On Solving Fredholm Integral Equations of the First Kind

S. C. SAHASRABUDHE AND A. D. KULKARNI

Indian Institute of Technology, Bombay, India

Abstract: A method for numerical solution of Fredholm integral equations of the first kind is derived and illustrated. The solution \( f(x) \) of the integral equation is assumed to be a sample function of a wide-sense stationary random process with known autocorrelation function. From the set of permissible solutions, the solution that "best" satisfies the statistical properties of the random process is admitted as the correct solution. With a kernel matrix \( A \), the search for this solution is carried out by introducing the orthogonal frame of reference of the symmetrized matrix \( A^T A \) and then suitably adjusting the components along the principal axes with small eigenvalues of \( A^T A \) (i.e., small singular values of \( A \)). The method is illustrated for an example first considered by Phillips and also for another problem from the area of image processing.

Key words and phrases: Fredholm integral equations of the first kind, numerical solution, random function, ill-conditioning, singular values, autocorrelation function

1. Introduction

In many physical problems, such as the measurement of spectral distribution, cosmic radiation, and the angular variation of scattered light, indirect sensing devices are used [3, 5, 6]. Here the relation between the quantity observed and the quantity to be measured is given by a Fredholm integral equation of the first kind:

\[
\int_a^b K(y, x)f(x)dx = g(y), \quad a \leq y \leq b. \tag{1}
\]

The degradation of an image caused by the point spread function of the optical system is also an example represented by the same equation.

In all such cases \( f(x) \) can be obtained by solving (1) when \( g(y) \) is the observation at various \( y \) values and the kernel \( K(y, x) \) is known. The problem of solving (1) is in many cases difficult because the solution may be extremely sensitive to the presence of noise in \( g(y) \), such as measurement errors or roundoff errors. With the presence of noise, (1) should be rewritten as

\[
\int_a^b K(y, x)f(x)dx = g(y) + \epsilon(y), \quad a \leq y \leq b. \tag{2}
\]

Here the error function \( \epsilon(y) \) is arbitrary except for some restriction on its magnitude. The family of allowed functions \( \epsilon(y) \) defines a family \( F \) of allowed functions \( f(x) \), but most of these solutions will not be physically meaningful.

Phillips [8] assumed the desired solution to be reasonably smooth and tried to find the function \( f(x) \in F \) that is smoothest in some sense. He chose as the smoothness criterion that the second derivative be minimum. Twomey [11, 12] simplified and generalized the method of Phillips; his method is simple because it involves only one matrix inversion rather than two. Twomey also showed how to generalize the method.
in order to use a priori constraints other than Phillips's smoothness constraint. However, in both cases no systematic method of determining the required amount of smoothing is given. Strand and Westwater [10] and Helstrom [41] have given similar procedures for solving (2). They assume \( f(x) \) and \( \varepsilon(y) \) to be independent stochastic Gaussian processes with zero mean and known covariances. Vemuri and Fang-Pai [13] have given an initial value method which gives a family of solutions, and to select one they have to use some smoothing constraint. In their method also, the amount of smoothing required is not determined. In the method described below the constrained function \( f(x) \) is assumed to be a sample function of a wide-sense stationary process with known autocorrelation function, which can be the case in many practical problems. The method automatically incorporates the amount of smoothing required. We also show how the constraints suggested by Phillips and Twomey can be interpreted in terms of the autocorrelation function.

2. Method

Let \( f(x) \) be a sample function of a wide-sense stationary process with known autocorrelation function in the interval \( a \leq x \leq b \) and let \( K(y, x) \) be a continuous or a piecewise continuous function of \( x \) and \( y \). To solve (2) numerically, it is necessary to make the variables discrete and replace the integral equation by a set of finite linear equations.

Continuous variables \( x \) and \( y \) can be replaced by sets of finite mesh points

\[
a \leq y_1 < y_2 < \cdots < y_{n_1-1} < y_{n_1} \leq b,
\]

\[
a \leq x_1 < x_2 < \cdots < x_{n_2-1} < x_{n_2} \leq b,
\]

where \( n_1 \geq n_2 \). The integral equation can then be replaced by a set of equations

\[
g(y_i) + \varepsilon(y_i) = \int_a^b K(y_i, x)f(x)dx = \sum_{j=1}^{n_2} w_j K(y_i, x_j)f(x_j),
\]

where \( i = 1, 2, \ldots, n_1 \) and \( w_1, w_2, \ldots, w_{n_2} \) are the weighting coefficients for the quadrature formula used. Vectors \( f, g, \) and \( \varepsilon \) and matrix \( A \) can be defined as

\[
f = f(x_i), \quad \varepsilon = \varepsilon(y_i), \quad g = g(y_i), \quad \text{and} \quad A = w_j K(y_i, x_j),
\]

where \( i = 1, 2, \ldots, n_1 \) and \( j = 1, 2, \ldots, n_2 \). With the above notation, (3) can be rewritten as

\[
Af = g + \varepsilon.
\]

Let \( H \) be the g-inverse of matrix \( A \); then by introduction of the orthogonal frame of reference of the symmetrized matrix \( A^T A \), \( H \) can be decomposed into \( U[D, 0]V \) [9], where \( U \) and \( V \) are square orthogonal matrices, the columns of \( U \) are eigenvectors of \( A^T A \), and the rows of \( V \) are eigenvectors of \( A A^T \). \([D, 0]\) represents an \( n_1 \times n_1 \) matrix and \( D = \text{diag}[1/\omega_1, 1/\omega_2, \ldots, 1/\omega_{n_2}] \), where \( \omega_1, \omega_2, \ldots, \omega_{n_2} \) are singular values of \( A \). Hence the “true” solution of (2) can be written as

\[
f_i = \sum_{j=1}^{n_2} U_{ij} \beta_j / \omega_j, \quad i = 1, 2, \ldots, n_2,
\]

where \( \beta = Vg \).

Here the components of vector \( \beta \) are not known to any desired degree of accuracy because of the presence of errors \( \varepsilon_i \) in measurement of \( g_i \). Even small inaccuracies in the components of \( \beta \) corresponding to small singular values cause considerable variations in the values of \( f \). However, \( f \) is relatively insensitive to the errors in the components of \( \beta \) corresponding to large singular values. Therefore these components may be presumed to be known on the basis of measurements of \( g_i \). In eq. (6) they are written as known constants \( C_1, C_2, \ldots, C_{n_2} \), but the components of \( \beta \) corresponding to small singular values are considered to be unknowns. They are to be determined on the basis of the a priori knowledge about \( f \) that we have in the known autocorrelation function of \( f \). Let the number of unknown \( \beta \) components be \( m \); then (5) can be rewritten as
The criterion used for "smallness" of a singular value is as follows. In solving the system of (4), if ε is the error in g, then η, the error in the solution f, is given by [14]

\[ \frac{\|\eta\|}{\|f\|} \leq \left( \frac{\omega_{\text{max}}}{\omega_{\text{min}}} \right) \left( \frac{\|\varepsilon\|}{\|g\|} \right), \]

where \( \omega_{\text{max}} \) and \( \omega_{\text{min}} \) are the largest and the smallest singular values of \( A \), respectively. If we choose \( \|\eta\| = \|\varepsilon\| \) and \( \omega_{\text{min}} \leq \epsilon \), then (7) may be rewritten as

\[ \frac{\|\eta\|}{\|f\|} \leq C(\epsilon /\|g\|), \]

where the condition number \( C \) is the ratio of \( \omega_{\text{max}} \) to \( \omega_{\text{min}} \). In (8) the ratio \( \|\eta\| /\|f\| \) may now be interpreted as the usual noise/signal ratio. In a physical situation, the upper bound to the noise/signal ratio is usually known. This gives us the permissible upper limit to the condition number \( C \). Equation (8) gives the upper bound of \( \|\eta\| \) when the error vector \( \varepsilon \) is oriented in favor of the principal axis of \( A^T A \) corresponding to the smallest singular value. If the error vector is oriented along the principal axis corresponding to any other \( \omega_i \), then the corresponding condition number would be \( \omega_{\text{max}} /\omega_i \). The criterion for the smallness of a singular value is that all \( \omega_i \) for which the condition number \( \omega_{\text{max}} /\omega_i \) exceeds the permissible limit should be considered as small and the corresponding terms in expansion (6) should be treated as unknown.

However, the maximum number of components of \( \beta \) that may be treated as unknowns is limited. The vector \( f \) is a finite sample of a stochastic process. For a good point estimate of the autocorrelation function, the number of lag points for which it is estimated should not, as a rule of thumb, exceed \( n_2/5 \), where \( n_2 \) is the total number of sample points. Preferably it should be limited to \( n_2/10 \) [1]. If the number of small singular values is less than \( n_2/5 \), then the system of (6) can be an overdetermined system.

On the other hand, if the number of small singular values is more than \( n_2/5 \), then the number of unknowns in (6) has to be restricted to terms corresponding to the smallest \( n_2/5 \) singular values. Since the error bound on \( f \) given by (7) is rather pessimistic [14], even in such a case the possibility of successful solution is quite good.

If the points \( x_i \) in (4) are chosen to be equally spaced, then the estimated autocorrelation function can be written as

\[ \hat{R}(i) = \frac{f_1 f_{i+1} + f_2 f_{i+2} + \cdots + f_{n_2-i} f_{n_2}}{f_1^2 + f_2^2 + \cdots + f_{n_2}^2}, \quad i = 1, 2, \ldots, k, \]

where \( m \leq k \leq n_2/5 \). By equating the estimated autocorrelation function \( \hat{R}(i) \) with the known autocorrelation function \( R(i) \), we get

\[ \hat{R}(i) - R(i) = 0, \quad i = 1, 2, \ldots, k. \]

Equation (10) can be rewritten with the help of (6) and (9) for each value of \( i \), yielding

\[ F_i(\beta_{n_2-m+1}, \ldots, \beta_{n_2-1}, \beta_{n_2}) = 0, \quad i = 1, 2, \ldots, k, \]

where \( F_1, F_2, \ldots, F_k \) represent quadratic equations in unknown \( \beta \) values which are then determined as solutions of (11). Substituting these \( \beta \) values into (6) gives the required \( f \).

Consider the function

\[ S(\beta_{n_2-m+1}, \ldots, \beta_{n_2-1}, \beta_{n_2}) = \sum_{i=1}^{k} [F_i(\beta_{n_2-m+1}, \ldots, \beta_{n_2-1}, \beta_{n_2})]^2. \]

It is obvious that each solution of system (11) reduces the function \( S \) to zero; contrariwise the \( \beta \) values for which the function \( S \) is zero are the roots of the system (11). Hence the solution of system (11) can be obtained by finding the minimum of the function \( S \) in \( m \)-dimensional space \( E_m = (\beta_{n_2-m+1}, \ldots, \beta_{n_2-1}, \beta_{n_2}) \). The system of (11)
could also be an overdetermined system. In the illustrative examples presented here, for minimizing $S$ a modified sequential simplex method has been employed [2].

It can easily be shown that Phillips's constraint for choosing an $f$ that satisfies the condition

$$\int_a^b [f''(x)]^p \, dx = \min_{f \in F} \int_a^b [f''(x)]^p \, dx$$

is essentially choosing an $f$ which satisfies the condition $\min_{f \in F} [3R(0) - 4R(1) + R(2)]$. The constraints of minimizing higher derivatives can also be expressed in an analogous manner.

3. Numerical Results

Example 1 (Phillips's problem).

$$g(y) = \int_{-6}^{6} K(x - y)f(x) \, dx,$$

$$K(z) = \begin{cases} 1 + \cos(\pi z/3) & \text{for } |z| \leq 3, \\ 0 & \text{for } |z| > 3, \end{cases}$$

$$g(y) = (6 + y)(1 - (1/2)\cos(\pi y/3)) - (9/2\pi)\sin(\pi y/3) \quad \text{for } |y| \leq 6,$$

$$= 0 \quad \text{for } |y| > 6.$$

The variables $x$ and $y$ were made discrete by using 25 equally spaced points. The values of $g_i$ were truncated at the third decimal point, yielding $|\epsilon_i| \leq 0.01$, which gives $e^2 = 25 \times 10^{-4}$. It was desired to keep the noise/signal ratio less than ten times the noise/signal ratio in the measurement $g_i$. This forces the condition number in (8) to be equal to 10. The number of small singular values, according to this criterion, is 18. This is far in excess of the maximum number of permissible unknowns, which is $n_2/5 = 5$. Hence the number of unknowns in (6) was chosen to be equal to 5. Since in this example $f$ is not a sample function of a stochastic process, its autocorrelation function was obtained from the known analytical solution. Estimated $f$ and the true solution and the solution obtained by simple inversion $f = A^{-1}(g + \epsilon)$ are shown in Figure 1. The closeness of the estimated solution to the true solution is indicative of the pessimistic error bound in (7).

Example 2.

$$g(y) = \int_{-1/3}^{1/2} K(y - x)f(x) \, dx,$$

where $K(z) = 1/C_1\cos(\sin^{-1}(z/C_2))$ and $C_1$, $C_2$ are constants.

The kernel $K(y - x)$ chosen in this example is a point spread function arising in certain image processing problems where the degradation of the image is due to vibratory motion of the imaging device. The variables $x$ and $y$ were made discrete by using 64 equally spaced points. The error in $g_i$ was such that $|\epsilon_i| \leq 0.25$, yielding $e = 4$. The noise/signal ratio criterion used was the same as in Example 1; as a consequence the number of unknowns in (6) is 8. This is less than the maximum permissible number of unknowns, which is 12 in the present case. Hence the system of (11) is overdetermined (12 equations). The sequence $f$ for this example was generated as an output of a difference equation driven by white noise [7], and hence the autocorrelation function of $f$ is a known function in terms of the constants of the difference equation. The estimated solution and the true solution and the solution due to numerical inversion $f = A^{-1}(g + \epsilon)$ are shown in Figure 2.

4. Summary

A method for solving Fredholm integral equations of the first kind in the presence of
measurement errors in \( g(y) \) has been derived and illustrated. To solve numerically, the integral equation is first made discrete, yielding a system of linear equations of the type \( Af = g + \epsilon \). The vector \( f \) is then expressed in terms of singular values of \( A \). The condition number \( C \) for the desired noise/signal ratio of the solution \( f \) is then evaluated. The
singular values for which the ratio $\omega_{\text{max}}/\omega_i$ exceeds the condition number $C$ are considered as small singular values, and the corresponding components of $g + \epsilon$ are treated as unknowns. These are then determined by minimizing $S = \sum_{i=1}^{k} [R(i) - R'(i)]^2$. The solution is computed subsequently from these obtained components.

REFERENCES