its Applications, Australian National University, 1991.

[14] V.A. Morozov, On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.* **7**(1966), 414-417.

[15] V.A. Morozov, *Methods for Solving Incorrectly Posed Problems.* Springer, New York, 1984.

[16] D.W. Nychka and D.D. Cox, Convergence rates for regularized solutions of integral equations from discrete noisy data. *Ann. Statist.* **17**(1989), 556-572.

[17] F. O'Sullivan and T. Wong, Determining a functional diffusion coefficient in the heat equation, Tech. Report 98, Department of Statistics, University of California, Berkeley, 1987.

[18] F. Utreras, Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comput.* **2**(1981), 349-362.

[19] G. Wahba, Smoothing noisy data by spline functions. *Numer. Math.* **24**(1975), 383-393.

[20] G. Wahba, Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14**(1977), 651-667.

[21] G. Wahba, A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**(1985), 1378-1402.

[22] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

School of Mathematical and Physical Sciences
Murdoch University
Murdoch W.A. 6150
Australia