<table>
<thead>
<tr>
<th>Title</th>
<th>A Practical Method to Examine the Stability of a Class of Numerical Procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Kitagawa, Takashi</td>
</tr>
<tr>
<td>Citation</td>
<td>数理解析研究所講究録 (1989), 703: 187-202</td>
</tr>
<tr>
<td>Issue Date</td>
<td>1989-10</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/2433/101577">http://hdl.handle.net/2433/101577</a></td>
</tr>
<tr>
<td>Type</td>
<td>Departmental Bulletin Paper</td>
</tr>
<tr>
<td>Textversion</td>
<td>publisher</td>
</tr>
</tbody>
</table>

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

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 regulariza-
tion 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

\begin{align*}
(2.1) & \quad \Delta u = 0 \quad \text{in} \quad \Omega \\
(2.2) & \quad u = g \quad \text{on} \quad \partial \Omega ,
\end{align*}
where

\[ \Omega = \{ \omega \in \mathbb{R}^2 \mid \|\omega\|_2 < \rho \}. \]

The fundamental solution method approximates the solution \( u(x) \) by

\[ u_n(x) = \sum_{k=1}^{n} c_k \ G(x, y_k) \quad x \in \Omega \]

where \( G(x, y) \) is the Green's function for \( (\Delta, \Omega) \),

\[ G(x, y) = -\frac{1}{2} \log \|x - y\|_2 \quad x, y \in \mathbb{R}^2. \]

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

\[ u_n(x_j) = g(x_j) \quad j = 1, 2, \ldots, n, \]

where \( x_j \)'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 \( x_k = \rho e^{\frac{2\pi}{n} (k-1)i} \) 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 \( u_n \) converges to the solution \( u \) 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 \( n \),

\[
(2.5) \quad \| u - u_n \|_\infty \lesssim \sup_{\|x\|_2 = r_0} \frac{2}{1 - \rho/r_0} \{ (1+C(R,\rho)) (\rho/r_0)^{n/3} + 4(\rho/R)^{n/3} \},
\]

where \( C(R,\rho) \) is a constant depends on \( R \) and \( \rho \) 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 \( \text{Cond}(n,R) \) can be estimated by

\[
\text{Cond}(n,R) \sim \frac{\log R}{2} n \left( \frac{R}{\rho} \right)^{n/2}.
\]

Since the estimate (2.5-6) follows from the fact that the coefficient matrix for the particular location of \( x_k \) 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

\[
\Gamma \, c = g
\]
for given data \( g \) which may be contaminated by some perturbation \( \Delta g \), where \( g \in Y = \mathbb{R}^m \), \( c \in X = \mathbb{R}^n \) and \( \Gamma : X \to Y \) with

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

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

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

\[
(3.2) \quad f = \Lambda \ c,
\]

where \( f \in X \) and \( \Lambda : X \to Y \) with

\[
f_j = \tilde{h}_n(x_j) = \sum_{k=1}^{n} c_k \ G(x_j, y_k) = (\Lambda \ c)_j, \ x_j \in C(\gamma), \ j = 1, 2, \ldots, n.
\]

Due to the ill-conditioning of (3.1), some 'large' perturbation \( \Delta c \) may be introduced to the intermediate solution \( c \). One may assume intuitively that the error \( ||\Delta f|| \) in the final result \( f \), where \( \Delta f = \Lambda \Delta c \), is on a level with \( ||\Delta c|| \) or as large as \( ||\Lambda|| \ ||\Delta c|| \). 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 $\tilde{g} = g + \Delta g$ and the intermediate solution $\tilde{c} = c + \Delta c$. We have $\Gamma \tilde{c} = \tilde{g}$ as well as (3.1). Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ be singular values of $\Gamma$ and $\{u_i\}_{i=1,2,\ldots,m}$, $\{v_j\}_{j=1,2,\ldots,n}$ be singular vectors of $\Gamma$. Reflecting the ill-conditioning of $\Gamma$, 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

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

Next, we suppose that the final result $\tilde{f}$ is given by $\tilde{f} = f + \Delta f$. We have $\tilde{f} = \Lambda \tilde{c}$ as well as (5.2). Let $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_n \geq 0$ be singular values of $\Lambda$ and $\{\hat{u}_i\}_{i=1,2,\ldots,m}$, $\{\hat{v}_j\}_{j=1,2,\ldots,n}$ be singular vectors of $\Lambda$. 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 $i = 1,2,\ldots,n$ be the orthonormal singular systems for $\Gamma$ and $\Lambda$ respectively. Then we have

$$(3.4) \quad \|\Delta f\|_2 \leq \|\Xi \ast \Theta\|_F \|\Delta g\|_2$$

where
(3.5a) \[ \Xi = (\xi_{ij}) \quad \xi_{ij} = \hat{\sigma}_i / \sigma_j \]

(3.5b) \[ \Theta = (\theta_{ij}) \quad \theta_{ij} = (v_i, v_j), \quad i, j = 1, 2, \ldots, n, \]

\( \Xi \ast \Theta \) represents the Hadamard product of the matrices \( \Xi \) and \( \Theta \) and \( ||\cdot||_F \) denotes the Frobenius norm.

Making use of singular value decomposition, we can construct the matrices \( \Theta \) and \( \Xi \) 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 \( \Xi \) and \( \Theta \)

The elements \( \xi_{ij} \) of \( \Xi \), which we called explosive factor matrix in [12-13], represent the upper bound of the magnification of the \( u_1 \)-component \( (\Delta g, u_1) \) of perturbation \( \Delta g \) to \( \hat{u}_j \)-component \( (\Delta f, \hat{u}_j) \) of \( \Delta f \). To clarify what "magnification of \( (\Delta g, u_1) \) to \( (\Delta f, \hat{u}_j) \)" means, we define it precisely.

**Definition 3.2.** We call the partial derivative

\[ \frac{\partial (\Delta f, \hat{u}_j)}{\partial (\Delta g, u_1)} \]
the magnification of \((\Delta g, u_i)\) to \((\Delta f, \dot{u}_j)\).

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 \(\{\dot{u}_i, \dot{v}_i; \dot{\sigma}_i\}\) for \(i = 1, 2, \ldots, n\) be the orthonormal singular systems for \(\Gamma\) and \(\Lambda\) respectively. Then the magnification of \((\Delta g, u_i)\) to \((\Delta f, \dot{u}_j)\) is given by \(\xi_{ij} \theta_{ij}\), where \(\xi_{ij}\) and \(\theta_{ij}\) represent the \(ij\)-th elements of the matrices \(\Xi\) and \(\Theta\) respectively defined in (3.5a-b).

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

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

and

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

By expanding \(\Delta c\) by \(\{v_i\}\) \(i=1,2,\ldots,n\), we have

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

Putting this into (3.7), we have
(3.8) \((\Delta f, \hat{u}_j) = \hat{\sigma}_j \left( \sum_{i=1}^{n} (\Delta c, v_i) v_i, \hat{v}_j \right)\).

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

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

or

\[
= \xi_{ij} \theta_{ij}
\]

from the definition of \(\Xi\) and \(\Theta\). Q.E.D.

For instance, the largest element \(\xi_{1n}\) gives the upper bound of \(\hat{\sigma}_1 / \sigma_n\) which coincides with the straightforward 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 \(\theta_{ij}\) of \(\Theta\), 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 $\Delta g$ to $\Delta f$ is given by the square root of the sum of
squares of $\xi_{ij} \times \theta_{ij}$, or $\|E \ast \Theta\|_{F}$ (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 $\Delta g$ can be written as

\begin{equation}
(\Gamma^{*} \Gamma + \mu \ I) \ c = \Gamma^{*}(g + \Delta g).
\end{equation}

We write the solution of (4.1) $c(\mu, \Delta g)$. To examine the effective-
ness 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 $\Delta f(\mu, \Delta g) = f(\mu, \Delta g) - f(0,0)$ in the
theorem.

Theorem 4.1. Let $\{u_{i}, v_{i}; \sigma_{i}\}$ and $\{\hat{u}_{i}, \hat{v}_{i}; \hat{\sigma}_{i}\}$ for $i = 1, 2, . . . , n$
be the orthonormal singular systems for $\Gamma$ and $\lambda$ respectively.

Then we have

\begin{equation}
\|\Delta f(\mu, \Delta g)\|_{2} \leq \|E_{\xi} \ast \Theta\|_{F} \|g\|_{2} + \|E_{\rho} \ast \Theta\|_{F} \|\Delta g\|_{2}
\end{equation}

where

\begin{align}
(4.3a) \ E_{\xi} = (\xi_{ij}^{\xi}), \ & \xi_{ij}^{\xi} = \sigma_{i} \ \mu / (\sigma_{j}^{2} + \mu) \\
(4.3b) \ E_{\rho} = (\xi_{ij}^{\rho}), \ & \xi_{ij}^{\rho} = \sigma_{i} \ \sigma_{j} / (\sigma_{j}^{2} + \mu)
\end{align}
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

\[(4.4) \quad \xi(\mu) = f(\mu, 0) - f(0, 0)\]

and

\[(4.5) \quad \rho(\mu, \Delta g) = f(\mu, \Delta g) - f(\mu, 0),\]

we have

\[(4.6) \quad \|\Delta f(\mu, \Delta g)\|_2 \leq \|\rho(\mu, \Delta g)\|_2 + \|\xi(\mu)\|_2.\]

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

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

\[(4.7) \quad \|\rho(\mu, \Delta g)\|_2 \leq \|E_p \cdot \Theta\|_F \|\Delta g\|_2\]

and

\[(4.8) \quad \|\xi(\mu)\|_2 \leq \|E_\xi \cdot \Theta\|_F \|g\|_2.\]

From Theorem 5.1, we also have

\[(4.9) \quad \|\Delta f\|_2 \leq \|E \cdot \Theta\|_F \|\Delta g\|_2.\]

If we compare the error due to \(\Delta g\) of (4.7) with that of \(f\) without regularization of (4.9), we can recognize when the regularization
is effective. Checking corresponding elements of $E_\rho$, $E$ and $\Theta$, 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 $\mu$ carefully.

We can actually construct the matrices $E_\xi$, $E_\rho$ and $\Theta$ and we examine how the method of regularization stabilizes the numerical process and how we should choose the regularization parameter.

4.2. Matrices $E_\rho$ and $E_\xi$

We first note that since the elements $\theta_{ij}$ of the matrix $\Theta$ is independent of the regularization parameter $\mu$, the distortion coefficients matrix $\Theta$ is common with that without regularization. We also note that the matrices $E_\xi$ and $E_\rho$ as well as $E$ and $\Theta$ do not depend on $g$ or $\Delta g$ at all and, accordingly, we do not have to construct these matrices for different functions of $g$.

First we examine the elements $E_{ij}^\rho$ of matrix $E_\rho$ due to perturbation $\Delta g$ 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 $E_\rho$ represents the explosive factor matrix with regularization. The elements $E_{ij}^\rho$ of critical part of lower right corner (i.e., and j.e) are significantly smaller than those of $E$ without regularization. This can be
understood very easily if we compare the elements $\xi_{ij}^0$ and $\xi_{ij}$ of $\Xi$ and $\Xi_\rho$. As we have seen in Section 3.2 the elements $\xi_{ij}$ grow large for large $i$ and $j$ mainly because the denominator $\sigma_j$ approaches to zero as $j \to n$.

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

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

Another factor of error $\xi(\mu)$ 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 $\xi(\mu)$ is given in (4.9), its interpretation is somewhat more delicate than the case of $\Xi_\rho$. The element $\xi_{ij}^0 \cdot \theta_{ij}$ of the matrix $\Xi_\xi \cdot \Theta$ involved in (4.9) represents the magnification of the propagation of
(g,u₁) to (Δf, u₂) due to introduction of the regularization parameter μ. The size of (Δg, u₁) may not differ much among different i's, but the fourier coefficients (g, u₁) of g may be quite different in size. This is because the function g is harmonic and very smooth, which may result in very rapid convergence of the coefficients (g, u₁) to zero.

The matrices Ξ, Ξ', 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 ξᵢj of Ξ whose corresponding elements of θᵢj of Θ are close to unity

ii) we avoid contaminating the elements ξᵢj of Ξ whose corresponding elements of θᵢj of Θ are close to unity and the corresponding j-th Fourier coefficients (g, uⱼ) are significant.

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

References


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


