

CARACTERISTICA, ALMACENAMIENTO DE UNA  
MATRIZ DE RIGIDEZ Y COMPARACION DE  
METODOS DEL GRADIENTE CONJUGADO  
Y DE CHOLESKY

Hector M. Mora E.

0. INTRODUCCION.

El cálculo de los esfuerzos (momento y cortante) en un pórtico plano (el esqueleto de un edificio) por el método de la rigidez se hace a partir de los desplazamientos resultantes de la aplicación de las cargas. El valor de los desplazamientos se obtiene resolviendo el sistema de ecuaciones  $Ax = b$  donde  $A$  es la matriz de rigidez. Usualmente se utiliza el método de Cholesky. Como se muestra en [4] también se puede utilizar el método del gradiente conjugado.

En este artículo se muestra un almacenamiento más compacto de la matriz de rigidez que disminuye el tiempo de cálculo del método del gradiente conjugado y se comparan los tres métodos (Cholesky y dos almacenamientos para el gradiente conjugado) desde el punto de vista de memoria, tiempo de cálculo y precisión.

## 1. DESCRIPCION DEL METODO DE CHOLESKY Y DEL METODO DEL GRADIENTE CONJUGADO.

El método de Cholesky (MCH) y el método del gradiente conjugado (MGC) sirven para resolver sistemas de ecuaciones lineales  $Ax = b$  donde la matriz  $A$  de tamaño  $n \times n$  es simétrica y definida positiva,  $x^t = \{x_1 \ x_2 \dots x_n\}$  indica las incógnitas y  $b^t = \{b_1 \ b_2 \dots b_n\}$  corresponde a los términos independientes.

**DEFINICION.** Una matriz  $A$  de tamaño  $n \times n$ , simétrica es definida positiva si  $x^t Ax > 0$  para todo  $x$  no nulo.

**TEOREMA.** Sea  $A$  una matriz de tamaño  $n \times n$ , simétrica; las siguientes proposiciones son equivalentes [2]:

(i)  $A$  es definida positiva.

(ii) Los valores propios de  $A$  son positivos.

(iii) Los menores principales  $\delta_1, \delta_2, \dots, \delta_n$  de  $A$  son positivos.

$$\delta_1 = \det[a_{11}], \quad \delta_2 = \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \dots, \delta_n = \det(A)$$

(iv) Existe una matriz  $U$  de tamaño  $n \times n$ , triangular superior, regular, tal que  $A = U^t U$ .

El MCH se basa en la proposición (iv): inicialmente se calcula la matriz  $U$  tal que  $A = U^t U$ ; en seguida para resolver  $U^t Ux = b$  se resuelve primero  $U^t y = b$ ; por último, conocido  $y$  se resuelve  $Ux = y$ ; la solución de estos dos sistemas es sencilla pues son sistemas triangulares.

Sea  $x^*$  la solución exacta del sistema  $Ax = b$ ; para un  $x$  cualquiera, aproximación mala, regular o buena de  $x^*$  se define:

$$\text{-vector residuo} = r(x) = r = Ax - b$$

$$\text{-vector error} = e(x) = e = x - x^*$$

Obviamente

$$r(x^*) = r^* = 0$$

$$e(x^*) = e^* = 0$$

y además

**PROPOSICION.** Dado el sistema  $Ax = b$ , las siguientes afirmaciones son equivalentes:

- i)  $\tilde{x}$  es solución del sistema.
- ii)  $r(\tilde{x}) = \tilde{r} = 0$
- iii)  $e(\tilde{x}) = \tilde{e} = 0$ .

El MGC también sirve para resolver  $Ax = b$ , con  $A$  definida positiva. Sin embargo su objetivo original es resolver otro problema equivalente [2]:

$$\text{minimizar } f(x) = \frac{1}{2} x^t A x - x^t b$$

o sea encontrar  $x^*$  tal que

$$f(x^*) \leq f(x) \quad \text{para todo } x.$$

En el MGC dado un  $x^{(0)}$  aproximación mala, regular o buena de la solución se construyen aproximaciones mejores  $x^{(1)}, x^{(2)}, \dots, x^{(n)}$  y se garantiza teóricamente [1] que la solución  $x^*$  se obtiene a más tardar con  $x^{(n)}$ , o sea que  $x^* = x^{(k)}$  para algún  $k$  no superior a  $n$ .

Las fórmulas que definen el MGC y que garantizan su convergencia finita son las siguientes:

$$d^{(0)} = -r^{(0)}$$

$$\left. \begin{aligned} \eta_k &= \frac{\|r(k)\|_2^2}{\|r(k+1)\|_2^2} \\ d(k) &= -r(k) + \eta_k d^{(k-1)} \end{aligned} \right\} k = 1, 2, \dots, n-1$$

$$\left. \begin{aligned} h_k &= \frac{\|r(k)\|_2^2}{d(k)^t A d(k)} \\ x^{(k+1)} &= x^{(k)} + h_k d(k) \end{aligned} \right\} k = 0, 1, 2, \dots, n-1$$

donde  $\|u\|_2$  es la norma euclidiana de  $u$  definida por:

$$\|u\|_2^2 = \sum_{i=1}^n u_i^2$$

De manera informal el algoritmo se puede describir así:

- sea  $x^{(0)}$  una aproximación inicial de  $x^*$
- sea  $\epsilon > 0$  número pequeño utilizado para fijar la precisión
- $r^{(0)} = Ax^{(0)} - b$
- si  $\|r^{(0)}\|_2^2 < \epsilon$  vaya a (F)
- $\eta_0 = 0$ .
- para  $k = 0$  hasta  $n-1$  hacer
  - $d(k) = -r(k) + \eta_k d^{(k-1)}$
  - $h_k = \frac{\|r(k)\|_2^2}{d(k)^t A d(k)}$

$$- x^{(k+1)} = x^{(k)} + h_k d^{(k)}$$

$$- r^{(k+1)} = Ax^{(k+1)} - b$$

$$- \text{si } \|r^{(k+1)}\|_2^2 < \varepsilon \text{ vaya a } \textcircled{F}$$

$$- \eta_{k+1} = \frac{\|r^{(k+1)}\|_2^2}{\|r^{(k)}\|_2^2}$$

- fin para  $k$

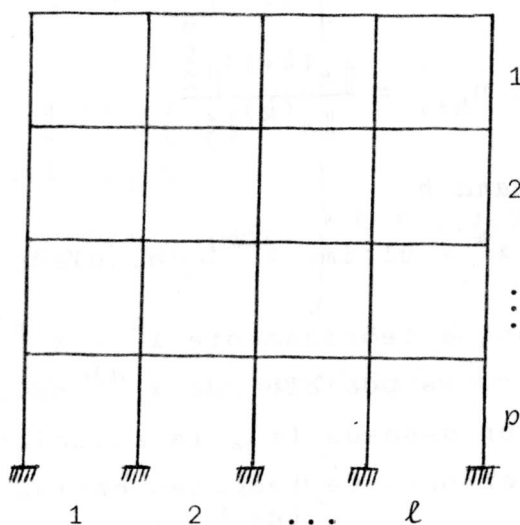
-  $\textcircled{F}$ :  $x^* = \text{último } x^{(k)} \text{ calculado.}$

Aunque teóricamente  $x^* = x^{(n)}$  o algún  $x^{(k)}$  anterior, es posible que  $x^{(n)}$  no sea, con la precisión deseada ( $\varepsilon$ ), la solución  $x^*$  a causa de los errores de redondeo en las operaciones, sin embargo este  $x^{(n)}$  será una aproximación muy buena de  $x^*$  para empezar un nuevo ciclo de  $n$  iteraciones.

## 2. CARACTERISTICAS Y ALMACENAMIENTO DE LA MATRIZ DE RIGIDEZ.

Cuando se estudia un pórtico plano de  $p$  pisos y  $\ell$  luces constituido por barras horizontales y verticales de inercia constante, considerando tres grados de libertad (desplazamiento horizontal, desplazamiento vertical y giro) en cada nudo (intersección de una ba

rra vertical y una barra horizontal), si las bases de las columnas inferiores están empotradas (apoyo que no permite desplazamientos ni giros)



y si se utiliza el método matricial de rigidez se obtiene un sistema de ecuaciones  $Ax = b$ , donde  $A$  es la matriz de rigidez de tamaño  $n \times n$  con  $n = 3p(\ell+1)$  simétrica y definida positiva [3],  $b$  es el vector de términos independientes (acciones) construido a partir de las cargas que soporta la estructura y la solución del sistema, el vector  $x^*$ , indica los desplazamientos (horizontal, vertical, giro) de los nudos.

Para el pórtico en consideración la matriz de rigidez  $A$  además de ser simétrica, presenta

otras características que veremos a continuación.

La matriz  $A$  es una matriz banda [3], [4] con ancho superior de banda =  $m = 3(\ell+1)$ , es decir

$$a_{ij} = 0 \quad \begin{array}{l} i = 1, 2, \dots, n \\ i+m \leq j \leq n. \end{array}$$

Esto permite (es lo que se hace usualmente) almacenar únicamente los elementos de la banda superior de  $A$  (matriz de  $n^2$  elementos) en un arreglo  $B = [\beta_{ij}]$  más pequeño de tamaño  $n \times m$ , con la siguiente correspondencia:

$$\beta_{ij} = a_{i, i+j-1} \quad \begin{array}{l} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \\ i+j-1 \leq n. \end{array}$$

$$\beta_{ij} = 0 \text{ (o sin valor) en los demás casos}$$

Este tipo de almacenamiento es el que se utiliza para el método de Cholesky, puesto que la matriz  $U$ , además de ser triangular superior, también es banda de ancho  $m$ ,

$$u_{ij} = 0 \quad \begin{array}{l} i = 1, 2, \dots, n \\ i+m \leq j \leq n \end{array}$$

y permite guardar los elementos de  $U$  donde estaban los elementos de  $A$ .



La observación del arreglo  $B$ , [4] permite detectar que únicamente ocho columnas de  $B$  tienen elementos no nulos, las número 1,2,3,4,5,  $m-4$ ,  $m-2$ ,  $m$ .

Utilizando la siguiente función  $\psi$

	$\psi$	
1	$\longrightarrow$	0
2	$\longrightarrow$	1
3	$\longrightarrow$	2
4	$\longrightarrow$	3
5	$\longrightarrow$	4
6	$\longrightarrow$	$m-5$
7	$\longrightarrow$	$m-3$
8	$\longrightarrow$	$m-1$

se puede obtener un almacenamiento de  $A$  (o de  $B$ ) más compacto en un arreglo  $\Omega = [\omega_{ij}]$  de tamaño  $n \times 8$  con la correspondencia

$$\begin{aligned} \omega_{ij} &= \beta_{i, \psi(j)+1} = a_{i, i+\psi(j)} & i &= 1, 2, \dots, n \\ & & j &= 1, 2, \dots, 8 \\ & & & i+\psi(j) \leq n \\ \omega_{ij} &= 0 & & \text{en los demás casos} \end{aligned}$$

Este tipo de almacenamiento permite resolver el sistema  $Ax = b$  utilizando el MGC [4], sin embargo no permite la utilización del método de Cholesky pues aunque inicialmente el



$\mu \backslash v$	1	2	3	4	5
1	0	2	3	$m-3$	$m-1$
2	0	1	3	4	$m-3$
3	0	2	3	$m-5$	$m-3$

se obtiene un almacenamiento de  $A$  más compacto en un arreglo  $c = [c_{ij}]$  de tamaño  $n \times 5$  mediante la siguiente correspondencia:

$$c_{ij} = a_{i, i + \varphi(\mu, j)}$$

$$i = 3k + \mu$$

$$\mu = 1, 2, 3$$

$$k = 1, 2, \dots, n/3$$

$$j = 1, 2, \dots, 5$$

$$i + \varphi(\mu, j) \leq n$$

$$c_{ij} = 0$$

en los demás casos.

La adecuación de este almacenamiento al MCC requiere hacer: 1) un subprograma que construye la matriz de rigidez directamente con esta forma de almacenamiento. 2) un subprograma que efectúe el producto de la matriz  $A$  por un vector columna cualquiera, teniendo en cuenta que los valores representativos de  $A$  están guardados en el arreglo  $c$ . 3) un subprograma que utilice el MCC valiéndose del subprograma 2).

Este método que llamaremos en adelante GC5

permite resolver el sistema  $Ax = b$ , mejorando en algunos aspectos las características de GC8 (método del gradiente conjugado utilizando el arreglo  $\Omega$  de tamaño  $n \times 8$ ) y de CH (método de Cholesky utilizando el arreglo  $B$  de tamaño  $n \times m$ , con  $m = 3(\ell + 1)$  ).

### 3. COMPARACION DE LOS METODOS.

Diferentes métodos para resolver un mismo problema se pueden comparar según las siguientes características:

- 1) Facilidad para hacer el programa o subprograma.
- 2) Cantidad necesitada de memoria principal del computador.
- 3) Tiempo de computador.
- 4) Precisión en los resultados.

**3.1. *Facilidad de programación.*** Esta característica es casi imposible de cuantificar, sin embargo se puede decir que la implementación de los tres métodos tiene características de facilidad o de dificultad semejantes, no es sencilla ni demasiado complicada.

**3.2. *Cantidad de memoria central del computa-***

**dor.** Fuera de unos pocos arreglos unidimensionales, el factor limitante de la memoria, es el almacenamiento de los datos representativos de la matriz de rigidez  $A$ . Para un pórtico de  $p$  pisos y  $\ell$  luces el número de elementos de  $B$  (en CH) es  $nm = 3p(\ell+1)3(\ell+1) = 9p(\ell+1)^2$ ; el número de elementos de  $\Omega$  (en GC8) es  $8n = 8.3p(\ell+1) = 24p(\ell+1)$  y el número de elementos del arreglo  $c$  (en GC5) es  $5n = 5.3p(\ell+1) = 15p(\ell+1)$ . Veamos algunos resultados.

	C H	G C 8	G C 5
	$9p(\ell+1)^2$	$24p(\ell+1)$	$15p(\ell+1)$
$p=6, \ell=5, n=108$	1944	864	540
$p=12, \ell=11, n=432$	15552	3456	2160

Los ejemplos muestran claramente la ventaja de GC5 con respecto a GC8 y sobre todo con respecto a CH.

**3.3. Tiempo de computador.** Como se están comparando métodos de resolución del sistema  $Ax = b$ , únicamente se tuvo en cuenta el tiempo de ejecución del subprograma o subprogramas que resuelven el sistema de ecuaciones. No se tuvo en cuenta el tiempo de compilación, ni el tiempo

po de edición de enlace (LINKED), ni el tiempo de ejecución gastado en entrada de datos, construcción de la matriz de rigidez y del vector  $b$  y escritura de resultados. Los resultados aparecen en el cuadro final y corresponden a ejemplos resueltos utilizando el computador IBM 360 de la Universidad Nacional, en Fortran y con precisión sencilla.

**3.4. Precisión en los resultados.** Cuando no se conoce la solución exacta de un problema es casi imposible medir exactamente el error de la aproximación de la solución, sin embargo en este caso se puede estudiar cualitativamente y mediante otros parámetros. La siguiente proposición da algunas ideas al respecto.

**PROPOSICION.** Sea el sistema  $Ax = b$ ,  $x^*$  su solución exacta,  $\{x^{(k)}\}$  una sucesión de vectores,  $r^{(k)} = Ax^{(k)} - b$ ,  $e^{(k)} = x^{(k)} - x^*$ . Las afirmaciones siguientes son equivalentes.

$$i) \quad \lim_{k \rightarrow \infty} x^{(k)} = x^*$$

$$ii) \quad \lim_{k \rightarrow \infty} r^{(k)} = 0$$

$$iii) \quad \lim_{k \rightarrow \infty} e^{(k)} = 0.$$

Dada una solución aproximada  $\tilde{x}$ , se desea ver qué tan pequeño es  $\tilde{e} = \tilde{x} - x^*$ , pero  $\tilde{e}$  no se puede conocer; sin embargo sí es posible conocer  $\tilde{\kappa} = A\tilde{x} - b$  y si  $\tilde{\kappa}$  es muy pequeño se puede suponer que  $\tilde{e}$  es muy pequeño y que  $\tilde{x}$  es buena aproximación de  $x^*$ ; el inconveniente de este análisis es la falta de cuantificación. El tamaño de  $e$  o de  $\kappa$  se mide mediante una norma y se usó  $\| \cdot \|_2$ , la norma euclídeana (ó  $\| \cdot \|_2^2$ ) y la norma del máximo  $\| \cdot \|_\infty$  definida por  $\|u\|_\infty = \max_{1 \leq i \leq n} |u_i|$ . Dada una norma  $\| \cdot \|$  para vectores, se genera una norma para matrices

$$\| \| A \| \| = \max_{x \neq 0} \frac{\| Ax \|}{\| x \|}$$

que resulta ser compatible con  $\| \cdot \|$ , [1], es decir

$$\| Ax \| \leq \| \| A \| \| \| x \| \quad \text{para toda matriz } A \text{ y todo vector } x.$$

Con base en estas propiedades se tiene que

$$\begin{aligned} \tilde{\kappa} &= A\tilde{x} - b \\ &= A\tilde{x} - b + 0 \\ &= A\tilde{x} - b + b - Ax^* \\ &= A(\tilde{x} - x^*) \\ &= A\tilde{e} \end{aligned}$$

luego

$$\|\tilde{x}\| \leq \|A\| \|\tilde{e}\|$$

Este resultado es muy útil pero *no permite* afirmar

$$\|\tilde{x}\| < \|\bar{x}\| \Rightarrow \|\tilde{e}\| < \|\bar{e}\|$$

para  $\tilde{x}$ ,  $\bar{x}$  aproximaciones de  $x^*$ . En el cuadro final de resultados es claro el ejemplo ( $p = 6$ ,  $\ell = 5$ ,  $n = 108$ , simulación de carga horizontal) donde

$$\|r_{CH}\| < \|r_{GC5}\| \quad \text{y} \quad \|e_{CH}\| > \|e_{GC5}\|$$

También se puede obtener un resultado para acotar  $\|e\|$  ya que

$$\tilde{x} = A\tilde{e}$$

$$\tilde{e} = A^{-1}\tilde{x}$$

$$\|\tilde{e}\| \leq \|A^{-1}\| \|\tilde{x}\|$$

pero su utilización supone conocer  $A^{-1}$  y calcular  $\|A^{-1}\|$ , lo cual generalmente es un trabajo mucho más largo que resolver  $Ax = b$ .

En la práctica la única manera de comparar la precisión de una aproximación es mediante  $\|r\|$ . En este trabajo para poder comparar la precisión en función de  $\|e\|$ , el procedimiento utilizado es el siguiente: **1)** Tomar una solución aproximada  $\tilde{x}^*$  del sistema  $Ax = b$  y fijar la como solución exacta. **2)** Construir  $\tilde{b}^* =$

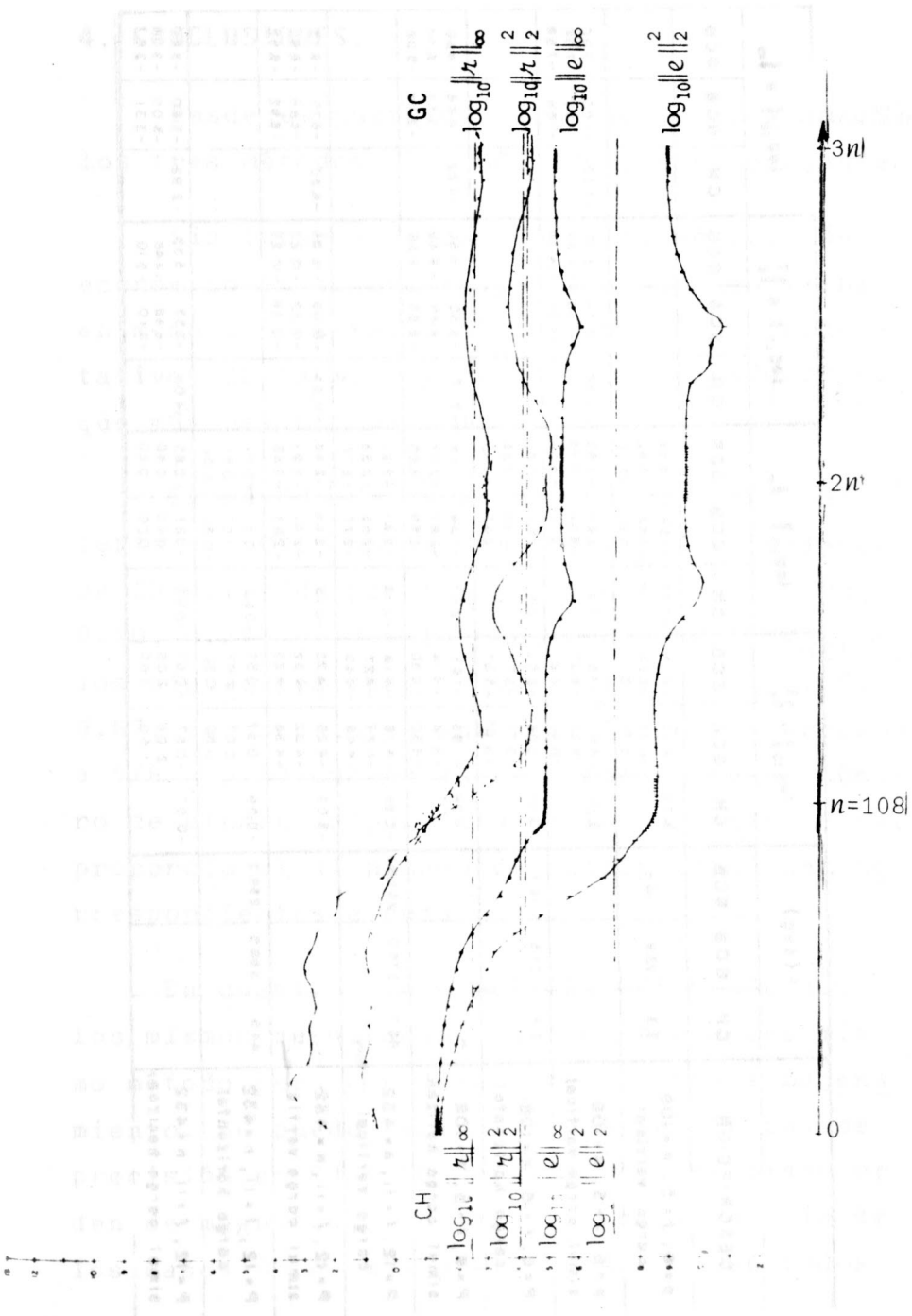


$A\tilde{x}^*$ . 3) Resolver el sistema  $Ax = \tilde{b}^*$ . 4) Comparar con  $\tilde{x}^*$  las soluciones de este segundo sistema.

Para comparar de manera más fácil la precisión de los métodos en los diferentes ejemplos se utilizó el valor  $\log_{10} \| \|$  en lugar del valor  $\| \|$ . Para el MGC (GC8 ó GC5), en cada ejemplo aparecen tres valores correspondientes a los resultados finales de tres ciclos consecutivos, siendo cada ciclo de  $n$  iteraciones.

Las curvas de la gráfica adjunta muestran la tendencia de la variación de  $\log_{10} \| \|$  a lo largo de tres ciclos consecutivos en el MGC, comparada con el valor correspondiente de  $\log_{10} \| \|$  en el MCH, en lo que podría ser un ejemplo típico, (en este caso  $p = 6$ ,  $\ell = 5$ ,  $n = 108$ , con simulación de carga horizontal).

\*



DESCRIPCION	$t \text{ (seg)}$			$\log_{10} \left  \frac{z}{z_0} \right $			$\log_{10} \left  \frac{r}{r_0} \right $			$\log_{10} \left  \frac{z}{z_0} \right $			$\log_{10} \left  \frac{r}{r_0} \right $		
	CH	GC8	GC5	CH	GC8	GC5	CH	GC8	GC5	CH	GC8	GC5	CH	GC8	GC5
$P=6, f=5, n=108$ carga vertical	33	239	164	-5.06	-6.25	-6.25	-2.93	-3.56	-3.56	-2.93	-3.56	-3.56			
$P=6, f=5, n=108$ simul. carga vertical				-5.10	-6.61	-6.27	-2.91	-2.69	-3.56	-2.69	-13.34	-13.34	-7.20	-7.00	-7.00
$P=6, f=5, n=108$ carga horizontal	34	235	167	-3.99	-2.52	-2.52	-2.47	-1.81	-1.81	-2.47	-14.33	-14.33		-7.58	-7.58
$P=6, f=5, n=108$ simul. carga horizontal				-4.06	-1.83	-1.83	-2.47	-1.34	-1.34	-2.47	-7.17	-8.50	-4.22	-4.86	-4.86
$P=12, f=11, n=432$ carga vertical	467	3750	2689	-2.95	-4.18	-4.18	-2.13	-2.61	-2.61	-2.13	-2.53	-2.53		-5.34	-5.34
$P=12, f=11, n=432$ simul. carga vertical				-3.09	-4.25	-4.25	-2.15	-2.66	-2.66	-2.15	-11.33	-12.05	-6.30	-5.08	-5.08
$P=12, f=11, n=432$ carga horizontal	449	3865	2593	-0.09	-0.37	-0.37	-0.69	-0.71	-0.71	-0.69	-12.22	-12.22		-6.75	-6.75
$P=12, f=11, n=432$ simul. carga horizontal				-0.07	2.06	2.06	-0.63	0.40	0.40	-0.63	-12.26	-12.26	-2.92	-6.92	-6.92
					1.60	1.60		0.20	0.20		-5.33	-5.33	-3.60	-3.60	-3.60
											-4.48	-4.48	-3.09	-3.09	-3.09
											-5.10	-5.10	-3.31	-3.31	-3.31

#### 4. CONCLUSIONES.

Desde el punto de vista de la programación los tres métodos (CH, GC8, GC5) son semejantes.

Sin lugar a dudas GC5 es el método más económico en lo que respecta a almacenamiento en memoria central de los elementos representativos de la matriz de rigidez, siendo CH el que más memoria necesita.

El método CH es el más rápido y el más lento es GC8. Los valores de  $t_{CH}/t_{GC5}$  (tiempo de CH dividido por tiempo de GC5) son: 0.20, 0.20, 0.17, 0.17 y la comparación  $t_{GC5}/t_{GC8}$  da los siguientes resultados: 0.69, 0.71, 0.72, 0.67; estos valores son relativamente cercanos a  $5/8 = 0.63$  que es la relación entre el número de elementos por fila de los arreglos  $c$  y  $\Omega$ , proporcional al número de multiplicaciones correspondientes a cada fila.

En cuanto a la precisión GC5 y GC8 dan los mismos resultados, pues se trata del mismo método pero con diferente tipo de almacenamiento. Se puede decir que los resultados de precisión en CH y GC dan valores del mismo orden de magnitud, sin embargo en la mayoría de los casos GC es más preciso que CH (en todos

los casos estudiados con  $n = 432$ ), en otros pocos casos los resultados de GC en el primer ciclo no son mejores que los de CH pero mejoran parcialmente o totalmente para el segundo ciclo de GC.

En algunos casos los resultados de GC en el segundo ciclo son un poco menos buenos que en el primer ciclo, y en los demás casos la mejora del segundo ciclo con respecto al primero no es demasiado importante. Esto aconsejaría, para valores de  $n$  semejantes a los estudiados, hacer un solo ciclo de  $n$  iteraciones.

Los valores de la precisión en función de  $\|r\|$  sí indican precisión en función de  $\|e\|$  pero es una indicación cualitativa y no existe proporción absoluta.

En resumen, las características de GC5 son siempre mejores o iguales a las de GC8. El método GC5 es el que necesita menos memoria y el método CH es el más rápido. El método GC5 presenta una tendencia a ser más preciso que el método CH.

## BIBLIOGRAFIA

- [1] Arminjon, P., "Analyse Numerique Matri-  
cielle", Les Presses de l'Université de Mon-  
treal, 1978.
- [2] Gastinel, N., *Análisis Numérico Lineal*,  
Reverté, Barcelona, 1975.
- [3] Gere, J.M., Weaver, W. Jr., *Análisis de  
Estructuras Reticulares*, Compañía Edito-  
rial Continental S.A. (CECSA), México,  
1967.
- [4] Mora, H., *Aplicación del Método del Gra-  
diente conjugado al programa PORTRECT*,  
Proyecto de grado en Ingeniería Civil,  
Universidad Nacional, Bogotá, 1984.

\* \*

Departamento de Matemáticas y Estadística  
Universidad Nacional de Colombia  
BOGOTÁ. D.F.

*Boletín de Matemáticas*  
Vol. XVIII, Nº 1,2,3, (1984)

**APUNTES**

## **A P U N T E S**

La sección de **APUNTES** del Boletín de Matemáticas ha sido creada para ayudar al profesor a motivar a sus alumnos para el estudio de un tema específico o/y proveer un desarrollo detallado y detenido de un tema tratado someramente por lo general en los libros de texto.

Igualmente en ella los alumnos encontrarán material que pueda complementar la presentación de un tema en el salón de clase.