UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

DIVISION DE ESTUDIOS DE POSGRADO

"RESOLUCION VIA BIDIAGONALIZACION DE LANCZOS DE PROBLEMAS DE MINIMOS CUADRADOS REGULARIZADOS DE MEDIANA ESCALA"

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRIA EN CIENCIAS (MATEMATICAS)

P R E S E N T A

ARMANDO SAAVEDRA ESPINOSA

MEXICO. D.F. 1993

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

PROLOGO

Para describir de una alguna manera las diferentes formas de obtener el grado de Maestría en Matemáticas en nuestro país un buen ejemplo, es analizar el caso de la Facultad de Ciencias de la U.N.A.M. La afirmación anterior se fundamenta en el hecho de ser ésta la pionera de las escuelas de matemáticas del país.

Siendo México un país que inicia su desarrollo al término de la revolución vemos que en ese entonces la población en general carece de una formación con sólidas bases educativas, y que el mismo Estado Mexicano no tenía una clara idea del camino a seguir para superar esta deficiencia, es por esta causa que éste adopta un esquema norteamericano para sus escuelas de posgrado.

Esta manera de resolver un problema interno adaptando una solución extranjera condena al país a depender tecnológicamente del exterior.

Es importante fundamentar la afirmación anterior: los primeros egresados en el área de matemáticas del país cursan sus estudios de posgrado en E.U., a su regreso traen consigo una formación sólida en la materia generando con ello un amplio desarrollo de las matemáticas en México; sin embargo, las áreas que abarcan son limitadas a la matemática teórica; es por esta razón que no inciden en la iniciativa privada para un desarrollo tecnológico del país y los temas de investigación que desarrollan no coinciden con las necesidades de investigación del país.

En la Facultad de Ciencias no es sino hasta mediados de la década de los ochentas cuando se inicia el ofrecimiento de manera

sistematica de cursos de posgrado en el área de Matemáticas aplicadas (particularmente en Análisis Numérico).

Esta nueva época trae consigo un mayor campo de trabajo para los matemáticos aplicados, cabe señalar que algunos de los egresados de esta nueva corriente han logrado integrarse a instituciones importantes en el país por ejemplo, Bancomer, IMP, ININ, Chevrolet, Aseguradoras, Cometro, etc.

Tijonov y Kostomárov en su libro titulado "Algo acerca de la matemática aplicada" Editorial MIR, Moscú, mencionan que un país desarrollado tecnológicamente debe tener un 70% de matemáticos aplicados y un 30% de matemáticos téoricos, en México se tiene en forma inversa, es decir, existen una gran cantidad de matemáticos téoricos y un pequeño porcentaje de matemáticos aplicados. Si existiera en México el porcentaje adecuado, el flujo de información acerca de las necesidades de investigacion en nuestro país sería mayor.

En la Facultad de Ciencias existen dos alternativas para obtener el grado de maestría: la primera de ellas es presentando examenes generales, es decir, presentando un examen oral de tres materias básicas (Algebra, Variable Compleja y Análisis Matemático) y dos materias optativas a escojer; la segunda es presentando en un examen oral el desarrollo de una tesis.

Desde mi punto de vista y recordando lo anterior, se decide realizar el presente trabajo para así obtener una mejor formación en investigación.

En esta tesis se combinan ideas de dos escuelas matemáticas diferentes como son la Norteamericana y la Sovietica: por un lado, la gran industria norteamericana necesita resolver problemas matemáticos de gran escala y por el otro lado los sovieticos debido a sus escasos recursos requieren métodos que sean susceptibles de realizarse con pocos recursos.

Opino que en el caso particular de México es deseable aprovechar la cercanía con E.U. y dada nuestra situación económica retomar las ideas sovieticas. Este hecho implica tener una gran cantidad de bibliografía por un lado y debido a problemas de lenguaje (Ruso) poca del otro, lo que originó que el trabajo tomara una gran cantidad de tiempo en su desarrollo.

Mi deseo es que esta tesis sea un grano de arena en el desierto, que ayude al desarrollo de México.

Quiero agradecer muy especialmente al DR. Jesus López Estrada el haber aceptado dirigir esta tesis, asegurando que he logrado obtener un gran aprovechamiento de la misma.

Así mismo deseo agradecer al Dr. Humberto Madrid de la Vega y al M. en C. José López Estrada el ayudarme a realizar esta tesis cuando por razones de trabajo el Dr. Jesus se ausentó del país.

Finalmente agradezco a:

M. en C. Hans Luis Fetter Natanski

M. en C. María Elena García Alvarez

Dra. Susana Gómez Gómez

Dra. Patricia Saavedra Barrera

Por haber revisado el presente, ya que con sus comentarios y sugerencias se logró tener una mejor presentación del mismo.

INDICE

En el presente trabajo se abordará el problema de resolver numéricamente vía el criterio de Mínimo de Cuadrados el sistema de ecuaciones lineales algebraicas sobre-determinado $A x = b$, donde $A \in R^{mxn}$ es moderada o grande, rala y mal condicionada. El objetivo es dar un algoritmo que aproveche, en su caso, la poca densidad de la matriz A para obtener una solución aceptable.

Para lograr esto se combinan dos métodos: El primero de ellos, tiene como propósito simplificar el sistema y es la Bidiagonalización Inferior de Lanczos, el cual transforma el sistema original en uno de la forma $B y = c$, donde B es una matriz bidiagonal inferior; este método será presentado en el capítulo 2 de este trabajo.

La matriz bidiagonal B hereda el mal condicionamiento de la matriz original, esto hace necesario aplicar un segundo método, el cual amortigua el mal comportamiento y es conocido como Regularización de Tijonov. En el capítulo 3 se presenta el método de Regularización de Tijonov y con el propósito de comparación, el método de Regularización Directa.

El capítulo 4 tiene como propósito presentar un algoritmo que combina la Regularización de Tijonov con la Bidiagonalización de Lanczos para obtener una solución aceptable de un sistema de ecuaciones sobre-determinado en donde se aproveche la poca densidad de la matriz A.

En el capítulo 5 se presenta el trabajo experimental desarrollado en esta tesis, así como algunos detalles

computacionales utilizados en los experimentos numéricos.

Finalmente, en el capítulo 6 se presentan las conclusiones obtenidas en el presente trabajo y la bibliografía.

1.1 PLANTEAMIENTO DEL PROBLEMA

El problema es el siguiente:

Dados $A \in \mathbb{R}^{m \times n}$, $m \geq n \gg 1$ y $b \in \mathbb{R}^{m \times 1}$ resolver:

$$A\,x = b \qquad \ldots(1.1)$$

Los términos "mal condicionada" y posiblemente "rala" son adjetivos subjetivos, por lo cual, es conveniente precisarlos.

i) Se dice que una matriz A es rala si al menos el 60% de sus elementos son iguales a cero.

ii) Se dice que una matriz A es mal condicionada si $(K(A))^{-1} = (\| A \| \, \| A^{-1} \|)^{-1} \approx \mu$ (siendo μ la unidad de redondeo).

iii) Resta aclarar el símbolo $n \gg 1$. En los últimos años se ha dado un gran desarrollo en la computación, es por esta causa que el término $n \gg 1$ (grande) depende de la máquina en que se trabaje (Medio Ambiente). Diremos que n es "grande" en una PC si n es del orden de 300 y en una máquina de mayor capacidad si n es del orden de 800 ó más.

1.2 PROBLEMA TIPICO QUE DA ORIGEN A MATRICES MAL CONDICIONADAS

Hoy en día una gran cantidad de problemas de la Matemática

Aplicada dan origen a sistemas de ecuaciones lineales de la forma $A x = b$, donde $A \in \mathbb{R}^{mxn}$ y mal condicionada. En esta sección se dará un ejemplo típico de uno de ellos, como lo es la discretización de la Ecuación Integral de Fredholm de primer orden. Con este propósito se darán algunas definiciones y resultados del Análisis Funcional, los cuales indican un posible mal condicionamiento de esta matriz, otros ejemplos son analizados en [1].

En la siguiente definición U y F son espacios de funciones (Normados o Métricos) y $A : U \longrightarrow F$ es un operador el cual no es necesariamente lineal y/o continuo.

Definición 1.2.1 (Hadamard) Dado $f \in F$, el problema de hallar $u \in U$ tal que $A u = f$, se dice bien planteado si éste tiene una única solución u_0, la cual depende continuamente de sus datos. Y se dice mal planteado, si no es bien planteado. Por datos en el sentido amplio se entiende A y f y en el sentido restringido se entiende usualmente solo a f.

A continuación se hará ver que para un operador $A: U \longrightarrow F$ lineal y compacto con U y F de dimensión infinita, cuando A^{-1} existe, A^{-1} no puede ser continuo. Para mayor detalle ver [2] (Decir continuo en el caso lineal significa que manda conjuntos acotados en acotados).

Definición 1.2.2 Sean X y Y espacios lineales normados, entonces un operador lineal $A : X \longrightarrow Y$ es compacto, si y sólo si, A manda conjuntos acotados de X en conjuntos precompactos de Y.

Teorema 1.2.3 Todo operador lineal compacto es continuo.

Demostración: Sea S un conjunto acotado contenido en el dominio de A, como A es compacto, la imagen de S bajo A, A (S) es precompacto, luego A (S) es acotado, por lo tanto, A es continuo. ■

Teorema 1.2.4 Si A es un operador lineal sobre un espacio normado de dimensión finita U, entonces A es compacto.

Demostración: Como A es lineal, si U es de dimensión finita, A (U) es de dimensión finita. Como todo operador lineal sobre un espacio de dimensión finita es acotado, entonces para todo conjunto S acotado de U, A (S) es acotado. Como A(S) ⊂ A (U) y A (U) es de dimensión finita, se tiene que $\overline{A(S)}$ es compacto. luego entonces, A (S) es precompacto; por lo tanto A es compacto. ■

Observación 1.2.5. Un operador acotado no necesariamente es compacto.

Como ejemplo supóngase que U es de dimensión infinita y tomando I: U——> U el operador identidad, claramente éste es un operador acotado y ahora se probará que no es compacto.

Como U es de dimensión infinita, se puede seleccionar $x_1, x_2, x_3, \ldots, x_n, \ldots$ vectores linealmente independientes y se desea que existan vectores $y_1, y_2, y_3, \ldots, y_n, \ldots$ tales que:

$$\| y_i \| = 1 \qquad , \qquad y_i \in M_i = \text{Span}\,(x_1, x_2, \ldots, x_i)$$

$$y \qquad \| y_{i+1} - x \| \geq 1/2 \quad \forall\, x \in M_i$$

para lograr esto, se toma $y_1 = x_1 / \| x_1 \|$ y sea $M_1 = \text{Span}(y_1)$, nótese que M_1 es un subespacio de dimensión finita, por lo cual es cerrado. Como x_1 y x_2 son linealmente independientes, la

contención $M_1 \subset M_2$ es propia y aplicando el teorema de Riesz [2] existe $y_2 \in M_2$ tal que: $\| y_2 - x \| \geq 1/2 \quad \forall x \in M_1$, considerando que $M_1 \subset M_2 \subset M_3 \subset \ldots \subset M_n \subset \ldots$ con el mismo razonamiento se puede obtener $y_1, y_2, \ldots, y_n \ldots$ que satisfagan las condiciones requeridas. Tomando la sucesión $S = \{ y_i \}$ para n y m tales que $n \neq m$ se tiene que $\| y_n - y_m \| \geq 1/2$, de esta sucesión no se puede extraer una subsucesión convergente, por lo tanto S no es precompacto. Tomando la imagen de S bajo el operador identidad, es claro que $Inf\ S) = S$ no es precompacto; así queda demostrado que $A: U \longrightarrow U$ no es compacto. ■

Teorema 1.2.6. Sean U, F, Z espacios normados, si $A: U \longrightarrow F$ y $B: Z \longrightarrow U$ son operadores lineales, donde A es compacto y B acotado, entonces AB es compacto.

Demostración: Sea S un conjunto acotado de Z y considérese $(A B) (S) = A (B (S))$. Como B es acotado, $B (S)$ es un conjunto acotado en U y como A es compacto, $A (B (S))$ es precompacto en F, por lo tanto AB es compacto ■

De manera análoga si $B: F \longrightarrow Z$ es lineal acotado y $A: U \longrightarrow F$ es lineal y compacto, entonces BA es lineal compacto.

Corolario 1.2.7. Sea U de dimensión infinita y A un operador lineal compacto de U en F, si $A^{-1}: F \longrightarrow U$ existe, entonces A^{-1} no puede ser acotado.

Demostración: Supóngase que A^{-1} es acotado; entonces, por el teorema anterior $I = A A^{-1}$ es compacto, lo cual contradice a la observación anterior; en vista de esto queda demostrado el corolario. ■

Utilizando estos resultados es posible ver que la ecuación integral de Fredholm de primera especie resulta ser, bajo ciertas condiciones, un problema mal planteado a la Hadamard. En efecto, el operador integral K: $L_2[a,b] \longrightarrow L_2[c,d]$ dado por:

$$K[u](x) = \int_a^b k(x,t)u(t)dt \qquad ...(1.2)$$

donde $K(x,t)$ está en $L_2([a,b]x[c,d])$, con frecuencia continuo, resulta ser un operador lineal compacto.

Así, el problema:

$$K[u](x) = f(x)$$

resulta ser mal planteado a la Hadamard.

Por lo tanto, la ecuación integral de Fredholm de primer orden:

$$\int_a^b k(x,t)u(t)dt = f(x), \qquad c \le x \le d \qquad ...(1.3)$$

es un problema mal planteado a la Hadamard.

Tomemos el sistema de ecuaciones $K\,u = f$, que resulta de la discretización de esta ecuación integral, usualmente es mal condicionado [1].

El corolario (1.2.7), es de gran importancia en nuestro estudio; nos indica que una fuente importante de sistemas mal condicionados es la discretización de problemas inversos, a saber: Inversa Generalizada, Ecuación Retrógrada del Calor, etc., los cuales tienen gran aplicación en áreas como Estadística, Geofísica, Tomografía, etc.

I.3.- DISCRETIZACION DE LA ECUACION INTEGRAL

En la sección anterior se mencionó que la discretización de una ecuación integral de Fredholm de primer orden usualmente es mal condicionada; sin embargo, no se presenta en forma explícita esta discretización, a continuación aclararemos este detalle.

Sea $\int_a^b k(x,t)u(t)dt = f(x), \quad c \le x \le d \quad \ldots(1.3.1)$

ecuación integral de Fredholm de primer orden

donde:

$k(x, t)$: núcleo de la ecuación integral.

$u(t)$: función a estimar.

$f(x)$: lado derecho de la ecuación integral.

definimos $\{ x_0, x_1, \ldots, x_m \}$ partición para el intervalo $[c, d]$

y $\{ t_0, t_1, \ldots, t_n \}$ partición para el intervalo $[a, b]$

sustituyendo x_i en (1.3.1) se obtiene:

$$\int_a^b k(x_i,t)u(t)dt = f(x_i), \quad c \le x \le d, \quad i=0,\ldots,m \quad \ldots(1.3.2).$$

A continuación tomando una regla de cuadratura del tipo siguiente:

$$\int_a^b g(t)dt \cong \sum_{j=0}^{n} \mu_j g(t_j)$$

siendo μ_j = pesos de la cuadratura, la cuadratura Gaussiana o Newton-Cotes.

Al aplicar esta regla de cuadratura en (1.3.2) da como resultado:

$$\int_a^b k(x_i,t)u(t)dt \cong \sum_{j=0}^{n} \mu_j k(x_i,t_j)u(t_j),$$

de donde

$$\sum_{j=0}^{n} \mu_j k(x_i,t_j)u(t_j) = f(x_i), \quad i=0,1,\ldots,m,$$

representando esta última expresión el sistema de ecuaciones siguiente:

$$\begin{pmatrix} \mu_0 k(x_0,t_0)\mu_1 k(x_0,t_1)\ldots\mu_n k(x_0,t_n) \\ \mu_0 k(x_1,t_0)\mu_1 k(x_1,t_1)\ldots\mu_n k(x_1,t_n) \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ \mu_0 k(x_m,t_0)\mu_1 k(x_m,t_1)\ldots\mu_n k(x_m,t_n) \end{pmatrix} \begin{pmatrix} u(t_1) \\ u(t_2) \\ \vdots \\ u(t_m) \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_m) \end{pmatrix}$$

CAPITULO 2 BIDIAGONALIZACION INFERIOR DE LANCZOS.

En el presente capitulo se presenta el algoritmo de bidiagonalización de Lanczos, el cual tiene su origen en el cálculo de los valores singulares de A, donde $A \in \mathbb{R}^{m \times n}$ es grande y rala. Este algoritmo está relacionado con el método de Iteraciones Minimizadas dado por Lanczos [3].

En la segunda sección de este capitulo se presenta la bidiagonalización estandar (superior) de Lanczos.

En la tercera sección se presenta el algoritmo de bidiagonalización inferior de Lanczos, el cual consiste en calcular dos matrices Q y U tales que $A\,Q = U\,B$, demostrando la ortogonalidad de las matrices Q y U .

En la cuarta y última sección se observan algunas ventajas numéricas de la bidiagonalización inferior en la solución de sistemas de ecuaciones y minimos cuadrados lineales.

2.1 ANTECEDENTES.

Empecemos con el problema de encontrar la descomposición singular de una matriz, el cual consiste en lo siguiente:

Dada una matriz $A \in \mathbb{R}^{m \times n}$, encontrar V y W matrices ortogonales (i.e. $V^* V = I$ y $W^* W = I$) y $\sum \in \mathbb{R}^{m \times n}$ que:

$$A = V \sum W^*$$

con

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & 0 \\ & & \ddots & \\ & 0 & & \sigma_n \\ \hline & & 0 & \end{pmatrix}, \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$$

Golub and Kahan [5] sugieren un algoritmo particular para obtener una matriz bidiagonal con los mismos valores singulares que A. Una forma equivalente de obtener una bidiagonal con los mismos valores singulares que A es usando transformaciones de Householder, las cuales son más estables; pero su aplicación en problemas de gran escala resulta ser costosa en tiempo de operación y requieren una gran cantidad de memoria adicional. A pesar de la conocida inestabilidad numérica del algoritmo de Lanczos [12], éste resulta ser extremadamente útil en el caso de matrices grandes y ralas con una importante reducción de cómputo y almacenamiento, además cobra una mayor fuerza en el caso particular de que no se requieran las matrices V y W explícitamente.

2.2 BIDIAGONALIZACION ESTANDAR (SUPERIOR) DE LANCZOS.

Golub and Kahan [5] proponen para la primera étapa en el cálculo de la descomposición en valores singulares de una matriz A, el cálculo de la bidiagonalización por el método de Lanczos, el cual consiste en lo siguiente:

Dada $A \in \mathbb{R}^{m \times n}$, encontrar U y Q tales que:

$$U^* A Q = B ,$$

en donde, $U \in \mathbb{R}^{m \times m}$ y $Q \in \mathbb{R}^{n \times n}$ son ortogonales, esto es,

$$U^{*} U = I \qquad y \qquad Q^{*} Q = I$$

donde $B \in \mathbb{R}^{m \times n}$ es de la forma siguiente:

$$B = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & & \\ & & & \alpha_{n-1} & \beta_{n-1} \\ & & & & \alpha_n \\ \hline & & 0 & & \end{pmatrix}$$

Dado que U y Q son matrices ortogonales es claro que en Aritmética Real los valores singulares de B son los valores singulares de A.

En la práctica el algoritmo puede truncarse sin completar toda la factorización, es decir tomar unicamente k pasos ($k \leq n$), este valor de k puede ser determinado por ensayo o por algún criterio de truncamiento, obteniendo:

$$U_k^{*} A Q_k = B_k$$

donde:

$U_k \in \mathbb{R}^{m \times k}$, $Q_k \in \mathbb{R}^{n \times k}$ tales que:

$$U_k^{*} U_k = I \quad , \quad Q_k^{*} Q_k = I$$

y $B_k \in \mathbb{R}^{l \times k}$ bidiagonal; esto es

$$B_k = \begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & & \\ & & & & \beta_{k-1} \\ 0 & & & & \alpha_k \\ \hline & & 0 & & \end{pmatrix}$$

Nota 2.2.1 Los valores singulares más grandes y más chicos de B_k están cercanos a los valores singulares más grandes y más chicos de A [6]. Por ello se tiene que $K(A) \cong K(B_k)$, es decir el mal condicionamiento de la matriz A se hereda a la matriz B_k ■

El algoritmo para efectuar esta bidiagonalización truncada se deriva de las siguientes relaciones:

$$A\,Q = U\,B \qquad y \qquad A^{*}\,U = Q\,B^{*}$$

Algoritmo 2.2.2 Dados $A \in \mathbb{R}^{m\times n}$, $z_1 \in \mathbb{R}^{n}$ $(z_1 \neq 0)$ y k

tomar $q_1 = z_1 / \| z_1 \|$, $y_1 = A\,q_1$

$\alpha_1 = \| y_1 \|$ y $u_1 = y_1 / \alpha_1$

Para $i = 1,2,\ldots,k-1$

$z_{i+1} = A^{*} u_i - \alpha_i q_i$, $\beta_i = \| z_{i+1} \|$

$q_{i+1} = z_{i+1} / \beta_i$

$y_{i+1} = A\,q_{i+1} - \beta_i u_i$, $\alpha_{i+1} = \| y_{i+1} \|$

$u_{i+1} = y_{i+1} / \alpha_{i+1}$

los vectores q_i y u_i son la i-ésima columna de la matriz Q y U respectivamente.

Supóngase que la matriz A tiene M elementos diferentes de cero, entonces el costo del algoritmo es $(2M + 8n)k + M + 4n$ flops; si se compara éste con el costo de la bidiagonalización obtenida usando transformaciones de Householder $(5/2)(n+1)m + 4)n$, encontramos que existe un importante ahorro computacional.

Para establecer la conexión entre la bidiagonalización de Lanczos y el método de Iteraciones Minimizadas (es decir, mediante transformaciones ortogonales transformar una matriz S simétrica en

una matriz T tridiagonal simétrica) se tomarán en cuenta algunas definiciones y resultados.

Definición 2.2.3 Sea $S \in \mathbb{R}^{n\times n}$ simétrica y $b \in \mathbb{R}^n$ tal que $b \neq 0$, entonces la sucesión $b, Sb, S^2b, S^3b, \ldots$ recibe el nombre de sucesión de pseudo-Krilov. ■

Observación. 2.2.4 Dada $S \in \mathbb{R}^{n\times n}$ y $b \in \mathbb{R}^n$, $b \neq 0$ existe $k \leq n$, tal que $b, Sb, S^2b, \ldots, S^kb$ es una sucesión de vectores linealmente independientes, siendo la sucesión $b, Sb, S^2b, \ldots, S^{k+1}b$ linealmente dependiente. ■

Definición. 2.2.5 Sea $S \in \mathbb{R}^{n\times n}$ y $b \in \mathbb{R}^n$, la matriz $[\, Sb, S^2b, \ldots, S^kb \,]$ ($Kri(S, b, k)$) recibe el nombre de matriz de Krilov de orden k. ■

Teorema. 2.2.6 Sea $S \in \mathbb{R}^{n\times n}$ y $q_1 \in \mathbb{R}^n \neq 0$,entonces $Kri(S, q_1, k) = Q\, Kri(T, e_1, k)$, donde Q y T son las matrices obtenidas de la tridiagonalización de Lanczos $S\,Q = Q\,T$.

Demostración: ver [7]. ■

Teorema 2.2.7 La matriz $Kri(T, e_1, k)$ es una matriz triangular superior.

Demostración: ver [7]. ■

De los teoremas anteriores se puede ver que:

$$Q\, Kri(T, e_1, k)$$

es la factorización $Q\,R$ de $Kri(S, q_1, k)$.

A continuación se probará un teorema del cual se desprende la conexión de la bidiagonalización de Lanczos con el método de Iteraciones Minimizadas (Tridiagonalización de una matriz $S \in \mathbb{R}^{n\times n}$ simétrica).

Teorema 2.2.8 Sea $A \in \mathbb{R}^{m\times n}$ y $q_1 \in \mathbb{R}^n \neq 0$, entonces $(A^* A)^j q_1 = Q\,(B^* B)^j e_1$, donde B y Q son las matrices

obtenidas de la bidiagonalización $U^* A Q = B$.

Demostración. Por inducción sobre j.

Sea $U^* A Q = B$ la bidiagonalización inferior de la matriz A, de esta se obtienen las relaciones siguientes:

$$A Q = U B \qquad y \qquad A^* U = Q B^*$$

Sea k = 1, entonces $A q_1 = A Q e_1$; multiplicando por A^* se obtiene que:

$$A^* (A q_1) = (A_1^* A) q_1 = A^* U B e_1 = Q (B^* B) e_1;$$

por lo tanto, $(A^* A) q_1 = Q (B^* B) e_1$.

Supóngase que el teorema es valido para k = j , es decir:

$$(A^* A)^j q_1 = Q (B^* B)^j e_1 \qquad (2.2.1)$$

Por demostrar que es valido para k = j+1, esto es:

$$(A^* A)^{j+1} q_1 = Q (B^* B)^{j+1} e_1 .$$

Multiplicando (2.2.1) primeramente por A, se obtiene:

$$A (A^* A)^j q_1 = A Q (B^* B)^j e_1 = U B (B^* B)^j e_1, ,$$

y posteriormente multiplicando por A^*:

$$A^* A (A^* A)^j q_1 = A^* U B (B^* B)^j e_1 = Q B^* B (B^* B)^j e_1$$

Por lo tanto:

$$(A^* A)^{j+1} q_1 = Q (B^* B)^{j+1} e_1 \qquad \blacksquare$$

Corolario 2.2.9 Si Krif(A^*A, q_1, k) es la matriz de Krilov generada por $A^* A$ y q_1, entonces Q Krif($B^* B$, e_1, k) es la factorización Q R de Krif($A^* A$, q_1, k).

Demostración: Usando el teorema anterior

$$Krif(A^* A, q_1, k) = Q Krif(B^* B, e_1, k) ,$$

como $B^* B$ es una matriz tridiagonal simétrica, al aplicar el teorema (2.2.7), Krif($B^* B$, e_1, k) resulta ser una matriz triangular superior, luego entonces Q Krif($B^* B$, e_1, k) es la

factorización Q R de Krif $A^* A, q_1, k$].

Paige [8] obtiene otra forma de ver esta relación y es de la manera siguiente:

Sean $W = [w_1, w_2, \ldots, w_{2k}] = \begin{pmatrix} u_1 & 0 & u_2 & 0 & \ldots & u_k & 0 \\ 0 & v_1 & 0 & v_2 & \ldots & 0 & v_k \end{pmatrix}$,

$$S = \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \quad \text{y} \quad T = \begin{pmatrix} 0 & \alpha_1 & & & & \\ \alpha_1 & 0 & \beta_2 & & 0 & \\ & \beta_2 & 0 & \alpha_2 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & \ddots & \\ & & & & \alpha_k & 0 \end{pmatrix}$$

obsérvese que:

$$S W = W T + \beta_{k+1} w_{2k} e_{2k}^t ,$$

$$W^* W = I_{2k} \quad \text{y} \quad W^* w_{2k+1} = 0$$

Esto es el resultado de aplicar 2k étapas del método de Iteraciones Minimizadas a S con vector inicial w_1.

De todo lo anterior se observa que la bidiagonalización de Lanczos está intimamente relacionada con la tridiagonalización de $A^* A$.

2.3 BIDIAGONALIZACION INFERIOR DE LANCZOS

Existe otra forma alternativa de efectuar una bidiagonalización de la matriz A, la cual fue dada por Paige [8] de la manera siguiente:

Dada $A \in \mathbb{R}^{mxn}$, obtener $U \in \mathbb{R}^{mxm}$ y $Q \in \mathbb{R}^{nxn}$ matrices ortogonales, tales que:

$$U^* A Q = B$$

donde

$$U^{*} U = I \quad , \quad Q^{*} Q = I$$

con

$$B = \left(\begin{array}{ccccc} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & 0 & \\ & \beta_3 & \alpha_3 & & \\ & & \ddots & \ddots & \\ & 0 & & \beta_n & \alpha_n \\ \hline & & 0 & & \end{array}\right)$$

Esta factorización se puede obtener mediante transformaciones de Householder; pero usando la idea de Lanczos [3] se aprovecha la poca densidad de la matriz A, la idea es la siguiente:

Dado u_1 tal que $\| u_1 \| = 1$, sean $U = \{u_1, u_2, \ldots, u_n\}$ y $Q = \{q_1, q_2, \ldots, q_n\}$ matrices cuyas columnas son $u_1, \ldots u_n$ y $q_1, q_2, \ldots, q_n$ respectivamente.

Utilizando las identidades $A^{*} U = Q B^{*}$ y $A Q = U B$ es decir:

$$A^{*} [u_1, u_2, \ldots, u_n] = [q_1, q_2, \ldots, q_n] \left(\begin{array}{ccccc} \alpha_1 & \beta_2 & & & \\ & \alpha_2 & \beta_3 & 0 & \\ & & \ddots & \ddots & \\ & 0 & & \ddots & \beta_n \\ & & & & \alpha_n \end{array}\right) \quad \text{y}$$

$$A [q_1, q_2, \ldots, q_n] = [u_1, u_2, \ldots, u_n] \left(\begin{array}{ccccc} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & 0 & \\ & \ddots & \ddots & & \\ & 0 & & \beta_n & \alpha_n \end{array}\right)$$

se tiene que:

$$A^{*} u_1 = \alpha_1 q_1 \quad \rightarrow \quad q_1 = A^{*} u_1 / \alpha_1$$

y que:

$$A q_1 = \alpha_1 u_1 + \beta_2 u_2 \quad \Rightarrow \quad u_2 = (A q_1 - \alpha_1 u_1) / \beta_2$$

en general se tiene que:

$$A q_i = \alpha_i u_i + \beta_{i+1} u_{i+1} \quad \Rightarrow \quad u_{i+1} = (A q_i - \alpha_i u_i) / \beta_{i+1}$$

y que:

$$A^{*} u_{i+1} = \beta_{i+1} q_i + \alpha_{i+1} q_{i+1} \Rightarrow \quad q_{i+1} = (A u_{i+1} - \beta_{i+1} q_i) / \alpha_{i+1}$$

Esto es las q_{i+1}, u_{i+1}, α_{i+1} y β_{i+1} se pueden obtener de manera iterativa conociendo las q_i, u_i, α_i y β_i anteriores.

El algoritmo termina en el momento de encontrar $\alpha_i = 0$ ó $\beta_i = 0$ y en a lo más n iteraciones, el algoritmo es el siguiente:

Algorithm 2.3.1. Dada $A \in \mathbb{R}^{m \times n}$ y $u_1 \in \mathbb{R}^{(m,1)}$ unitario, tomar

$$y_1 = A^{*} u_1 , \qquad \alpha_1 = \| y_1 \| , \qquad q_1 = y_1 / \alpha_1$$

mientras $\alpha_{i+1} \neq 0$ y $\beta_{i+1} \neq 0$ y $i+1 \leq n$

$$z_{i+1} = A q_i - \alpha_i u_i$$

$$\beta_{i+1} = \| z_{i+1} \|$$

$$u_{i+1} = z_{i+1} / \beta_{i+1}$$

$$y_{i+1} = A^{*} u_{i+1} - \beta_{i+1} q_i$$

$$\alpha_{i+1} = \| y_{i+1} \|$$

$$q_{i+1} = y_{i+1} / \alpha_{i+1}$$

Nuevamente supóngase que A tiene M elementos diferentes de cero, entonces el costo de este algoritmo es el mismo que el algoritmo de la bidiagonalización superior((2M + 8n) k + M + 4n) flops).

Obsérvese que los escalares α_i y β_i son no negativos, ya que estos fueron calculados de tal forma que $\alpha_i = \| y_i \|$ y $\beta_i = \| z_i \|$ además $\| q_i \| = 1$ y $\| u_i \| = 1$.

Usando un poco de algebra en las relaciones $A Q = U B$ y $A^* U = Q B^*$ se obtienen 3 relaciones importantes en el análisis de la k-ésima etapa de este algoritmo, en base a las cuales se demostrará que Q y U son de columnas ortonormales.

A partir de la primera relación $A Q = U B$, es decir:

$$A [\, q_1, q_2, \ldots, q_k \mid q_{k+1}, \ldots, q_n \,] =$$

$$[u_1, u_2, \ldots, u_k | u_{k+1}, \ldots, u_m] \left(\begin{array}{c|c} \begin{array}{ccccc} \alpha_1 & & & & \\ \beta_2 & \alpha_2 & & 0 & \\ & \beta_3 & \alpha_3 & & \\ 0 & & \ddots & & \\ & & \beta_k & \alpha_k & \end{array} & 0 \\ \hline \begin{array}{ccc} & & \beta_{k+1} \end{array} & \begin{array}{ccc} \alpha_{k+1} & & \\ & \ddots & \\ 0 & & \end{array} \\ \hline & \begin{array}{cc} \ddots & \\ \beta_n & \alpha_n \end{array} \end{array}\right)$$

así, tomando $Q = [\, Q_k \mid Q_k^c \,]$ y $U = [\, U_k \mid U_k^c \,]$ respectivamente, se tiene que

$$A [\, Q_k , Q_k^c \,] = [\, U_k , U_k^c \,] \left(\begin{array}{c|c} B_k & 0 \\ \hline \begin{array}{ccc} \beta_{k+1} e_k^* & \end{array} & \begin{array}{ccc} \alpha_{k+1} & & \\ & \ddots & \\ 0 & & \end{array} \\ \hline & \begin{array}{cc} \ddots & \\ \beta_n & \alpha_n \end{array} \end{array}\right)$$

de donde se obtiene la relación

$$A Q_k = U_k B_k + \beta_{k+1} u_{k+1} e_k^* \quad \ldots(2.3.1).$$

Tomando la segunda relación $A^* U = Q B^*$, es decir:

$$A^{*}\,[U_k \mid U_k^{c}] \;=\; [Q_k \mid Q_k^{c}] \left(\begin{array}{ccc|c} \alpha_1 & \beta_2 & 0 & \\ & \alpha_2 & \beta_3 & 0 \\ & & \ddots & \\ 0 & & \alpha_k & \beta_{k+1} \\ \hline & & & \ddots \\ & 0 & & \begin{array}{cc} \beta_n & \\ 0 & \alpha_n \end{array} \end{array}\right)$$

se obtiene que

$$A^{*}\,U_k = Q_k\,B_k^{*} \qquad \ldots\ldots(2.3.2),$$

transponiendo esta relación se obtiene

$$U_k^{*}\,A = B_k\,Q_k^{*}$$

y multiplicando por Q_k en ambos lados de esta última da como resultado:

$$U_k^{*}\,A\,Q_k = B_k\,Q_k^{*}\,Q_k \qquad \ldots\ldots(2.3.3).$$

Si $\alpha_{k+1} \neq 0$ una etapa más de la bidiagonalización puede ser calculada con $\overline{U_{k+1}} \equiv [\,U_k,\, u_{k+1}\,]$ y $\overline{B_{k+1}}^{*} \equiv [\,B_k^{*},\, \beta_{k+1}\, e_{k+1}^{*}\,]$ obteniendo

i) $A^{*}\,\overline{U}_{k+1} = Q_k\,\overline{B}_{k+1}^{*} + \alpha_{k+1}\,q_{k+1}\,e_{k+1}^{*}$

ii) $A\,Q_k = \overline{U}_{k+1}\,\overline{B}_{k+1}$

de i) se sigue que:

$$Q_k^{*}\,A^{*}\,\overline{U}_{k+1} = Q_k^{*}\,Q_k\,\overline{B}_{k+1}^{*} + \alpha_{k+1}\,Q_k^{*}\,q_{k+1}\,e_{k+1}^{*} \qquad \ldots(2.3.4)$$

y de ii)

$$Q_k^{*}\,A^{*}\,\overline{U}_{k+1} = \overline{B}_{k+1}^{*}\,\overline{U}_{k+1}^{*}\,\overline{U}_{k+1}.$$

A continuación se probará que las matrices Q y U obtenidas en esta bidiagonalización son de columnas ortonormales.

En efecto, tomando en cuenta las relaciones anteriores (2.3.1) y (2.3.4) se tiene que:

$$U_k^{*}\,A\,Q_k = U_k^{*}\,U_k\,B_k + \beta_{k+1}\,U_k^{*}\,u_{k+1}\,e_k^{*} = B_k\,Q_k^{*}\,Q_k$$

y

$$Q_k^* A^* \bar{U}_{k+1} = Q_k^* Q_k \bar{B}_{k+1}^* + \alpha_{k+1} Q_k^* q_{k+1} e_{k+1}^* = \bar{B}_{k+1}^* \bar{U}_{k+1}^* \bar{U}_{k+1}$$

suponiendo que $\alpha_{k+1} \neq 0$, $\beta_{k+1} \neq 0$ y que $Q_k^* Q_k = I_k$, $U_k^* U_k = I_k$, entonces de la primera relación se desprende que $U_k^* u_{k+1} = 0$ y de la segunda se desprende que $Q_k^* q_{k+1} = 0$.

2.4. PROPIEDADES NUMERICAS DE LA BIDIAGONALIZACION

En esta sección se ven algunas ventajas numéricas de la bidiagonalización inferior de Lanczos en comparación con la superior (estandar) al resolver un sistema de ecuaciones o un problema de mínimos cuadrados lineales. Como se observó en la sección (2.3), la bidiagonalización inferior de Lanczos puede truncarse en k (≤ n) iteraciones, esto es, se tiene una bidiagonalización truncada

$$A Q_k = U_k B_k + \beta_{k+1} u_{k+1} e_k^*.$$

Supóngase que se desea resolver el sistema $A x = b$ con $A \in \mathbb{R}^{m \times n}$ y $b \in \mathbb{R}^m$.

Sean $U_k^* A Q_k = B_k$, $k \leq n$, la bidiagonalización truncada inferior de Lanczos de $A \in$ y $\in \mathbb{R}^m$ tal que:

$$Q_k y = x_k \qquad (2.4.1),$$

entonces

$$A x_k = b$$

es equivalente a

$$A Q_k y = b,$$

multiplicando por U_k^* se obtiene que:

$$U_k^* A Q_k y = U_k^* b,$$

por lo tanto

$$B_k\, y = U_k^{*}\, b \quad .$$

Ahora, si se toma $u_1 = b \,/\, \| b \|$ en el algoritmo (2.3.1), entonces

$$U_k^{*}\, b = \begin{pmatrix} \beta_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad ,$$

de donde el vector y puede obtenerse en la forma siguiente; (sustitución hacia adelante)

$$\begin{aligned} & y_1 = \beta_1 \,/\, \alpha_1 \\ & \text{para} \quad i = 2,\ 3,\ \ldots,\ k \\ & \quad y_i = - \beta_i\, y_{i-1} \,/\, \alpha_i \quad , \end{aligned}$$

luego sustituyendo las y_i en (2.4.1) tenemos

$$x_k = y_1\, q_1 + y_2\, q_2 + \ldots + y_k\, q_k = Q_k\, y \ .$$

Todo lo anterior se puede llevar a cabo sin necesidad de almacenar las matrices Q_k y U_k en memoria, usando el algoritmo siguiente:

Algoritmo 2.4.1 dada $A \in \mathbb{R}^{m \times n}$ y $b \in \mathbb{R}^m$

tomar $\beta = \| b \|$, $u = b \,/\, \beta$, $q = A^{*}\, u$,

$\alpha = \| q \|$, $q = q \,/\, \alpha$, $y = \beta \,/\, \alpha$, $x = y\, q$

mientras $\beta \neq 0$ y $\alpha \neq 0$

- $u = A\, q - \alpha\, u$, $\beta = \| u \|$
- $u = u \,/\, \beta$
- $q = A^{*}\, u - \beta\, q$ $\alpha = \| q \|$
- $q = q \,/\, \alpha$
- $y = - \beta\, y \,/\, \alpha$
- $x = x + y\, q$

Ventajas sobre la bidiagonalización superior

En el caso de la bidiagonalización superior el sistema

$$B_k \quad y = U_k^* b$$

no puede ser resuelto utilizando la información de la etapa anterior, ya que se requiere realizar una sustitución hacia atrás, lo cual obliga a mantener en memoria la matriz Q_k y realizar el producto de $U_k^* b$ de forma explícita; además, si se utiliza bidiagonalización superior, entonces $U_k^* b$ es diferente para cada valor de k.

Es muy importante que el lector observe que en comparación a la bidiagonalización superior, en la bidiagonalización inferior Q_k no se necesita mantener en memoria y que $U_k^* b = \beta_1 e_1^*$ para todo valor de k, como se desprende del algoritmo (2.4.1).

Ahora, supóngase que se desea resolver el problema de mínimos de cuadrados siguiente:

$$\min_x \| A x - b \|_2^2 \qquad \ldots(2.4.2)$$

cuyas ecuaciones normales son

$$A^* A x = A^* b \qquad \ldots\ldots(2.4.3).$$

Sea $A Q_k \cong U_k B_k$ la bidiagonalización truncada inferior de Lanczos de A y sea $x_k = Q_k y$ la solución aproximada del problema (2.4.3); sustituyendo esta última en (2.4.3) se obtiene:

$$A^* A x_k \cong A^* b \quad \Leftrightarrow \quad A^* A Q_k y \cong A^* b$$

Utilizando la relación $A Q_k \cong U_k B_k$ en esta última expresión se obtiene:

$$A^* U_k B_k y \cong A^* b$$

como $A^* U_k = Q_k B_k^*$, la última expresión es equivalente a:

$$Q_k B_k^* B_k y = A^* b$$

multiplicando por Q^* da como resultado

$$Q_k^* Q_k B_k^* B_k y = Q_k^* A^* b = (A Q_k)^* b$$

recordando que Q_k tiene columnas ortogonales y que $A Q_k = U_k B_k$ se obtiene la relación:

$$B_k^* B_k y = (U_k B_k)^* b$$

la cual es equivalente a:

$$B_k^* B_k y = B_k^* U_k^* b \qquad \ldots\ldots (2.4.4)$$

observando que $U_k^* b = \beta_1 e_1^*$ se obtiene:

$$B_k^* B_k y = B_k^* c,$$

donde $c = \beta_1 e_1^*$, esta última expresión son las ecuaciones normales del problema de mínimos cuadrados siguiente:

$$\min_y \| B_k y - c \|_2^2 \qquad \ldots\ldots (2.4.5),$$

es decir, encontrar una solución aproximada del problema original es equivalente a resolver otro problema de mínimos cuadrados más sencillo.

Un método para resolver (2.4.5) es mediante la factorización QR de B mediante el uso de rotaciones de Givens, la idea es la siguiente:

Dada $[\, B_k \mid \beta_1 e_1^* \,]$ matriz aumentada del problema (2.4.5), tomar $G_{1,2} [\, B_k \mid \beta_1 e_1^* \,]$ de tal forma que:

$$\left(\begin{array}{cc|c} c_1 & -s_1 & \\ s_1 & c_1 & 0 \\ \hline & & \\ 0 & & I \end{array}\right) \left(\begin{array}{ccc|c} \alpha_1 & & & \beta_1 \\ \beta_2 & \alpha_2 & 0 & 0 \\ & \beta_3 & \alpha_3 & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & \beta_k & \alpha_k \\ \hline & 0 & & 0 \end{array}\right) = \left(\begin{array}{ccc|c} \rho_1 & 0_1 & & \phi_1 \\ & \rho_2 & 0 & \phi_2 \\ & \beta_3 & \alpha_3 & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & \beta_k & \alpha_k \\ \hline & 0 & & 0 \end{array}\right),$$

posteriormente multiplicando por $G_{2,3}$ el resultado anterior se

obtiene:

$$\left(\begin{array}{c|c|c} 1 & & 0 \\ \hline & \begin{matrix} c_2 & -s_2 \\ s_3 & c_3 \end{matrix} & \\ \hline 0 & & 1 \end{array}\right)\left(\begin{array}{c|c|c} \rho_1 & \theta_1 & \\ & \rho_2 & 0 \\ \hline & \beta_3 & \alpha_3 \\ & & \ddots \\ 0 & & \beta_k \quad \alpha_k \\ \hline & 0 & \end{array}\right)\left(\begin{array}{c} \phi_1 \\ \bar{\phi}_2 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{array}\right) = \left(\begin{array}{c|c|c} \rho_1 & \theta_1 & \\ & \rho_2 & \theta_2 \quad 0 \\ \hline & & \rho_3 \\ & & \ddots \\ 0 & & \beta_k \quad \alpha_k \\ \hline & 0 & \end{array}\right)\left(\begin{array}{c} \phi_1 \\ \phi_2 \\ \bar{\phi}_3 \\ \vdots \\ \vdots \\ 0 \end{array}\right)$$

y en forma análoga se multiplica por $G_{3,4}$, ... $G_{k-1,k}$ hasta obtener:

$$\left(\begin{array}{c|c|c} 1 & & 0 \\ \hline & \begin{matrix} c_{k-1} & -s_{k-1} \\ s_{k-1} & c_{k-1} \end{matrix} & \\ \hline 0 & & 1 \end{array}\right)\left(\begin{array}{c|c|c} \rho_1 & \theta_1 & \\ & \rho_2 \theta_2 & 0 \\ \hline & & \rho_{k-1} \\ & & \\ 0 & & \beta_k \quad \alpha_k \\ \hline & 0 & \end{array}\right)\left(\begin{array}{c} \phi_1 \\ \phi_2 \\ \vdots \\ \bar{\phi}_k \\ \bar{\phi}_{k-1} \\ 0 \end{array}\right) = \left(\begin{array}{c|c|c} \rho_1 & \theta_1 & \\ & \rho_2 \quad \theta_2 & 0 \\ \hline & & \rho_{k-1} \quad \theta_{k-1} \\ 0 & & \rho_k \\ \hline & 0 & \end{array}\right)\left(\begin{array}{c} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_k \\ \bar{\phi}_{k+1} \end{array}\right)$$

Sea $G^* = G_{k-1,k} \ldots G_{1,2}$, como G^* es el producto de matrices ortogonales, entonces G^* es ortogonal.

La matriz $R_k = G^* B_k$ resulta ser una matriz triangular superior, de donde $B_k = G R_k$ es la factorización QR de B_k; con lo cual el problema (2.4.5) es equivalente a resolver el sistema siguiente:

$$\begin{pmatrix} \rho_1 & \theta_1 & & \\ & \rho_2 & \theta_2 & 0 \\ & & \ddots & \\ & & \rho_{k-1} & \theta_{k-1} \\ 0 & & & \rho_k \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_k \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \\ \phi_k \end{pmatrix} .$$

La solución de este sistema puede ser llevada a cabo de manera iterativa tomando

$$y_j = R_j^{-1} \phi_j = W_j \phi_j ,$$

las columnas de $W_j = [\, w_1, \ldots, w_j \,]$ pueden ser calculadas usando el esquema siguiente: ($W_j R_j = I_j$)

$$\left(\begin{array}{c|c|c|c|c} w_1 & w_2 & \cdots & w_j \end{array} \right) \begin{pmatrix} \rho_1 & \theta_1 & & \\ & \rho_2 & \theta_2 & 0 \\ & & \ddots & \ddots \\ & & & \rho_{k-1} \; \theta_{k-1} \\ 0 & & & \rho_k \end{pmatrix} = \left(\begin{array}{c|c|c|c|c} e_1 & e_2 & \cdots & e_j \end{array} \right)$$

a partir de este se observa que:

$$\theta_{i-1} \; w_{i-1} + \rho_i \; w_i = e_i$$

resolviendo para w_i se obtiene (tomando $w_0 = 0$):

$$w_i = (1 / \rho_i) (e_i - \theta_{i-1} \; w_{i-1}) \quad \ldots.(2.4.6).$$

Este método tiene dos ventajas numéricas importantes a saber: la estimación del residual y la condición numérica de A.

Sea $r_j = b - A x_j$ el residual para la solución j-ésima del problema (2.4.2), recordando que

$$x_j = Q_j y_j \quad , \quad U_j^* b = \beta_1 e_1 \quad y \quad U_j^* A Q_j = B_j$$

al multiplicar el residual anterior por U_j^* y utilizando las relaciones recordadas se obtiene la expresión siguiente:

$$t_j = \beta_1 e_1 - B_j y_j \; , \quad \text{donde} \quad t_j = U_j^* r_j ;$$

es decir, la norma del residual original coincide con

$$\| t_j \| = \| \beta_1 e_1 - B_j y_j \|$$

y la norma del residual de (2.4.5) resulta ser

$$\| t_j \| = \| \bar{\phi}_{j+1} Q_j e_{j+1} \| = \bar{\phi}_{j+1}$$

donde

$$\bar{\phi}_{j+1} = \beta_1 s_1 \cdots s_j \; .$$

A partir de $U_k^* A Q_k = B_k$ se tiene que:

$$B_k^* B_k = Q_k^* A^* A Q_k \; ,$$

utilizando el teorema minimax de Courant-Ficher [2], los valores singulares de $B_k^* B_k$ están intercalados por los de $A^* A$ y acotados superior e inferiormente por el mayor y menor respectivamente de estos. Lo mismo podemos decir de los de B_k y

A obteniendo:

$$\| B_k \| \leq \| A \|$$

en forma análoga, también implica que:

$$\| R_k^{-1} \| = \| B_k^{+} \| \leq \| A^{+} \|$$

por lo tanto

$$1 \leq \| B_k \| \, \| B_k^{+} \| \leq \| A \| \, \| A^{+} \| \qquad (2.4.7)$$

la norma $\| B_j \|_F$ puede ser calculada directamente, ya que

$$\| B_j \|_F = \alpha_1 + \sum_{i=2}^{j} \alpha_i^2 + \beta_1^2 .$$

Para estimar $\| B_j^{+} \|$ obsérvese la identidad siguiente:

$$B_j^{+} B_j = R_j^{-1} G_j^{-1} G_j R_j = R_j^{-1} R_j ,$$

de donde se obtiene $\| B_j^{+} \|_F = \| R_j^{-1} \|$ y utilizando la identidad $R_j^{-1} = W_j$ da como resultado:

$\| B_j^{+} \|_F = (\sum_{i=1}^{j} \| w_i \|^2)^{1/2}$, la cual puede ser fácilmente calculada a partir de (2.4.6).

La subestimación del condicional numérico de A dada por (2.4.7) en términos de B_k, ésta será utilizada como criterio de truncamiento del algoritmo de bidiagonalización, véase secc 4.6 y secc 5.2.

Una de las características importantes de una matriz es su condición numérica K (A). El que una matriz sea mal condicionada indica que ésta es muy susceptible a amplificar los errores por redondeo en la resolución de sistemas de ecuaciones, de problemas de mínimo de cuadrados, etc. Es por esta causa que requieren de un tratamiento adecuado, con el fin de amortiguar este mal condicionamiento.

En el primera sección de este capítulo se presenta el método de Regularización Directa (truncamiento), el cual es analizado en forma completa debido a la "forma sencilla" de llevarla a cabo.

En la segunda sección de este capítulo se presenta la regularización de Tijonov, la cual es un método más global que el anterior. Esta regularización requiere de la estimación de un parámetro γ conocido como parámetro de regularización de Tijonov. Nuestra forma de cómo escoger este parámetro será tratado en la tercera sección de este capítulo. Otros métodos o variantes de éstos se discuten en [4].

3.1. METODO DE REGULARIZACION DIRECTA

La solución numérica de un sistema de ecuaciones lineales de la forma

$$A\,x = b \qquad \text{.....(3.1.1.)}$$

puede ser analizada en términos de la decomposición singular de la matriz A (SVD (A)), esto es:

Dada $A \in \mathbb{R}^{m\times n}$, $m \geq n$, existen $W \in \mathbb{R}^{m\times m}$ y $V \in \mathbb{R}^{n\times n}$

ortogonales tales que:

$$A = W\ D\ V^{*} = \sum_{i=1}^{n} \sigma_i w_i v_i^{*}$$

donde $W = [w_1, w_2, \ldots, w_m]$, $V = [v_1, v_2, \ldots, v_n]$,

$$D = \text{diag}\ [\sigma_1, \sigma_2, \ldots, \sigma_n],$$

$W^{*} W = I_m$, $V^{*} V = I_n$ y $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_n = 0$.

A la matriz $D^{+} = \text{diag}\ [1/\sigma_1, 1/\sigma_2, \ldots, 1/\sigma_k, 0, \ldots, 0]$ se le conoce como inversa generalizada Moore-Penrose de D [19]; se define la inversa generalizada Moore-Penrose de A como:

$$A^{+} = V\ D^{+}\ W^{*}$$

y en términos de ésta, la solución generalizada Moore-Penrose de (3.1.1) está dada por:

$$x^{+} = A^{+}\ b = \sum_{i=1}^{k} \left(\left(w_i^{*}\ b \right) / \sigma_i \right) v_i \qquad \ldots\ldots (3.1.2).$$

Si la matriz es mal condicionada, entonces σ_i decrece rápidamente a cero conforme crece i; a partir de (3.1.2) es fácil ver que a pequeñas modificaciones de b estas pueden generar grandes cambios en la solución x.

El método de regularización directa consiste en truncar la expresión (3.1.2) en $r \leq k$ términos, conocido como "Regularización Directa" o "Truncamiento de la Descomposición Singular", esto es, tomar:

$$x^{(r)} = \sum_{i=1}^{r} \left(\left(w_i^{*}\ b \right) / \sigma_i \right) v_i$$

y cuya norma residual es:

$$\| b - A\ x^{(r)} \| = \sum_{i=r+1}^{k} \left(w_i^{*}\ b \right)^2$$

El fundamento de este procedimiento descansa en los dos

hechos siguientes:

i) $\| A - A' \|_2 = \sigma_{r+1}$ ya que $A - A' = \sum_{i=r+1}^{k} \sigma_i w_i v_i^*$

ii) $K_2(A') \leq K_2(A)$ ya que $K_2(A) = \sigma_1/\sigma_k$ y $K_2(A') = \sigma_1/\sigma_r$

3.2. REGULARIZACION DE TIJONOV.

Un método, el cuál ha tenido una gran aceptación en los últimos años, es el método "Regularización de Tijonov". Esto se debe a su mayor generalidad en comparación con la regularización directa.

Para ello considérese la funcional regularizante

$$J (x_\gamma ; b) = \| b - A x_\gamma \|_2^2 + \gamma \| x_\gamma \|_S^2 \qquad (3.2.1)$$

donde $\| x \|_S^2 = x^* S x$ con S simétrica y positiva definida.

El tomar $S = I$ tal vez no sea un elección óptima; en [1] se propone tomar S como la matriz que resulta de discretizar por diferencias divididas la norma de Sobolev:

$$\|u\|_{m,2}^2 = \int_b^a (| u^{(m)} (t)|^2 + \alpha | u (t) |^2) dt \quad \text{con } m = 0,1,2 \quad (3.2.6),$$

para el caso de $m = 0$ se toma $\alpha = 0$, para $m = 1,2$ se toma $1 \gg \alpha > 0$; esto es, para $m = 0$ se obtiene que:

$$\|u\|_{m,2}^2 = \int_b^a | u (t) |^2 dt \quad .$$

Utilizando la regla de cuadratura del rectángulo compuesto con $h = (b - a) / n$ tenemos que:

$$\int_b^a | u (t) |^2 dt \cong h \sum_{i=1}^{n} u_i^2$$

$$h\,(u_1, u_2, \ldots, u_n)^{*} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = h\,u^{*}\,S\,u$$

donde $t_i = a + h * i$ y $u(I) = u(t_i)$.

Obsérvese que para $m = 0$ se tiene el caso particular de $S = hI$.

Para $m = 1$ la expresión (3.2.6) toma la forma siguiente:

$$\int_b^a \left\{ \left| u^{(1)}(t) \right|^2 + \alpha \left| u(t) \right|^2 \right\} dt$$

$$\cong \int_b^a \left| u^{(1)}(t) \right|^2 dt + \alpha\, h\, u^{*}\, I\, u$$

utilizando diferencias divididas hacia adelante y nuevamente la regla de cuadratura del rectángulo compuesto se obtiene:

$$h \sum_{i=1}^{n} \frac{(u_{i+1} - u_i)^2}{h^2} + \alpha\, h\, u^{*}\, I\, u$$

$$= \frac{1}{h} \sum_{i=1}^{n} (u_{i+1} - u_i)^2 + \alpha\, u^{*}\, hI\, u$$

$$= \alpha\, h\, u^{*}\, I\, u + h^{-1}\, u^{*}\, L\, L^{*}\, u = u^{*}\, S\, u,$$

en donde

$$S = \alpha\, hI + h^{-1}\, L\, L^{*}$$

y

$$L^{*} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \\ & & & & 1 \end{pmatrix} \quad \blacksquare$$

Para $m \geq 2$ se procede en forma totalmente análoga.

Si x_γ minimiza $J(x_\gamma ; b)$ dada por (3.2.1), entonces tomando en cuenta que:

$$\| b - A x_\gamma \|_2^2 + \gamma \| x_\gamma \|_S^2 = (b - A x_\gamma)^* (b - A x_\gamma) + \gamma x_\gamma^* S x_\gamma$$

derivando con respecto a x_γ e igualando a cero se obtiene que:

$$\gamma S x_\gamma - A^* (b - A x_\gamma) = 0$$

o bien, que x_γ es la solución del sistema:

$$(A^* A + \gamma S) x = A^* b \qquad (3.2.2).$$

Esta última expresión son las ecuaciones normales del problema:

$$\underset{x_\gamma}{\text{Min}} \left\| \begin{pmatrix} A \\ \cdots\cdots \\ \sqrt{\gamma}\, R^* \end{pmatrix} x_\gamma - \begin{pmatrix} b \\ \cdots \\ 0 \end{pmatrix} \right\|_2^2 \qquad (3.2.3),$$

siendo R la matriz obtenida de la factorización de Cholesky de S.

Realizando un poco de algebra el sistema (3.2.3) puede ser simplificado de la manera siguiente:

Sea $S = R^* R$ la factorización de Cholesky de S, se tiene que:

$$(A^* A + \gamma S) x = A^* b \quad \Leftrightarrow \quad (A^* A + \gamma R^* R) x = A^* b$$

$$\Leftrightarrow \quad (R^* R^{-*} A^* A R^{-1} R + \gamma R^* R) x = A^* b$$

$$\Leftrightarrow \quad R^* (R^{-*} A^* A R^{-1} + \gamma I) R x = A^* b$$

$$\Leftrightarrow \quad (R^{-*} A^* A R^{-1} + \gamma I) R x = R^{-*} A^* b.$$

Ahora, si se toma $X = A R^{-1}$ y $y = R x$, entonces

$$(A^* A + \gamma S) x = A^* b$$

se reescribe como sigue:

$$(X^* X + \gamma I) y = X^* b \qquad (3.2.4)$$

lo que corresponde a las ecuaciones normales del problema de mínimo de cuadrados siguiente:

$$\underset{y_\gamma}{\text{Min}} \left\| \begin{pmatrix} A \\ \cdots\cdots \\ \sqrt{\gamma}\ I \end{pmatrix} y_\gamma - \begin{pmatrix} b \\ \cdots \\ 0 \end{pmatrix} \right\|_2^2 \qquad (3.2.5),$$

Esta reducción del problema de mínimo de cuadrados (3.2.3) al (3.2.5) es de gran relevancia práctica y juega un papel clave en el desarrollo de nuestro algoritmo.

La flexibilidad de este método se ejemplifica claramente en el caso de $S = I$, es decir, al tomar el problema

$$(A^* A + \gamma I) x_\gamma = A^* b \qquad (3.2.6)$$

sea $V \sum W^* = A$ la descomposición singular de A, con $\sum = \text{diag}(\sigma_1 , \sigma_2 , \ldots , \sigma_r , 0, \ldots, 0)$ (siendo r = rango de A), al sustituir y simplificar en (3.2.6) se obtiene:

$$(W \Sigma^* \Sigma W^* + \gamma I) x_\gamma = W \Sigma V^* b \qquad (3.2.7),$$

como $W W^* = I$, la expresión (3.2.6) es equivalente a:

$$(W \Sigma^* \Sigma W^* + \gamma W W^*) x_\gamma = W \Sigma V^* b,$$

factorizando W y W^* de esta última expresión da como resultado:

$$W (\Sigma^* \Sigma + \gamma I) W^* x_\gamma = W \sum V^* b,$$

resolviendo para x_γ se obtiene:

$$x_\gamma = \sum_{i=1}^{r} (\sigma_i / (\sigma_i^2 + \gamma)) z_i w_i$$

donde w_i y v_i son las i-ésimas columnas de W y V respectivamente y $z_i = v_i^* b$.

Ahora el condicional numérico de (3.2.6) resulta ser:

$$(\sigma_1^2 + \gamma) / (\sigma_r^2 + \gamma) ;$$

Desde un punto de vista numérico este procedimiento de regularización nos conduce al sistema (3.2.6) para el cual

$$K_2(A^* A + \gamma I) = \frac{\sigma_1^2 + \gamma}{\sigma_r^2 + \gamma} < K_2(A) = \sigma_1/\sigma_r ,$$

siempre y cuando $\sigma_1 \sigma_r < 2$, esto es, con un condicional numerico menor que el correspondiente al problema de mínimo de cuadrados clásico

$$\min_x \| A x - b \|_2^2 .$$

3.3. ELECCION DEL PARAMETRO DE REGULARIZACION

El método de regularización de Tijonov no dice que valor tomar para el parámetro γ. Existen varias estrategias para esto como se puede ver en [4].

En nuestro caso particular se toma el principio de optimalidad y pseudo-optimalidad propuesto por López E. Je. en [1].

Para presentar el principio se tomará, sin pérdida de generalidad, el caso particular de $S = I$ partiendo del problema de resolver

$$\min_x \| A x - b \|_2^2 \quad \ldots (3.3.1)$$

en forma estable a perturbaciones del lado derecho. Esto es, dada $\delta > 0$, resolver:

$$\min_x \| A x - b_\delta \|_2^2$$

con b tal que:

$$\| b_\delta - b \|_2 < \delta \qquad (3.3.2).$$

Sean

$$R_\gamma (b_\delta) = (A^* A + \gamma I)^{-1} A^* b_\delta \qquad (3.3.3)$$

y

$$x^+ = \sum_{i=1}^{r} \frac{v_i^* b}{\sigma_i} w_i \quad \ldots (3.3.4),$$

siendo r = rango (A) .

Se desea encontrar γ de tal forma que minimize

$$\| R_\gamma (b_\delta) - x^+ \|.$$

Como:

$$\| R_\gamma (b_\delta) - x^+ \| = \| R_\gamma (b_\delta) - R_\gamma (b) + R_\gamma (b) - x^+ \|,$$

al aplicar desigualdad del triángulo y simplificar se obtiene:

$$\| R_\gamma (b_\delta) - x^+ \| \leq \| R_\gamma (b_\delta) - R_\gamma (b) \| + \| R_\gamma (b) - x^+ \|$$

$$= \| R_\gamma (b_\delta - b) \| + \| R_\gamma (b) - x^+ \|,$$

utilizando (3.3.2) da como resultado:

$$\| R_\gamma (b_\delta) - x^+ \| \leq \| R_\gamma \| \delta + \| R_\gamma (b) - x^+ \|.$$

En términos de la descomposición singular de A se tiene:

$$R_\gamma = (A^* A + \gamma I)^{-1} A^*$$

$$= W (\Sigma^* \Sigma + \gamma I)^{-1} \Sigma^* V^*$$

$$R_\gamma (b) = \sum_{i=1}^{r} \left(\frac{\sigma_i \langle b , v_i \rangle}{\sigma_i^2 + \gamma} \right) w_i \quad \blacksquare$$

Con el objeto de simplificar esta expresión introducimos la siguiente definición.

Definición.- Para $A \in \mathbb{R}^{m \times n}$, sea $\| A \|_{F,C}^2 = tr (A^* C A)$, donde C es una matriz simétrica y positiva definida.

En particular, para un vector se obtiene:

$$\| z \| = tr (z^* C z) = z^* C z = \| z \|_{2,C}^2 .$$

En términos de esta definición, se tiene que:

$$\| R_\gamma \|_{F,C}^2 = tr(V \Sigma (\Sigma^* \Sigma + \gamma I)^{-*} W^* C W (\Sigma^* \Sigma + \gamma I)^{-1} \Sigma^* V^*).$$

Ahora, tomando $C = W D W^*$, con $D = Diag (\mu_1, \mu_2, \ldots, \mu_r)$, $(\mu_i > 0)$

se obtiene:

$$\| R_\gamma \|_{F,C} = \left\{ \sum_{i=1}^{r} \mu_i \frac{\sigma_i^2}{(\sigma_i + \gamma)^2} \right\}^{1/2}$$

y

$$\| R_\gamma (b) - x^+ \|_{2,C}^2 = \left\{ \sum_{i=1}^{r} \frac{\gamma^2 \mu_i z_i^2}{\sigma_i^2 (\sigma_i^2 + \gamma)^2} \right\}$$

donde $z_i = \langle b , v_i \rangle$.

Sean $v(\gamma) = \| R_\gamma \|_{F,C}$ y $b (\gamma) = \| R_\gamma (b) - x^+ \|_{2,C}$

es decir,

$$v (\gamma) = \left\{ \sum_{i=1}^{r} \frac{\mu_i \sigma_i^2}{(\sigma_i^2 + \gamma)} \right\}^{1/2}$$

y

$$b (\gamma) = \left\{ \sum_{i=1}^{r} \frac{\mu_i \gamma^2 z_i^2}{\sigma_i^2 (\sigma_i^2 + \gamma)^2} \right\}^{1/2} .$$

En conclusión:

$$\varphi (\gamma) = \delta \, v (\gamma) + b (\gamma) \qquad ...(3.3.5) ,$$

resulta ser una cota superior para $\| R_\gamma (b_\delta) - x^+ \|$ inmejorable.

Por lo tanto, el problema de encontrar γ de tal forma que minimize $\| R_\gamma (b_\delta) - x^+ \|$ se reduce en hallar γ de tal forma que minimize a su cota $\varphi (\gamma)$. En esto consiste el principio de optimalidad para la elección del parámetro de regularización.

Otra alternativa para la elección del parámetro de regularización es el principio de pseudo-optimalidad, el cual consiste en elegir a γ de tal manera que resuelve el problema

$$\delta^2 v (\gamma) - \tau b (\gamma) = 0$$

con τ dada. Para mayor detalle consultar [1] .

CAPITULO 4.- ALGORITMO BIDIAGONALIZACION-REGULARIZACION

En el presente capítulo se presenta un algoritmo que combina la Bidiagonalización Inferior de Lanczos con la Regularización de Tijonov. El objetivo central es discutir aspectos algoritmicos relativos a ambos métodos, así como los procedimientos usados en el cálculo de los elementos que intervienen en ambos (δ^2, γ, σ_i, etc.); lo cual se resume en un algoritmo presentado en lenguaje informal. Conviene recordar que desde un punto de vista estadistico nuestro problema consiste en hallar una "buena" estimación de $\bar{x}$ para x en el modelo lineal $A\,x + \bar{\varepsilon} = b$, donde el vector $\bar{\varepsilon} \sim (0, \delta^2 I)$, bajo el supuesto que A es grande, rala y mal condicionada (i.e. en términos estadísticos bajo presencia de colinealidad)

4.1. SOLUCION GAUSS-MARKOV

Uno de los problemas centrales de la regresión lineal consiste en obtener una "buena" estimación $\bar{x}$ para x en el modelo lineal estandar

$$A\,x + \bar{\varepsilon} = b \qquad \text{con} \quad \bar{\varepsilon} \sim (0, \delta^2 I)$$

donde $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\bar{\varepsilon} \in \mathbb{R}^m$ vector aleatorio de errores (no observable de media 0 y matriz de covarianzas $\delta^2 I$) y $x \in \mathbb{R}^n$ vector de parámetros a estimar.

Cuando A es de rango máximo ($R(A) = n$) y $\bar{\varepsilon}$ tiene distribución normal con media 0 y desviación estandar $\delta^2 I$, la estimación más frecuentemente usada debido principalmente a sus propiedades estadísticas de ser un estimador insesgado de mínima varianza, máxima verosimilitud, consistencia y eficiencia, es la estimación Gauss-Markov

$$\bar{x} = (A^* A)^{-1} A^* b \quad (= A^+ b),$$

para la cual se conoce que es un vector aleatorio con media x y matriz de covarianzas $\delta^2 (A^* A)^{-1}$, además que es minimax con respecto a cualquier función de pérdida $L_C(\bar{x})=(x-\bar{x})C(x-\bar{x})$, donde C es una matriz simétrica y positiva definida. Así, el error cuadrático medio (ECM ($\bar{x}$)) está dado por:

$$ECM_C(\bar{x}) = E(L_C (\bar{x})) = \delta^2 \, tr[C (A^* A)^{-1}] \geq \frac{\alpha \, \delta^2}{\| A \|} K_2(A)^2,$$

donde $\alpha = \min \{ w^* C w \mid \| w \| = 1 \}$.

Si la matriz A está cercana a una matriz de rango deficiente, o sea mal condicionada en términos del análisis numérico, o bien bajo presencia de colinealidad en términos estadísticos, entonces el cálculo numérico de $\bar{x}$ resulta ser no sólo insatisfactorio, sino que hasta inservible, lo cual se suele manifestar con la aparición de componentes de $\bar{x}$ muy grandes o bien de signos inaceptables.

Una alternativa a la estimación Gauss-Markov clásica bajo presencia de colinealidad la sugiere de manera natural (por ser un estimado sesgado) el método de regularización de Tijonov.

Ahora, con respecto al modelo lineal general

$$A x + \bar{\varepsilon} = b \qquad \bar{\varepsilon} \sim (0, \delta^2 I),$$

la idea consiste en tomar la siguiente funcional regularizante

$$J_\gamma [\bar{x} ; b] = \| b - A \bar{x} \|_2^2 + \gamma \| \bar{x} \|_S^2$$

para definir:

$$\bar{x}_\gamma = \arg \min_{\bar{x}} J_\gamma [\bar{x} ; b]$$

como solución regularizada Gauss-Markov.

4.2 Primera simplificación via cambio de base

Como se acaba de mencionar el método de Regularización de Tijonov consiste en sustituir el problema original $\underset{x}{Min} \| A x - b \|_2^2$ por un nuevo problema de mínimos cuadrados a saber:

$$\underset{x,\gamma>0}{Min} \| A x - b \|_2^2 + \gamma \| x \|_S^2 \qquad ...(4.2.1)$$

en donde S es una matriz simétrica , positiva definida, bien condicionada, usualmente dada por el investigador.

Las ecuaciones normales de (4.2.1) son:

$$(A^* A + \gamma S) x = A^* b \qquad ...(4.2.2).$$

Este problema puede resolverse directamente; sin embargo, es conveniente simplificarlo transformandolo en otro donde S se reduce a la identidad.

Sea R la matriz que resulta de la factorización de cholesky de S ($S = R^* R$), al sustituir ésta en (4.2.2) se obtiene:

$$(A^* A + \gamma R^* R) x = A^* b$$

factorizando de ésta última expresión a R^* y R da como resultado:

$$R^* (R^{-*} A^* A R^{-1} + \gamma I) R x = A^* b$$

multiplicando por R^{-*} por la izquierda se obtiene:

$$(R^{-*} A^* A R^{-1} + \gamma I) R x = R^{-*} A^* b,$$

utilizando algunas sustituciones, esta expresión puede ser simplificada de la manera siguiente:

Sean $y = R x$ y $T = A R^{-1}$, sustituyendo en la expresión anterior da como resultado:

$$(T^* T + \gamma I) y = T^* b \qquad ...(4.2.3) .$$

4.3 Segunda simplificación vía bidiagonalización

Una vez efectuada la simplificación vista en la sección anterior, se procede a calcular la bidiagonalización inferior de T.

En el capítulo 2 se discutió la bidiagonalización inferior de una matriz.

Sea $T\,Q_k \cong U_k\,B_k$ la bidiagonalización truncada inferior de T donde Q y U son matrices con columnas ortonormales y B_k matriz bidiagonal inferior de la forma:

$$B = \begin{pmatrix} \alpha_1 & & & 0 \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ 0 & & \beta_k & \alpha_k \end{pmatrix}$$

y sea $y = Q_k\,z$, sustituyendo ésta en (4.2.3) se obtiene la relación siguiente:

$$(T^* T + \gamma I)\, Q_k\, z = T^* b$$

la cual es equivalente a:

$$(T^* T\, Q_k + \gamma\, Q_k)\, z = T^* b\,,$$

multiplicando por Q_k^* se obtiene:

$$(Q_k^* T^* T\, Q_k + \gamma\, Q_k^* Q_k)\, z = Q_k^* T^* b\,,$$

usando la relación $T\,Q_k \cong U_k\,B_k$, da como resultado

$$(B_k^* U_k^* U_k B_k + \gamma\, Q_k^* Q_k)\, z \cong B_k^* U_k^* b,$$

ahora, recordando que tanto U_k como Q_k son matrices con columnas ortonormales y además que $U_k^* b = c = \beta_1 e_1^*$, al simplificar la última expresión se obtiene:

$$(B_k^* B_k + \gamma I)\, z = B_k^* c \qquad (4.3.1)$$

con $c = \beta_1 e_1^*$, resultando ser esta expresión las ecuaciones

normales del problema de mínimos cuadrados siguiente:

$$\underset{z}{\text{Min}} \left\| \begin{bmatrix} B_k \\ \hline \gamma^{1/2} I \end{bmatrix} z - c \right\|_2^2 \qquad \ldots(4.3.2)\ .$$

Todas las transformaciones hasta aqui utilizadas dan como resultado un problema de mínimos cuadrados en el cual se ha logrado una simplificación importante, ya que B_k es una matriz bidiagonal, obteniendo una importante reducción de requerimientos de memoria adicional para obtener una solución aceptable del problema original.

Para obtener la bidiagonalización de T se utiliza el algoritmo (2.3.1) el cual requiere de multiplicaciones de T y T^* por un vector, para realizar éstas no es conveniente obtener T en forma explicita ya que al calcular AR^{-1} explicitamente se pierde una cualidad importante de la matriz A, a saber la poca densidad; sin embargo, este problema puede ser resuelto en una forma sencilla ya que calcular T q es equivalente a resolver primeramente el sistema R a = q y posteriormente multiplicar por A el vector a, en forma análoga, para calcular T^* u primero se multiplica y = A^* u y luego se resuelve el sistema R a = y (en el código a e y son vectores auxiliares de trabajo).

En el capitulo 2 se discutió como calcular el condicional numérico de una matriz usando la Bidiagonalización Inferior de Lanczos, esto es, para calcular $K_F(T) = \| T \|_F \| T^{-1} \|_F$ se utiliza la identidad (2.4.7)

$$K_F(T) = \| B \| \| W^{-1} \|_F = \left[\left(\alpha_1^2 + \sum_{i=2}^{n} \alpha_i^2 + \beta_i^2 \right) \cdot \left(\sum_{i=1}^{n} \| w_i \|^2 \right) \right]^{1/2} \qquad \ldots(4.3.3)\ ,$$

asimismo, se observó que éste cálculo puede realizarse en forma iterativa al mismo tiempo en que se efectua la bidiagonalización;

por otro lado se mencionó que esta bidiagonalización puede truncarse en k-pasos ($k \leq n$) obteniendo una matriz B_k cuyos valores singulares están cercanos a los valores singulares de la matriz T. El criterio utilizado para parar el proceso iterativo de la bidiagonalización es tomar un límite superior (usualmente dado por el investigador) para el condicional numérico de B_k. Para mayor detalle véase las secciones 4.6 y 5.2.

4.4.- Estimación de γ

En esta sección se discute un método para estimar el parámetro γ de regularización que tiene lugar en la regularización de Tijonov.

El método de Regularización de Tijonov no dice que γ utilizar; sin embargo, recordemos que nuestro problema central consiste en resolver el problema $\underset{x}{\text{Min}} \| A x - b \|_2^2$ en forma estable y que este se sustituye por el problema de mínimos cuadrados

$$\underset{z}{\text{Min}} \left\| \begin{bmatrix} B_k \\ \gamma^{1/2} I \end{bmatrix} z - c \right\|_2^2 ,$$

cuyas ecuaciones normales son

$$(B_k^* B_k + \gamma I) z = B_k^* c .$$

Sea $\hat{z}_\gamma$ la solución calculada de este problema, entonces se define la función de pérdida para $\hat{z}_\gamma$ como:

$$L_C (\hat{z}_\gamma) = (z - \hat{z})^* C (z - \hat{z}_\gamma)$$

donde $C = I$ ó $C = B_k^* B_k$ y se define el error cuadrático medio de $\hat{z}_\gamma$ como:

$$ECM (\hat{z}_\gamma) = E (L_C (\hat{z}_\gamma)) ,$$

entendiendo por $E(L_c(\hat{z}_\gamma))$ la esperanza de la función de pérdida para $\hat{z}_\gamma$. Recuérdese que el error cuadrático medio $ECM(z_\gamma)$ para $\gamma = 0$ es el error cuadrático medio para el estimador clásico $\bar{z}$ de Gauss-Markov.

Teorema.- Existe $\gamma_0 > 0$, tal que $ECM(\hat{z}_{\gamma_0}) < ECM(\hat{z})$

Demostración [1]

Este teorema es el fundamento del criterio de optimalidad para la elección de γ:

$$\gamma\text{-óp} = \operatorname{Arg}\min_{\gamma \geq 0} ECM(\hat{z}_\gamma) \qquad ...(4.4.1),$$

Como el $ECM(\hat{z}_\gamma)$ depende de δ^2 y z desconocidos de antemano siendo z el vector a estimar, se propone sustituir δ^2 por s^2 y $\hat{z}$ por z en la expresión (3.3.5) obtenida en el capítulo 3 para así obtener una sobreestimación para el $ECM(\hat{z}_\gamma)$ donde s^2 es la estimación insesgada usual para δ^2, obteniendo

$$\varphi(\gamma) = s^2 \sum_{i=1}^{k} \frac{p_i \sigma_i^2}{(\sigma_i^2 + \gamma)^2} + \sum_{i=1}^{k} p_i \left(\frac{d_i}{\sigma_i}\right)^2 \left(\frac{\gamma}{\sigma_i^2 + \gamma}\right)^2 \qquad ...(4.4.2)$$

donde $d_i = w_i^* z$, $i=1,2,...,k$, y σ_i $i=1,2,...,k$ son los valores singulares de B_k, $p_i = \begin{cases} 1 & \text{si } C = I \\ \sigma_i^2 & \text{si } C = B^* B \end{cases}$

y $s^2 = \begin{cases} Res^2/(m-n) & \text{si } m > n \\ Res^2 & \text{si } m = n \end{cases}$,

siendo aquí $Res = \| B_k \hat{z} - c \|_2^2$, para mayor detalle consultar [1].

Desde un punto de vista práctico es importante dejar asentado que en el cálculo de $\| B_k^+ \|$ se calcula a la vez la estimación Gauss-Markov $\hat{z}$, esto es, se resuelve el problema de mínimo de cuadrados siguiente:

$$\underset{z}{\text{Min}} \; \| B_k z - c \|_2^2 \qquad \ldots(4.4.3) \; ,$$

Golub-Kahan proponen como primera etapa del cálculo de los valores singulares de una matriz T el obtener la bidiagonalización superior de T y en la segunda etapa utilizar el algoritmo QR con esta matriz bidiagonal. En el presente trabajo se utiliza la bidiagonalización inferior de T (B_k) por esta razón como segunda etapa se propone utilizar el algoritmo QR con B_k^*, es decir:

$$SVD \; (B_k^*) = V^* \; \Sigma \; W \; ,$$

si SVD (B_k) = $W^* \; \Sigma \; V$, en otras palabras, los vectores singulares derechos de B_k son los vectores singulares izquierdos de B_k^* y en forma análoga los vectores singulares izquierdos de B_k son los vectores singulares derechos de B_k^* ; por otra parte, en el cálculo de γ-óptima se requiere del calculo de $d_i = \langle w_i, c \rangle$, como en este trabajo se utilizó la matriz transpuesta, entonces se usará $d_i = \langle v_i, c \rangle$.

Luego entonces el criterio de selección de γ óptima consiste en minimizar la función φ_γ dada por (4.4.2) en vez de la obtenida por (4.4.1).

Una alternativa para el principio de selección de γ óptima es el criterio de γ pseudo-óptima el cual consiste en hallar la solución de la ecuación:

$$s^2 \sum_{i=1}^{k} \frac{p_i \sigma_i^2}{(\sigma_i^2 + \gamma)^2} - \tau \sum_{i=1}^{k} p_i \left(\frac{d_i}{\sigma_i}\right)^2 \left(\frac{\gamma}{\sigma_i^2 + \gamma}\right)^2 = 0 \qquad \ldots(4.3.4)$$

con τ un parámetro de amortiguamiento, en este trabajo se toma $\tau = 27/32$ para mayor detalle consultar [1].

4.5 Solución del problema de mínimos cuadrados y recuperación de transformaciones ortogonales

En la sección 2.4 se utilizó un método basado en rotaciones planas para resolver el problema

$$\min_{y} \| B_k y - c \|_2^2 .$$

Ahora, al efectuar la regularización de Tijonov se da lugar al problema de mínimos cuadrados siguiente:

$$\min_{z} \left\| \begin{pmatrix} B_k \\ \gamma^{1/2} I \end{pmatrix} z - c \right\|_2^2 \qquad \ldots(4.5.1)$$

para resolver este problema se utiliza una modificación al método anterior, la idea es la siguiente:

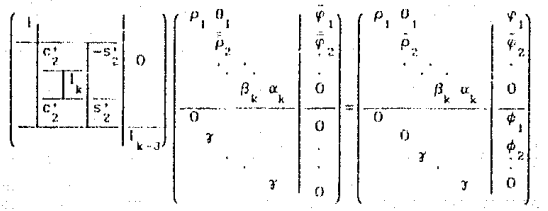
Dada $\left[\begin{array}{c|c} B_k & \\ \hline \gamma^{1/2} I & \end{array}\right.$ $\left.\begin{array}{c} \\ \hline \end{array}\right|$ $\beta_1 \; e_1^*$ $\Big]$ matriz aumentada del problema (4.4.1) se multiplica por $\tilde{G}_{1,2}$ de tal forma que:

$$\left(\begin{array}{cc|c} c_1' & -s_1' & \\ & I_k & 0 \\ s_1' & c_1' & \\ \hline 0 & & I_{k-2} \end{array}\right)\left(\begin{array}{cccc|c} \alpha_1 & & & & \beta_1 \\ \beta_2 & \alpha_2 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \beta_k & \alpha_k & 0 \\ \hline \gamma & & & & 0 \\ & \gamma & & & \vdots \\ & & \ddots & & \\ & & & \gamma & 0 \end{array}\right) = \left(\begin{array}{cccc|c} \tilde{\rho}_1 & & & & \tilde{\varphi}_1 \\ \beta_2 & \alpha_2 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \beta_k & \alpha_k & 0 \\ \hline 0 & & & & \tilde{\phi}_1 \\ & \gamma & & & 0 \\ & & \ddots & & \vdots \\ & & & \gamma & 0 \end{array}\right),$$

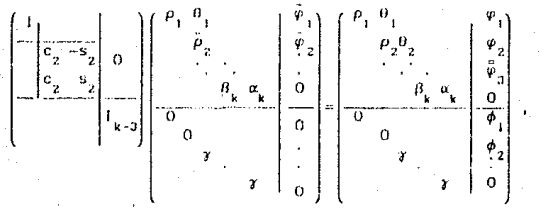
a continuación se multiplica $G_{1,2}$ al resultado anterior para eliminar β_2, es decir:

$$\left(\begin{array}{cc|c} c_1 & -s_1 & \\ s_1 & c_1 & 0 \\ \hline & & I_{2k-2} \\ 0 & & \end{array}\right)\left(\begin{array}{cccc|c} \tilde{\rho}_1 & & & & \tilde{\varphi}_1 \\ \beta_2 & \alpha_2 & & & 0 \\ & \ddots & \ddots & & \vdots \\ & & \beta_k & \alpha_k & 0 \\ \hline 0 & & & & \tilde{\phi}_1 \\ & \gamma & & & 0 \\ & & \ddots & & \vdots \\ & & & \gamma & 0 \end{array}\right) = \left(\begin{array}{cccc|c} \rho_1 & \theta_1 & & & \varphi_1 \\ & \tilde{\rho}_2 & & & \tilde{\varphi}_2 \\ & & \ddots & & \vdots \\ & & \beta_k & \alpha_k & 0 \\ \hline 0 & & & & \tilde{\phi}_1 \\ & \gamma & & & 0 \\ & & \ddots & & \vdots \\ & & & \gamma & 0 \end{array}\right)$$

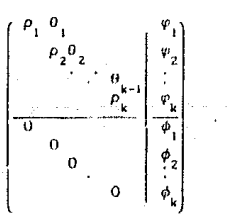
de manera análoga al multiplicar primero por $\bar{G}_{2,3}$ se obtiene:



y posteriormente por $\bar{G}_{2,3}$ da como resultado:



y así, sucesivamente se multiplica por $\bar{G}_{3,4}\bar{G}_{3,4}\cdots\bar{G}_{k-1,k}\bar{G}_{k-1,k}$ hasta obtener:



El vector solución $\hat{z}_{\gamma}$ puede ser obtenido en forma análoga a como se obtuvó en la sección (2.4) a partir de la resolución del sistema $R_k\ z_k = \bar{\varphi}_k$ donde $\bar{\varphi}^* = (\varphi_1, \varphi_2, \ldots \varphi_k)$ y R_k es una matriz bidiagonal superior. Si este sistema se resuelve directamente, entonces es necesario conocer de antemano toda la bidiagonalización; sin embargo, es importante señalar que este sistema se puede resolver de manera recursiva como se vió en la

sección 2.4 del capítulo 2 utilizando el esquema siguiente:

$z_k = R_k^{-1}\,\tilde{\theta}_k = w_k\,\tilde{\theta}_k$ siendo $\tilde{\theta}^* = (\theta_1, \theta_2, \ldots, \theta_k)$, $w_0 = 0$ y $w_i = (1/\rho_i)(e_i - \theta_{i-1}\,w_{i-1})$.

Finalmente, habiendo obtenido la solución del problema (4.5.1), resta únicamente recuperar las transformaciones ortogonales utilizadas en la simplificación del problema original (4.2.1), para ello se utiliza la identidad $y = Q_k\,z$ y se resuelve el sistema $R\,x = y$.

4.6 Algoritmo

El objetivo central de esta sección es presentar un algoritmo en lenguaje informal que permita resolver el problema al cual ésta dedicado el presente trabajo, es decir, resolver en forma estable el sistema $A\,x = b$ donde $A \in \mathbb{R}^{m\times n}$ y $b \in \mathbb{R}^m$ en el sentido de mínimos cuadrados.

El algoritmo a grandes rasgos consta de tres etapas centrales además de lectura de datos y escritura de resultados: la primera de ellas es la bidiagonalización de la matriz T, truncando ésta en base al criterio de tolerancia para el condicional numérico de B_k, además incluye el cálculo del número de condición $K_F(B_k)$, el residual y la estimación Gauss-Markov. Cabe recordar que estos cálculos son realizados en forma iterativa al mismo tiempo que se realiza la bidiagonalización.

En la segunda etapa del algoritmo se calculan los elementos que intervienen el la estimación de γ-óptima y γ-pseudo-óptima, así como la estimación de estos parámetros.

En la tercera y última etapa se resuelve el problema de mínimos cuadrados simplificado y se recupera todas las transformaciones ortogonales utilizadas en la simplificación del

problema original.

De esta manera, el algoritmo puede resumirse de la forma siguiente:

Etapa uno: *Lectura de datos.*

n, m, conlim, A, b, orden de regularización, pesos y criterio para la estimación de la gamma.

Etapa dos: *Bidiagonalización de T.*

-Factorización de Cholesky de la matriz S ($S = R^* R$) según el orden de discretización elegido (véase sección uno).

-Bidiagonalización inferior de Lanczos de T; para la obtención de esta bidiagonalización se utiliza el algoritmo (2.3.1) con la modificación mencionada en la sección anterior, obteniendo además K_F (B_k), la estimación Gauss-Markov y su correspondiente residual.

Toda esta etapa en lenguaje informal toma la forma siguiente:

-Tomar $\beta_1 = \| b \|$, $u_1 = b / \beta_1$, aux2 = $A^* u_1$

-Resolver R^* aux = aux2

-Luego tomar $\alpha_1 = \|$ aux $\|$, q_1 = aux / α_1, $w_0 = 0$, $x_0 = 0$, $\bar{\rho}_2 = \alpha_1$, $\bar{\varphi}_1 = \beta_1$, res = β_1, BNORM2 ($\| B_k \|^2$) = α_1^2, DNORM2 ($\| D \|^2$) = 0, $\bar{\theta}_1 = 0$, k = 1.

Mientras $\beta_k \neq 0$ y $\alpha_k \neq 0$ y $k \leq n$ y Cond $\leq$ conlim

(Bidiagonalización de T)
Resolver R aux = q_k
aux2 = A^* aux
z = aux2 - $\alpha_k u_k$
$\beta_{k+1} = \| z \|$
$u_{k+1} = z / \beta_{k+1}$
aux2 = $A^* u_{k+1}$

$$
\begin{aligned}
&\text{resolver } R^{*}\ \text{aux} = \text{aux2}\\
&y = \text{aux} - \beta_{k-1}\, q_{k}\\
&\alpha_{k+1} = \| y \|\\
&q_{k+1} = y \,/\, \alpha_{k+1}\\
&\text{Estimación del condicional: Resuelve Min}_{x} \| T x - c \|\\
&d = (\bar{\rho}_{k}^{2} + \beta_{k+1}^{2})^{1/2}\\
&c = \bar{\rho}_{k} \,/\, d\\
&s = \beta_{k+1} \,/\, d\\
&\rho_{k} = c\, \bar{\rho}_{k} + s\, \beta_{k+1}\\
&\theta_{k} = \bar{\theta}_{k}\\
&\bar{\theta}_{k+1} = s\, \alpha_{k+1}\\
&\bar{\rho}_{k+1} = c\, \alpha_{k+1}\\
&\varphi_{k} = c\, \bar{\varphi}_{k}\\
&\bar{\varphi}_{k+1} = s\, \bar{\varphi}_{k}\\
&res = res * s\\
&w_{k} = (1/\rho_{k})\, (\theta_{k}\, w_{k-1} + e_{k})\\
&x_{k} = x_{k-1} + \varphi_{k}\, w_{k}\\
&BNORM2 = BNORM2 + \alpha_{k+1}^{2} + \beta_{k+1}^{2}\\
&DNORM2 = DNORM2 + \| w_{k} \|^{2}\\
&Cond = BNORM * DNORM
\end{aligned}
$$

Etapa tres: *Estimación de γ-óptima y γ-pseudo-óptima*

Cálculo de la descomposición singular de B_k

-Cálculo del $s^2 = \begin{cases} res^2/(m-n) & \text{si } m \neq n \\ res^2 & \text{si } m = n \end{cases}$

-Cálculo de $d_i = \langle v_i, \beta_i e_i \rangle$

-$p_i = \begin{cases} \sigma_i^2 & \text{si pesos} = 1 \\ 1 & \text{si pesos} = 0 \end{cases}$

-Estimación de γ-óptima. Resolviendo:

$$\min_{\gamma} \; s^2 \sum_{i=1}^{k} (p_i \sigma_i^2)/(\sigma_i^2 + \gamma)^2 + \tau \sum_{i=1}^{k} p_i (d_i/\sigma_i)^2 [\gamma/(\sigma_i^2 + \gamma^2)]^2$$

-Estimación de γ-pseudo-óptima. Resolviendo:

$$s^2 \sum_{i=1}^{k} (p_i \sigma_i^2)/(\sigma_i^2 + \gamma)^2 - \tau \sum_{i=1}^{k} p_i (d_i/\sigma_i)^2 [\gamma/(\sigma_i^2 + \gamma^2)]^2 = 0$$

Etapa cuatro: *Solución del problema de mínimos cuadrados* simplificado y recuperación de transformaciones ortogonales.

-Resolver $\min_{z} \left\| \begin{pmatrix} B_k \\ \gamma^{1/2} I \end{pmatrix} z - c \right\|_2^2$

-Recupera transformaciones ortogonales

-$y = Q_k z$

-$x = R^{-1} y$

Etapa cinco: *Impresión de resultados*

-Imprime x.

Capítulo 5 Experimentos Numéricos

En el capítulo 1 se mencionó que una fuente importante de problemas de mínimo de cuadrados mal condicionados es la ecuación integral de Fredholm de primer orden; sin embargo, en la bibliografía se encontraron pocos problemas cuyo núcleo (kernel) sea esencialmente diferente, además, todos excepto uno (Phillips) generan matrices llenas, lo cual rompe con la idea de este trabajo (i.e. que la matriz sea rala), no obstante, cabe mencionar que se reportan experimentos numéricos sparse, los cuales se generaron con estructura de manera aleatoria.

5.1 Aspectos computacionales

Empecemos por mencionar que la función

$$\varphi(\gamma) = s^2 \sum_{i=1}^{n} \frac{\rho_i \sigma_i}{(\sigma_i + \gamma)^2} + \tau \sum_{i=1}^{n} p_i \left(\frac{d_i}{\sigma_i}\right)^2 \left(\frac{\gamma}{\sigma_i + \gamma}\right)^2 \qquad (5.1.1)$$

con frecuencia resulta ser una función demasiado plana alrededor de su punto de mínimo por, ello es de esperar que se presenten dificultades prácticas para encontrar el mínimo.

Dificultades similares se observan también al resolver $\varphi(\gamma) = 0$

donde

$$\psi(\gamma) = s^2 \sum_{i=1}^{n} \frac{\rho_i \sigma_i}{(\sigma_i + \gamma)^2} - \tau \sum_{i=1}^{n} p_i \left(\frac{d_i}{\sigma_i}\right)^2 \left(\frac{\gamma}{\sigma_i + \gamma}\right)^2 \qquad (5.1.2).$$

Con respecto a la ecuación $\varphi(\gamma)$ se cuenta con un método "seguro" para resolver este problema, este método es el método de bisección; sin embargo, éste resulta tener una convergencia extremadamente lenta si el valor buscado se encuentra "cerca" de

un extremo del intervalo inicial. Por esta razón se realiza una modificación al método de bisección con el objeto de acelerar su convergencia.

Un método "seguro" para resolver $\varphi(\gamma) = 0$ resulta de combinar la idea del método de bisección con la filosofía del método de regla falsa modificada, la idea es la siguiente:

Se efectúan 2 iteraciones del método de bisección y se observa que los puntos medios obtenidos "caen del mismo lado del intervalo; en la tercera iteración se trisecta el intervalo, si el punto obtenido "cae" del mismo lado, entonces en la siguiente iteración se divide el intervalo en 4, etc., hasta obtener un punto del otro lado del intervalo. En el fondo, lo que se está haciendo es dividir el intervalo en una razón dada, es por esto que el método recibe el nombre de "Método de la Razón" y su algoritmo es el siguiente:

Datos:

Dados $\varphi(x) \in C[a,b]$ tal que $\varphi(a)\,\varphi(b) < 0$, nmax = número máximo de iteraciones permitidas, tol= tolerancia para la solución aproximada.

Variables internas:

ia, ib ,i : contadores, r = razón para dividir el intervalo.

Inicio:

ia = 0, ib = 0, i = 0 , r = 1.

Proceso iterativo:

$$(1) \qquad m = \frac{a + r\,b}{1 + r}$$

```
i = i+1

si φ(m) = 0,   m = solución del problema
no | si φ(a) φ(m) < 0
   |    | b = m , φ(b) = φ(m), ib = ib +1, ia = 0
   |    | si ib ≥ 2
   |    |    | r = ib
   |    |  no | r = 1
   |  no | a = m, φ(a) = φ(m), ia = ia + 1, ib = 0
   |    | si ia ≥ 2
   |    |    | r = 1/ia
   |    |  no | r = 1
   |
   | si | |b - a| / |a| < tol,   m solución aproximada del problema
   |  no | si | i ≥ imax, se realizaron todas las iteraciones permitidas y no se obtuvo la tolerancia deseada
   |     | no | ve a ( 1 )
```

En los ejemplos que se realizaron con bisección se requería aproximadamente de 100 iteraciones para obtener una tolerancia de 10^{-8}, utilizando esta modificación se requiere de aproximadamente 15 iteraciones, con lo que se obtuvó un importante ahorro computacional en esta parte del método.

Por otro lado, en la evaluación de la expresión (5.1.2) se está trabajando con valores de σ_i^2, como σ_i tiende "rápidamente" a cero conforme i crece, esto resulta ser extremadamente delicado. Una alternativa para resolver este problema es obtener una expresión equivalente de (5.1.2) con un mejor comportamiento. Para el primer término de la expresión se utiliza la igualdad

$$\frac{\sigma_i}{\sigma_i^2 + \gamma} = \frac{1}{(\sigma_i + \gamma / \sigma_i)} \qquad (5.1.3)$$

y para el segundo término

$$\frac{\gamma}{\sigma_i^2 + \gamma} = \frac{\gamma}{\sigma_i \left(\sigma_i + \gamma / \sigma_i \right)} \qquad (5.1.4)$$

Por lo que respecta a la sumatoria, ésta se realiza en sentido inverso, ya que de esta forma se está realizando una suma cuyos elementos van en forma creciente y por lo tanto generan un menor error.

5.2 Ejemplos y resultados numéricos

Como se mencionó anteriormente, una fuente importante de problemas que dan lugar a sistemas de ecuaciones y problemas de mínimo de cuadrados mal condicionados es la ecuación integral de Fredholm de primer orden. En esta sección se presentan brevemente, destacando sus características, los ejemplos que se trabajaron experimentalmente, así como los resultados obtenidos en su solución.

Desde un punto de vista numérico las ecuaciones integrales de Fredholm pueden clasificarse de acuerdo a su núcleo en tres tipos: blando, moderado y duro.

Se dice que la ecuación:

$$\int_a^b K (x, t) u (t) dt = f (x) \quad c \le x \le d \qquad ...(5.2.1)$$

es de tipo *blando* si k(x, t) es una función poco suave, por ejemplo, continua y derivable a trozos; dentro de esta categoría está el ejemplo clásico (aparece frecuentemente en la literatura) de Phillips (ejemplo i1). Para este tipo de casos se tomó CONLIM = m * 10^5.

Diremos que la ecuación (5.2.1) es de tipo *moderado* si su función núcleo K (x, t) es algebraico y esencialmente de clase C^{∞}, como lo es el ejemplo I5 (Fox-Goodwin). En este tipo de

problemas se tomó CONLIM = m * 10^9.

Finalmente, diremos que la ecuación (5.2.1) es de tipo *duro*, si K (x, t) es una función analítica, dentro de esta categoría cae el ejemplo 1. Para este tipo de casosse tomó CONLIM = m 10^{13}.

Esto es, a mayor suavidad del núcleo K (x, t) en (5.2.1) sus valores singulares tienden más rapidamente a cero conforme i crece, es de esperar que si K (x, t) es analítico, entonces

$$\sigma_i = O(a^i) , \quad 0 < a < 1$$

y a mayor suavidad del núcleo más cercano a cero el valor de a.

observación 1.-

En la expresión para obtener el parámetro de regularización γ se utiliza la estimación

$$s^2 = \begin{cases} res^2/ (m-n) & \text{si } m \neq n \\ res^2 & \text{si } m = n \end{cases} ;$$

sin embargo para la clase de los ejemplos *duros* este valor resulta ser extremadamente "pequeño", lo cual conduce a obtener valores "pequeños" para la γ, lo que va en contra a necesidad de tener una mayor perturbación del problema de mínimos cuadrados. Es por esta razón que se tomó la heuristica siguiente:

$$s^2 = \begin{cases} s^2 & \text{si el problema es } blando \\ (s^2)^{1/2} & \text{si el problema es } moderado \\ (s^2)^{1/4} & \text{si el problema es } duro \end{cases} ,$$

obsérvese que si $s^2 < 1$, esta modificación lo agranda.

observación 2.-

Los errores son muy aceptables en la mayoría de los casos aún cuando éstos muestran un crecimiento en los extremos del intervalo como lo muestran las gráficas de los errores usando ($Log_{10} | x_i - x_i^* | / | x_i |$).

observación 3.-

En términos generales el número de iteraciones realizadas en el proceso de bidiagonalización no superan en la mayoría de los casos el 30% (n / 3), obteniendo una buena reducción de cómputo.

A continuación se presentan los ejemplos trabajados.

Ejemplo 1.-(Lewis [1]) Este ejemplo forma parte de una familia más general que se presentará en el ejemplo 7, está dentro de la categoría de los *duros* y su solución es $u(t) = e^{t}$.

$$\int_0^1 e^{xt}\ u(t)\,dt = \frac{e^{x+1}-1}{x+1} \qquad 0 \le x \le 1 .$$

Ejemplo 2.-(Miller [12]) Este ejemplo es un ejemplo típico dentro de la clase de los *duros* y su solución es $u(t) = t$.

$$\int_0^\infty e^{-xt}\ u(t)\,dt = \frac{1}{x^2} \qquad 0 \le x \le \infty .$$

Ejemplo 3.- (Miller [12]) Este ejemplo del tipo clásico de los *duros* y tiene como solución $u(t) = \dfrac{e^{-t^2/16}}{4\sqrt{\pi\, t^2}}$.

$$\int_0^\infty e^{-xt}\ u(t)\,dt = e^{-\sqrt{x}/2} \quad 0 \le x \le \infty .$$

Ejemplo 4.- (Miller [12]) El núcleo de este ejemplo es separable y la imagen del operador integral que define es de dimensión 3. Este cae dentro de la categoría de los *duros* y su solución es $u(t) = t$.

$$\int_0^\infty (x-t)^2\, u(t)\,dt = 1/2\, x^2 - 2/3\, x + 1/4 \qquad 0 \le x \le 1.$$

Ejemplo 5.-(Marti [11]) Este ejemplo es de los clásicos *blandos*, su núcleo es continuo y derivable a trozos, su solución es $u(t) = t^4 - 2t^3 + t$.

$$\int_0^1 K(x,t)\, u(t)\,dt = \frac{-x^6+3x^5-5x^3+3x}{30} \qquad 0 \le x \le 1.$$

$$\text{con } k(x,t) = \begin{cases} x(1-t) & x \le t \\ t(1-x) & x > t \end{cases}.$$

Ejemplo 6.- (Fox-Goodwin [14]) El núcleo de este ejemplo es separable y la imagen del operador integral que define es de dimensión 2. Este cae dentro de la categoría de los *duros* y su solución es $u(t) = t$.

$$\int_0^1 (x-t)\,u(t)\,dt = 1/2\,x + 1/3 \qquad 0 \le x \le 1 .$$

Ejemplo 7.-(Lewis [15]) Este ejemplo genera una familia de problemas cuya solución es $u(t) = e^t$, la familia es de tipo *duro*, en nuestro caso particular se tomó $n = 2$.

$$\int_0^1 e^{x^n t}\,u(t)\,dt = \frac{e^{x^n+1}+1}{x^n+1} \qquad 0 \le x \le 1 .$$

Ejemplo 8.-(Dubner [9]) Este ejemplo es una ecuación de Volterra y es del tipo de los *duros* cuya solución es $u(t) = 1$.

$$\int_0^1 \cos(x-t)\,u(t)\,dt = \operatorname{sen} x \qquad 0 \le x \le 1 .$$

Ejemplo 9.- (Dubner [9]) Este ejemplo es de tipo duro y surge de calcular la inversa de la transformada de Laplace, su solución es $u(t) = 1/2^*$ sen t.

$$\int_0^\infty e^{-xt}\,u(t)\,dt = \frac{x}{(x^2+1)^2} \qquad 0 \le x \le \infty .$$

Ejemplo 10.-(Dubner [9]) Similar al anterior es, del tipo duro y su solución es $u(t) = 2/3\ e^{-t/2}$ sen $(3t/2)$.

$$\int_0^\infty e^{-xt}\,u(t)\,dt = \frac{1}{(x+1)^2} \qquad 0 \le x \le \infty .$$

Ejemplo 11.-(Phillips [1]) Este ejemplo es clásico en la literatura y es conocido como la ecuación integral de Phillips, cabe señalar que es el único ejemplo de los trabajados que da origen a una matriz sparse con estructura de banda, su tipo es *blando.*

$$\int_{-6}^{6} K(x,t)\,u(t)\,dt = f(x) \qquad -6 \le x \le 6$$

donde

$$f(x)=(6-|x|)(1+1/2\cos(\pi x/3))+\operatorname{sgn}(x).9/2\pi\operatorname{sen}(\pi x/3),$$

$$k(x,t)=\begin{cases} 1+\cos(\pi(x-t)/3) & \text{si } |x-t| \le 3 \\ 0 & |x-t| > 3 \end{cases}$$

y su solución es

$$u(t)=\begin{cases} 1+\cos(\pi(t)/3) & \text{si } |t| \le 3 \\ 0 & |t| > 3 \end{cases}.$$

Ejemplo 12.-(Fox-Goodwin [14]) Este ejemplo es del tipo de los *duros* y su solución es u (t) = sen t.

$$\int_{0}^{1} \cos(x-t)\,u(t)\,dt = 1/2\cos x + \pi/4 \operatorname{sen} x \qquad 0 \le x \le 1.$$

Ejemplo 13.-(Fox-Goodwin [14]) Este ejemplo es del tipo de los *duros*, su solución es u(t) = cos t + sen t.

$$\int_{0}^{\pi/2} \operatorname{sen}(x+t)\,u(t)\,dt = \frac{\cos x + \operatorname{sen} x}{1/2 + \pi/4} \qquad 0 \le x \le \pi/2 .$$

Ejemplo 14.-(Fox-Goodwin [14]) Análogo al anterior y su solución es u (t) = cos t - sen t.

$$\int_{0}^{\pi/2} \operatorname{sen}(x+t)\,u(t)\,dt = \frac{\cos x - \operatorname{sen} x}{1/2 - \pi/4} \qquad 0 \le x \le \pi/2 .$$

Ejemplo 15.-(Fox-Goodwin [14]) Este ejemplo es del tipo clásico de los moderados con núcleo algebraico y su solución es u(t)=t.

$$\int_{0}^{1} \sqrt{x^2+t^2}\;u(t)\,dt = 1/3(1+x^2)\left(\sqrt{1+x^2}-x^3\right) \qquad 0 \le x \le 1.$$

Ejemplo 16.-(López [1]) Este ejemplo de la categoría de los *moderados*, es de nueva creación y su solución es u(t)=t(10t+1).

$$\int_{0}^{1} \frac{1}{x+t+.1}\,u(t)\,dt = 5-10x+(10x+1)\ln\left(\frac{10x+11}{10x+1}\right) \qquad 0 \le x \le 2 .$$

Ejemplo 17.-(Varah [1]) Es un ejemplo que pertenece a la categoría de los *duros*, corresponde a la transformada inversa del

coseno y su solución es u (t) = t.

$$\int_0^1 \cos(xt)\, u(t)\, dt = \operatorname{sen}x/x+(\cos(x-1)/x^2 \qquad 0\le x\le 1 .$$

Ejemplo 18.-(Graves [13]) Este ejemplo es del tipo *duro* y su solución es u(t) = t.

$$\int_0^1 (x-t)^2\, u(t)\, dt = \frac{6x^2 - 8x + 3}{12} \qquad 0 \le x \le 1.$$

Ejemplo 19.-(Graves [13]) Análogo al anterior y su solución es u (t) = 16 t^2 - 16 t + 3.

$$\int_0^1 (x-t)^2\, u(t)\, dt = \frac{5x^2 - 5x + 3}{15} \qquad 0 \le x \le 1 ,$$

Ejemplo 20.-(Graves [13]) Nuevamente el núcleo de esta ecuación es similar al anterior y su solución es u (t) = cos (5t).

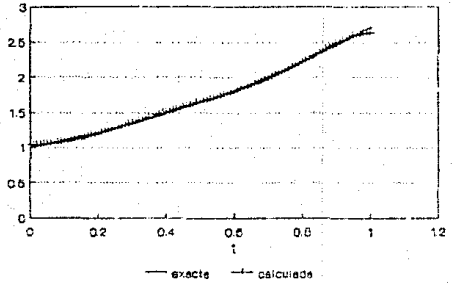
$$\int_0^1 (x-t)^2\, u(t)dt=\frac{\operatorname{sen}5x^2}{5}-2\left(\frac{5\operatorname{sen}5+\cos5-1}{25}\right)x-\frac{10\cos5+23\operatorname{sen}5}{125} \quad 0\le x\le 1 .$$

A continuación se presentan los resultados obtenidos, así como las gráficas de éstos.

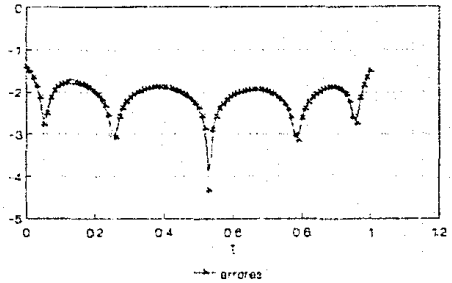
observación 4.-

El caso de los ejemplos que no se presentan las gráficas de sus resultados se debe a que se presentaron problemas en el cálculo de los valores singulares.

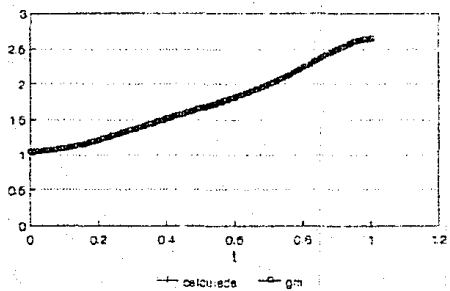
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mark



Ejemplo 1

n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 23

Limite para el condicional = 3.e+14

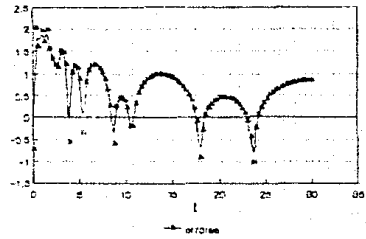
s cuadrada = 1.5e-09

Gamma optima = 4.9e-09

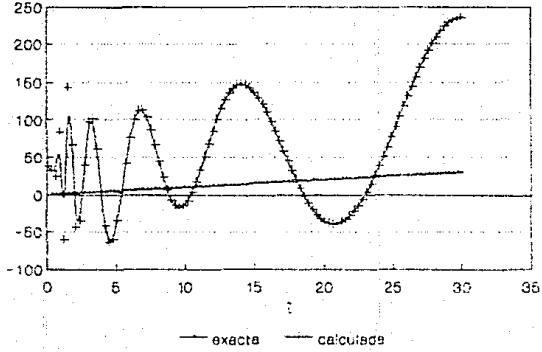
Gamma pseudo-optima = 7.2e-08

pesca = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 2

n = 101 m = 301

tipo duro

No. de iteraciones realizadas = 101

Limite para el condicional = 3.e-14

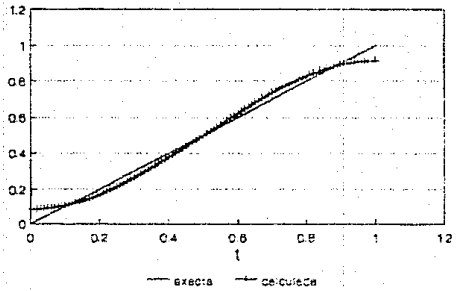
e cuadrados = 6.8e-04

Gamma optima = 3.7e-09

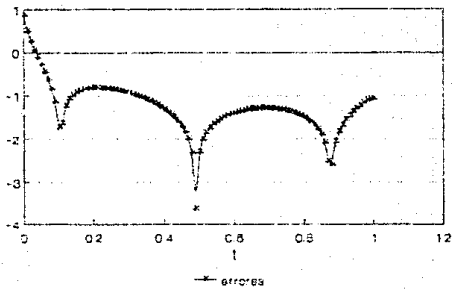
Gamma pseudo-optima =5.4e-12

pesos = 1 orden = 1

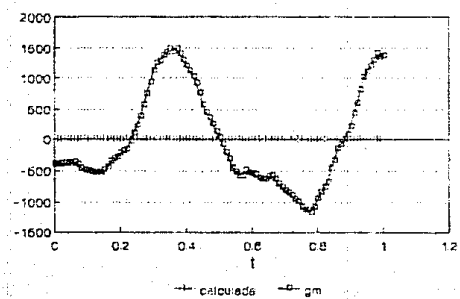
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mar



Ejemplo4

n = 101 m 301

tipo: duro

No. de iteraciones realizadas = 16

Limite para el condicional = 3.e+14

a cuadrada = 8.0e-05

Gamma optima = 1.0e-01

Gamma pseudo-optima = 1.6e-01

pesos = 1 orden = 1

Comentarios sobre la gráfica

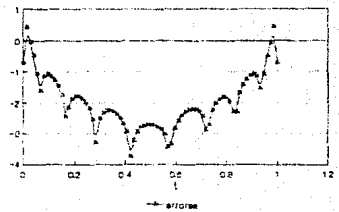
En este ejemplo tomamos n = 101, m = 301; el número de iteraciones realizadas en el proceso de Bidiagonalización fue de k = 16, como se mencionó en el trabajo, el número máximo de iteraciones que se pueden efectuar es k = n = 101, es por esta razón que se observa un ahorro del 84% en los cálculos computacionales.

Por lo que respecta al tipo de problema es un ejemplo clásico de la clase de los "*duros*", por lo tanto se tomó como límite para el condicional CONLIM = $3.e^{14}$. Como se mencionó, en este tipo de problemas la solución Gauss-Markov puede tomar valores inadmisibles, esto se observa en la gráfica, la solución Gauss-Markov toma valores que varían entre $-1.163e^{3}$ y $1.495e^{3}$.

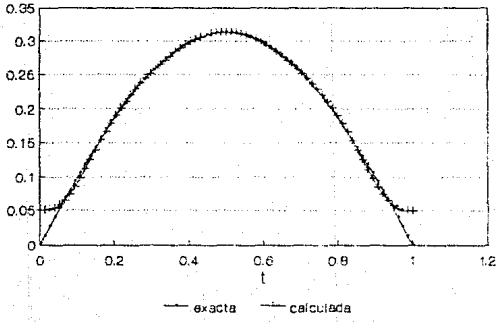
Observando la gráfica de la solución exacta contra la gráfica de la solución calculada por nuestro método propuesto se confunden ambas soluciones siendo apenas apreciable una pequeña diferencia en los extremos. Esto se ve reflejado en la gráfica de los errores relativos: usando Log_{10} los errores varían entre $-3.5e^{0}$ y $8.7e^{-1}$; así mismo, éstos toman su valor máximo en el extremo inicial del intervalo y presentan un descenso en la parte central con un pequeño repunte en el extremo final. Cabe destacar que este mismo fenómeno se observó en la mayoría de los ejemplos trabajados, es decir, los errores tienden a disminuir en la parte central y aumentan en los extremos.

Del párrafo anterior, de observar la gráfica y el intervalo de los errores, podemos decir que en promedio se tiene 2 cifras de aproximación.

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 5

n = 75 m = 75

tipo pendo

No. de iteraciones realizadas = 10

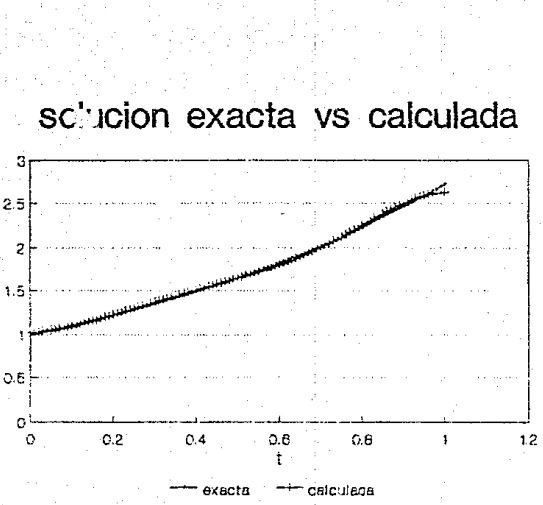
Limite para el condicional = 7.5e-06

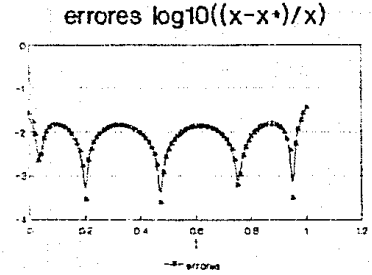
s cuadrada = 1.8e-07

Gamma optima = 7.8e-06

Gamma pseudo-optima = 9.5e-07

pesos = 1 orden = 1





Ejemplo7

n = 101 m = 301

tipo curv

No. de iteraciones realizadas = 49

Limite para el condicional = 3.0e-14

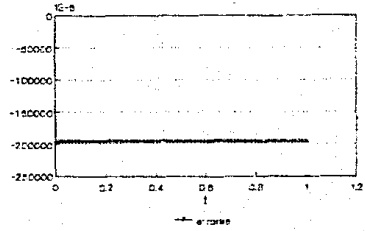
s cuadrada = 3.1e-06

Gamma optima = 2.6e-06

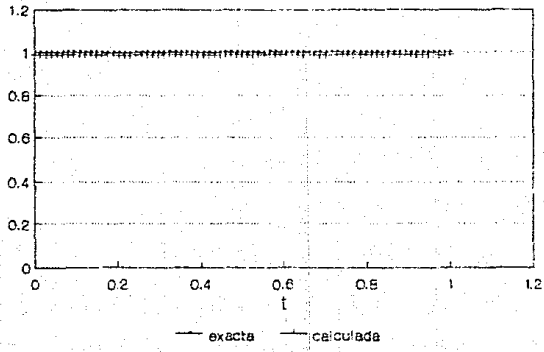
Gamma pseudo-optima = 1.1e-09

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada

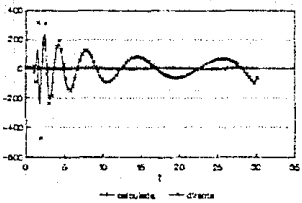


Ejemplo8

n = 75 m = 75
tipo. duro
No. de iteraciones realizadas = 75
Limite para el condicional = 1.0e+12
e cuadrada = -7.7-01 ?
Gamma optima = 5.9e-09
Gamma pseudo-optima =5.7e-04
pesos = 1 orden = 1

Ejemplo 9

solucion calculada vs directa



n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 101

Linha para el condicional = 3.0e-14

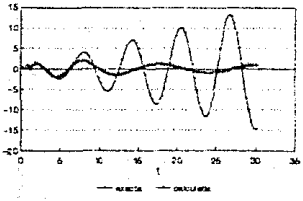
a cuadrada = 1.5e-04

Gamma optima = 3.2e-06

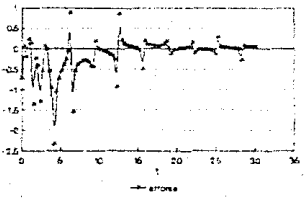
Gamma pseudo-optima = 2.4e-09

pesos = 1 orden = 1

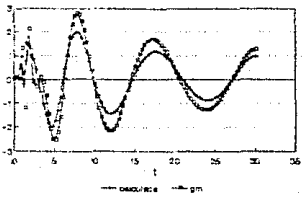
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mark



Comentarios sobre la gráfica

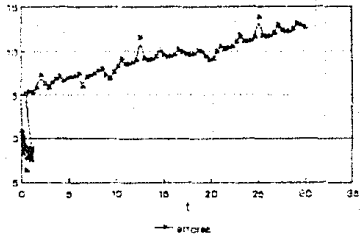
En este ejemplo de la categoria de los "*duros*" se tomó n = 101, m =301 y CONLIM = $3.e^{14}$, el número de iteraciones realizadas fue de K = 101; es decir, no se obtuvó ahorro computacional con el criterio de truncamiento.

El comportamiento de la solución Gauss-Markov es muy similar al de la solución calculada.

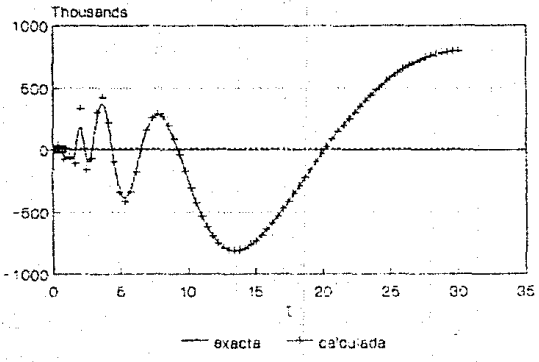
Por lo que respecta a la solución usando el método de "Regularización Directa" ésta presenta valores inaceptables; cabe destacar que sólo en un ejemplo trabajado este método dio resultados aceptables.

Finalmente, observando la gráfica de los errores, éstos varían entre $-2.3e^{0}$ y $8.31e^{-1}$ presentando una pequeña zona de "buena" aproximación y una "mala" aproximación en el resto del intervalo. En conclusión, este ejemplo no dio resultados aceptables.

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 10

n = 75 m = 75

tipo: duro

No. de iteraciones realizadas = 75

Límite para el condicional = 7.5e+12

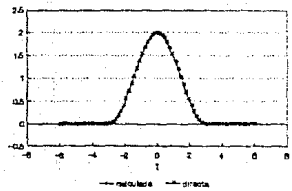
a cuadrado = 1.1e-01

Gamma optima = 1.0e-00

Gamma pseudo-optima = 6.2e-12

pesos = 1 orden = 1

solucion calculada vs directa



Ejemplo 11

n = 75 m = 75

tipo: blando

No. de iteraciones realizadas = 31

Limite para el condicional = 7.5e-06

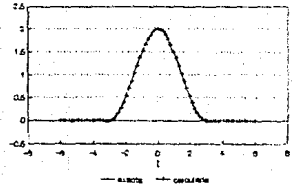
s cuadrada = 7.0e-08

Gamma optima = 5.8e-06

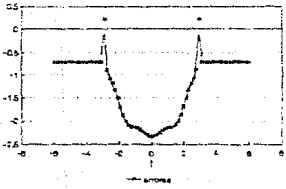
Gamma pseudo-optima = 6.0e-06

pasos = 1 orden = 1

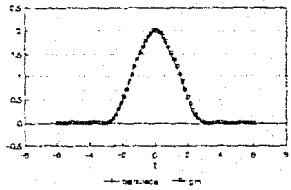
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mark



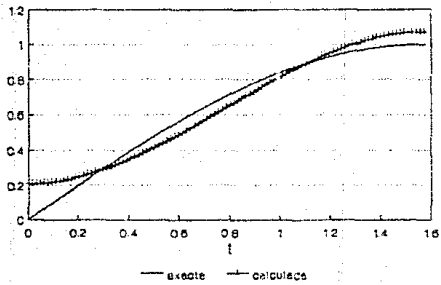
Comentarios sobre la gráfica

En este ejemplo que pertenece a la categoría de los "*blandos*" se tomó n = 75, m = 75 y CONLIN = $7.5e^6$; el número de iteraciones realizadas fue de k = 31 obteniendo un ahorro del 58% en el trabajo realizado .

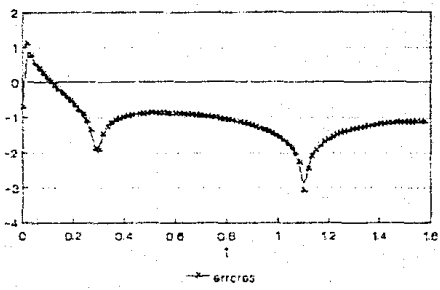
Se observa que no existe una diferencia sustancial entre las soluciones obtenidas por los métodos que se implementaron. Cabe destacar que el método de Regularización Directa sólo en este ejemplo dio resultados aceptables.

Por lo que respecta a los errores éstos presentan 2 picos en los puntos de discontinuidad de la solución (x = -3 y x = 3). En conclusión, en promedio se tiene más de una cifra exacta en la solución calculada.

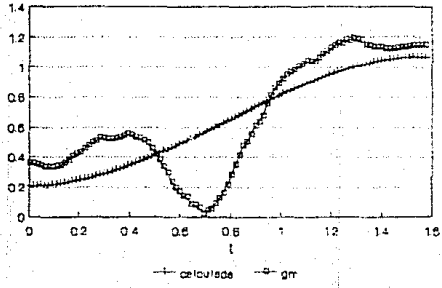
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mark



Ejemplo 12

n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 10

Limite para el condicional = 3.0e+14

s cuadrada = 9.2e-05

Gamma optima = 1.7e-05

Gamma pseudo-optima = 6.3-02

pesos = 1 orden = 1

Comentarios sobre la gráfica

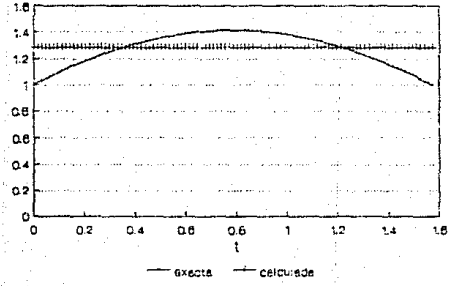
En este ejemplo de la categoría de los "*duros*" se tomó n = 101, m = 301 y CONLIM = $3.e^{14}$.

Es importante señalar que al utilizar el criterio de truncamiento de la Bidiagonalización únicamente se realizaron k = 10 iteraciones con un ahorro del 90% en los cálculos computacionales, es decir, se tiene una importante reducción en los requerimientos de cómputo.

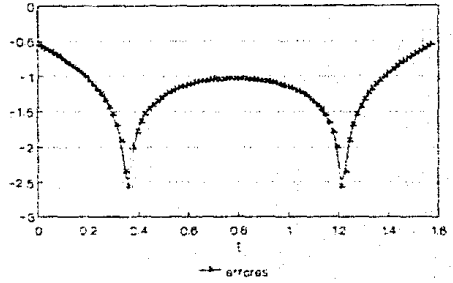
La solución Gauss-Markov si bien no toma valores inaceptables presenta oscilaciones apreciables con respecto a la solución calculada por nuestro método.

Por lo que respecta a los errores éstos presentan un franco descenso en la parte central del intervalo, como se puede observar, varían entre $-3.08e^{0}$ y $1.09e^{0}$, es decir, no se obtuvó ninguna cifra significativa en el extremo inicial del intervalo; sin embargo, se tiene en promedio 2 cifras de aproximación en el resto del intervalo.

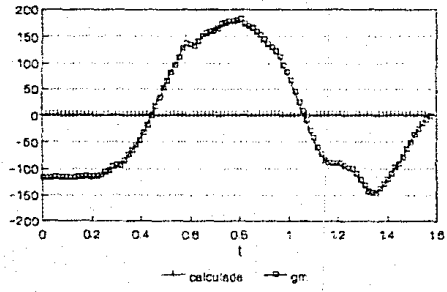
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs Gauss-Mark



Ejemplo 13

n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 8

Limite para el condicional = 3.0e+14

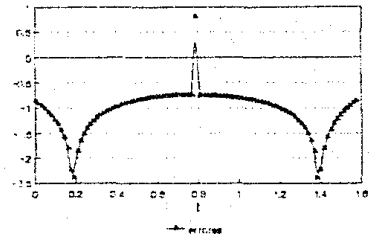
s cuadrada = 1.0e-04

Gamma optima = 3.1e-05

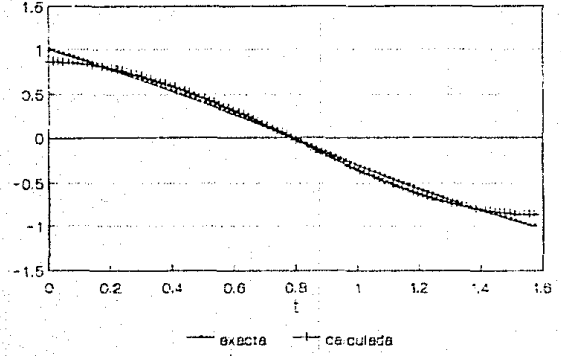
Gamma pseudo-optima = 3.1-05

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 14

n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 8

Limite para el condicional = 3.0e+14

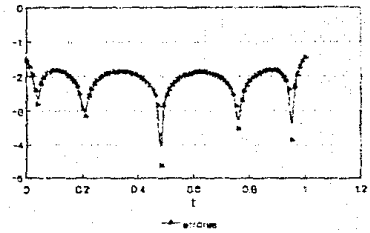
s cuadrada = 8.2e-05

Gamma optima = 2.0e-06

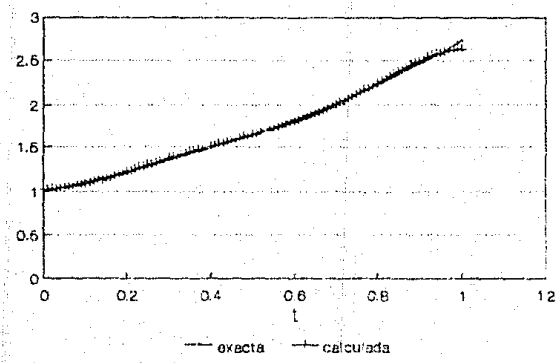
Gamma pseudo-optima = 2.2-02

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo15

n = 101 m = 301

tipo redondo

No. de iteraciones realizadas = 5

Limite para el condicional = 3.0e-11

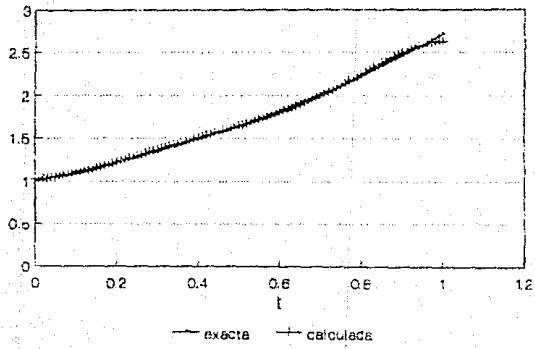
s cuadrado = 0.0e+00

Gamma optima = 0.0e+00

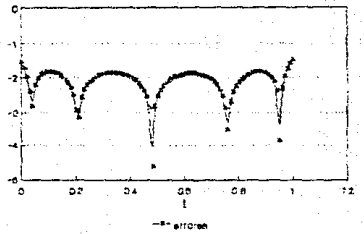
Gamma pseudo-optima = 0.0+00

pesos = 1 orden = 1

solucion exacta vs calculada



errores log10((x-x*)/x)



Ejemplo 16

n = 101 m = 301

tipo duro

No. de iteraciones realizadas = 41

Limite para el condicional = 3.0e+14

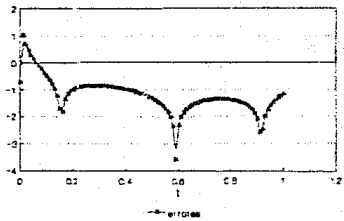
s cuadrada = 1.2e-06

Gamma optima = 1.3e-08

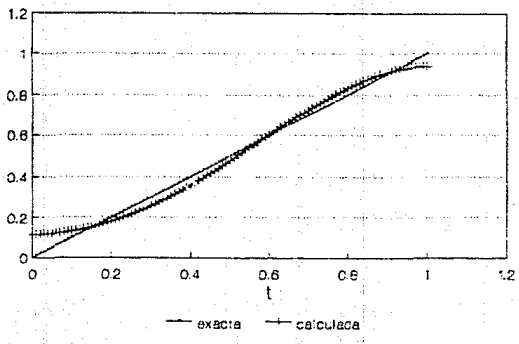
Gamma pseudo-optima = 6.1-10

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 17

n = 101 m = 301

tipo dun

No. de iteraciones realizadas = 38

Limite para el condicional = 3.0e-14

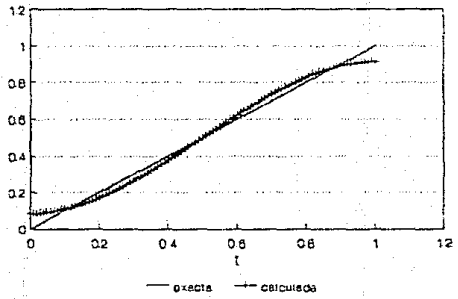
σ cuadrada = 1.2e-02

Gamma optima = 1.2e-02

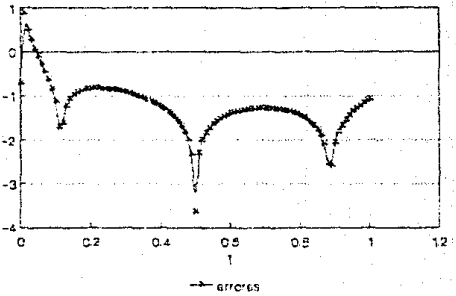
Gamma pseudo-optima = 4.4-05

pesos = 1 orden = 1

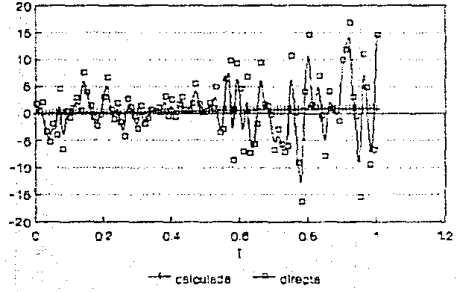
solucion exacta vs calculada



errores log10((x-x*)/x)



solucion calculada vs directa



Ejemplo 18

n = 101 m = 301

tipo: duro

No. de iteraciones realizadas = 18

Limite para el condicional = 3.0e+14

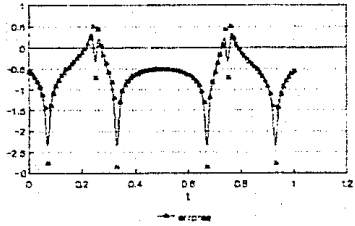
s cuadrada = 8.1e-05

Gamma optima = 5.6e-02

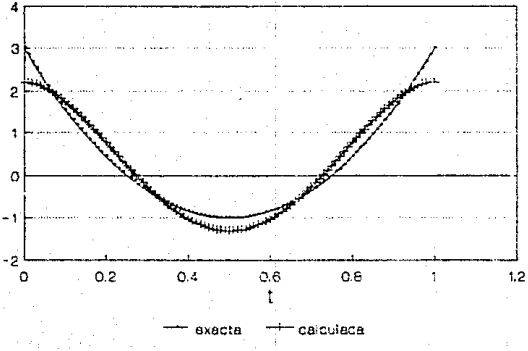
Gamma pseudo-optima = 1.3-01

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 19

n = 101 m = 301

tipo: euro

No. de iteraciones realizadas = 14

Limite para el condicional = 2.0e-14

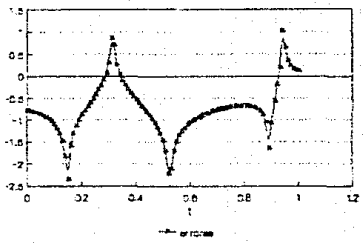
s cuadrada = 5.2e-05

Gamma optima = 4.8e-05

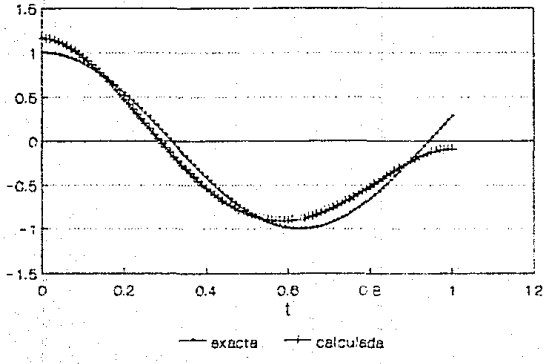
Gamma pseudo-optima = 1.7-04

pesos = 1 orden = 1

errores log10((x-x*)/x)



solucion exacta vs calculada



Ejemplo 20

n = 101 m = 301

tipo dato

No. de iteraciones realizadas = 16

Limite para el condicional = 3.0e+14

e cuadrada = 1.3e-04

Gamma optima = 5.6e-07

Gamma pseudo-optima = 6.1-04

pesos = 1 orden = 1

CAPITULO 6.- CONCLUSIONES

El problema de resolver numéricamente un sistema mal condicionado de mediana ó gran escala tiene características muy importantes: por un lado, el tamaño de la matriz obliga a tener un método que optimice los recursos de memoria y por el otro, el mal comportamiento de la matriz requiere métodos que sean estables a pequeñas perturbaciones.

Los métodos clásicos para resolver este problema requieren de memoria adicional o modifican internamente la matriz original generando elementos diferentes de cero en lugares donde había ceros, es decir, llenan la matriz. Este hecho genera una contradicción.

La principal conclusión al respecto en el presente trabajo es el haber logrado una combinación de dos métodos, a saber el método de Bidiagonalización de Lanczos y el método de Regularización de Tijonov con la elección automática del parámetro de Regularización, resolviendo en forma razonable las dos grandes dificultades prácticas antes mencionadas.

Dicho de otra forma, en esta tesis se presenta un método nuevo, estable y eficiente para la resolución de problemas de mínimo de cuadrados muy mal condicionados con pocos requerimientos de memoria y cómputo.

Cabe mencionar que existen trabajos en la literatura en la misma dirección. Paige-Saunders [16] presentan el algoritmo LSQR para resolver

$$\underset{x}{Min} \left\| \begin{pmatrix} A \\ \gamma I \end{pmatrix} x - \begin{pmatrix} b \\ \hline 0 \end{pmatrix} \right\|_2^2$$

conocido como mínimos cuadrados amortiguados, lo cual es equivalente a la regularización de Tijonov de orden 0 con γ dada de antemano. Nuestro algoritmo es una generalización de LSQR en dos direcciones: por un lado, nuestra propuesta admite ordenes de regularización mayores que cero, en nuestro trabajo experimental se contemplaron ordenes = 0, 1 y 2 ;por otro lado nuestra propuesta estima automáticamente el parámetro γ de regularización evitando el ensayo. Este último aspecto adquiere gran relevancia en comparación con LSQR debido a que este último recalcula toda la bidiagonalización y factorizaciones involucradas cada vez que cambia el parámetro γ de regularización.

O' Leary- Simmons[17] presentan un algoritmo que combina la bidiagonalización Superior de Lanczos con el método de Regularización Directa. En el capitulo 2 del presente trabajo se hace ver las ventajas de la bidiagonalización inferior con relación a la superior. En este mismo articulo se da un criterio muy burdo para truncar la bidiagonalización en base al cual se presenta un método de regularización directa.

Este método no dio resultados aceptables para los ejemplos de ecuaciones integrales de la categoria de los *duros* que en esta tesis se estudiaron. En nuestro estudio experimental con el método de Regularización Directa propuesto por O'Leary-Simmons se logró reproducir los resultados reportados para el ejemplo de Phillips. Es importante dejar asentado que éste fue el único de nuestros ejemplos para el cual dicho método dió resultados aceptables. Así pues, el método propuesto en esta tesis resulta ser muy superior

al método de regularización directa vía bidiagonalización propuesto por O'Leary-Simmons.

La tercera conclusión es relativa a los principios de optimalidad y pseudo-optimalidad para la estimación del parámetro de regularización y consiste en la obtención de una versión matricial de éstos en términos de la norma Frobenius mejorando las formulas obtenidas en [1] para la varianza y el sesgo cuadrado, además, se obtiene una expresión equivalente y se desarrolla un método para calcular de manera más estable las γ's.

Para finalizar, no podemos dejar de mencionar que se requiere mantener en memoria los vectores $q_1, q_2, \ldots, q_k$, los cuales se utilizan para recuperar las transformaciones ortogonales utilizadas. El problema (abierto) radica en obtener un método "barato" y de manera iterativa para el cálculo de la descomposición singular o bien para la estimación de las γ's a partir directamente de la α's y β's de la bidiagonalización.

BIBLIOGRAFIA

[1] López Jc., *Principio de seudo-optimalidad en la Resolución de Problemas mal Planteados por el método de Regularización de Tijonov*, Tesis Doctoral U.N.A.M., 1988.

[2] Bachman G., Narici L., *Functional analysis*, Academic Press (1966).

[3] Lanczos C., *Solution of Systems of Linear Equations by Minimized Iterations*, Journal of Research of the National Bureau of Standars, vol 49, No.1, July 1952, 33-53.

[4] Elden L., Björck A., *Algorithms for the Regularization of ill-conditioned Least Squares Problems*, Bit 17(1977), 134-145.

[5] Golub G., Kahan W., *Calculating the Singular Values and Pseudo-inverse of a Matrix*, Siam Numer. Anal. Ser B, vol 2, No.2 (1986), 205-224.

[6] Golub G., Luk F., Overton M., *A Block Lanczos Method for Computing the Singular Values and Corresponding Singular Vectors of a Matrix*, ACM Transantions on Mathematical Software, vol 17, No. 2, June 1981, 149-169.

[7] Saavedra A. *Método de Lanczos para Sistemas Lineales*

Grandes, Huecos y Simétricos, Tesis Profesional U.N.A.M. (1986).

[8] Paige C., *Bidiagonalization of Matrices and Solution of Linear Equations*, Siam J. Numer Anal. Vol 11, No. 1, March 1974, 197-208.

[9] Dubner H., Abate J., *Numerical Inversion of Laplace Transforms by Relating them to the Finite Fourier Cosine Transform*, Journal of the Associations for Computing Machinery, vol 15, No. 1, January 1968, 115-123.

[10] Arthur D., *The Solution of Fredholm Integral Equations Using Splines Funtions*, Journal Inst. Maths. Applics. (1973) 11,121-129.

[11] Marti J., *An Algorithm for Computing Minimum Norm Solutions of Integral Equations of the First Kind*, Siam J. Numer Anal., vol 15, No. 6, Dec 1978, 1071-1076.

[12] Miller M., Guy W., *Numerical Inversion of the Laplace Transform by use Jacobi Polynomials*, Siam J. Numer. Anal., vol 3, No. 4, 1966, 624-635.

[13] Graves J., Prenter P., *Numerical Iterative Filters Applied of First Kind Fredholm Integral Equations*, Numer. Math. 30, 281-299(1978).

[14] Fox L., Goodwin E., *The Numerical Solution of Non-Singular Linear Equations*, Phil. Trans. Roy. Soc. 241(1953).

[15] Babolian E., Delves L., *An Augmented Galerkin Method for First Kind Fredholm Equations*, J. Inst. Maths. Applics. 24, 157-174, (1979).

[16] Paige C., Saunders M., *LSQR: An Algorith for Sparse Linear Equations and Sparse Least Squares*, ACM Transactions on Mathematical Software vol 8, No.1, March 1982, 43-71.

[17] O'Leary D., Simmons J., *A Bidiagonalization-Regularization Procedure for Large Scale Discretizations of Ill-Posed Problems*, Siam J. Sci. Stat. Comput. vol 2, NO. 4, 1981, 474-489.

[18] Paige C., *Error analysis of the Lanczos Algorithm for Tridiagonalizing a Symmetric Matrix*, J. Inst. Maths. Applics. 18(1976), 341-349.

[19] Golub G., Van Loan Ch., *Matrix Computations*, The Johns Hopkins, University Press (1989).

[20] Bulirsch R., Stoer J., Introducción al Análisis Numérico, Springer Verlag (1980).