Cálculo Numérico (I) y Analisis Numérico I
(521230 – 521348 – 529242)
Apuntes de Sistemas Lineales

12 de abril de 2002

Índice General

1 Acerca del condicionamiento de matrices

La definición de condicionamiento de matrices surge de manera natural al hacer un análisis de error de los métodos numéricos que permiten resolver $Ax = b$, examinando la estabilidad de la solución x relativa a una pequeña perturbación del miembro derecho b. En efecto, sea A una matriz no-singular ; sea x solución de $Ax = b$; sea $\hat{x}$ solución del problema perturbado $A\hat{x} = \hat{b}$. Restando estos dos sistemas de ecuaciones, se tiene

$$A(x - \hat{x}) = b - \hat{b}$$

$$x - \hat{x} = A^{-1}(b - \hat{b})$$

Luego de la submultiplicidad de la norma matricial se obtiene

$$\|x - \hat{x}\| = \|A^{-1}(b - \hat{b})\| \leq \|A^{-1}\|\|b - \hat{b}\|$$

y dividiendo por $\|x\|$ se tiene que

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{\|A^{-1}\|\|b - \hat{b}\|}{\|x\|} = \|A\|\|A^{-1}\|\frac{\|b - \hat{b}\|}{\|A\|\|x\|}$$

Luego usando la desigualdad $\|b\| = \|Ax\| \leq \|A\|\|x\|$, obtenemos

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|b - \hat{b}\|}{\|b\|}$$

Más aún, para una elección adecuada de b y $\hat{b}$, esta última desigualdad puede llegar a ser una igualdad. El número

$$\text{cond}(A) = \|A\|\|A^{-1}\|$$

se denomina *condicionamiento* de A. Cuando este número es muy grande, la solución de $Ax = b$ puede ser extremadamente sensible a pequeñas perturbaciones de b, y se dice que el sistema está *mal condicionado* (ver el ejemplo de la matriz de Hilbert en la última página de este texto). Inversamente, cuando el número de condicionamiento es pequeño, se dice que el sistema está *bien condicionado.*

Por otro lado, sabemos que (ya lo demostramos!) que si consideramos una perturbación de A, entonces al comparar las soluciones de $Ax = b$ con $\hat{A}\hat{x} = b$, se tiene la siguiente estimación :

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\frac{\|A - \hat{A}\|}{\|A\|}}\frac{\|A - \hat{A}\|}{\|A\|}$$

siempre y cuando $\|A - \hat{A}\| < 1/\|A^{-1}\|$. De modo más general, al comparar $Ax = b$ con $\hat{A}\hat{x} = \hat{b}$ se tiene

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\frac{\|A - \hat{A}\|}{\|A\|}}\left(\frac{\|A - \hat{A}\|}{\|A\|} + \frac{\|b - \hat{b}\|}{\|b\|}\right)$$

2 Criterio de detención para métodos iterativos

A veces se tiende a detener un método iterativo $x^{(k+1)} = Mx^{(k)}$ (caso lineal con matriz de iteación M) cuando $\|x^{(k)} - x^{(k+1)}\| < tolerancia$. Veremos aquí como mejorar este *pseudo-criterio* de detetención, dando uno más adecuado, y que utilice una verdadera estimación de la medida del error $\|e^{(k)}\| = \|x^{(k)} - x\|$. Primero que nada, si el método converge, es decir si $x^{(k)} \to x$, se tiene entonces que

$$\sum_{j=1}^{\infty} \left(x^{(k+j)} - x^{(k+j+1)}\right) = \lim_{J\to\infty} \left((x^{(k+1)} - x^{(k+2)}) + (x^{(k+2)} - x^{(k+3)}) + \cdots + (x^{(k+J)} - x^{(k+J+1)})\right)$$

$$= \lim_{J\to\infty} \left(x^{(k+1)} - x^{(k+J+1)}\right) = x^{(k+1)} - x = e^{(k+1)}$$

Por otro lado como $x^{(k+1)} = Mx^{(k)}, \ldots, x^{(k+j)} = M^j x^{(k)}$, y $x^{(k+j+1)} = M^j x^{(k+1)}$, entonces se deduce que $e^{(k+1)} = \left(\sum_{j=1}^{\infty} M^j\right) \left(x^{(k)} - x^{(k+1)}\right)$ y podemos hacer la estimación siguiente :

$$\begin{aligned} \|e^{(k+1)}\| &\leq \left\|\sum_{j=1}^{\infty} M^j\right\| \|x^{(k)} - x^{(k+1)}\| \leq \left(\sum_{j=1}^{\infty} \|M\|^j\right) \|x^{(k)} - x^{(k+1)}\| \\ &\leq \frac{\|M\|}{1 - \|M\|} \|x^{(k)} - x^{(k+1)}\| \end{aligned}$$

Todo consiste entonces en poder calcular $\|M\|$. Por ejemplo, para el método de Jacobi, es sencillo pues se conoce explicitamente la matriz de iteración

$$M = \begin{pmatrix} 0 & & -\frac{a_{ij}}{a_{ii}} \\ & \ddots & \\ -\frac{a_{ij}}{a_{ii}} & & 0 \end{pmatrix}$$

Así, por ejemplo si consideramos la norma $\|\cdot\|_\infty$, tenemos $\|M\|_\infty = \max_{1\leq i\leq n} \frac{1}{|a_{ii}|} |\sum_{j=1, j\neq i}^{n} |a_{ij}|$. Pero para otros métodos iterativos (como por ejemplo Gauss-Seidel o relajación) el cálculo directo de $\|M\|$ no es fácil. Para estos casos se hace entonces la siguiente estimación :

$$\|x^{(k+1)} - x^{(k)}\| = \|M(x^{(k)} - x^{(k-1)})\| \leq \|M\| \|x^{(k)} - x^{(k-1)}\|$$

con lo cual

$$\|M\| \geq \frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)} - x^{(k-1)}\|}$$

Normalmente estos cocientes se estabilizan en un valor que llega a ser una buena aproximación de $\|M\|$. En cualquier caso, siempre podremos estimar la norma de M mediante

$$\|M\| \approx m_k = \max_{1\leq j\leq k} \frac{\|x^{(j+1)} - x^{(j)}\|}{\|x^{(j)} - x^{(j-1)}\|}$$

luego usar este valor estimado de $\|M\|$ para a su vez poder estimar el error :

$$\|e^{(k)}\| < \frac{m_k}{1 - m_k} \|x^{(k)} - x^{(k+1)}\| =: \varepsilon_k$$

y detener el programa cuando $\varepsilon_k < tolerancia$. Se ve claramente la importancia de considerar el factor $\frac{m_k}{1 - m_k}$ especialmente cuando m_k es cercano a 1.

3 El método SOR o de Relajación

Los métodos de Jacobi, Gauss-Seidel, o de Relajación (SOR), consisten en resolver el sistema $Ax = b$, mediante la descomposición de la matriz A bajo la forma $A = N - P$ donde N es una matriz invertible. Sea entonces D la matriz diagonal cuyos elementos coinciden con la diagonal de A, sea E la matriz triangular inferior y F triangular superior tal que $A = D - E - F$. La siguiente tabla resume la descripción de estos tres métodos iterativos :

Nombre del método	Descomposición $A = N - P$	Matriz $M = N^{-1}P$ del método iterativo	Descripción de una iteración
Jacobi	$A = D - (E + F)$	$D^{-1}(E + F) = I - D^{-1}A$	$Dx^{(k+1)} = (E + F)x^{(k)} + b$
Gauss-Seidel	$A = (D - E) + F$	$(D - E)^{-1}F$	$(D - E)x^{(k+1)} = Fx^{(k)} + b$
SOR	$A = \left(\frac{D}{\omega} - E\right) - \left(\frac{1-\omega}{\omega}D + F\right)$	$\left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$	$\left(\frac{D}{\omega} - E\right)x^{(k+1)} = \left(\frac{1-\omega}{\omega}D + F\right)x^{(k)} + b$

Si $\omega = 1$, el método SOR (Successive Overrelaxation) o de relajación, se convierte en el método de Gauss-Seidel. Un teorema debido a Kahan muestra que SOR falla si se encuentra fuera del intervalo $(0, 2)$. Si la matriz es simétrica y definida positiva entonces el método converge para todo $\omega \in (0, 2)$. En general no es posible calcular ω por adelantado. Frecuentemente se utiliza en elementos finitos una estimación heurística $\omega = 2 - O(h)$ donde h es el tamaño del espaciamiento entre las mallas de una discretización para un dominio físico (ver libro de Atkinson parrafo 8.8).

4 Métodos de descenso

Si la matriz A es **simétrica definida positiva** entonces el problema $Ax = b$ **es equivalente a** resolver el problema de minimización siguiente

$$\min_{x \in \mathbf{R}^n} J(x)$$

con $J : \mathbf{R}^n \mapsto \mathbf{R}$ forma cuadrática definida por $J(x) = \frac{1}{2}x \cdot Ax - b \cdot x$. En efecto este último problema es equivalente a anular el gradiente de J, es decir, $\nabla J(x) = Ax - b = 0$. Para resolver este problema de minimización se pueden utilizar algunos métodos de descenso como

4.1 El método del gradiente

El gradiente de la forma cuadratica $J(x)$ está dado por $\nabla J(x) = Ax - b$. Luego el método del gradiente consiste en minimizar J descendiendo en la dirección $-\nabla J$ y con *paso constante* en cada iteración. Aplicado a la resolución de $Ax = b$ es :

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \alpha d^{(k)}, \qquad k = 0, 1, \dots \\ d^{(k)} &= -\nabla J(x^{(k)}) = -(Ax^{(k)} - b) \end{aligned}$$

con α una constante positiva convenientemente elegida (suficientemente pequeña).

El método con *paso variable* consiste en escoger α_k de modo de que el paso sea óptimo cuando $J(x)$ es igual a la forma cuadrática. Así :

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k(Ax^{(k)} - b), \qquad k = 0, 1, \dots \\ \alpha_k &= \frac{r^{(k)} \cdot r^{(k)}}{r^{(k)} \cdot Ar^{(k)}}, \qquad r^{(k)} = Ax^{(k)} - b \end{aligned}$$

4.2 El método del gradiente conjugado

Este método es un mejoramiento del método del gradiente y consiste en el siguiente algoritmo : Sea $x^{(0)}$ dado, y $d^{(0)} = -r^{(0)} = -(Ax^{(0)} - b)$

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)}, \\ \alpha_k &= -\frac{r^{(k)} \cdot d^{(k)}}{d^{(k)} \cdot Ad^{(k)}}, \\ d^{(k+1)} &= -r^{(k+1)} + \beta_k d^{(k)}, \\ \beta_k &= \frac{r^{(k+1)} \cdot Ad^{(k)}}{d^{(k)} \cdot Ad^{(k)}}, \qquad k = 0, 1, \ldots \end{aligned}$$

donde $r^{(k)} = Ax^{(k)} - b$. Este método converge si J convexo y coercivo – qué es coercivo ? (ver P.-G. Ciarlet : *Introduction à l'analyse numérique matricielle et à l'optimization*, Masson, Paris (1985)). Si $J(x) = \frac{1}{2}x \cdot Ax - b \cdot x$ entonces hay convergencia a la solución de $Ax = b$ (cuando A es simétrica definida positiva). Pero además de todo esto el método posee sorprendentes propiedades reflejadas en los siguientes teoremas :

Teorema 1. Este método converge a la solución de $Ax = b$ en a lo más n iteraciones (!!!).

Demostración. Tarea para los estudiantes de Ingeniería Matemática, Licenciatura en Matemáticas, y Todos aquellos estudiantes que también deseen hacerla : demuestre primero (por inducción) que para todo k, el subespacio generado por los vectores $\{d^{(0)}, \ldots, d^{(k)}\}$ es igual al subespacio generado por $\{r^{(0)}, \ldots, r^{(k)}\}$ e igual al generado por $\{r^{(0)}, Ar^{(0)}, \ldots, A^k r^{(0)}\}$; luego demuestre (también por inducción) que $r^i \cdot r^j = 0$ para $i \neq j$, y $d^i \cdot Ad^j = 0$ para $i \neq j$; concluya que necesariamente $r^{(k)} = Ax^{(k)} - b = 0$, para algún $k \leq n$.

Observación. De este último teorema que dice que el método de Gradiente Conjugado converge en un tiempo finito (!!!) se puede deducir además que se necesita en la resolución de $Ax = b$, de a lo más del orden de n^3 sumas, n^3 multiplicaciones, y $2n$ divisiones. Es decir tantas operaciones elementales o menos que el método de Cholewsky (o descomposición LU). Pero cuidado! el método solo funciona (por desgracia) para matrices simétricas y definidas positivas.

Teorema 2. Para el método del Gradiente Conjugado se tiene la siguiente estimación del error

$$\|x^{(k)} - x\|_A \leq 2 \left(\frac{\sqrt{\mathrm{cond}(A)_*} - 1}{\sqrt{\mathrm{cond}(A)_*} + 1} \right)^k \|x^{(0)} - x\|_A$$

donde $\|x\|_A = \sqrt{x \cdot Ax}$ (es una norma si A es sim. def. pos.), y $\mathrm{cond}(A)_* = \dfrac{\max_{i=1,\ldots,n} |\lambda_i|}{\min_{i=1,\ldots,n} |\lambda_i|}$.

Demostración. Ver O. Axelsson : *A class of iterative methods for finite element equations.* Computation Methods in Applied Mechanic and Engeenering. 9 (1976).

Observación. Esta estimación del error obviamente no es óptima, pero asegura que el método converge muy rapidamente si el sistema está *bien condicionado*.

4.3 El método del gradiente conjugado precondicionado (SSOR)

Para un problema específico como por ejemplo una matriz proveniente de un problema de elementos finitos, puede convenir en algunos casos *precondicionar* la matriz y resolver el siguiente sistema mediante gradiente conjugado aplicado a la matriz simétrica definida positiva $B = C^{-1/2}AC^{-1/2}$ de menor condicionamiento que la matriz A

$$C^{-1/2}AC^{-1/2}y = C^{-1/2}b, \qquad x = C^{-1/2}y$$

con $C^{-1/2}$ matriz simétrica definida positiva tal que $C^{-1/2}C^{-1/2} = C^{-1}$ es la inversa de

$$C = \left(\frac{D}{\omega} - E\right) D^{-1} \left(\frac{D}{\omega} - E\right)$$

y ω es escogido de modo de minimizar el condicionamiento de la matriz B.

RAD/RRA/MSC/msc
http://www.ing-mat.udec.cl/pregrado/asignaturas/521230/
22/03/02