CAPITULO XVIII. ESTIMACIONES DE ERROR Y REFINAMIENTO ITERATIVO

1. ESTIMACIONES DE ERROR

Parece razonable intuitivamente que si $\tilde{\mathbf{x}}$ es una aproximación a la solución $\mathbf{x}$ de $A\ \mathbf{x} = \mathbf{b}$ y el vector residual $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}}$ tiene la propiedad de que $||\mathbf{r}||$ es pequeño, entonces $||\mathbf{x} - \tilde{\mathbf{x}}||$ será también peque˜no. Aún cuando éste es frecuentemente el caso, ciertos sistemas especiales, que aparecen bastante en la práctica, no tienen esta propiedad.

Ejemplo 1.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} ,$$

tiene la solución única $\mathbf{x} = (1,1)^t$. La aproximación a esta solución $\tilde{\mathbf{x}} = (3,0)^t$ tiene vector residual

$$\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = \begin{pmatrix} 3 \\ 3.0001 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.0002 \end{pmatrix} ,$$

así que $||\mathbf{r}||_\infty = 0.0002$.

Aunque la norma del vector residual es pequeña, la aproximación $\tilde{\mathbf{x}} = (3,0)^t$ es obviamente bastante pobre; en realidad, $||\mathbf{x} - \tilde{\mathbf{x}}||_\infty = 2$.

Esta dificultad se puede explicar muy simplemente si se observa que la solución del sistema representa la intersección de las rectas

$$l_1 : \quad x_1 + 2\ x_2 = 3 \quad \text{y} \quad l_2 : \quad 1.0001\ x_1 + 2\ x_2 = 3.0001\ .$$

El punto $(3,0)$ se encuentra en l_1 y las rectas son casi paralelas. Esto implica que $(3,0)$ se encuentra también cerca de l_2, aún cuando difiere significativamente del punto de intersección $(1,1)$. Si las rectas no hubieran sido casi paralelas, se esperería que un vector residual peque˜no implicara una aproximación precisa.

En general, no podemos depender de la geometría del sistema para obtener una indicación de cúando pueden presentarse problemas. Sin embargo, podemos extraer esta información considerando las normas de la matriz A y de su inversa.

Definición. El **número de condición** $K(A)$ de la matriz no singular A relativo a la norma $||\ \cdot\ ||$ se define como

$$K(A) = ||A||\ ||A^{-1}||\ .$$

Teorema XVIII.1

Si $\tilde{\mathbf{x}}$ es una aproximación a la solución de $A\ \mathbf{x} = \mathbf{b}$ y A es una matriz no singular, entonces para cualquier norma natural,

$$||\mathbf{x} - \tilde{\mathbf{x}}|| \leq ||\mathbf{r}||\ ||A^{-1}|| = K(A) \frac{||\mathbf{r}||}{||A||} \qquad (XVIII.1)$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \le ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||} = K(A)\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ , \qquad (XVIII.2)$$

siempre que $\mathbf{x} \neq 0$ y $\mathbf{b} \neq 0$, donde $\mathbf{r}$ es el vector residual de $\tilde{\mathbf{x}}$ con respecto al sistema $A\ \mathbf{x} = \mathbf{b}$.

Demostración: como $\mathbf{r} = \mathbf{b} - A\ \tilde{\mathbf{x}} = A\ \mathbf{x} - A\ \tilde{\mathbf{x}}$ y A no es singular:

$$||\mathbf{x}-\tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \le ||A^{-1}||\ ||\mathbf{r}||\ .$$

Además, como $\mathbf{b} = A\ \mathbf{x}$, $||\mathbf{b}|| \le ||A||\ ||\mathbf{x}||$; así que

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||}{||\mathbf{x}||} \le ||A^{-1}||\ ||A||\ \frac{||\mathbf{r}||}{||\mathbf{b}||}\ .$$

c.q.d.

Las desigualdades $(XVIII.1)$ y $(XVIII.2)$ implican que las cantidades $||A^{-1}||$ y $K(A) = ||A||\ ||A^{-1}||$ pueden ser usadas para dar una indicación de la conexión entre el vector residual y la precisión de la aproximación. En general, el error relativo $||\mathbf{x} - \tilde{\mathbf{x}}||/||\mathbf{x}||$ es de mayor interés y por la desigualdad $(XVIII.2)$ este error está acotado por el producto del número de condición $K(A) = ||A||\ ||A^{-1}||$ con el residual relativo para esta aproximación $||\mathbf{r}||/||\mathbf{b}||$. Para esta aproximación puede usarse cualquier norma que sea conveniente, el único requisito es que se use consistentemente desde el principio hasta el final.

Ya que para cualquier matriz no singular A

$$1 = ||I|| = ||A \cdot A^{-1}|| \le ||A||\ ||A^{-1}|| = K(A)\ ,$$

se espera que la matriz A tenga un buen comportamiento (llamada formalmente una **matriz bien condicionada**) si $K(A)$ está cerca de uno y un comportamiento defectuoso (llamada una **matriz mal condicionada**) cuando $K(A)$ sea significativamente mayor que uno. El comportamiento en esta situación se refiere a la relativa seguridad de que un vector residual peque˜no implique correspondientemente una solución aproximada precisa.

Ejemplo 2.

La matriz del sistema considerado en el ejemplo 1 es

$$A = \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}\ ,$$

que tiene $||A||_\infty = 3.0001$. Esta norma no se considera grande, sin embargo

$$A^{-1} = \begin{pmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{pmatrix}\ ,$$

y $||A^{-1}||_\infty = 20000$ y para la norma infinita $K(A) = 20000 \times 3.0001 = 60002$. El tama˜no del número de condición para este ejemplo seguramente nos detendría al tomar decisiones apresuradas acerca de la precisión, basadas en el residual de la aproximación.

Mientras que, en teoría, el número de condición de una matriz depende totalmente de las normas de la matriz y de su inversa, en la práctica, el cálculo de la inversa está sujeto a errores de redondeo y es dependiente de la exactitud con la que se estén haciendo los cálculos. Si hacemos la suposición de que la solución aproximada al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ se determina usando aritmética de t dígitos y eliminación Gaussiana, se puede demostrar que el vector residual $\mathbf{r}$ para la aproximación $\tilde{\mathbf{x}}$ tiene la propiedad

$$||\mathbf{r}|| \approx 10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}||\ . \qquad (XVIII.3)$$

De esta ecuación aproximada, se puede obtener una estimación del número de condición efectivo para la aritmética de t dígitos, sin la necesidad de invertir la matriz A. [La aproximación en la ecuación $(XVIII.3)$ supone que todas las operaciones aritméticas en la técnica de eliminación Gaussiana se efectúan usando aritmética de t dígitos, pero que las operaciones que se necesitan para determinar el residual se hacen en doble precisión, es decir, $2t$ dígitos, para eliminar la pérdida de precisión involucrada en la sustracción de números casi iguales que ocurre en los cálculos del residual].

La aproximación del número de condición $K(A)$ a t dígitos viene de considerar el sistema lineal $A\ \mathbf{y} = \mathbf{r}$. La solución de este sistema puede aproximarse fácilmente ya que los multiplicadores para el método de eliminación Gaussiana han sido ya calculados y supuestamente retenidos. De hecho $\tilde{\mathbf{y}}$, la solución aproximada de $A\ \mathbf{y} = \mathbf{r}$, satisface que

$$\tilde{\mathbf{y}} \approx A^{-1}\ \mathbf{r} = A^{-1}\ (\mathbf{b} - A\ \tilde{\mathbf{x}}) = A^{-1}\ \mathbf{b} - A^{-1}\ A\ \tilde{\mathbf{x}} = \mathbf{x} - \tilde{\mathbf{x}}\ ; \qquad (XVIII.4)$$

así que $\tilde{\mathbf{y}}$ es una estimación del error cometido al aproximar la solución del sistema original. Consecuentemente la ecuación $(XVIII.3)$ puede usarse para deducir que

$$\begin{aligned} ||\tilde{\mathbf{y}}|| &\approx ||\mathbf{x} - \tilde{\mathbf{x}}|| = ||A^{-1}\ \mathbf{r}|| \leq \\ &\leq ||A^{-1}||\ ||\mathbf{r}|| \approx ||A^{-1}||\ (10^{-t}\ ||A||\ ||\tilde{\mathbf{x}}||) = 10^{-t}\ ||\tilde{\mathbf{x}}||\ K(A)\ . \end{aligned}$$

Esto proporciona una aproximación para el número de condición involucrado en la solución del sistema $A\ \mathbf{x} = \mathbf{b}$ usando eliminación Gaussiana y el tipo de aritmética de t dígitos descrito anteriormente:

$$K(A) \approx 10^t\ \frac{||\tilde{\mathbf{y}}||}{||\tilde{\mathbf{x}}||}\ . \qquad (XVIII.5)$$

Ejemplo 3.

El sistema lineal $A\ \mathbf{x} = \mathbf{b}$ dado por

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix} ,$$

tiene la solución exacta $\mathbf{x} = (1, 1, 1)^t$.

Usando eliminación Gaussiana y aritmética de redondeo de 5 dígitos llegamos sucesivamente a las matrices ampliadas

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 & | & 15913 \\ 0 & -10596 & 16.501 & | & -10580 \\ 0 & -7451.4 & 6.525 & | & -7444.9 \end{pmatrix}$$

y

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 & | & 15913 \\ 0 & -10596 & 16.501 & | & -10580 \\ 0 & 0 & -5.079 & | & -4.7 \end{pmatrix} .$$

La solución aproximada a este sistema es

$$\tilde{\mathbf{x}} = (1.2001, 0.99991, 0.92538)^t .$$

El vector residual correspondiente a $\tilde{\mathbf{x}}$ calculado con doble precisión (y luego redondeado a cinco dígitos) es

$$\begin{aligned} \mathbf{r} = \mathbf{b} - A\,\tilde{\mathbf{x}} = & \\ = & \begin{pmatrix} 15913 \\ 28.544 \\ 8.4254 \end{pmatrix} - \begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} 1.2001 \\ 0.99991 \\ 0.92538 \end{pmatrix} = \\ = & \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} ; \end{aligned}$$

así que

$$||\mathbf{r}||_\infty = 0.27413 .$$

La estimación del número de condición dada en la discusión anterior se obtiene resolviendo primero el sistema $A\,\mathbf{y} = \mathbf{r}$:

$$\begin{pmatrix} 3.3330 & 15920 & -10.333 \\ 2.2220 & 16.71 & 9.612 \\ 1.5611 & 5.1791 & 1.6852 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -0.0051818 \\ 0.27413 \\ -0.18616 \end{pmatrix} ,$$

lo cual implica que $\tilde{\mathbf{y}} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t$. Usando la estimación dada por la ecuación $(XVIII.5)$:

$$K(A) \approx 10^5\, \frac{||\tilde{\mathbf{y}}||_\infty}{||\tilde{\mathbf{x}}||_\infty} = \frac{10^5\ (0.20008)}{1.2001} = 16672 .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x} - \tilde{\mathbf{x}}||_\infty \leq K(A) \frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(16672)(0.27413)}{15934} = 0.28683$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \le K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(16672)(0.27413)}{15913} = 0.28721\ .$$

Para determinar el número de condición exacto de A, necesitamos construir primero A^{-1}. Usando aritmética de redondeo de 5 dígitos para los cálculos se obtiene la aproximación:

$$A^{-1} = \begin{pmatrix} -1.1701\times10^{-4} & -1.4983\times10^{-1} & 8.5416\times10^{-1} \\ 6.2782\times10^{-5} & 1.2124\times10^{-4} & -3.0662\times10^{-4} \\ -8.6631\times10^{-5} & 1.3846\times10^{-1} & -1.9689\times10^{-1} \end{pmatrix}\ .$$

El Teorema XIII.13 puede usarse para demostrar que $||A^{-1}||_\infty = 1.0041$ y $||A||_\infty = 15934$. Como consecuencia la matriz A mal condicionada tiene

$$K(A) = (1.0041)\ (15934) = 15999\ .$$

La aproximación que habíamos obtenido antes está bastante cerca de este $K(A)$ y ha requerido un esfuerzo computacional considerablemente menor.

Como la solución real $\mathbf{x} = (1,1,1)^t$ de este sistema es conocida, podemos calcular ambos

$$||\mathbf{x}-\tilde{\mathbf{x}}||_\infty = 0.2001$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} = 0.2001\ .$$

Las cotas de error dadas en el Teorema XVIII.1 para estos valores son

$$||\mathbf{x}-\tilde{\mathbf{x}}||_\infty \le K(A)\frac{||\mathbf{r}||_\infty}{||A||_\infty} = \frac{(15999)(0.27413)}{15934} = 0.27525$$

y

$$\frac{||\mathbf{x}-\tilde{\mathbf{x}}||_\infty}{||\mathbf{x}||_\infty} \le K(A)\ \frac{||\mathbf{r}||_\infty}{||\mathbf{b}||_\infty} = \frac{(15999)(0.27413)}{15913} = 0.27561\ .$$

2. REFINAMIENTO ITERATIVO

En la ecuación $(XVIII.4)$ usamos la estimación $\tilde{\mathbf{y}} \approx \mathbf{x}-\tilde{\mathbf{x}}$, en la que $\tilde{\mathbf{y}}$ es la solución aproximada al sistema $A\ \mathbf{y} = \mathbf{r}$. Sería razonable sospechar, a partir de este resultado, que $\tilde{\mathbf{x}}+\tilde{\mathbf{y}}$ fuese una mejor aproximación a la solución del sistema lineal $A\ \mathbf{x} = \mathbf{b}$ que la aproximación inicial $\tilde{\mathbf{x}}$.

El método que usa esta suposición se llama **refinamiento iterativo**, o mejora iterativa y consiste en llevar a cabo iteraciones sobre el sistema cuyo lado derecho es el vector residual para las aproximaciones sucesivas, hasta que se obtiene una precisión satisfactoria. El procedimiento se usa generalmente sólo en los sistemas en que se sospecha que la matriz involucrada es mal condicionada, debido a que esta técnica no mejora mucho la aproximación para un sistema bien condicionado.

Algoritmo de refinamiento iterativo.

==

Para aproximar la solución al sistema lineal $A\ \mathbf{x} = \mathbf{b}$ cuando se sospecha que A sea mal condicionada.

Entrada: número de incógnitas y de ecuaciones n; las componentes de la matriz $A = (a_{ij})$ donde $1 \leq i, j \leq n$; las componentes b_i, con $1 \leq i \leq n$, del término no homogéneo $\mathbf{b}$; la tolerancia TOL; el número máximo de iteraciones N_0.

Salida: solución aproximada $\mathbf{xx} = (xx_1, xx_2, \ldots, xx_n)$ ó mensaje de que el número de iteraciones fue excedido.

Paso 0: Resolver el sistema $A\ \mathbf{x} = \mathbf{b}$ para $x_1, x_2, \ldots, x_n$ por eliminación Gaussiana guardando los multiplicadores m_{ji}, $j = i+1, i+2, \ldots, n$, $i = 1, 2, \ldots, n-1$ y haciendo notar los intercambios de filas.

Paso 1: Tomar $k = 1$.

Paso 2: Mientras que $k \leq N_0$ seguir los pasos 3–8.

Paso 3: Para $i = 1, 2, \ldots, n$ (*calcular* $\mathbf{r}$*, realizando los cálculos con doble precisión aritmética*), tomar

$$r_i = b_i - \sum_{j=1}^{n} (a_{ij}\ x_j).$$

Paso 4: Resolver el sistema lineal $A\ \mathbf{y} = \mathbf{r}$ usando eliminación Gaussiana en el mismo orden que en el paso 0.

Paso 5: Para $i = 1, 2, \ldots, n$ tomar

$$xx_i = x_i + y_i.$$

Paso 6: Si $||\mathbf{x} - \mathbf{xx}|| < TOL$ entonces SALIDA $(xx_1, xx_2, \ldots, xx_n)$; *(procedimiento completado satisfactoriamente)* PARAR.

Paso 7: Tomar $k = k + 1$.

Paso 8: Para $i = 1, 2, \ldots, n$ tomar $x_i = xx_i$.

Paso 9: SALIDA *(número máximo de iteraciones excedido)*; *(procedimiento completado sin éxito)* PARAR.

==

Si se está usando aritmética de t dígitos, un procedimiento recomendable para parar en el paso 6 consiste en iterar hasta que $|y_i^{(k)}| \leq 10^{-t}$ para cada $i = 1, 2, \ldots, n$.

Debe enfatizarse que la técnica de refinamiento iterativo no da resultados satisfactorios para todos los sistemas que contienen matrices mal condicionadas. En particular, si $K(A) \geq 10^t$, es probable que el procedimiento falle y que la única alternativa sea el uso de mayor precisión en los cálculos.

Ejemplo 4.

En el ejemplo 3 encontramos que la aproximación al problema que habíamos estado considerando, usando aritmética de cinco dígitos y la eliminación Gaussiana, era

$$\tilde{\mathbf{x}}^{(1)} = (1.2001, 0.99991, 0.92538)^t$$

y que la solución a $A\ \mathbf{y}^{(1)} = \mathbf{r}^{(1)}$ era

$$\tilde{\mathbf{y}}^{(1)} = (-0.20008, 8.9989 \times 10^{-5}, 0.074607)^t\ .$$

Usando el paso 5 del algoritmo, tenemos que

$$\tilde{\mathbf{x}}^{(2)} = \tilde{\mathbf{x}}^{(1)} + \tilde{\mathbf{y}}^{(1)} = (1.0000, 1.0000, 0.99999)^t$$

y el error real en esta aproximación es

$$||\mathbf{x} - \tilde{\mathbf{x}}^{(2)}||_\infty = 1.0 \times 10^{-5} \ .$$

Usando la técnica de paro sugerida para el algoritmo, calculamos $\mathbf{r}^{(2)} = \mathbf{b} - A\ \tilde{\mathbf{x}}^{(2)}$, y resolvemos el sistema $A\ \mathbf{y}^{(2)} = \mathbf{r}^{(2)}$, obteniéndose

$$\tilde{\mathbf{y}}^{(2)} = (-2.7003 \times 10^{-8}, 1.2973 \times 10^{-8}, 9.9817 \times 10^{-6})^t \ .$$

Puesto que $||\tilde{\mathbf{y}}^{(2)}||_\infty \leq 10^{-5}$, concluímos que

$$\tilde{\mathbf{x}}^{(3)} = \tilde{\mathbf{x}}^{(2)} + \tilde{\mathbf{y}}^{(2)} = (1.0000, 1.0000, 1.0000)^t$$

es suficientemente preciso. De hecho es claramente correcto.

EJERCICIOS.

1. Resolver los siguientes sistemas lineales usando eliminación Gaussiana y refinamiento iterativo (con aritmética a dos dígitos para el problema a) y a tres dígitos para b) y c)). Calcular $K(A)$.

$$a)\quad \begin{aligned} 3.9\,x_1 &+ 1.6\,x_2 = 5.5\,, \\ 6.8\,x_1 &+ 2.9\,x_2 = 9.7\,. \end{aligned}$$

$$b)\quad \begin{aligned} 4.56\,x_1 &+ 2.18\,x_2 = 6.74\,, \\ 2.79\,x_1 &+ 1.38\,x_2 = 4.13\,. \end{aligned}$$

$$c)\quad \begin{aligned} x_1 &+ \tfrac{1}{2}\,x_2 + \tfrac{1}{3}\,x_3 = \tfrac{11}{6}\,, \\ 5\,x_1 &+ \tfrac{10}{3}\,x_2 + \tfrac{5}{2}\,x_3 = \tfrac{65}{6}\,, \\ \tfrac{100}{3}\,x_1 &+ 25\,x_2 + 20\,x_3 = \tfrac{235}{3}\,. \end{aligned}$$

2. Calcular los números de condición de las siguientes matrices, relativos a $||\cdot||_\infty$.

$$a)\quad \begin{pmatrix} 1 & 2 \\ 1.0001 & 2 \end{pmatrix}\,,$$

$$b)\quad \begin{pmatrix} 3.9 & 1.6 \\ 6.8 & 2.9 \end{pmatrix}\,,$$

$$c)\quad \begin{pmatrix} 4.56 & 2.18 \\ 2.79 & 1.38 \end{pmatrix}\,,$$

$$d)\quad \begin{pmatrix} 1.003 & 58.09 \\ 5.550 & 321.8 \end{pmatrix}\,.$$

3. El sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 1.0001 & 2. \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.0001 \end{pmatrix}$$

tiene solución $(1,1)^t$. Cambiar la matriz A por

$$\begin{pmatrix} 1. & 2. \\ 0.9999 & 2. \end{pmatrix}$$

y considerar el sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 0.9999 & 2. \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.0001 \end{pmatrix}\,.$$

Calcular la nueva solución $\tilde{\mathbf{x}}$ usando aritmética de cinco dígitos.

4. El sistema lineal

$$\begin{pmatrix} 1. & 2. \\ 1.00001 & 2. \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3. \\ 3.00001 \end{pmatrix}$$

tiene solución $(1,1)^t$. Encontrar la solución del sistema lineal perturbado

$$\begin{pmatrix} 1. & 2. \\ 1.000011 & 2. \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3.00001 \\ 3.00003 \end{pmatrix}\,,$$

usando aritmética de siete dígitos.