数理解析研究所講究録
703巻 1989年 187-202

A Practical Method to Examine the Stability of a Class of Numerical Procedures

愛媛大・理　北川 高嗣 (Takashi Kitagawa)

§1. Introduction

In this paper, we developed a theory to examine the stability of a class of numerical procedures. We analyze the numerical stability and the effectiveness of the method of regularization. We give several theorems to illustrate error propagation for the class of the numerical procedures which include solving an ill-posed problem as the first step. We also study how the situation is changed by applying the method of regularization. We also mention the selection of the regularization parameter.

§2. Fundamental solution method

To illustrate the class of the numerical procedures, we consider the fundamental solution method to approximate the solution of the Dirichlet problem of Laplace equation of the form

$$(2.1) \qquad \Delta \mathbf{u} = 0 \quad \text{in} \quad \Omega$$

$$(2.2) \qquad \mathbf{u} = g \quad \text{on} \quad \partial\Omega ,$$

where

$$\Omega = \{ \omega \in \mathbf{R}^2 \mid \|\omega\|_2 < \rho \}.$$

The fundamental solution method approximates the solution $\mathbf{u}(x)$ by

$$u_n(x) = \sum_{k=1}^{n} c_k \; G(x, y_k) \quad , \qquad x \in \Omega \tag{2.3}$$

where $G(x,y)$ is the Green's function for (Δ, Ω),

$$G(x,y) = - \frac{1}{2} \log \|x - y\|_2 \quad , \; x, y \in \mathbf{R}^2.$$

Points y_k's, called charge points, are chosen appropriately and c_k's are constants to be determined. The vector $c = (c_1, c_2, \ldots, c_n)^t \in \mathbf{R}^n$ is called charge and determined in such a way that $u_n(x)$ satisfies the boundary condition

$$u_n(\hat{x}_j) = g(\hat{x}_j) \qquad j = 1, 2, \ldots, n, \tag{2.4}$$

where $\hat{x}_j$'s are properly chosen n collocation points on the boundary. Let the charge points $y_1, y_2, \ldots y_n$ be on the auxiliary boundary which is the outer circle with radius R (with "outer" we imply $R > \rho$).

With the collocation points $\hat{x}_k = \rho\, e^{\frac{2\pi}{n}(k-1)i}$ and the charge points $y_k = R\, e^{\frac{2\pi}{n}(k-1)i}$, $k=1,2,\ldots,n$, the following results stating that the approximate solution $\mathbf{u}_n$ converges to the solution $\mathbf{u}$ exponentially with respect to n are known .

Theorem 2.1.(Katsurada[10]) a) Suppose that the harmonic extension of u exists in

$$\Omega_{r_0} = \{\omega \mid \|\omega\|_2 < r_0\} \text{ with } \rho < r_0 ,$$

then we have, for sufficiently large n,

$$(2.5) \qquad \| \mathbf{u} - \mathbf{u}_n \|_\infty \leq$$

$$\sup_{\|x\|_2 = r_0} |\mathbf{u}(x)| \frac{2}{1 - \rho/r_0} \{(1+C(R,\rho))\ (\rho/r_0)^{n/3} +$$

$$4(\rho/R)^{n/3}\},$$

where $C(R,\rho)$ is a constant depends on R and ρ of the form

$$C(R,\rho) = \max \{ 1, \log(R^n + \rho^n)/|\log(R^n - \rho^n)| \}.$$

We call n "sufficient large" if $(\rho/R)^n \leq 1/2$, $n \log R \geq 4(\rho/R)^n$ and $(\rho/R)^{2n/3} \leq n \log R$.

b) The condition number of the coefficients matrix of the equation (2.4) which determines the charge c grows exponentially with respect to n. Approximately the condition number Cond(n,R) can be estimated by

$$(2.6) \qquad \mathrm{Cond}(n,R) \sim \frac{\log R}{2}\, n \left(\frac{R}{\rho}\right)^{n/2} .$$

Since the estimate (2.5-6) follows from the fact that the coefficient matrix for the particular location of $\hat{x}_k$ and y_k is circulant, Theorem 2.1 is only valid for the circular domain and cannot be applied to more complicated regions. The result (b) is also obtained by Christiansen [1]. For the properties of circulant matrices, see e.g. Davis[2]. Numerical stability of this method is studied in Kitagawa[12].

§3. Stability of the numerical procedures

3.1 Formulation and Basic Results

The method of Section 3 reduces to the numerical process of the following two steps:

1) We first solve an ill-conditioned linear system to determine the charge c in the form of

$$(3.1) \qquad \Gamma c = g$$

for given data g which may be contaminated by some perturbation Δg, where $g \in \mathbf{Y} = \mathbf{R}^m$, $c \in \mathbf{X} = \mathbf{R}^n$ and $\Gamma : \mathbf{X} \to \mathbf{Y}$ with

$$(\Gamma\, c)_j \equiv \sum_{k=1}^{n} c_k\, \mathbf{G}(w_j, y_k) = g(w_j),\ w_j \in C(\rho),\ j = 1,2,\ldots,n,$$

where $C(\rho)$ denotes the disc with radius ρ.

2) We use the intermediate solution c to obtain the final result f by

$$(3.2) \qquad f = \boldsymbol{\Lambda}\, c\ ,$$

where $f \in \mathbf{X}$ and $\boldsymbol{\Lambda} : \mathbf{X} \to \mathbf{Y}$ with

$$f_j \equiv \bar{h}_n(x_j) = \sum_{k=1}^{n} c_k\, \mathbf{G}(x_j, y_k) \equiv (\boldsymbol{\Lambda}\, c)_j,\quad x_j \in C(\gamma),\ j = 1,2,\ldots,n.$$

Due to the ill-conditioning of (3.1), some 'large' perturbation Δc may be introduced to the intermediate solution c. One may assume intuitively that the error $\|\Delta f\|$ in the final result f, where $\Delta f \equiv \boldsymbol{\Lambda}\, \Delta c$, is on a level with $\|\Delta c\|$ or as large as $\|\boldsymbol{\Lambda}\|\ \|\Delta c\|$. If this is the case, the method of regularization (Groetsch[7] and Tikhonov et al.[16]), applied to (3.1) may be very effective. But this is not always true. Even if the error $\|\Delta c\|$ is very large, $\|\Delta f\|$ can be very small. In this case, we do not necessarily have to use the method, or in some cases, we may have worse result by using the method. To examine whether the method of regularization is effective or not for this class of numerical procedures, we have the following results.

We assume that given data $\bar{g} = g + \Delta g$ and the intermediate solution $\bar{c} = c + \Delta c$. We have $\Gamma \bar{c} = \bar{g}$ as well as (3.1). Let $\sigma_1 \geq \sigma_2 \geq k1.. \geq \sigma_n \geq 0$ be singular values of Γ and $\{u_i\}$ i=1,2,...,m, $\{v_j\}$ j=1,2,...n be singular vectors of Γ. Reflecting the ill-conditioning of Γ, we assume that $\sigma_n \to 0$ as $n \to \infty$. We can see that $\|\Delta c\|_2 \to \infty$ as $\sigma_n \to 0$ from

$$(3.3) \qquad \Delta c = \sum_{i=1}^{n} \frac{1}{\sigma_i} (\Delta g, u_i)\, v_i .$$

Next, we suppose that the final result $\bar{f}$ is given by $\bar{f} = f + \Delta f$. We have $\bar{f} = \Lambda \bar{c}$ as well as (5.2). Let $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_n \geq 0$ be singular values of Λ and $\{\hat{u}_i\}$ i=1,2,...,m, $\{\hat{v}_j\}$ j=1,2,...n be singular vectors of Λ. As for $\|\Delta f\|$, we have the following result from Theorem 4.1 in Kitagawa[12].

Theorem 3.1. Let $\{u_i, v_i; \sigma_i\}$ and $\{\hat{u}_i, \hat{v}_i; \hat{\sigma}_i\}$ for i= 1,2,...,n be the orthonormal singular systems for Γ and Λ respectively. Then we have

$$(3.4) \qquad \|\Delta f\|_2 \leq \|\Xi * \Theta\|_F \, \|\Delta g\|_2$$

where

(3.5a) $\Xi = (\xi_{ij})$, $\xi_{ij} = \hat{\sigma}_i / \sigma_j$

(3.5b) $\Theta = (\theta_{ij})$, $\theta_{ij} = (\hat{v}_i, v_j)$, $i,j=1,2,\ldots,n$,

$\Xi * \Theta$ represents the Hadamard product of the matrices Ξ and Θ and $\|\cdot\|_F$ denotes the Frobenius norm.

Making use of singular value decomposition, we can construct the matrices Θ and Ξ numerically. Then we can examine the numerical stability of the procedures which consist of two steps denoted in the form of (3.1) and (3.2).

3.2. On the matrices Ξ and Θ

The elements ξ_{ij} of Ξ, which we called explosive factor matrix in [12-13], represent the upper bound of the magnification of the u_i-component $(\Delta g, u_i)$ of perturbation Δg to $\hat{u}_j$- component $(\Delta f, \hat{u}_j)$ of Δf. To clarify what "magnification of $(\Delta g, u_i)$ to $(\Delta f, \hat{u}_j)$" means, we define it precisely.

Definition 3.2. We call the partial derivative

$$\frac{\partial\,(\Delta f, \hat{u}_j)}{\partial\,(\Delta g, u_i)}$$

the magnification of $(\Delta g, u_i)$ to $(\Delta f, \hat{u}_j)$.

The following fundamental theorem about the magnification is the basis of the further discussion.

Theorem 3.3. Let $\{u_i, v_i; \sigma_i\}$ and $\{\hat{u}_i, \hat{v}_i; \hat{\sigma}_i\}$ for i= 1,2,...,n be the orthonormal singular systems for Γ and Λ respectively. Then the magnification of $(\Delta g, u_i)$ to $(\Delta f, \hat{u}_j)$ is given by $\xi_{ij}\ \theta_{ij}$, where ξ_{ij} and θ_{ij} represent the ij-th elements of the matrices Ξ and Θ respectively defined in (3.5a-b).

Proof. Note that the singular vectors $\{u_i\}$ and $\{\hat{u}_i\}$ form orthonormal bases of **X**, and $\{v_i\}$ and $\{\hat{v}_i\}$ form those of **Y**. First, we have the following relations of the Fourier coefficients of $\Delta f \in \mathbf{Y}$, $\Delta c \in \mathbf{X}$ and $\Delta g \in \mathbf{Y}$ from the relations of $\Gamma\ \Delta c = \Delta g$ and $\Lambda\ \Delta c = \Delta f$:

$$(3.6) \qquad (\Delta c,\ v_i) = \frac{1}{\sigma_i} (\Delta g,\ u_i) \quad \text{for } i = 1,2,\ldots,n$$

and

$$(3.7) \qquad \hat{\sigma}_j\ (\Delta c,\ \hat{v}_j) = (\Delta f,\ \hat{u}_j) \quad \text{for } j = 1,2,\ldots,n.$$

By expanding Δc by $\{v_i\}$ i=1,2,...,n, we have

$$\Delta c = \sum_{i=1}^{n} (\Delta c,\ v_i)\ v_i .$$

Putting this into (3.7), we have

$$(3.8) \qquad (\Delta f, \hat{u}_j) = \hat{\sigma}_j \left(\sum_{i=1}^{n} (\Delta c, v_i)\, v_i, \hat{v}_j \right).$$

Substituting (3.6) into (3.8), we obtain

$$(\Delta f, \hat{u}_j) = \hat{\sigma}_j \left(\sum_{i=1}^{n} \frac{1}{\sigma_i} (\Delta g, u_i)\, v_i, \hat{v}_j \right).$$

Thus we have

$$\frac{\partial\, (\Delta f, \hat{u}_j)}{\partial\, (\Delta g, u_i)} = \hat{\sigma}_j \left(\frac{1}{\sigma_i} v_i, \hat{v}_j \right) \text{ for } i,j = 1,2,\ldots,n,$$

or

$$= \xi_{ij}\, \theta_{ij}$$

from the definition of Ξ and Θ. **Q.E.D.**

For instance, the largest element ξ_{1n} gives the upper bound of $\hat{\sigma}_1 / \sigma_n$ which coincides with the straight forward upper bound with the spectral norm $\|\cdot\|_s$ of matrix given by $\|\Delta f\|_2 \leq \|\Gamma^{-1}\|_s \|\Lambda\|_s \|\Delta g\|_2$, since $\|\Gamma^{-1}\|_s \|\Lambda\|_s = \hat{\sigma}_1 / \sigma_n$. On the other hand, the elements θ_{ij} of Θ, which we call distortion coefficients matrix, represents the actual ratio of propagation of $(\Delta g, u_i)$ to $(\Delta f, \hat{u}_j)$. The actual magnification of propagation of $(\Delta g, u_i)$ to $(\Delta f, \hat{u}_j)$ is given by $\xi_{ij} \times \theta_{ij}$ (Theorem 3.3) and the upper bound of the total propagation of Δg to Δf is given by the square root of the sum of squares of $\xi_{ij} \times \theta_{ij}$, or $\|\Xi * \Theta\|_F$ (Theorem 3.1).

§4. Effectiveness of the Method of Regularization

4.1. Basic Result

The method of regularization applied to the equation (3.1) with perturbation Δg can be written as

$$(4.1) \qquad (\Gamma^t\Gamma + \mu I)\,\bar{c} = \Gamma^t(g + \Delta g) .$$

We write the solution of (4.1) $c(\mu,\Delta g)$. To examine the effectiveness of the method of regularization, we have the next result from Theorem 3.1 in Kitagawa[13]. We use the notations of $f(\mu,\Delta g) = \Lambda\, c(\mu,\Delta g)$ and $\Delta f(\mu,\Delta g) \equiv f(\mu,\Delta g) - f(0,0)$ in the theorem.

Theorem 4.1. Let $\{u_i, v_i; \sigma_i\}$ and $\{\hat{u}_i, \hat{v}_i; \hat{\sigma}_i\}$ for $i = 1,2,\dots,n$ be the orthonormal singular systems for Γ and Λ respectively. Then we have

$$(4.2) \qquad \|\Delta f(\mu,\Delta g)\|_2 \le \|\Xi_\zeta * \Theta\|_F\, \|g\|_2 + \|\Xi_\rho * \Theta\|_F\, \|\Delta g\|_2$$

where

$$(4.3a) \qquad \Xi_\zeta = (\xi^\zeta_{ij}) , \quad \xi^\zeta_{ij} = \hat{\sigma}_i\, \mu / (\sigma_j(\sigma_j^2 + \mu))$$

$$(4.3b) \qquad \Xi_\rho = (\xi^\rho_{ij}) , \quad \xi^\rho_{ij} = \hat{\sigma}_i\, \sigma_j / (\sigma_j^2 + \mu)$$

and the rest of the symbols are the same as Theorem 5.1.

Based on the Theorems 3.1 and 4.1, we can examine the effectiveness of the method of regularization very clearly. Letting

$$\zeta(\mu) = f(\mu,0) - f(0,0) \tag{4.4}$$

and

$$\rho(\mu,\Delta g) = f(\mu,\Delta g) - f(\mu,0), \tag{4.5}$$

we have

$$\|\Delta f(\mu,\Delta g)\|_2 \leq \|\rho(\mu,\Delta g)\|_2 + \|\zeta(\mu)\|_2. \tag{4.6}$$

$\rho(\mu,\Delta g)$ defined by (4.5) represents the error due to Δg to the solution f with regularization.

From the proof of Theorem 3.1 of [13], we have

$$\|\rho(\mu,\Delta g)\|_2 \leq \|\Xi_\rho * \Theta\|_F \, \|\Delta g\|_2 \tag{4.7}$$

and

$$\|\zeta(\mu)\|_2 \leq \|\Xi_\zeta * \Theta\|_F \, \|g\|_2. \tag{4.8}$$

From Theorem 5.1, we also have

$$\|\Delta f\|_2 \leq \|\Xi * \Theta\|_F \, \|\Delta g\|_2. \tag{4.9}$$

If we compare the error due to Δg of (4.7) with that of f without regularization of (4.9), we can recognize when the regularization is effective. Checking corresponding elements of Ξ_ρ, Ξ and Θ, we can examine the effectiveness of the regularization. The inequality (4.8) suggests that we should avoid using the method of regularization when it is not effective and we should choose the regularization parameter μ carefully.

We can actually construct the matrices Ξ_ς ,Ξ_ρ and Θ and we examine how the method of regularization stabilizes the numerical process and how we should choose the regularization parameter.

4.2. Matrices Ξ_ρ and Ξ_ς

We first note that since the elements θ_{ij} of the matrix Θ is independent of the regularization parameter μ, the distortion coefficients matrix Θ is common with that without regularization. We also note that the matrices Ξ_ς and Ξ_ρ as well as Ξ and Θ do not depend on g or Δg at all and, accordingly, we do not have to construct these matrices for different functions of g.

First we examine the elements ξ^ρ_{ij} of matrix Ξ_ρ due to perturbation Δg to study how the method of regularization stabilizes the numerical process 1) and 2) of Section 3.1. The elements are again given by rounding off the fractions of logarithm with basis 10. The matrix Ξ_ρ represents the explosive factor matrix with regularization. The elements ξ^ρ_{ij} of critical part of lower right corner ($i \approx n$ and $j \approx n$) are significantly smaller than those of Ξ without regularization . This can be understood very easily if we compare the elements ξ^{ρ}_{ij} and ξ_{ij} of Ξ and Ξ_{ρ}. As we have seen in Section 3.2 the elements ξ_{ij} grow large for large i and j mainly because the denominator σ_j approaches to zero as $j \to n$.

On the contrary, the denominator $(\sigma_j^2 + \mu)$ of the elements ξ^{ρ}_{ij} do not approach to zero even if $j \to n$ and σ_j approaches to zero as far as the regularization parameter $\mu > 0$. Since the numerator of the elements ξ^{ρ}_{ij} are independent from μ, the elements ξ^{ρ}_{ij} for large j's do not grow large as in the case of ξ_{ij} of Ξ without regularization. Accordingly, the corresponding elements $\xi^{\rho}_{ij} * \theta_{ij}$ in lower right corner of matrix $\Xi_{\rho} * \Theta$ are much smaller than those of the matrix $\Xi * \Theta$.

Moreover the Frobenius norm of $\Xi_{\rho} * \Theta$ may much smaller than that of $\Xi * \Theta$. This explains that the method of regularization significantly reduces the magnification of the propagation of the perturbation Δg to the final approximation f.

Another factor of error $\varsigma(\mu)$ which is defined by (4.4), however, shall be inevitably introduced when we employ the method of regularization. Though the upper bound of the error $\varsigma(\mu)$ is given in (4.9), its interpretation is somewhat more delicate than the case of Ξ_{ρ} . The element $\xi^{\varsigma}_{ij} \times \theta_{ij}$ of the matrix $\Xi_{\varsigma} * \Theta$ involved in (4.9) represents the magnification of the propagation of

(g,u_i) to $(\Delta f,\hat{u}_j)$ due to introduction of the regularization parameter μ. The size of $(\Delta g,u_i)$ may not differ much among different i's, but the fourier coefficients (g,u_i) of g may be quite different in size. This is because the function g is harmonic and very smooth, which may results in very rapid convergence of the coefficients (g,u_i) to zero.

The matrices Ξ_ζ ,Ξ_ρ and Θ give us an idea on the choice of the regularization parameter. We should choose μ in such a way that

i) we reduce the size of element ξ^{ρ}_{ij} of Ξ_ρ whose corresponding elements of θ_{ij} of Θ are close to unity

ii)we avoid contaminating the elements ξ^{ζ}_{ij} of Ξ_ζ whose corresponding elements of θ_{ij} of Θ are close to unity and the corresponding j-th Fourier coefficients (g,u_j) are significant.

A detailed and illustrative examples for above discussions are given in [14] which studies two contrastive cases.

References

[1] Christiansen S., Condition number of matrices derived from two classes of integral equations, Mat. Meth. Appl. Sci., **3**(1981), 364-392

[2] Davis P.J. , Circulant Matrices, John Wiley & Sons, 1979.

[3] Fairweather G.F.and Johnston R.L., The method of fundamental solutions for problems in potential theory, in Treatment of Integral Equations by Numerical Methods.(eds. C.T.H.Baker and G.F.Miller), Academic Press,1982.

[4] Franklin J.N., Well-posed stochastic extensions of ill-posed linear problems, J. Math. Anal. Appl., 31(1970), 682-716

[5] Golub G.H., Singular value decomposition and least squares solutions, Numer. Math. 14(1970), 206-216

[6] Golub G.H. and Van Loan C.F., Matrix Computations. Johns Hopkins University Press, 1983.

[7] Groetsch C.W., The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind, Pitman 1984.

[8] Henrici P., Applied and Computational Complex Analysis, Vol.1, John Wiley & Sons, New York, 1974.

[9] Hsiao M. and McCamy R.C., Solution for boundary value problems by integral equations of the first kind, SIAM Rev., 15(1973), 687-705

[10] Katsurada M., An Analysis on Charge Simulation Method.(In Japanese) Master thesis, Department of Mathematics, Tokyo University, 1986.

[11] Kitagawa T., A deterministic approach to the optimal regularization, Japan J. Appl. Math., 4(1987), 371-391

[12] Kitagawa T., On the numerical stability of the method of fundamental solution applied to Dirichlet problem, Japan J. Appl. Math., 5(1988), 123-133

[13] Kitagawa T., On the effectiveness of the method of regularization for a certain class of numerical procedures, Japan J. Appl. Math.5(1988),305-311

[14] Kitagawa T., A practical method to examine the numerical stability of a class of numerical procedures, - In the case of numerical harmonic continuation -, Submidtted to publication

[15] Lavrent'ev M.M. et al., Ill-Posed Problems of Mathematical Physics and Analysis, American Mathematical Society, 1986.

[16] Murashima S., Charge Simulation Method and its application.(In Japanese) Morikita Shuppan, Tokyo, 1983.

[17] Tikhonov A.N. and Arsenin V.W., Solutions of Ill-Posed Problems. Winston-Wiley, New York, 1977.

[18] Varah J.M., On the numerical solution of ill-conditioned linear systems with applications to ill-posed problems, SIAM J. Numer. Anal.(1973),257-267