CHAPTER 8

Fredholm Integral Equations of the First Kind

8.1 Existence of solutions, the method of regularization.

An equation of the form

$$g(t) = \int_\Omega K(t,s)f(s)\,ds,\ t \in \mathcal{T} \tag{8.1.1}$$

is known as a Fredholm integral equation of the first kind. Rewriting (8.1.1) as

$$g = \mathcal{K}f, \tag{8.1.2}$$

we have that Picard's theorem (Courant and Hilbert (1965, p.160)) says that $\mathcal{K}(\mathcal{L}_2) = \mathcal{H}_{K*K}$, where

$$K * K(u,v) = \int_\Omega K(u,s)K(v,s)\,ds.$$

Therefore a solution f in $\mathcal{L}_2$ for (8.1.2) exists provided $g \in \mathcal{H}_{K*K}$. Nashed and Wahba (1974) showed that $\mathcal{K}(\mathcal{H}_R) = \mathcal{H}_Q$, where

$$Q(u,v) = \iint_{\Omega\times\Omega} K(u,s)R(s,t)K(v,t)\,dsdt. \tag{8.1.3}$$

Therefore $g \in \mathcal{H}_Q$ ensures that there exist at least one solution $f \in \mathcal{H}_R$. (If K has a nontrivial null space in $\mathcal{H}_R$ then the solution is not unique but a unique generalized inverse can be defined; see Nashed and Wahba (1974).) The numerical solution of such equations has been a subject of intensive study over a long period of time. Until 1969 however (see Wahba (1969)) the literature concerned the case where g was presumed known exactly.

Brute force discretization of (8.1.1) to obtain a matrix equation and then solving the resultant linear system was doomed to failure because under very mild smoothness conditions on $K(\cdot,\cdot)$ adjacent rows of the matrix will become closer and closer as the discretization becomes finer and finer, and the calculation will become numerically unstable. Pretending to have infinite precision arithmetic, and infinitely accurate data, one could ask for $f_0 \in \mathcal{H}_R$ to minimize $\|P_1 f_0\|$ subject to

$$g(t_i) = \int K(t_i,s)f_0(s)\,ds,\ i = 1,2,\ldots,n.$$

In the computation of f_0, the matrix M of (1.3.9) is replaced by Σ (since $\lambda = 0$), which has ijth entry $Q(t_i, t_j)$. The condition number of Σ is the ratio of the largest to smallest eigenvalue. If the t_i's are roughly uniformly distributed, then the eigenvalues of Σ will (roughly) behave as n times the eigenvalues of Q. If K and R behave as Green's functions for kth order and $2m$th order linear differential operators, respectively, then the eigenvalues of Q will decay at the rate $\nu^{-(2m+2k)}$ and the estimate of the condition number of Σ is $O(n^{-(2m+2k)})$. If K is analytic, the eigenvalues will decay exponentially. Even with double precision on supercomputers, such an exact solution is numerically unstable for even moderate n.

Tikhonov (1963) suggested solving

$$g\left(\frac{i}{n}\right) = \int K\left(\frac{i}{n}, s\right) f(s)\, ds$$

approximately by (roughly), finding $(f(1/n), \ldots, f(n/n))$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(g\left(\frac{i}{n}\right) - \frac{1}{n}\sum_{j=1}^{n} K\left(\frac{i}{n}, \frac{j}{n}\right) f\left(\frac{j}{n}\right)\right)^2 + \lambda \sum_{j=2}^{n-1}\left(f\left(\frac{j+1}{n}\right) - 2f\left(\frac{j}{n}\right) + f\left(\frac{j-1}{n}\right)\right)^2. \tag{8.1.4}$$

He suggested choosing λ by trial and error. Phillips (1962) and Cook (1963) suggested a similar approach, and these and related methods are sometimes called Tikhonov–Phillips regularization.

The minimization of (1.3.4) was proposed in Wahba (1969) and the use of GCV to choose λ in this context appears in Wahba (1977a).

We mention only a few other references, which are directly related to the approach discussed here: Wahba (1980a, 1981a, 1981c, 1982c), Merz (1980), Crump and Seinfeld (1982), Mendelsson and Rice (1982), Nychka, Wahba, Goldfarb, and Pugh (1984), Cox and O'Sullivan (1990), O'Sullivan (1986a), Nychka and Cox (1989), Girard (1987a,b, 1989). See also the books of Groetsch (1984) and the books edited by Anderson, deHoog, and Lukas (1980), and Baker and Miller (1982). A recent translation from the Russian (Tikhonov and Goncharsky (1987)), gives a number of applications of Tikhonov–Phillips regularization to ill-posed problems in various fields. The discrepancy method and trial and error for choosing λ are used in the examples reported there.

8.2 Further remarks on ill-posedness.

The typical integral equation arising in practice can be very ill-posed; basically this means that very large data sets can contain a surprisingly small amount of information about the desired solution. As an example, consider Fujita's equation considered in Wahba (1979b) and references cited there. It is

$$g(t) = \int_0^{s_{\max}} \frac{\theta s e^{-\theta s t}}{[1 - e^{-\theta s}]} f(s)\, ds \tag{8.2.1}$$

where θ is given. This equation relates optical density along a tube after centrifugation $(g(t), t =$ distance along the tube) to $f(s)$, the molecular weight distribution of the contents of the tube. One should always look at the eigenvalues $\{\lambda_{\nu n}\}$ of the problem

$$g_i = L_i f = < \eta_i, f >,$$

which consist of M ones and the $n - M$ eigenvalues of $Q_2' \Sigma Q_2$, where Σ is the $n \times n$ matrix with ijth entry $< P_1\eta_i, P_1\eta_j >$. For the example of Fujita's equation considered in Wahba (1979b), with $n = 41$ and $M = 0$, we obtained the first five eigenvalues as 1, $10^{-3.5}, 10^{-7}, 10^{-10.5}$, and 10^{-14}. The remaining eigenvalues were "computer zero." Loosely speaking, even with extremely accurate data (say to eight figures), there are still only at most three linearly independent pieces of information in the data for this problem. The "number of independent pieces of information in the data" was considered in Wahba (1980a). Let

$$y_i = < \eta_i, f > + \epsilon_i, \ i = 1, \ldots, n$$

and let the $n \times n$ matrix with ijth entry $< \eta_i, \eta_j >$ satisfy

$$\{< \eta_i, \eta_j >\} = \Sigma = \Gamma D \Gamma'.$$

Let

$$\begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix} = D^{-1/2}\Gamma' \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_n \end{pmatrix},$$

and let z be the transformed data

$$z = \Gamma' y.$$

Then

$$z_\nu = \sqrt{\lambda_\nu} < \psi_\nu, f > + \tilde{\epsilon}_\nu,$$

where λ_ν is the ith diagonal entry of D and $\tilde{\epsilon} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n)' \sim \mathcal{N}(0, \sigma^2 I)$. If $\lambda_\nu < \psi_\nu, f >^2$ is large compared to σ^2, then one can obtain a good estimate of $< \psi_\nu, f >$, and, if it is not, one cannot. One might identify "the number of independent pieces of information in the data" with the number of ν's for which $\lambda_\nu/\sigma^2 >> 1$.

We note that

$$K(t, s) = \frac{\theta s e^{-\theta s t}}{[1 - e^{-\theta s}]}$$

of (8.2.1) is infinitely differentiable in t. The "smoother" $K(\cdot, \cdot)$ is, the more ill-posed the integral equation. $K(t, s) = (t - s)_+^{m-1}/(m - 1)!$ corresponds to $g^{(m)} = f$ for m some positive integer. The larger m is, the more ill-posed the problem. For $m < 1$ we have Abel's equations (see Anderssen and Jakeman (1975), Nychka et al. (1984)). A plot of the relevant eigenvalues appears in Nychka et al. Abel's equations are only mildly ill-posed. The equations arising in computerized tomography (Radon transforms; see e.g., Wahba (1981c) and references cited there) are also only mildly ill-posed.

8.3 Mildly nonlinear integral equations.

Remote sensing experiments frequently involve the observation of data on mildly nonlinear functions. For example, upwelling radiation above the atmosphere is related to the atmospheric vertical temperature distribution, and this radiation is measured from sensors aboard satellites (see, e.g., Fritz et al. (1972)). It is desired to recover the vertical temperature distribution from this data for the purpose of estimating initial conditions for numerical weather prediction. With some idealizations the relation is

$$R_\nu(T) = \int_{\text{surface}}^{\text{top}} \mathcal{B}_\nu(T(x))\tau_\nu'(x)\,dx \tag{8.3.1}$$

where $T(x)$ is the temperature at vertical coordinate x along a column of the atmosphere in the line of sight of the satellite sensor, R_ν is upwelling radiance at wavenumber ν, and $\mathcal{B}_\nu$ is Planck's function

$$\mathcal{B}_\nu[T(x)] = \frac{c_1\nu^3}{e^{c_2\nu/T(x)} - 1} \tag{8.3.2}$$

where c_1 and c_2 are known physical constants and τ_ν is the transmittance (usually assumed known). The data model is

$$y_\nu = R_\nu(T) + \epsilon_\nu,\ \nu = 1, \ldots, n.$$

The following approach was proposed in O'Sullivan and Wahba (1985).

Let

$$T(x) \simeq \sum_{k=1}^{N} c_k B_k(x)$$

where the B_k are B-splines, and let

$$R_\nu(T) \simeq N_\nu(c) = \int \mathcal{B}_\nu \left(\sum_{k=1}^{N} c_k B_k(x) \right) \tau_\nu'(x)\,dx. \tag{8.3.3}$$

Find $c = (c_1, \ldots, c_N)'$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - N_i(c))^2 + \lambda c'\Sigma c \tag{8.3.4}$$

where $\Sigma = \{\sigma_{ij}\}$, $\sigma_{ij} =< P_1B_i, P_1B_j >$. Fix λ, and use a Gauss–Newton iteration to find $c = c(\lambda)$: For $c = c^{(l)}$, the lth iterate, we have

$$N_i(c) \approx N_i(c^{(l)}) + \sum_{k=1}^{N} \left.\frac{\partial N_i}{\partial c_k}\right|_{c=c^{(l)}} (c_k - c_k^{(l)}). \tag{8.3.5}$$

Let $X^{(l)}$ be the $n \times N$ matrix with ikth entry $\partial N_i/\partial c_k|_{c=c^{(l)}}$, and let the "pseudodata" $y^{(l)}$ be

$$y^{(l)} = y - \begin{pmatrix} N_1(c^{(l)}) \\ \vdots \\ N_n(c^{(l)}) \end{pmatrix} + X^{(l)}c^{(l)}. \tag{8.3.6}$$

The minimization problem becomes: Find $c^{(l+1)}$ to minimize

$$\frac{1}{n}\|y^{(l)} - X^{(l)}c\|^2 + \lambda c'\Sigma c, \tag{8.3.7}$$

and

$$c^{(l+1)} = (X^{(l)'}X^{(l)} + n\lambda\Sigma)^{-1}X^{(l)'}y^{(l)}.$$

This iteration is run to convergence, say until $l = L = L(\lambda)$. Then the quadratic approximation to the original optimization problem (8.3.4) in the neighborhood of $c^{(L)}$ has the influence matrix $A^{(L)}(\lambda)$,

$$A^{(L)}(\lambda) = X^{(L)}(X^{(L)'}X^{(L)} + n\lambda\Sigma)^{-1}X^{(L)'} \tag{8.3.8}$$

and the GCV function can be evaluated for this λ as

$$\frac{\frac{1}{n}\,\mathrm{RSS}(\lambda)}{\frac{1}{n}\,\mathrm{Tr}\,(I - A^{(L)}(\lambda))^2},$$

and the process repeated for a new λ.

8.4 The optimal λ for loss functions other than predictive mean-square error.

The GCV estimate of $\hat{\lambda}$ has been shown to be good for estimating the λ that minimizes

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(L_i f - L_i f_\lambda)^2.$$

Suppose that one is really interested in choosing λ to minimize some other loss function, for example,

$$D(\lambda) = \int (f_\lambda(t) - f(t))^2\,dt.$$

If $T(\lambda)$ and $D(\lambda)$ have to a good approximation the same minimizer, then it is sufficient to use $\hat{\lambda}$. Examples where this is so numerically appear in Craven and Wahba (1979). Suggestions for modifying the GCV function for other loss functions have been made, notably in O'Sullivan (1986a). In the comments to that paper I argued that, if the original problem is ill-conditioned, then the computation of the modified GCV function is also likely to be ill-conditioned. In any case it would be nice to know when T and D (or other loss functions) are likely to have (approximately) the same minimizer. A number of authors have provided convergence rate calculations that contribute to an answer to this question, including those in Section 4.5.

In Wahba and Wang (1987) we examined a simple family of cases that suggests the range of results that might be obtainable in general. The result is that under a range of circumstances the optimal rate of decay of λ is the same for a variety of loss functions, and under other circumstances the optimal rate

of decay is different for different loss functions. We summarize the results here. We let

$$\begin{aligned} g(t) &= \int_0^1 h(t-s)f(s)\,ds,\ t \in [0,1], \\ y_i &= g\left(\frac{i}{n}\right) + \epsilon_i \end{aligned}$$

and we assumed that

$$\begin{aligned} g(t) &= \sum_{\nu=1}^{\infty} 2g_\nu \cos 2\pi\nu t, \\ h(t) &= \sum_{\nu=1}^{\infty} 2h_\nu \cos 2\pi\nu t, \\ f(t) &= \sum_{\nu=1}^{\infty} 2f_\nu \cos 2\pi\nu t, \end{aligned}$$

thus $g_\nu = h_\nu f_\nu$. f is estimated as a periodic function that integrates to zero and minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \int_0^1 h\left(\frac{i}{n} - s\right) f(s)\,ds\right)^2 + \frac{\lambda}{(2\pi)^{2m}}\int_0^1 (f^{(m)}(t))^2 dt.$$

Letting

$$\tilde{g}_\nu = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} y_i \cos 2\pi\nu\left(\frac{i}{n}\right),$$

then, to an approximation good enough for our purposes

$$f_\lambda(s) \approx 2\sum_{\nu=1}^{n} \hat{f}_\nu \cos 2\pi\nu s$$

where

$$\hat{f}_\nu = \frac{h_\nu \tilde{g}_\nu}{h_\nu^2 + \lambda\nu^{2m}};$$

furthermore,

$$g_\lambda(t) = \int_0^1 h(t-s)f_\lambda(s)\,ds \simeq 2\sum_{\nu=1}^{n} \hat{g}_\nu \cos 2\pi\nu t$$

with

$$\hat{g}_\nu = h_\nu \hat{f}_\nu.$$

Then the mean-square error in the solution is

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

the mean-square prediction error is

$$\frac{1}{n}\sum_{\nu=1}^{n}\left(g_\lambda\left(\frac{i}{n}\right)-g\left(\frac{i}{n}\right)\right)^2 \approx \int_0^1 (g_\lambda(t)-g(t))^2 dt \approx \sum_{\nu=1}^{n}(\hat{g}_\nu - g_\nu)^2$$

$$= \sum_{\nu=1}^{n} h_\nu^2(\hat{f}_\nu - f_\nu)^2,$$

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

$$\int_0^1 (f_\lambda^{(l)}(s) - f^{(l)}(s))^2 ds \approx \sum_{\nu=1}^{n}(2\pi\nu)^{2l}(\hat{f}_\nu - f_\nu)^2.$$

Now

$$\tilde{g}_\nu = \frac{\sqrt{2}}{n}\sum_{i=1}^{n}(g(\frac{i}{n}) + \epsilon_i)\cos 2\pi\nu(\frac{i}{n}) \approx g_\nu + \tilde{\epsilon}_\nu,$$

where $\tilde{\epsilon}_\nu \sim \mathcal{N}(0, \sigma^2/n)$ giving

$$E(\hat{f}_\nu - f_\nu)^2 \approx \left(\frac{\lambda\nu^{2m}}{h_\nu^2 + \lambda\nu^{2m}}\right)^2 f_\nu^2 + \frac{h_\nu^2\sigma^2}{(h_\nu^2 + \lambda\nu^{2m})^2}. \tag{8.4.1}$$

Wahba and Wang (1987) considered loss functions of the form

$$T_q(\lambda) = \sum_{\nu=1}^{n} q_\nu(\hat{f}_\nu - f_\nu)^2 \tag{8.4.2}$$

for

$$q_\nu \approx \nu^\gamma,\ f_\nu \approx \nu^{-\alpha},\ h_\nu \approx \nu^{-\beta},\ \alpha, \beta > 0. \tag{8.4.3}$$

Thus $\gamma = 0$ corresponds to mean-square solution error, $\gamma = -2\beta$ corresponds to mean-square prediction error, and $\gamma = 2m$ corresponds to $\|f - f_\lambda\|_R^2$. Substituting (8.4.3) and (8.4.1) into (8.4.2), one obtains

$$ET_q(\lambda) \approx \lambda^2 \sum_{\nu=1}^{n} \frac{\nu^{4(m+\beta)+\gamma-2\alpha}}{(1+\lambda\nu^{2(m+\beta)})^2} + \frac{\sigma^2}{n}\sum_{\nu=1}^{n}\frac{\nu^{\gamma+2\beta}}{(1+\lambda\nu^{2(m+\beta)})^2}, \tag{8.4.4}$$

and the optimal λ for the interesting combinations of α, β, γ, and m were found. We only consider the case $\beta > 0$ (which guarantees a bona fide convolution equation), $\alpha + \beta > 1$ (which guarantees that g is in a reproducing kernel space), and $m > \frac{1}{4}$. We only repeat the results for $\gamma = 0$ and -2β here. Let λ_D minimize (8.4.4) for $\gamma = 0$ (mean-square solution error (domain error)) and λ_* be the minimizer for $\gamma = -2\beta$ (mean-square prediction error). The results are:

(A) Suppose $\frac{1}{2} < \alpha \leq 2m + 1$. Then

$$\lambda_* \approx \lambda_D \approx n^{-(m+\beta)/(\alpha+\beta)}$$

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

$$\lambda_* \approx \lambda_D \approx n^{-(m+\beta)/(\alpha+\beta)}$$

(B.2) Suppose $2m + 1 < \alpha$ and $\beta \leq (\alpha - (2m + \frac{1}{2}))$. Then λ_* does not (otherwise) depend on α and

$$\lambda_* \approx o(\lambda_D).$$