COMMUN. STATIST.-THEORY METH., 19(5), 1685-1700 (1990)

WHEN IS THE OPTIMAL REGULARIZATION PARAMETER INSENSITIVE TO THE CHOICE OF THE LOSS FUNCTION ?

Grace Wahba

Department of Statistics
University of Wisconsin-Madison
Madison, WI53706

Yonghua Wang

Department of Statistics
University of California-Berkeley
Berkeley, CA94720

Key Words and Phrases: convergence rates; method of regularization; deconvolution; optimal smoothing parameter.

ABSTRACT

We investigate the behavior of the optimal regularization parameter in the method of regularization for solving first kind integral equations with noisy data, under a range of definitions of "optimal", varying from mean square error in higher derivatives of the solution, to mean square error in the predicted data. We study how the optimal regularization parameter changes when the optimality criteria changes, under a broad range of smoothness assumptions on the solution, the kernel of the integral operator, and the penalty functional. Although some of the calculations we present have been given elsewhere, we organize the results with a specific goal in mind. That is, we study a certain class of problems within which we can identify conditions on the solution, the kernel of the operator and the penalty functional for which the rate at which the optimal regularization parameter goes to zero is the same for both predictive mean square error and solution mean square error optimality criteria, and for which it is different. The former circumstances are of interest because then data based estimates of the regularization parameter such as generalized cross-validation, which are known to be optimal for predictive mean square error, will also go to zero at the optimal rate for solution mean square error.

1. INTRODUCTION

Consider the Fredholm integral equation of the first kind

$$\int_0^1 K(t,s)f(s)ds = g(t), \quad t \in [0,1] \tag{1.1}$$

where g is measured discretely and with error. We are given data

$$y_i = g(t_i) + \varepsilon_i \qquad i = 1,2,\ldots,n \tag{1.2}$$

where

$$E\,\varepsilon_i = 0, \qquad E\varepsilon_i\varepsilon_j = \sigma^2\delta_{ij}, \quad i,j=1,2,\ldots,n \tag{1.3}$$

and σ^2 is unknown, $\delta_{ij} = 0, i \neq j; 1, i=j$.

In the method of regularization, the estimate $\hat{f}_{n\lambda}$ of f is taken as the minimizer of

$$\frac{1}{n}\sum_{j=1}^{n}[(Kf)(t_j) - y_j]^2 + \lambda \|f\|^2$$

where

$$(Kf)(t) = \int_0^1 K(t,s)f(s)ds,$$

and $\|\cdot\|$ is a norm or seminorm in a Hilbert space in which $L_t f = (Kf)(t)$ is a bounded linear functional for each $t \in [0,1]$ (see, for example Wahba [20] and the bibliography there).

As is well known, the solution (both in theory and practice) can be very sensitive to the choice of λ. As $n \to \infty$, the optimal $\lambda \to 0$ under rather general assumptions. Rates of decay for the optimal λ in this problem and related smoothing problems have been obtained by many authors under various assumptions (see for example [1, 2, 4-6, 8, 10, 14-20]).

Generalized cross validation (GCV) has been a popular method for choosing λ from the data, see for example [3, 9, 11]. This method has optimality properties for a predictive mean square error criteria, in particular, if $\hat{\lambda}$ is the GCV estimate of λ and

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[g(t_i) - \int_0^1 K(t_i,s)\hat{f}_{n\lambda}(s)ds]^2$$

then $R(\hat{\lambda})/\min_{\lambda} R(\lambda)$ is known to decrease to 1 as $n \to \infty$ in various senses, see, for

example, Wahba [17], Speckman [16], Li [7]. Recently there has been some discussion in the literature raising the question of what to do if one were more interested in minimizing a different and somewhat more natural loss function, such as

$$D(\lambda) = \int_0^1 (\hat{f}_{n,\lambda}(s) - f(s))^2 ds$$

(see e.g. Rice [14], O'Sullivan [12]). Since the properties of $\hat{\lambda}$ with respect to $R(\lambda)$ are fairly well established, the question arises of determining the circumstances, if any, under which the minimizer of $R(\lambda)$ also comes close to minimizing $D(\lambda)$. It has been the first author's observation based on a number of realistic numerical simulations in different contexts, that the λ that minimizes $R(\lambda)$ also comes close to minimizing $D(\lambda)$. Further discussion and references concerning this point may be found in Wahba [19] (p.1361). Thus, we thought it appropriate to collect in one place as general as possible a set of rate calculations so that this issue of whether or not $\hat{\lambda}$ is good for minimizing $D(\cdot)$ can be examined theoretically.

We consider only convolution equations

$$g(t) = \int_0^1 h(t-s) f(s) ds$$

with periodic kernels and periodic solutions.

Modest generalizations when the kernels K and R commute may be obtained by the methods in Wahba [19], but we do not discuss them here.

We develop rates of convergence of the optimal λ under a wide range of conditions, and then specialize to the criteria $ER(\lambda)$ and $ED(\lambda)$, and let $\|f\|^2 = \int_0^1 (f^{(m)}(u))^2 du$. In section 5, we give the following results: Let the kth Fourier coefficients of f and h go to zero at the rates $k^{-\alpha}$ and $k^{-\beta}$ with $\beta > 0$ and $\alpha+\beta > 1$. Let λ_S and λ_P be the minimizers of $ED(\lambda)$ and $ER(\lambda)$, respectively.

Suppose $2m + \frac{1}{2} \geq \alpha \geq \frac{1}{2}$, $m > \frac{1}{4}$, then as $n \to \infty$,

$$\lambda_S \asymp \lambda_P \asymp n^{-\frac{m+\beta}{\alpha+\beta}}$$

for any β, where "$U \asymp V$" means that there are two constants c and d such that $cU \leq V \leq dU$ for n sufficiently large. Suppose $\alpha > 2m + \frac{1}{2}$, $m > \frac{1}{4}$. Then

$$\lambda_P \asymp \lambda_S \asymp n^{-\frac{m+\beta}{\alpha+\beta}}$$

or

$$\lambda_P = o(\lambda_S)$$

according as $\beta > \alpha - (2m + \frac{1}{2})$ or $\beta \leq \alpha - (2m + \frac{1}{2})$.

2. THE STATISTICAL MODEL AND THE BASIC LEMMA

We consider the convolution equation

$$g(t) = \int_0^1 h(t-s)f(s)ds, \quad t \in [0,1] \text{ and } t_i = \frac{i}{n}, \quad i = 1,2,\ldots,n. \tag{2.1}$$

To avoid cumbersome notation, we assume that h and f have cosine series expansions. It is clear that the results go through if a general Fourier series expansion is allowed. Let the cosine series expansions of g, h, f be

$$g(t) = \sum_{k=1}^{\infty} 2g_k \cos 2\pi kt, \quad \text{with} \quad g_k = \int_0^1 \cos 2\pi kt \; g(t)dt;$$

$$h(t) = \sum_{k=1}^{\infty} 2h_k \cos 2\pi kt, \quad \text{with} \quad h_k = \int_0^1 \cos 2\pi kt \; h(t)dt;$$

$$f(t) = \sum_{k=1}^{\infty} 2f_k \cos 2\pi kt, \quad \text{with} \quad f_k = \int_0^1 \cos 2\pi kt \; f(t)dt$$

and $g_k = h_k f_k$. The relationship among g, h, f is described by the convolution $g = h * f$.

The method of regularization is to estimate f by the solution of the following problem: find the minimizer $\hat{f}_{n,\lambda}$ of

$$\frac{1}{n}\sum_{j=1}^{n} \left(y_j - (h*f)(t_j)\right)^2 + \frac{\lambda}{(2\pi)^{2m}} \int_0^1 \left(f^{(m)}(t)\right)^2 dt. \tag{2.3}$$

We will make use of the finite Fourier transform as Rice [14] did. Let

$$y_{kn} = \sum_{j=1}^{n} y_j \cos 2\pi jk/n,$$

$$g_{kn} = \sum_{j=1}^{n} g(t_j) \cos 2\pi jk/n$$

etc.

Now $n^{-1}g_{kn} = g_k + r_{kn}$ where $r_{kn} = \frac{1}{n}\sum_{j=1}^{n} g(t_j)\cos 2\pi jk/n - \int_0^1 g(t)\cos 2\pi kt\,dt$. If $g(\cdot)$ is a continuous function, r_{kn} tends to zero. In the remainder of this work we will proceed as though $g_{kn} = ng_k = nh_k f_k$ etc.

Assume that $\hat{f}_{n\lambda}$ has Fourier coefficients $\hat{f}_k$'s. To estimate f by the method of regularization essentially is to find $\hat{f}_k$'s, which are the minimizers of

$$\sum_{k=1}^{n}\left(\frac{y_{kn}}{n} - h_k f_k\right)^2 + \lambda\sum_{k=1}^{n} k^{2m} f_k^2. \tag{2.4}$$

Note that

$$\int_0^1 (f^{(m)}(t))^2\,dt = \frac{1}{2}\sum_{k=1}^{\infty}(2\pi)^{2m}k^{2m}f_k^2$$

$$\approx \frac{1}{2}(2\pi)^{2m}\sum_{k=1}^{n} k^{2m}f_k^2.$$

Then

$$\hat{f}_k = \frac{y_{kn}}{n}\,\frac{h_k}{h_k^2 + \lambda k^{2m}}.$$

The resulting estimate $\hat{g}_k$ of g_k is

$$\hat{g}_k = \hat{f}_k\,h_k = \frac{y_{kn}}{n}\,\frac{h_k^2}{h_k^2 + \lambda k^{2m}}.$$

In the rest of the paper, we will use the following lemma which was proved in Cox [1] and Lukas [8].

Lemma: Let

$$D(\lambda,s) = \sum_{k=1}^{\infty} k^s(1+\lambda k^r)^{-2}.$$

Then if $s < 2r-1$, $D(\lambda,s) < \infty$, for all $\lambda > 0$. Furthermore, when $\lambda \to 0^+$,

$$D(\lambda,s) \approx \begin{cases} \lambda^{-(s+1)/r} & \text{if } -1 < s < 2r-1 \\ \log(1/\lambda) & \text{if } s = -1 \\ 1 & \text{if } s < -1 \end{cases}$$

3. THE ESTIMATION ERRORS UNDER DIFFERENT NORMS

The mean square error in the solution

$$L(\hat{f}_{n\lambda} - f) = \int_0^1 (\hat{f}_{n\lambda}(t) - f(t))^2 dt, \tag{3.1}$$

is typically the loss function of interest in most practical cases.

By the discrete Fourier transformation and Parseval's indentity,

$$L(\hat{f}_n - f) = \sum_{k=1}^{\infty} (\hat{f}_k - f_k)^2. \tag{3.2}$$

That is

$$\begin{aligned} \text{MSSE} &= EL(\hat{f}_{n\lambda} - f) = E\sum_{k=1}^{\infty} (\hat{f}_k - f_k)^2 \\ &\approx E\sum_{k=1}^{n} (\hat{f}_k - f_k)^2. \end{aligned} \tag{3.3}$$

The mean square error in the lth derivative of the solution, if it exists, is

$$\int_0^1 (\hat{f}_{n\lambda}^{(l)}(t) - f^{(l)}(t))^2 dt \approx E\sum_{k=1}^{n} (2\pi k)^{2l} (\hat{f}_k - f_k)^2.$$

The mean square error for $\hat{g}_{n\lambda} = h * \hat{f}_{n\lambda}$, i.e. the mean square prediction error for the problem is

$$\begin{aligned} \text{MSPE} &= EL(\hat{g}_{n\lambda} - g) \\ &\approx E\sum_{k=1}^{n} (\hat{g}_k - g_k)^2 \\ &= E\sum_{k=1}^{n} h_k^2 (\hat{f}_k - f_k)^2. \end{aligned} \tag{3.4}$$

Let $\| \cdot \|_q$ be defined by:

$$\| f \|_q^2 = \sum_{k=1}^{\infty} q_k |f_k|^2, \quad q_k > 0, \ k = 1,2,\ldots, \tag{3.5}$$

where f_k's are the Fourier coefficients of f. (This is the general form of the norms defined in Cox [1] and Lukas [8]). Let H_q be the Hilbert space of functions for which (3.5) is finite, with inner product

$$\langle f, g \rangle = \sum_{k=1}^{\infty} q_k f_k g_k.$$

(See [1], [8], [10]).

The expected mean square error under this norm is defined as

$$\mathrm{MSE}_q(\lambda) = E\sum_{k=1}^{\infty} q_k(\hat{f}_k - f_k)^2 \tag{3.6}$$

for the f_k and $\hat{f}_k$ defined in Section 2.

MSSE, MSPE and mean square error in the derivative are all special cases of MSE_q. When $q_k = 1$, $k = 0,1,\ldots$, MSE_q is the mean square solution error, when $q_k = h_k^2$, $k=0,1,\ldots$, MSE_q is the mean square prediction error, and when $q_k = (2\pi k)^{2l}$, MSE_q is the mean square error in the lth derivative of f, if it exists. Now

$$E\left(\frac{y_{kn}}{n}\right) = g_k = h_k f_k,$$

$$\mathrm{cov}\left(\frac{y_{jn}}{n}, \frac{y_{kn}}{n}\right) = \frac{\sigma^2}{n}\delta_{jk},$$

$$E\hat{f}_k = f_k \frac{h_k^2}{h_k^2 + \lambda k^{2m}}.$$

Thus,

$$\begin{aligned}\mathrm{MSE}_q(\lambda) &\approx E\sum_{k=1}^{n} q_k(\hat{f}_k - f_k)^2 \\ &= \sum_{k=1}^{n} q_k\left(E\hat{f}_k - f_k\right)^2 + \sum_{k=1}^{n} q_k E\left(\hat{f}_k - E\hat{f}_k\right)^2 \\ &\approx \sum_{k=1}^{n} q_k f_k^2\left[\frac{\lambda k^{2m}}{h_k^2 + \lambda k^{2m}}\right]^2 + \frac{\sigma^2}{n}\sum_{k=1}^{n}\frac{q_k h_k^2}{(h_k^2 + \lambda k^{2m})^2}.\end{aligned}$$

We will use the definition of the optimal λ as follows:

Let λ_0 be the minimizer of $\mathrm{MSE}_q(\lambda)$. If λ has the property that

$$\mathrm{MSE}_q(\lambda) \approx \mathrm{MSE}_q(\lambda_0)$$

we will say λ is optimal. This definition corresponds to the weak optimality defined by Davies and Anderssen [5].

4. CONVERGENCE RATES OF THE OPTIMAL λ

In order to investigate the convergence properties of the optimal λ for MSE_q we will assume that

$$f_k \approx k^{-\alpha}, \quad h_k \approx k^{-\beta}, \quad q_k \approx k^{\gamma} \tag{4.1}$$

$$\alpha, \beta \geq 0, \quad k=1,2,\cdots.$$

Then

$$\mathrm{MSE}_q(\lambda) \approx \sum_{k=1}^{n} k^{\gamma-2\alpha}\left[\frac{\lambda k^{2m}}{k^{-2\beta}+\lambda k^{2m}}\right]^2 + \frac{\sigma^2}{n}\sum_{k=1}^{n}\frac{k^{\gamma-2\beta}}{(k^{-2\beta}+\lambda k^{2m})^2}$$

$$= \lambda^2 \sum_{k=1}^{n}\frac{k^{4m+\gamma-2\alpha}}{(k^{-2\beta}+\lambda k^{2m})^2} + \frac{\sigma^2}{n}\sum_{k=1}^{n}\frac{k^{\gamma+2\beta}}{(1+\lambda k^{2(m+\beta)})^2}$$

$$= \lambda^2 \sum_{k=1}^{n}\frac{k^{4(m+\beta)+\gamma-2\alpha}}{(1+\lambda k^{2(m+\beta)})^2} + \frac{\sigma^2}{n}\sum_{k=1}^{n}\frac{k^{\gamma+2\beta}}{(1+\lambda k^{2(m+\beta)})^2}. \tag{4.2}$$

Letting $s_1 = 4(m+\beta)+\gamma-2\alpha$, $s_2 = \gamma+2\beta$ and $r = 2(m+\beta)$, the asymptotic behavior of (4.2) as $\lambda \to 0$ is given by the Lemma in Section 2 provided $s_1 < 2r-1$, $s_2 < 2r-1$. We have to consider each of the interesting combinations of $s_i \in (-1, 2r-1)$, $s_i = -1$ and $s_i < -1$ separately.

Note that

$$\mathrm{MSE}_q = \sum_{k=1}^{\infty} k^{\gamma+2\beta}(\hat{g}_k - g_k)^2.$$

If $\gamma+2\beta < 0$ then this is a weaker norm than the L_2 norm and corresponds to putting less penalty on high frequency rather than low frequency errors. Thus it is of little practical interest and we omit the straight forward but tedious calculations for the special cases $s_2 = \gamma+2\beta \le -1$. In what follows, then, we are assume $-1 < s_2 < 2r-1$, which results in the conditions

$$\beta > \max\left\{-\frac{1}{2}(1+\gamma),\ \frac{1}{2}(1+\gamma)-2m\right\}. \tag{4.3}$$

The three cases below then correspond, respectively, to $s_1 \in (-1, 2r-1)$, $s_1 = -1$, and $s_1 < -1$.

Case I : $-1 < 4(m+\beta)+\gamma-2\alpha < 4(m+\beta)-1$

This gives

$$\gamma < 2\alpha-1, \tag{4.4}$$

$$\beta > \frac{1}{2}(\alpha-(2m+\tfrac{1}{2}))-\frac{1}{4}\gamma, \tag{4.5}$$

$$\beta > \max\left\{-\frac{1}{2}(1+\gamma),\ \frac{1}{2}(1+\gamma)-2m\right\}. \tag{4.6}$$

Thus under the condition

$$\beta > \max\left\{\frac{1}{2}(\alpha-(2m+\tfrac{1}{2}))-\frac{1}{4}\gamma\ ,\ -\frac{1}{2}(1+\gamma)\ ,\ \frac{1}{2}(1+\gamma)-2m\right\}, \tag{4.7}$$

we have

$$\mathrm{MSE}_q(\lambda) \approx \lambda^2 \cdot \lambda^{-\frac{4(m+\beta)+\gamma-2\alpha+1}{2(m+\beta)}} + \frac{\sigma^2}{n}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}}$$

$$= \lambda^{\frac{2\alpha-\gamma-1}{2(m+\beta)}} + \frac{\sigma^2}{n}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}}. \tag{4.8}$$

The optimal λ_q satisfies

$$\lambda^{\frac{2\alpha-\gamma-1}{2(m+\beta)}-1} - \frac{\sigma^2}{n}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}} \approx 0.$$

That is

$$\lambda^{\frac{2\alpha-\gamma-1+\gamma+2\beta+1}{2(m+\beta)}} \approx n^{-1},$$

$$\lambda_q \approx n^{-\frac{m+\beta}{\alpha+\beta}} \tag{4.9}$$

which does not depend on γ provided the condition (4.4), (4.5), (4.6) are satisfied.

Case II : $4(m+\beta)+\gamma-2\alpha = -1$

Combining this with (4.3) gives, if

$$\beta = \frac{1}{2}(\alpha-(2m+\frac{1}{2}))-\frac{1}{4}\gamma,$$

$$\beta > \max\left\{-\frac{1}{2}(1+\gamma),\ \frac{1}{2}(1+\gamma-2m)\right\},$$

then

$$\mathrm{MSE}_q(\lambda) \approx \lambda^2 \log\frac{1}{\lambda} + \frac{\sigma^2}{n}\lambda^{\frac{\gamma+2\beta+1}{2(m+\beta)}}. \tag{4.10}$$

The optimal λ_q satisfies

$$2\lambda\log\frac{1}{\lambda} - \lambda - \frac{\sigma^2}{n}\frac{\gamma+2\beta+1}{2(m+\beta)}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}-1} \approx 0.$$

That is

$$\lambda^{\frac{4(m+\beta)+\gamma+2\beta+1}{2(m+\beta)}}\left(2\log\frac{1}{\lambda} - 1\right) \approx n^{-1}, \tag{4.11a}$$

$$\lambda_q = o\left(n^{-\frac{2(m+\beta)}{4m+6\beta+\gamma+1}}\right). \tag{4.11b}$$

Case III : $4(m+\beta)+\gamma-2\alpha < -1$

These conditions, combined with (4.3) give

$$\beta < \frac{1}{2}(\alpha-(2m+\frac{1}{2}))-\frac{1}{4}\gamma,$$

$$\beta > \max\left\{-\frac{1}{2}(1+\gamma),\ \frac{1}{2}(1+\gamma-2m)\right\},$$

then

$$\mathrm{MSE}_q(\lambda) \approx \lambda^2+\frac{\sigma^2}{n}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}}, \tag{4.12}$$

and the optimal λ_q satisfies

$$2\lambda - \frac{\sigma^2}{n}\frac{\gamma+2\beta+1}{2(m+\beta)}\lambda^{-\frac{\gamma+2\beta+1}{2(m+\beta)}-1} \approx 0, \tag{4.13a}$$

$$\lambda_q \approx n^{-\frac{2(m+\beta)}{4m+6\beta+\gamma+1}}. \tag{4.13b}$$

5. CONVERGENCE RATES OF THE OPTIMAL λ FOR MSSE AND MSPE

We are especially interested in the convergence rates of the optimal λ's chosen by minimizing MSSE and MSPE.

$$\mathrm{MSE}_q(\lambda) = \mathrm{MSSE}, \quad with\ \gamma = 0 \tag{5.1}$$

and

$$\mathrm{MSE}_q(\lambda) = \mathrm{MSPE}, \quad with\ \gamma = -2\beta. \tag{5.2}$$

Let λ_S and λ_P be the minimizers of MSSE and MSPE, respectively. We have the following

Theorem:

(A) Suppose $2m+\frac{1}{2} \geq \alpha > \frac{1}{2}$, $m > \frac{1}{4}$. Then for any allowed β,

$$\lambda_P \approx \lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}}.$$

(B) Suppose $\alpha > (2m+\frac{1}{2})$, $m > \frac{1}{4}$, and $\beta > (\alpha-(2m+\frac{1}{2}))$. Then

$$\lambda_P \approx \lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}}.$$

(C) Suppose $\alpha > (2m+\frac{1}{2})$, $m > \frac{1}{4}$ and $\beta \le (\alpha-(2m+\frac{1}{2}))$. Then λ_P does not otherwise depend on α and

$$\lambda_P = o(\lambda_S).$$

(D) Under the conditions of (A) or (B),

$$\mathrm{MSSE}(\lambda_S) \approx \mathrm{MSSE}(\lambda_P) \approx n^{-\frac{\alpha-\frac{1}{2}}{\alpha+\beta}}.$$

If

$$0 < \frac{1}{2}(\alpha-(2m+\frac{1}{2})) < \beta \le (\alpha-(2m+\frac{1}{2})),$$

then

$$\mathrm{MSSE}(\lambda_S) \approx n^{-\frac{\alpha-\frac{1}{2}}{\alpha+\beta}}.$$

whereas

$$\frac{\mathrm{MSSE}(\lambda_S)}{\mathrm{MSSE}(\lambda_P)} = o(1).$$

If

$$0 < \beta < \frac{1}{2}(\alpha-(2m+\frac{1}{2})),$$

then

$$\mathrm{MSSE}(\lambda_S) \approx n^{-\frac{4(m+\beta)}{4m+6\beta+1}},$$

and

$$\frac{\mathrm{MSSE}(\lambda_S)}{\mathrm{MSSE}(\lambda_P)} = o(1),$$

and the rate of convergence of MSSE for this problem could be increased to $n^{-\frac{\alpha-\frac{1}{2}}{\alpha+\beta}}$ by increasing m so that $\beta > \frac{1}{2}(\alpha-(2m+\frac{1}{2}))$. (Note that $\frac{4(m+\beta)}{4m+6\beta+1}$ is increasing with m and equals $(\alpha-\frac{1}{2})/(\alpha+\beta)$ if $\beta = \frac{1}{2}(\alpha-(2m+\frac{1}{2}))$.)

Proof:

(A) $2m+\frac{1}{2} \ge \alpha > \frac{1}{2}$, $m > \frac{1}{4}$.

Under these conditions the inequality of (4.7) holds for $\gamma = 0$ and $\gamma = -2\beta$ so that the conditions of Case I are satisfied for any $\beta > 0$ and so

$$\lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}}, \tag{5.3}$$

$$\lambda_P \approx n^{-\frac{m+\beta}{\alpha+\beta}}. \tag{5.4}$$

(B) $\alpha > 2m+\frac{1}{2}$, $\beta > (\alpha-(2m+\frac{1}{2}))$, $m > \frac{1}{4}$.

Again, the inequality (4.7) holds for $\gamma = 0$ and $\gamma = -2\beta$. Thus

$$\lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}},$$

$$\lambda_P \approx n^{-\frac{m+\beta}{\alpha+\beta}}.$$

(C) The condition $\beta \le (\alpha-(2m+\frac{1}{2}))$ will be divided into four cases, (Ci) - (Civ).

(Ci) $\beta = (\alpha-(2m+\frac{1}{2})) > 0, \quad m \ge 0.$

When $\gamma = 0$, the Case I conditions are still satisfied,

$$\lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}}.$$

However, for $\gamma = -2\beta$, the conditions of the Case II are satisfied, and we have

$$\lambda_P = o(n^{-\frac{2(m+\beta)}{4m+4\beta+1}}). \tag{5.5}$$

Substituting $2\alpha = 4m+2\beta+1$ into (5.5) gives

$$\lambda_P = o(n^{-\frac{m+\beta}{\alpha+\beta}}). \tag{5.6}$$

(Cii) $0 < \frac{1}{2}(\alpha-(2m+\frac{1}{2})) < \beta < \alpha-(2m+\frac{1}{2})$.

If $\gamma = 0$, the Case I conditions are still satisfied,

$$\lambda_S \approx n^{-\frac{m+\beta}{\alpha+\beta}},$$

but $\gamma = -2\beta$ falls into Case III and

$$\lambda_P \approx n^{-\frac{2(m+\beta)}{4(m+\beta)+1}}. \tag{5.7}$$

Here $4(m+\beta)+1 < 2(\alpha+\beta)$, which gives

$$\lambda_P = o(\lambda_S).$$

(Ciii) $\beta = \frac{1}{2}(\alpha-(2m+\frac{1}{2})) > 0$, Here $4(m+\beta) = 2\alpha-1$.

$$\lambda_S = o(n^{-\frac{2(m+\beta)}{4m+6\beta+1}}) = o(n^{-\frac{m+\beta}{\alpha+\beta}})$$

and

$$\lambda_P \approx n^{\frac{2(m+\beta)}{4m+4\beta+1}} = n^{-\frac{(m+\beta)}{\alpha}}. \tag{5.8}$$

Thus,

$$\lambda_P = o(\lambda_S). \tag{5.9}$$

(Civ) $0 < \beta < \frac{1}{2}(\alpha-(2m+\frac{1}{2}))$.

Here

$$\lambda_S \approx n^{-\frac{2(m+\beta)}{4m+6\beta+1}}, \tag{5.10}$$

$$\lambda_P \approx n^{-\frac{2(m+\beta)}{4(m+\beta)+1}} = o(\lambda_S). \tag{5.11}$$

(D) These results are obtained by a straight forward substituting (5.3), (5.4), (5.6), (5.8), (5.9), (5.10), (5.11) into (4.8) and (4.12) under the different conditions.

We make some remarks concerning these results. Let W_θ be the Hilbert space of periodic functions whose Fourier coefficients g_k's satisfy $\sum_k k^{2\theta}|g_k|^2 < \infty$. W_θ is a reproducing kernel space for any $\theta > \frac{1}{2}$, and if $|g_k| \approx k^{-(\alpha+\beta)}$, then g will be in W_θ if $(\alpha+\beta) > \theta + \frac{1}{2}$. Thus the assumption $(\alpha+\beta) > 1$ insures that g is in some reproducing kernel space, i.e. $g(t)$ is well defined pointwise, not just as an L_2-function. The condition $\alpha > \frac{1}{2}$ is what is needed to guarantee that f is square integrable. Note that although the estimation procedure involves the W_m norm, the square bias plus variance for MSSE ($\gamma = 0$) is finite even though $f \notin W_m$.

We remark in passing that questions about f's not satisfying high order periodic boundary conditions can be answered by studying f's which are Bernoulli polynomials, since f_μ for the jth Bernoulli polynomial is a multiple of μ^{-j}.

Thus the following "theoretical" advice (as opposed to "practical" advice) obtains:

To obtain the best possible MSSE, for a given β and α, choose m large enough so that $\beta \geq \frac{1}{2}(\alpha - (2m + \frac{1}{2}))$. If after doing this, β is still less than $\alpha - (2m + \frac{1}{2})$ the λ chosen by an MSPE criteria will not result in the optimal convergence rate. If it is desired to use an MSPE estimate for λ and obtain the optimal MSSE rate, then choose m large enough so that $\beta > \alpha - (2m + \frac{1}{2})$.

Of course with small, medium and even large data sets, many other factors come into play including the hidden constants, and information such as the solution contains at most 300 or at least 10 continuous derivatives is not really meaningful. The authors take no responsibility for anyone following this theoretical advice, but do remark that GCV can also be used to choose m. The best practical advice remains: make as realistic a simulation as possible of the problem at hand, and test the available estimation methods against simulated truth.

ACKNOWLEDGEMENTS

We wish to thank John Rice for some helpful correspondence during which he sent us some of the $m = 2$ results in this section.

This research supported in part by AFOSR under grant AFOSR-87-0171 and NASA under contract NAG5-316.

BIBLIOGRAPHY

[1] D.D. Cox (1983) Approximation of method of regularization estimators, *Technical Report No. 723*, Dept. of Statistics, University of Wisconsin.

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

[3] J. Crump and J. Seinfeld (1982) A new algorithm for inversion of aerosol size distribution data, *Aerosol Science and Technology,1*, pp 15-34.

[4] A.R. Davies and R.S. Anderssen (1985) Improved estimates of statistical regularization parameters in Fourier differentiation and smoothing, *Research Report CMA-R01-85 Contre for Mathematical Analysis*, Anstralian National University.

[5] A.R. Davies and R.S. Anderssen (1986) Optimalization in the regularization of ill-posed problem, to appear in *Journal of Australian Mathematical Society* (Series B), special issue on optimization.

[6] D. Donoho(1985) The minimum complexity formalism and the notion of maximum reliable structure, *manuscript*.

[7] K. Li (1986) Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing, *Ann. Statist.,14*, pp 1101-1112.

[8] M.A. Lukas (1984) *Convergence rates for regularized solutions*, to be published.

[9] P. Merz (1980) Determination of adsorption energy distribution by regularization and a characterization of certain adsorption isotherms, *Journal of Computational Physics, 38*, pp 64-85.

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

[11] D. Nychka, G. Wahba, S. Goldfarb and T. Pugh (1984) Cross-validation spline methods for the estimation of three dimensional tumor size distributions from observations on two dimensional cross sections, *J. Am. Stat. Assoc., 79*, pp 832-846.

[12] F. O'Sullivan (1987) A statistical perspective on ill-posed inverse problems, *Statistical Science, 1*, pp 502-527.

[13] D. Ragozin (1983) Error bounds for derivative estimates based on spline smoothing of exact or noisy data, *J. Approx. Theory, 37*, pp335-355.

[14] J. A. Rice (1985) Choice of smoothing parameter in deconvolution problems. *In Function Estimates, Contemporary Mathematics 59*, S. Marron (ed.). American Mathematical Society, pp 137-151.

[15] J. A. Rice and M. Rosenblatt (1983) Smoothing splines: regression, derivatives and deconvolution, *Ann. Math. Statist., 11*, pp 141-156.

[16] P. Speckman (1985) Spline Smoothing and optimal rates of convergence in nonparametric regression models, *Ann. Math. Statist., 13*, pp 970-983.

[17] G.Wahba (1977) Practical approximate solutions to linear operator equations when the data are noisy, *SIAM J. Numer. Anal. 14*, pp 645-667.

[18] G. Wahba (1979) Smoothing and ill-posed problems, *Solution Methods for Integral Equations with applications*, M. Golberg (ed.). Plenum Press.

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

[20] G. Wahba (1990) *Spline Models for Observational Data*, v. 59, CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Received March 1990.