

Universitat Politècnica de València
Departamento de Matemática Aplicada



Métodos Numéricos para la Resolución de la Ecuación
de la Difusión Neutrónica Dependiente del Tiempo

TESIS DOCTORAL

Presentada por:

JOSÉ MARÍN MATEOS-APARICIO

Dirigida por:

RAFAEL BRU GARCÍA

DAMIÁN GINESTAR PEIRÓ

Universidad Politécnica de Valencia

Departamento de Matemática Aplicada



Métodos Numéricos para la Resolución de la Ecuación de la Difusión Neutrónica Dependiente del Tiempo

Memoria presentada para optar al grado de doctor en Ciencias Matemáticas por JOSÉ MARÍN MATEOS-APARICIO, subvencionada por el proyecto de investigación DGES PB97-0334.

Valencia, Junio de 2000

*A mis padres, Eleuterio y Antonia,
a Toñi y Luis, mis hermanos, y a
Asun: por mimarme.*

La realización de esta tesis no habría sido posible sin la ayuda de muchos compañeros y amigos, a los que quiero agradecer su desinteresado apoyo.

A los profesores Rafael Bru y Damián Ginestar, por su decisiva tutela y orientación. Al profesor Gumersindo Verdú, por su asesoramiento en cuestiones de física. A los profesores José Mas, Félix Martínez y Néstor Thome, por su ayuda en la edición de esta memoria. A las profesoras Ángeles Martínez, Juana Cerdán y Ana Urbano, por su colaboración en la docencia y en la investigación. Y finalmente, quiero agradecer los consejos y sugerencias recibidos de los profesores Rafael Bru, José Mas y Michele Benzi.

A todos ellos, gracias.

Índice General

Prólogo	xix
1 Introducción	1
1.1 Funcionamiento de una planta nuclear	1
1.2 Notación y definiciones previas	4
1.3 Teoría de matrices no negativas y particiones de matrices	8
1.4 Métodos iterativos y preconditionadores	10
1.4.1 Métodos iterativos estacionarios clásicos	10
1.4.2 Métodos iterativos basados en subespacios de Krylov	15
1.4.3 Métodos multinivel	19
1.4.4 Preconditionadores	22
1.5 Antecedentes y objetivos de la memoria	25
2 Discretización de la ec. de la difusión neutrónica	29
2.1 Introducción	29
2.2 Ecuación de la difusión neutrónica	30
2.3 Discretización espacial	32
2.3.1 Método en diferencias finitas centradas	32
2.3.2 Método de colocación nodal	36

Índice General

2.4	Discretización temporal	46
2.4.1	Método en diferencias hacia atrás de 1 paso	48
2.4.2	Método en diferencias hacia atrás de 2 pasos	50
2.4.3	Método en diferencias hacia atrás de 4 pasos	51
2.5	Problemas modelo	53
2.5.1	Transitorios bidimensionales	53
2.5.2	Transitorio tridimensional de Langenbuch	57
2.6	Experimentos numéricos	59
2.6.1	Transitorio bidimensional	60
2.6.2	Transitorio de Langenbuch	63
3	Esquema iterativo de segundo grado	67
3.1	Introducción	67
3.2	Resultados y conceptos preliminares	69
3.3	Métodos iterativos de grado \hat{s}	73
3.4	Métodos iterativos de segundo grado	76
3.4.1	Método iterativo de segundo grado A.	98
3.4.2	Método iterativo de segundo grado B.	99
3.4.3	Resultados de convergencia para el método AGS. . . .	104
3.5	Experimentos numéricos	111
4	Método variacional	119
4.1	Introducción	119
4.2	Preliminares	120
4.3	Método variacional	122

Índice General

4.3.1	Interpretación como método de proyección	124
4.3.2	Propiedades de convergencia	126
4.3.3	Implementación práctica	133
4.4	Método de segundo grado acelerado	136
4.5	Experimentos numéricos	137
4.5.1	Transitorio bidimensional	138
4.5.2	Transitorio de Langenbuch	147
5	Métodos multinivel	151
5.1	Introducción	151
5.2	Métodos multinivel	152
5.2.1	Operadores de interpolación y restricción	155
5.2.2	Multinivel algebraico	159
5.2.3	Multinivel geométrico	161
5.3	Experimentos numéricos	161
5.3.1	Transitorios bidimensionales	163
5.3.2	Transitorio de Langenbuch	167
	Conclusiones y líneas futuras	171
	Bibliografía	177

Índice de Tablas

2.1	Coeficientes de los métodos en diferencias hacia atrás.	48
2.2	Secciones eficaces de los materiales en el reactor Seed-Blanket.	55
2.3	Parámetros de los precursores de neutrones en el reactor Seed-Blanket.	55
2.4	Secciones eficaces de los materiales en el reactor Langenbuch.	57
2.5	Parámetros de los precursores de neutrones en el reactor Langenbuch.	57
2.6	Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio 1 del TWIGL.	61
2.7	Discretizaciones espaciales utilizadas para la simulación del transitorio de Langenbuch.	64
3.1	Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio 1 del TWIGL.	112
3.2	Simulación mediante el método iterativo B. Discretización mediante el método nodal con 4 polinomios de Legendre.	113
3.3	Simulación mediante el método iterativo A.	114
3.4	Simulación mediante el método iterativo B.	114

Índice de Tablas

3.5	Simulación mediante el método iterativo A. Método en diferencias con tamaño de malla $h_x = h_y = 3 \text{ cm}$	115
3.6	Simulación mediante el método iterativo B. Método en diferencias con tamaño de malla $h_x = h_y = 3 \text{ cm}$. ω es el factor de extrapolación.	116
3.7	Potencia alcanzada por los métodos de segundo grado A y B en $t = 0, 2$ segundos. Potencia de referencia 2, 17.	116
4.1	Tiempo de simulación del método $ASD(1.5, r, q)$ para diferentes valores de r y q . Discretización utilizada: método nodal con 4 polinomios.	140
4.2	Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización $Nodal(3)$	141
4.3	Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización $Nodal(4)$	141
4.4	Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización en diferencias finitas.	141
4.5	Resultados de simulación del transitorio 1 del TWIGL con el método $ASD^*(1.5, 5, 1)$ para diferentes discretizaciones.	142
4.6	Resultados del método BiCGSTAB para la discretización $Nodal(3)$	143
4.7	Resultados del método BiCGSTAB para la discretización $Nodal(4)$	143
4.8	Resultados del método BiCGSTAB para la discretización en diferencias finitas.	143
4.9	Resultados del método GMRES(10) para la discretización $Nodal(3)$	144

Índice de Tablas

4.10 Resultados del método GMRES(20) para la discretización <i>Nodal</i> (4).	
.....	144
4.11 Resultados del método GMRES(10) para la discretización en diferencias finitas.	144
4.12 Resultados del método TFQMR para la discretización <i>Nodal</i> (3).	
.....	145
4.13 Resultados del método TFQMR para la discretización <i>Nodal</i> (4).	
.....	145
4.14 Resultados del método TFQMR para la discretización en difer- encias finitas.	145
4.15 Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio de Langenbuch.	148
4.16 Resultados de simulación del transitorio de Langenbuch para la discretización <i>Nodal</i> (2).	148
4.17 Resultados de simulación del transitorio de Langenbuch para la discretización <i>Nodal</i> (3).	149
4.18 Resultados de simulación del transitorio de Langenbuch para la discretización <i>Nodal</i> (4).	149
5.1 Tiempo CPU (en segundos) utilizado por NOBACK1 para simular el transitorio 1.	163
5.2 Tiempo CPU (en segundos) utilizado por NOBACK1 para simular el transitorio 2.	164
5.3 Tiempo de CPU (en segundos) utilizado por GML para simular el transitorio 1. Número de polinomios de Legendre, $P(k) =$ $\{2, 3, 4, 5\}$.	165

Índice de Tablas

5.4	Tiempo de CPU (en segundos) utilizado por AML para simular el transitorio 1. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$	165
5.5	Tiempo de CPU (en segundos) utilizado por GML para simular el transitorio 2. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$	166
5.6	Tiempo de CPU (en segundos) utilizado por AML para simular el transitorio 2. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$	166
5.7	Tiempo de CPU (en segundos) utilizado por el algoritmo GML para simular el transitorio de Langenbuch. Número de polinomios de Legendre, $P(k) = \{2, 3, 4\}$	168
5.8	Tiempo de CPU en segundos utilizado por los códigos GML, NOBACK1 y NEM para simular el transitorio de Langenbuch.	168

Índice de Figuras

1.1	Estructura de un reactor BWR.	2
1.2	Malla de 2 niveles: Ω^h línea discontinua, Ω^{2h} línea continua. .	20
2.1	Discretización en diferencias finitas centradas. Ordenamiento natural.	33
2.2	Patrones de cinco puntos: (a) patrón estándar (b) patrón diagonal.	34
2.3	Posición de los nodos adyacentes al nodo e	41
2.4	Cuadrante del reactor Seed-Blanket.	54
2.5	Evolución de la potencia relativa para el transitorio 1 en el reactor Seed-Blanket.	55
2.6	Evolución de la potencia relativa para el transitorio 2 en el reactor Seed-Blanket.	56
2.7	Geometría del reactor Langenbuch.	58
2.8	Evolución de la potencia relativa para el transitorio en el reactor Langenbuch.	59
2.9	Evolución de la potencia relativa discretizando mediante el método de colocación nodal $Nodal(p)$ con $p = 2, 3, 4$	62

Índice de Figuras

2.10 Evolución de la potencia relativa discretizando mediante el método en diferencias finitas. Se muestra también la curva para el método de colocación nodal, <i>Nodal</i> (4).	62
2.11 Evolución de la potencia relativa discretizando mediante el método de colocación nodal <i>Nodal</i> (p) con $p = 2, 3, 4$	65
2.12 Evolución local de la potencia para el transitorio de Langenbuch en el nodo P11.	65
3.1 Gráfica de las ecuaciones $m(\lambda) = \lambda^2$ y $g(\lambda) = \omega\mu\lambda + (1 - \omega)\mu$, para $\mu = \bar{\mu} = 0,7995$ y diferentes valores de ω	89
5.1 Acción de los operadores prolongación (\mathcal{I}_i^j) y restricción (\mathcal{I}_j^i) sobre los coeficientes en el nodo e	156

Prólogo

La modelización matemática de muchos fenómenos físicos implica la utilización de sistemas de ecuaciones en derivadas parciales (EDP) para los cuales raramente se conoce una solución analítica. Los fenómenos físicos relacionados con la difusión de partículas en un medio constituyen una importante fuente de problemas formulados en términos de ecuaciones en derivadas parciales. De entre estos fenómenos destaca la modelización de la difusión de neutrones en el interior del núcleo de un reactor nuclear, uno de los fenómenos físicos que mayor atención ha merecido por su repercusión en la mejora de la seguridad de las centrales nucleares. La resolución numérica de la ecuación de la difusión neutrónica constituye de hecho una herramienta fundamental para la simulación y la prevención de riesgos nucleares.

A partir de la ecuación de la difusión neutrónica se realizan dos tipos de cálculos distintos aunque complementarios. Un primer tipo de cálculos estáticos consistentes en la determinación de la configuración estática del reactor en un instante de tiempo, y que constituye de hecho un problema de valores propios generalizado. El otro tipo de cálculos son los que se realizan para el estudio de un transitorio a partir de una perturbación efectuada sobre una configuración estática del reactor, haciendo uso para ello de la ecuación de la difusión neutrónica dependiente del tiempo.

La forma típica de resolver la ecuación de la difusión neutrónica es discretizándola, es decir, aproximando las EDP mediante ecuaciones algebraicas que involucran un número finito de incógnitas. La utilización de numerosos

Prólogo

métodos de discretización permite reducir el problema original a la obtención de la solución de sistemas de ecuaciones lineales cuya matriz de coeficientes es generalmente de gran tamaño y con pocos elementos no nulos (vacía). Además, estas matrices presentan una marcada estructura por bloques.

Tradicionalmente se han utilizado métodos directos para la solución de sistemas de ecuaciones lineales por su robustez y comportamiento predecible. Sin embargo, con el desarrollo en las últimas décadas de computadores de altas prestaciones y arquitecturas paralelas, para la solución de sistemas de ecuaciones lineales cuya matriz de coeficientes es de gran tamaño y vacía se ha asistido a una creciente utilización de métodos iterativos frente a los métodos directos. El motivo fundamental de este hecho es que el proceso de eliminación gaussiana raras veces puede explotar el carácter vacío de las matrices. Por contra, la multiplicación matriz-vector (operación básica en un método iterativo) puede ser optimizada para beneficiarse de ésta y otras características especiales de las matrices.

El objetivo principal de la presente memoria es estudiar distintos métodos para la resolución numérica de los sistemas de ecuaciones lineales que aparecen en la integración de la ecuación de la difusión neutrónica dependiente del tiempo, centrándose por tanto en el análisis de transitorios.

Los contenidos de la memoria se han estructurado en cinco capítulos. En el primer capítulo, se revisan algunos conceptos del Álgebra Lineal utilizados en capítulos posteriores. También se describen tanto los métodos iterativos clásicos como los modernos basados en subespacios de Krylov, así como su preconditionamiento. Se introducen brevemente las bases de las modernas técnicas multinivel que serán utilizadas en el capítulo quinto. Finalmente, se revisan los antecedentes de esta línea de trabajo y se plantean los objetivos que han motivado la presente memoria.

En el capítulo segundo, se detallan los métodos de discretización utilizados para la resolución numérica de la ecuación de la difusión neutrónica, así como la estructura de los sistemas de ecuaciones lineales que se obtienen.

Prólogo

Concretamente para la parte espacial de las ecuaciones se utilizan un método de colocación nodal y el método de las diferencias finitas centradas, y para la discretización temporal se hace uso de diferentes métodos en diferencias hacia atrás. También se presentan los transitorios bidimensionales y tridimensionales que serán utilizados en los experimentos numéricos encaminados a la evaluación de las prestaciones de los diferentes métodos que se proponen. Además, se estudian experimentalmente los dos métodos de discretización de la parte espacial presentados.

En el capítulo tercero, se estudian las propiedades de convergencia de métodos iterativos por bloques de segundo grado para la solución de los sistemas de ecuaciones lineales que se obtienen en la integración de la ecuación de la difusión neutrónica. El estudio se completa con los resultados de los experimentos numéricos correspondientes a la simulación de transitorios bidimensionales y tridimensionales.

En el capítulo cuarto, se presenta una técnica variacional para acelerar la convergencia de los métodos de segundo grado presentados en el capítulo tercero. La técnica variacional consiste en un método de proyección del residuo sobre un subespacio de dimensión dos, con el objetivo de obtener un método de fácil implementación y de bajo coste por iteración. Las prestaciones del método de segundo grado acelerado son evaluadas y comparadas con otras técnicas basadas en subespacios de Krylov al final del capítulo.

Por último, en el capítulo quinto se presenta una técnica multinivel basada en el número de polinomios utilizado en el método de colocación nodal. El objetivo que se persigue con esta técnica es la reducción del tiempo de computación cuando se desea obtener una elevada precisión en la solución. También se presentan los resultados de los experimentos numéricos de la simulación de transitorios bidimensionales y tridimensionales.

Al final de la memoria se exponen las conclusiones más significativas y se esquematizan algunas líneas futuras de trabajo.

Capítulo 1

Introducción

En este capítulo se presentan y revisan algunos conceptos y resultados que serán utilizados en capítulos posteriores. En la sección 1.1 se describe escuetamente el comportamiento de un reactor nuclear. A continuación, en las secciones 1.2 y 1.3 se revisan conceptos básicos sobre normas y resultados de la teoría de matrices no negativas. En la sección 1.4 se describen diferentes métodos iterativos para la resolución de sistemas de ecuaciones lineales (clásicos y basados en subespacios de Krylov) y diferentes técnicas de preconditionado. También se da una breve introducción a los métodos multinivel. Finalmente, en la sección 1.5 se plantean los objetivos que han motivado la presente memoria.

1.1 Funcionamiento de una planta nuclear

Un reactor nuclear es un sistema que utiliza la energía liberada al fisiónarse los átomos de uranio para calentar agua. Este agua ya en la fase de vapor se utiliza para generar electricidad mediante una turbina [80]. En particular, la estructura de un reactor de agua en ebullición (BWR) se muestra en la figura 1.1.

Los problemas de estabilidad de estos reactores han sido una preocupación

1.1. Funcionamiento de una planta nuclear

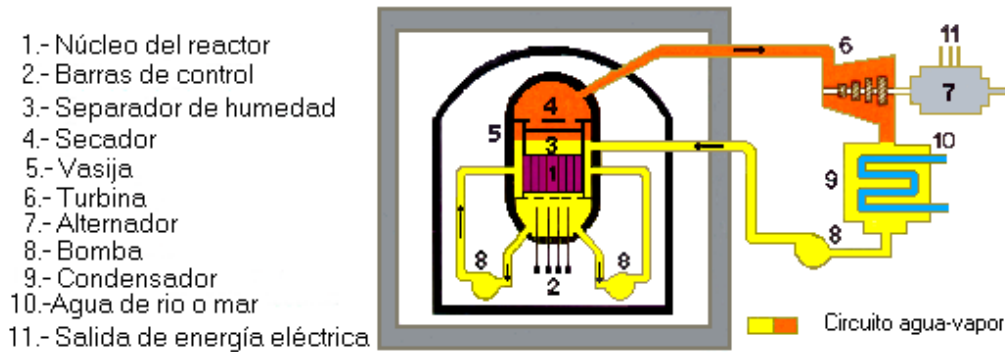


Figura 1.1: Estructura de un reactor BWR.

constante ya desde los primeros experimentos de diseño de los mismos alrededor de 1950. Generalmente, en condiciones de bajo caudal y potencia alta [65], que se pueden alcanzar en el arranque o en la parada de estos reactores se producen oscilaciones de la potencia y del caudal. Tanto en pruebas especiales de estabilidad de los reactores y en algunos sucesos de inestabilidad durante la operación comercial de este tipo de reactores se han observado dos tipos de oscilaciones. En el primer tipo de oscilaciones detectadas la potencia de todo el reactor varía de la misma forma con una frecuencia cercana a los 0.5 Hz y se conocen con el nombre de *oscilaciones en fase*. En el otro tipo de oscilaciones se observa que la potencia de una parte del reactor crece mientras que en la otra parte del reactor decrece manteniéndose la potencia media aproximadamente constante, y se denominan *oscilaciones fuera de fase* [66]. La detección y supresión de estas y, en la medida de lo posible, la predicción de estas oscilaciones es de vital importancia para mejorar la seguridad de los reactores BWR.

Para el estudio de estas oscilaciones, un camino ha consistido en tratar de estudiar y predecir este tipo de oscilaciones desarrollando modelos matemáticos para el reactor, cuya integración permita predecir el comportamiento del mismo en diversas situaciones.

Los modelos matemáticos de un reactor nuclear suelen dividirse en dos partes o módulos. Una parte que modeliza el comportamiento de la población

Capítulo 1. Introducción

neutrónica en el núcleo del reactor, que se conoce como módulo neutrónico, y otro módulo que da cuenta de la dinámica del líquido y el vapor en el reactor, la transferencia de calor y el cambio de fase. Este es el módulo termohidráulico. Los modelos correspondientes a estos dos módulos se suelen tratar de modo separado y el “acople” entre ellos depende del tipo de esquema de integración que se utilice para el modelo global. En esta memoria se trata exclusivamente la parte correspondiente al módulo neutrónico.

La ecuación que modeliza la evolución de la población neutrónica en el interior del núcleo del reactor es la ecuación del transporte de Boltzman. No obstante, asumiendo una serie de aproximaciones sobre el comportamiento de los neutrones en el interior del núcleo, se puede utilizar la ecuación de la difusión neutrónica para modelizar el comportamiento de los neutrones en el núcleo [102, 50]. En particular se suele utilizar la ecuación de la difusión neutrónica en la aproximación de dos grupos de energía y se supone que los neutrones se producen en el grupo rápido de energía y no hay trasvase de neutrones del grupo térmico al rápido [102]. Este modelo consiste en un sistema de ecuaciones en derivadas parciales lineales con coeficientes variables que hay que integrar para la geometría del núcleo del reactor. Su descripción detallada se dará en el capítulo 2.

Un primer problema asociado con este modelo es la resolución de la ecuación de los modos lambda [50], que es un problema parcial de valores propios generalizado asociado a un operador diferencial no autoadjunto. La obtención de los valores propios dominantes y las correspondientes funciones propias asociadas tanto a este problema como al problema adjunto es de interés ya que el valor propio dominante, conocido como la constante efectiva del reactor, nos da una idea de la distancia de la configuración estudiada del reactor de una configuración crítica, en la que la reacción en cadena se mantiene. La correspondiente función propia (el modo fundamental) describe el estado estacionario del reactor para la configuración dada y, por tanto, es el punto de partida de cualquier transitorio que se quiera estudiar.

1.2. Notación y definiciones previas

Es precisamente la simulación del comportamiento dinámico de un reactor nuclear el motivo principal de esta memoria. Para su estudio hay que ser capaz de integrar la ecuación de la difusión neutrónica dependiente del tiempo. En el capítulo 2 se describe el proceso de discretización de la parte espacial de esta ecuación y su integración temporal. Se mostrará cómo su resolución numérica se reduce a la obtención de la solución de sistemas de ecuaciones lineales. La resolución de estos sistemas de ecuaciones constituye el objetivo primero de esta tesis.

1.2 Notación y definiciones previas

A continuación, se introduce la notación que será utilizada con posterioridad para matrices reales. Mientras que no se indique lo contrario, se trabajará con matrices de $\mathbb{R}^{n \times n}$, es decir, del espacio de las matrices reales de tamaño $n \times n$, y con vectores de \mathbb{R}^n . Normalmente, se utilizarán letras minúsculas para vectores, y letras mayúsculas para matrices.

Se dice que una matriz A es *positiva* (respectivamente, *no negativa*) y se denota $A > O$ (respectivamente, $A \geq O$), si tiene todas sus componentes estrictamente mayores que 0 (respectivamente, mayores o iguales que 0). Mediante la letra O representamos a la matriz nula en $\mathbb{R}^{n \times n}$.

Dadas dos matrices A y B , se dice que $A > B$ (respectivamente, $A \geq B$) si la matriz $A - B$ es positiva (respectivamente, no negativa).

Dada una matriz A , $|A|$ denota la matriz cuyas componentes son los valores absolutos de las correspondientes componentes de A . De aquí se obtiene que $|A| \geq O$, y que $|AB| \leq |A||B|$, para dos matrices de tamaño apropiado. Se dice que una matriz es *monótona* si su inversa es no negativa, es decir, si $A^{-1} \geq O$.

Dada $A = [a_{ij}]$, se dice que A es una *M-matriz* invertible si $a_{ij} \leq 0$, $i \neq j$, y $A^{-1} \geq O$. Se dice que es *diagonal dominante* si $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$, $i = 1 \dots n$,

Capítulo 1. Introducción

y *estrictamente diagonal dominante* si se da la desigualdad estricta para todo i . La matriz A es *reducible* si existe una matriz de permutación, denotada π , tal que

$$\pi A \pi^T = \begin{bmatrix} B_{11} & B_{12} \\ O & B_{22} \end{bmatrix}.$$

Se dice que A es *irreducible* si no es reducible. A es *irreducible diagonal dominante* si es diagonal dominante, irreducible y además la desigualdad estricta se da para al menos un i .

Dada una matriz A , $\sigma(A)$ denota el *espectro* de A , es decir, el conjunto formado por todos los valores propios de la matriz A . El máximo de los módulos de los valores propios, es decir, *el radio espectral* de A , se denotará mediante $\rho(A)$. La siguiente definición guarda una estrecha relación con el espectro de una matriz.

Definición 1 *El campo de valores de una matriz A es el conjunto dado por*

$$\mathcal{F}(A) = \{y^T A y : y \in \mathbb{R}^n, y^T y = 1\}$$

Se denomina radio numérico al mayor elemento de $\mathcal{F}(A)$ en valor absoluto, y se denota $\nu(A)$.

El producto escalar en un espacio euclídeo E se denotará mediante la función \langle, \rangle . De especial interés en \mathbb{R}^n es el producto escalar canónico, o *producto euclídeo*, y que se define como,

$$\langle x, y \rangle = x^T y = \sum_{i=1}^n x_i y_i. \quad (1.1)$$

Los conceptos de norma vectorial y norma matricial son de gran utilidad en el análisis numérico. En concreto, se utilizan para definir la convergencia de una sucesión de vectores o matrices.

1.2. Notación y definiciones previas

Definición 2 En un espacio vectorial E sobre \mathbb{R} , la función

$$\|\cdot\| : E \longrightarrow \mathbb{R}$$

es una norma vectorial si satisface las siguientes propiedades:

1. $\|x\| \geq 0$, $\forall x \in E$, $y \quad \|x\| = 0$ si, y sólo si $x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\|$, $\forall x \in E$, $\forall \alpha \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in E$.

A partir del producto euclídeo (1.1), se define la norma euclídea como,

$$\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}, \quad (1.2)$$

y que es el caso particular para $p = 2$ de la norma

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}},$$

conocida como l_p -norma. Los casos $p = 1, 2$ son las normas más comunes. El caso límite cuando $p \rightarrow \infty$ se conoce como norma infinito,

$$\|x\|_\infty = \max_{i=1, \dots, n} |x_i|.$$

A continuación se recuerda el concepto de norma matricial y se muestran algunos resultados importantes.

Definición 3 La función $\|\cdot\| : \mathbb{R}^{n \times n} \longrightarrow \mathbb{R}$ es una norma matricial, si para todo par de matrices $A, B \in \mathbb{R}^{n \times n}$, satisface las siguientes propiedades,

1. $\|A\| \geq 0$, $y \quad \|A\| = 0$ si, y sólo si $A = O$.
2. $\|kA\| = |k| \|A\|$, $\forall k \in \mathbb{R}$.

$$3. \|A + B\| \leq \|A\| + \|B\|.$$

Si además cumple la propiedad submultiplicativa,

$$\|AB\| \leq \|A\|\|B\|,$$

se dice que la norma matricial es *consistente*.

Un caso particular de normas matriciales consistentes muy utilizadas en el análisis numérico son las normas matriciales inducidas.

Definición 4 Sea $\|\cdot\|$ una norma vectorial sobre \mathbb{R}^n . Se llama norma matricial inducida por dicha norma vectorial, a la función definida sobre $\mathbb{R}^{n \times n}$ dada por,

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

Toda norma matricial inducida por una norma vectorial es *compatible*, en el sentido en que se verifica la igualdad

$$\|Ax\| \leq \|A\|\|x\|, \quad x \in \mathbb{R}^n,$$

además, si denotamos por I la matriz identidad en $\mathbb{R}^{n \times n}$, se tiene $\|I\| = 1$. Entre las normas matriciales inducidas nos será de gran utilidad la norma matricial infinito, definida para una matriz $A = [a_{ij}]_{1 \leq i, j \leq n}$ como,

$$\|A\|_{\infty} = \max_{1 \leq i \leq n} \left(\sum_{j=1}^n |a_{ij}| \right).$$

El siguiente resultado muestra que cualquier norma matricial es una cota superior del radio espectral. Además siempre es posible encontrar una norma matricial que calculada sobre una matriz dada, esté tan próxima como queramos a su radio espectral (ver por ejemplo [74]).

1.3. Teoría de matrices no negativas y particiones de matrices

Teorema 5 Sea $\|\cdot\|$ una norma matricial y $A \in \mathbb{R}^{n \times n}$, entonces $\rho(A) \leq \|A\|$.

Teorema 6 Sea $A \in \mathbb{R}^{n \times n}$. Entonces dado un $\epsilon > 0$, existe al menos una norma matricial inducida $\|\cdot\|$ tal que $\|A\| \leq \rho(A) + \epsilon$.

La convergencia de una sucesión de vectores o matrices se puede definir en términos de normas como sigue.

Teorema 7 Sea E el espacio vectorial \mathbb{R}^n ($\mathbb{R}^{n \times n}$). La sucesión de vectores (matrices) v_1, v_2, \dots converge a $v \in E$ si, y sólo si para cualquier norma $\|\cdot\|$ definida en E ,

$$\lim_{k \rightarrow \infty} \|v_k - v\| = 0. \quad (1.3)$$

Dado que en un espacio vectorial de dimensión finita todas las normas son equivalentes, la convergencia en una norma implica la convergencia en cualquier norma definida en E .

1.3 Teoría de matrices no negativas y particiones de matrices

A continuación se revisan algunos resultados de la teoría de matrices no negativas y particiones de matrices monótonas. Un resultado fundamental sobre matrices no negativas es el teorema de Perron-Frobenius.

Teorema 8 [13, teoremas 2.1.1 y 2.1.4] Sea A una matriz de tamaño $n \times n$.

Si A es positiva, entonces $\rho(A)$ es un valor propio simple, mayor que cualquier otro en módulo.

Capítulo 1. Introducción

Si $A \geq O$ e irreducible, entonces $\rho(A)$ es un valor propio simple, y cualquier otro valor propio de A del mismo módulo es también simple. Además, A tiene un vector propio positivo x asociado a $\rho(A)$.

Si $A \geq O$, entonces $\rho(A)$ es un valor propio de A . Además, A tiene un vector propio no negativo ($x \geq 0$) asociado a $\rho(A)$.

Entre dos matrices no negativas, se puede establecer el siguiente resultado.

Teorema 9 [97, teorema 2.8] Sean A y B dos matrices de tamaño $n \times n$ no negativas, tales que $O \leq |B| \leq A$. Entonces

$$\rho(B) \leq \rho(A) .$$

A continuación, se define el término partición de una matriz, y se muestran algunos resultados para particiones de matrices.

Definición 10 Sea A una matriz de tamaño $n \times n$, tal que

$$A = M - N . \tag{1.4}$$

Se dice que la expresión (1.4) representa una partición de la matriz A .

Mientras que no se indique lo contrario se asumirá que la matriz M en (1.4) es invertible.

Definición 11 Se dice que la partición (1.4) es regular si $M^{-1} \geq O$ y $N \geq O$. Se dice que (1.4) es una partición débilmente regular si $M^{-1} \geq O$ y $M^{-1}N \geq O$. Se dice que (1.4) es una partición convergente si $\rho(M^{-1}N) < 1$.

Se observa que una partición regular es débilmente regular. Se tienen los siguientes teoremas.

1.4. Métodos iterativos y preconditionadores

Teorema 12 [13, teorema 5.6] Sea A una matriz invertible con $A^{-1} \geq O$ y sea $A = M - N$ una partición débilmente regular. Entonces

$$\rho(M^{-1}N) < 1 .$$

Teorema 13 [96, 97] Sea A una matriz invertible con $A^{-1} \geq O$ y sean $A = M_1 - N_1 = M_2 - N_2$ dos particiones regulares. Si

$$N_1 \leq N_2 ,$$

entonces

$$\rho(M_1^{-1}N_1) \leq \rho(M_2^{-1}N_2) < 1 .$$

En particular, si $A^{-1} > O$ y

$$N_1 < N_2 , \quad N_1 \neq N_2 ,$$

entonces

$$\rho(M_1^{-1}N_1) < \rho(M_2^{-1}N_2) < 1 .$$

La teoría sobre particiones regulares y débilmente regulares es extensa. Resultados similares y extensiones de los anteriores pueden encontrarse, por ejemplo, en [96, 97, 76, 94, 25, 67, 104, 81, 79].

1.4 Métodos iterativos y preconditionadores

1.4.1 Métodos iterativos estacionarios clásicos

Dado un sistema de ecuaciones lineales de la forma

$$Ax = b , \tag{1.5}$$

Capítulo 1. Introducción

donde A es una matriz invertible de $\mathbb{R}^{n \times n}$, el vector de incógnitas, x , y el vector b , son vectores de \mathbb{R}^n , se pretende obtener mediante el uso de un método iterativo una aproximación “satisfactoria” de la única solución del sistema, $x = A^{-1}b$. Si se considera una partición de la matriz $A = M - N$, un método iterativo estacionario obtiene una sucesión de vectores x_0, x_1, \dots de la forma,

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b = Hx_k + c. \quad (1.6)$$

La convergencia del esquema iterativo (1.6) a la solución del sistema (1.5), no está garantizada para cualquier tipo de matrices, aunque existe una amplia y rica teoría al respecto [97, 105]. Generalmente, la convergencia se estudia a partir del análisis del error. En efecto, se verifica

$$e_{k+1} = H^k e_0,$$

donde $e_k = x_k - x$ es el error en la etapa k . Se tiene el siguiente lema.

Lema 14 [13, lema 7.3.6] *Sea $A = M - N$ una partición. Entonces el método iterativo (1.6) converge a la solución del sistema, $x = A^{-1}b$, para todo vector inicial x_0 si, y sólo si $\rho(M^{-1}N) < 1$. En tal caso se dice que el método iterativo es convergente.*

Se llama *matriz de iteración* a la matriz $H = M^{-1}N$. De la definición 11 y del lema 14 se sigue que un método es convergente si, y sólo si su partición es convergente. Por ello, ambos términos se emplearán indistintamente.

Para comparar diferentes métodos iterativos se utiliza a menudo la velocidad asintótica de convergencia.

Definición 15 [13, definición 7.3.7] *Para el esquema iterativo (1.6), asumiendo $\rho(H) < 1$, se define la velocidad asintótica de convergencia como*

$$R_\infty(H) = -\log \rho(H).$$

1.4. Métodos iterativos y preconditionadores

Entre los métodos iterativos básicos (clásicos) de la forma (1.6) para resolver el sistema de ecuaciones lineales (1.5) se encuentran los métodos de Jacobi, Gauss-Seidel y el método de sobrerelajación o SOR.

Considérese la matriz de coeficientes $A = [a_{ij}]$, $1 \leq i, j \leq n$, del sistema de ecuaciones (1.5). Asumiendo que los elementos de la diagonal son no nulos, a partir de una solución inicial x_0 el método de Jacobi genera una sucesión de vectores de la forma

$$(x_{k+1})_i = - \sum_{j=1, j \neq i}^n \frac{a_{ij}}{a_{ii}} (x_k)_j + \frac{b_i}{a_{ii}}, \quad 1 \leq i \leq n, \quad k = 0, 1, \dots, \quad (1.7)$$

donde $(x_k)_i$ representa la i -ésima componente del vector x_k . A partir de la iteración (1.7), si para el cálculo de la nueva componente i -ésima del vector x_k se utilizan las nuevas componentes ya calculadas, se obtiene el método de Gauss-Seidel cuya iteración es por tanto de la forma,

$$(x_{k+1})_i = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} (x_{k+1})_j - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} (x_k)_j + \frac{b_i}{a_{ii}}, \quad 1 \leq i \leq n, \quad k = 0, 1, \dots \quad (1.8)$$

Para acelerar la convergencia de un método iterativo de la forma (1.6) es frecuente la introducción de un parámetro de relajación, denotado como $\omega \in \mathbb{R}$, $\omega \neq 0$, obteniéndose una nueva sucesión de vectores de la forma

$$x_{k+1} = \omega x_{k+1}^{GS} + (1 - \omega) x_k.$$

En particular, la iteración dada por

$$x_{k+1} = \omega x_{k+1}^{GS} + (1 - \omega) x_k, \quad (1.9)$$

donde x_{k+1}^{GS} es la aproximación obtenida por el método de Gauss-Seidel, corresponde al método de sobrerelajación o SOR.

Para cada uno de los métodos anteriores se puede identificar una partición de la matriz A distinta. En efecto, considérese la matriz A como la suma de matrices

$$A = D - E - F,$$

Capítulo 1. Introducción

donde D es una matriz diagonal formada por los elementos diagonales de A , $-E$ y $-F$ son las partes estrictamente triangular inferior y superior de A , respectivamente. Asumiendo que D es invertible, es decir, todos los elementos diagonales de A son distintos de 0, las siguientes particiones definen a cada uno de los métodos anteriores:

Jacobi:

$$M = D, \quad N = E + F, \quad (1.10)$$

Gauss-Seidel:

$$M = D - E, \quad N = F, \quad (1.11)$$

SOR:

$$M = \omega^{-1}D - E, \quad N = \omega^{-1}(1 - \omega)D + F. \quad (1.12)$$

Las matrices de iteración para estos métodos, de la forma $M^{-1}N$, responden a las expresiones:

Jacobi:

$$B = D^{-1}(E + F), \quad (1.13)$$

Gauss-Seidel:

$$\mathcal{L} = (D - E)^{-1}F, \quad (1.14)$$

SOR:

$$\mathcal{L}_\omega = (D - \omega E)^{-1}[(1 - \omega)D + \omega F]. \quad (1.15)$$

Existe abundante bibliografía en la que se estudian las propiedades de convergencia de estos métodos y variantes suyas (por ejemplo [75, 74, 13] y la bibliografía citada en cada referencia). Cabe citar los textos clásicos de Varga

1.4. Métodos iterativos y preconditionadores

[97] y Young [105]. En ellos se estudian propiedades de convergencia de los métodos anteriores fundamentalmente para M-matrices, matrices simétricas y definidas positivas, y matrices consistentemente ordenadas. Un resultado clásico que compara la velocidad de convergencia de los métodos SOR y Jacobi es el conocido como teorema de Stein y Rosenberg, una de cuyas versiones se da a continuación.

Teorema 16 [13, teorema 7.5.21] *Sea $A = I - L - U \in \mathbb{R}^{n \times n}$ con $L \geq O$ y $U \geq O$ estrictamente triangular inferior y superior, respectivamente. Entonces para $0 < \omega \leq 1$,*

1. $\rho(B) < 1$ si, y sólo si $\rho(\mathcal{L}_\omega) < 1$.
2. $\rho(B) < 1$ (y $\rho(\mathcal{L}_\omega) < 1$) si, y sólo si A es una M-matriz invertible, y en tal caso $\rho(\mathcal{L}_\omega) \leq 1 - \omega + \omega\rho(B)$.
3. Si $\rho(B) \geq 1$ entonces $\rho(\mathcal{L}_\omega) \geq 1 - \omega + \omega\rho(B) \geq 1$.

En [46] Hadjidimos introduce el método de sobrerrelajación acelerado o AOR, como una generalización de los métodos iterativos conocidos hasta la fecha. La matriz de iteración del método AOR viene dada por

$$\mathcal{L}_{\omega,\tau} = (I - \omega L)^{-1}[(1 - \tau)I + (\tau - \omega)L + \tau U], \quad (1.16)$$

donde ω es un factor de aceleración, y τ es un factor de relajación. Se observa que $\mathcal{L}_{0,1}$ coincide con el método de Jacobi, $\mathcal{L}_{1,1}$ con el método de Gauss-Seidel y $\mathcal{L}_{\omega,\omega}$ con el método SOR. Además, $\mathcal{L}_{\omega,1}$ coincide con el método de Gauss-Seidel acelerado [90], método que aparecerá en el capítulo 3.

Los métodos iterativos clásicos vistos, aunque normalmente son menos eficientes que los métodos basados en subespacios de Krylov y los métodos multinivel, todavía tienen un importante papel como preconditionadores.

1.4.2 Métodos iterativos basados en subespacios de Krylov

A diferencia de los métodos iterativos anteriores los métodos basados en subespacios de Krylov no tienen una matriz de iteración y, por tanto, no están basados en la iteración (1.6) que caracteriza a los métodos estacionarios.

Considérese el sistema de ecuaciones lineales (1.5). Los métodos basados en subespacios de Krylov tratan de obtener la “mejor” aproximación posible del *subespacio de Krylov de orden k* ,

$$\text{Env}\{b, Ab, \dots, A^{k-1}b\}, \quad (1.17)$$

asociado al vector b y la matriz A , y que denotaremos mediante $\mathcal{K}_k(b, A)$. Si se desea utilizar el conocimiento de una aproximación inicial de la solución, denotada x_0 , como espacio de soluciones se utiliza el espacio afín

$$x_0 + \mathcal{K}_k(r_0, A), \quad (1.18)$$

donde $r_0 = b - Ax_0$ es el *vector residuo* inicial. Como “mejor” se entiende aquella solución que sea óptima en el sentido de minimización de un funcional dado.

Los métodos basados en subespacios de Krylov, también referidos en la bibliografía como *métodos del tipo gradiente conjugado*, obtienen una solución del subespacio (1.18) mediante una recurrencia que se puede escribir como,

$$x_k = x_{k-1} + a_{k-1}p_{k-1}, \quad (1.19)$$

$$p_k = Ap_{k-1} - \sum_{j=k-s+1}^{k-1} b_{k-1,j}p_j, \quad (1.20)$$

donde s es un entero menor que n , a_{k-1} y $b_{k-1,j}$, $j = k - s + 1, \dots, k - 1$ son coeficientes que se eligen para que la solución sea óptima. Se puede comprobar que los vectores de dirección p_0, \dots, p_{k-1} , con $p_0 = r_0$, forman una base del subespacio de Krylov $\mathcal{K}_k(r_0, A)$.

1.4. Métodos iterativos y preconditionadores

Así por ejemplo, cuando la matriz del sistema A es simétrica y definida positiva, una opción posible consiste en elegir como solución, en la etapa k -ésima, el vector $x_k \in x_0 + \mathcal{K}_k(r_0, A)$ que minimiza la *norma- A* del vector *vector error* $e_k = x_k - x$, es decir,

$$\|e_k\|_A = \sqrt{e_k^T A e_k} . \quad (1.21)$$

A este método se le conoce como método del *gradiente conjugado* (abreviado CG), introducido en primera instancia por Hestenes y Stiefel como un método directo [51], y posteriormente utilizado como método iterativo. Este método también fue formulado independientemente y de una forma diferente por Lanczos [62].

Cuando la matriz es simétrica pero indefinida la ecuación (1.21) ya no define una norma, y por tanto como “mejor” solución se puede tomar alternativamente el vector $x_k \in x_0 + \mathcal{K}_k(r_0, A)$ que minimiza la norma euclídea del vector residuo $r_k = b - Ax_k$, es decir,

$$\|b - Ax_k\|_2 .$$

El método del *residuo mínimo* o MINRES genera esta aproximación [78].

Una importante característica de los dos métodos anteriores es que son capaces de generar aproximaciones óptimas a la solución, en el sentido de minimización de la norma del error, utilizando recurrencias cortas de tres términos, es decir, tomando $s = 3$ en (1.19) y (1.20). Este tipo de recurrencias conllevan un coste aritmético, a parte del producto matriz-vector, de tan sólo $O(n)$ operaciones por iteración. Además, los requerimientos de memoria para almacenar vectores son también muy bajos. Este hecho hace del gradiente conjugado y del MINRES los métodos “preferidos” para matrices simétricas.

En cambio, los métodos conocidos para matrices no simétricas, o bien no pueden extraer soluciones óptimas del subespacio (1.18) mediante recurrencias cortas (presentan costes aritméticos y de almacenamiento del orden $O(nk)$ en k iteraciones), o bien no obtienen soluciones óptimas. De forma

Capítulo 1. Introducción

más precisa, Faber y Manteuffel en [26] muestran que las únicas matrices para las que se puede obtener una solución del subespacio (1.18) mediante una recurrencia con $s < \sqrt{n}$ (ecuación (1.20)), y que a su vez minimice el error $e_k = x_k - x$ en alguna norma asociada a un producto escalar independiente del vector inicial, son matrices obtenidas de una rotación y un desplazamiento de una matriz simétrica.

Atendiendo a la observación anterior, para matrices no simétricas los métodos más utilizados se pueden agrupar en dos grupos: métodos que generen aproximaciones óptimas pero utilizando recurrencias largas, o métodos que utilizan recurrencias cortas pero generan aproximaciones no óptimas en el sentido de Faber y Manteuffel.

Entre los primeros, uno de los métodos más utilizados es el método del *residuo mínimo generalizado* o GMRES [88, 101]. El método GMRES puede ser visto como una extensión del método MINRES para matrices no simétricas. Por tanto, GMRES también obtiene del subespacio afín (1.18) aquella aproximación x_k que minimiza $\|b - Ax_k\|_2$. A diferencia del método MINRES, que se basa en el método de Lanczos para la construcción de una base ortonormal del subespacio de Krylov $\mathcal{K}_k(r_0, A)$, el método GMRES utiliza el algoritmo de Arnoldi para construir dicha base (ver [1, 62, 87, 42]). Básicamente, el método de Arnoldi consiste en la aplicación del método de ortogonalización de Gram-Schmidt a la base del subespacio $\mathcal{K}_k(r_0, A)$ dada por,

$$\mathcal{B} = \{r_0, Ar_0, \dots, A^{k-1}r_0\}. \quad (1.22)$$

En principio, si $\mathcal{V} = \{p_0, p_1, \dots, p_{k-1}\}$ es la base ortonormal obtenida a partir de la base \mathcal{B} , el elemento k -ésimo de la base se obtiene mediante una recurrencia de la forma (1.20) con $s = k$. Se puede demostrar ([87, teorema 6.2]) que para matrices simétricas la recurrencia (1.20) se reduce a tan sólo 3 términos ($s = 3$), que es la base del algoritmo de Lanczos (razón por la cual los métodos del gradiente conjugado y MINRES utilizan recurrencias

1.4. Métodos iterativos y preconditionadores

cortas). Cuando el número de iteraciones necesario para resolver el sistema de ecuaciones (1.5) es alto, las necesidades tanto de memoria como de trabajo para almacenar y ortogonalizar la base \mathcal{B} en (1.22) aumentan hasta el punto de hacer inviable la aplicación del método GMRES. En la práctica se utiliza una variación de éste método, el método GMRES reiniciado o GMRES(k) [88]. El método GMRES(k) se define simplemente reiniciando el método GMRES cada k iteraciones, tomando el último vector iterado como solución inicial para el próximo ciclo del GMRES. Como resultado se obtiene un método que ya no genera soluciones óptimas.

Entre los métodos no óptimos para matrices no simétricas pero que utilizan recurrencias cortas se encuentran los métodos BiCG [28], CGS [93], TFQMR [29, 30], BiCGSTAB [95], y algunas variantes del GMRES como el método GMRES(k) antes mencionado. De entre éstos métodos, son el BiCGSTAB, TFQMR y GMRES(k) los más utilizados. Los métodos BiCGSTAB y TFQMR se basan en el método de biortogonalización de Lanczos [61, 62]. Como se ha dicho, el método de Lanczos genera una base ortonormal del subespacio de Krylov $\mathcal{K}_k(r_0, A)$ mediante una recurrencia corta de tan sólo 3 términos. En un intento de extender este algoritmo al caso no simétrico, el método de biortogonalización de Lanczos utiliza dos recurrencias cortas de 3 términos para generar dos bases biortogonales entre sí, una para el subespacio de Krylov $\mathcal{K}_k(r_0, A)$ y la otra para el subespacio $\mathcal{K}_k(\hat{r}_0, A^T)$. Uno de los problemas de los métodos basados en este algoritmo es el trabajo extra que supone trabajar con la matriz traspuesta de A . En este sentido, la popularidad tanto del método BiCGSTAB como del TFQMR se debe a que no requieren la multiplicación por la matriz A^T .

Una extensa bibliografía sobre éstos y otros métodos relacionados se puede encontrar por en [45, 5, 44, 57, 87, 14, 42], donde se muestran diferentes énfasis y perspectivas.

1.4.3 Métodos multinivel

En esta sección se describen brevemente las ideas y conceptos básicos de los métodos multinivel¹. Los métodos multinivel surgen para intentar mejorar la eficiencia de los métodos iterativos clásicos vistos en la sección 1.4.1. Estos métodos también se utilizan como preconditionadores de métodos basados en subespacios de Krylov (sección 1.4.2). Una excelente exposición de las ideas fundamentales de los métodos multinivel y su aplicación puede encontrarse, por ejemplo, en [15], [45], [56] y [68].

Considérese el sistema de ecuaciones lineales

$$Au = f, \quad (1.23)$$

donde A es una matriz de tamaño $n \times n$. Frecuentemente la matriz de coeficientes A se obtiene de la discretización de una ecuación lineal en derivadas parciales en un dominio que denotaremos con Ω . Por simplicidad, se asumirá que se ha discretizado el dominio Ω mediante una malla uniforme de tamaño h (figura 1.2). El dominio discretizado se denota por Ω^h . De esta manera, cada elemento u_i , $i = 1, \dots, n$ representa el vector u en un punto de la malla Ω^h . Para simplificar la notación, Ω^h también se utilizará para representar el espacio vectorial de los vectores definidos en la malla Ω^h , es decir, $\Omega^h \equiv \mathbb{R}^n$.

Resolver (1.23) mediante un método iterativo supone el cálculo de sucesivas aproximaciones a la solución exacta $u = A^{-1}f$ de la forma

$$u_{k+1} \leftarrow G(A, u_k, f), \quad (1.24)$$

donde G es función de la solución anterior u_k (solución inicial para la primera vez). El proceso (1.24) se conoce con el nombre de *relajación*. Entre los métodos iterativos más sencillos para obtener las sucesivas aproximaciones en (1.24) se encuentran, por ejemplo, los métodos iterativos de Jacobi y

¹Se podría decir que los métodos multinivel surgen de la aplicación de las ideas de los métodos “multigrid” a las técnicas de preconditionamiento basadas en la descomposición de dominios. En muchos casos se pueden emplear ambos términos indistintamente.

1.4. Métodos iterativos y preconditionadores

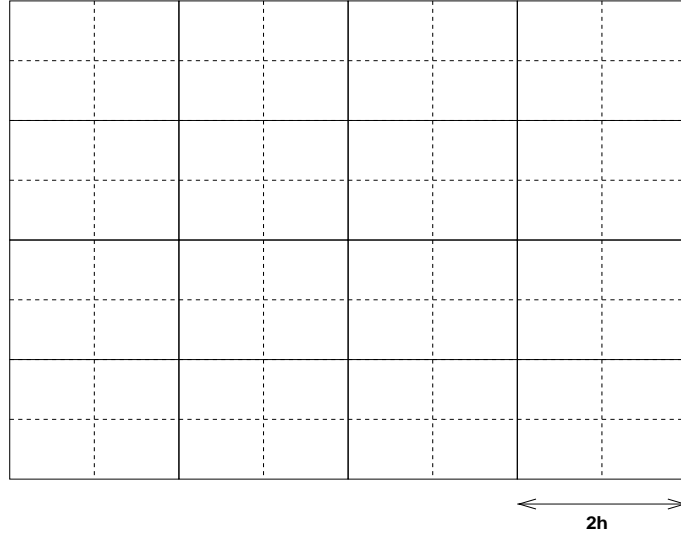


Figura 1.2: Malla de 2 niveles: Ω^h línea discontinua, Ω^{2h} línea continua.

Gauss-Seidel (ecuaciones (1.7) y (1.8), respectivamente). Estos métodos se caracterizan por la *propiedad de suavizado*², y que de hecho constituye una limitación. Esta propiedad se explica a partir de un análisis de Fourier del error [15, 68]. En este tipo de análisis el error se expresa como suma de funciones seno, de frecuencias diferentes, también llamadas *modos de Fourier*. En la terminología utilizada en los métodos multinivel, se denominan *modos oscilatorios* y *modos suaves* a los modos de alta y baja frecuencia, respectivamente. Normalmente, los métodos iterativos básicos eliminan con mayor o menor rapidez los modos oscilatorios. Desafortunadamente, una vez que estas componentes del error han sido eliminadas el proceso iterativo pierde eficacia en la eliminación de los modos suaves restantes de forma que parecen estancarse a partir de un cierto número de iteraciones. La siguiente observación es la base de los métodos multinivel [15],

modos suaves en una malla aparecen como oscilatorios en otra malla más gruesa,

²En la literatura, smoothing property.

Capítulo 1. Introducción

donde por malla gruesa se entiende aquella malla obtenida con un tamaño de discretización mayor. Este hecho sugiere que la aplicación sucesiva del método iterativo en mallas cada vez más gruesas debería eliminar todas las componentes del error. Por tanto, la idea básica de los métodos multinivel es el uso de diferentes mallas para mejorar la eficiencia de los métodos iterativos.

Otra posible forma de mejorar la eficiencia de un método iterativo es mediante la obtención de una solución inicial u_0 próxima a la solución del sistema (1.23) a un coste tan bajo como sea posible. Esto se lleva a cabo mediante la generación de la secuencia anidada de mallas, Ω^{ph} , $p = m, \dots, 2, 1$, donde ph indica el tamaño de la discretización. La obtención de una solución en cualquiera de éstas mallas resulta en principio menos costosa que en la malla más fina. A su vez, esta solución puede utilizarse como solución inicial mejorada para la siguiente malla. Por ejemplo, la figura 1.2 muestra el caso de 2 mallas anidadas. En un método de 2 niveles, la solución obtenida en la malla Ω^{2h} se utilizaría como solución inicial en la malla más fina, Ω^h . Para ello se necesita un operador de *interpolación o prolongación*. Si $I_{2h}^h : \Omega^{2h} \rightarrow \Omega^h$ denota este operador, la solución inicial para la malla Ω^h viene dada por

$$u_0^h = I_{2h}^h u^{2h}.$$

A esta estrategia se le denomina *iteración anidada*³. De forma equivalente, $I_h^{2h} : \Omega^h \rightarrow \Omega^{2h}$ define el operador de *restricción*.

En la práctica, muchos de los algoritmos multinivel/multimalla más utilizados realizan una iteración anidada para obtener una estimación inicial, no de la solución sino del error inicial cometido, que es utilizada para corregir la solución. A esta estrategia se le denomina habitualmente como *corrección de malla gruesa*⁴.

En el capítulo 5 se utilizará el esquema de iteración anidada para desarrollar un método multinivel para la solución de los sistemas de ecuaciones

³En la literatura, *nested iteration*.

⁴En la literatura, *coarse grid correction*.

1.4. Métodos iterativos y preconditionadores

lineales que aparecen en la integración de la ecuación de la difusión neutrónica.

1.4.4 Precondicionadores

Como se ha mencionado en la sección 1.4.2, los métodos basados en subespacios de Krylov obtienen soluciones aproximadas del subespacio (1.18). Estos métodos convergen rápidamente a la solución del sistema de ecuaciones (1.5) cuando la matriz de coeficientes A es próxima a la identidad en algún sentido. Desafortunadamente en muchas aplicaciones esto no es así, pero se podría pensar en reemplazar el sistema original por el sistema de ecuaciones modificado

$$M^{-1}Ax = M^{-1}b \quad \text{o} \quad AM^{-1}\hat{x} = b, \quad x = M^{-1}\hat{x}, \quad (1.25)$$

de forma que el nuevo subespacio de Krylov dado por

$$\text{Env}\{M^{-1}r_0, (M^{-1}A)M^{-1}r_0, \dots, (M^{-1}A)^{k-1}M^{-1}r_0\},$$

o

$$\text{Env}\{r_0, (AM^{-1})r_0, \dots, (AM^{-1})^{k-1}r_0\},$$

permita la obtención de una solución satisfactoria con un coste menor. Esta técnica se conoce con el nombre de preconditionamiento a la izquierda y a la derecha, respectivamente. Si M es simétrica y definida positiva, se puede preconditionar simétricamente resolviendo el sistema lineal modificado

$$L^{-1}AL^{-T}y = L^{-1}b, \quad x = L^{-T}y,$$

donde $M = LL^T$. L puede ser el factor triangular inferior de Choleski, la raíz cuadrada de M , o cualquier otra matriz que cumpla $M = LL^T$. En cualquiera de los casos, sólo se necesita resolver sistemas de ecuaciones cuya matriz de coeficientes es M , sin necesidad de calcular M^{-1} .

Capítulo 1. Introducción

Habitualmente el preconditionador M se elige de forma que los sistemas lineales con matriz de coeficientes M , sean fáciles de resolver, y que la matriz $M^{-1}A$, AM^{-1} o $L^{-1}AL^{-T}$ aproxime la identidad. El sentido en el que la matriz de coeficientes del sistema preconditionado debe estar próxima a la identidad depende del método iterativo empleado para resolverlo. Si se pretende utilizar un método basado en subespacios de Krylov, generalmente se busca que el *número de condición*, definido para una matriz A y una norma matricial $\|\cdot\|$ como

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

sea próximo a uno, o que la distribución de los valores propios sea favorable (posiblemente con los valores propios cercanos y localizados alrededor de algún punto lejano del origen).

Cabría, incluso, añadir un tercer requisito para la elección de un preconditionador. En los últimos años, y debido al desarrollo de máquinas de computación paralela cada vez más potentes y eficientes, se ha asistido a un creciente interés por el desarrollo de algoritmos paralelos que permitan aprovechar la potencia de cálculo que ofrece la nueva tecnología en la solución de los sistemas de ecuaciones lineales. Por tanto, que un preconditionador sea fácil de paralelizar es una característica deseable.

Una forma habitual de obtener un preconditionador es mediante una partición de la matriz $A = M - N$ (definición 1.4). Ejemplo de este tipo de preconditionadores son los preconditionadores basados en los métodos estacionarios básicos de Jacobi, Gauss-Seidel, SOR y algunas variantes de éstos (sección 1.4.1), [4]. Es bastante habitual utilizar estos métodos en sus versiones por bloques [21, 5].

También es posible obtener una partición mediante una factorización incompleta de la matriz A . Si, por ejemplo, A es simétrica y definida positiva, la factorización incompleta de Choleski induce una partición de la forma $A = LL^T - R$, donde L es el factor incompleto de Choleski [69, 64]. Este pre-

1.4. Métodos iterativos y preconditionadores

condicionador se engloba dentro de los preconditionadores conocidos como *factorizaciones incompletas LU* (abreviadamente ILU) [87]. Básicamente, estos algoritmos realizan una eliminación gaussiana de la matriz A , eliminando aquellos elementos situados en determinadas posiciones fuera de la diagonal según un patrón de llenado que puede ser fijo o dinámico. Entre las factorizaciones incompletas con patrón de llenado fijo se encuentra el preconditionador ILU0, que utiliza la estructura de elementos no nulos de la matriz de coeficientes. Otros preconditionadores, conocidos como ILU(k), permiten un cierto llenado adicional en los factores L , U hasta un nivel k [84]. La idea de no fijar *a priori* un patrón se basa en eliminar aquellos elementos cuyo módulo no supere cierta magnitud relativa. Por tanto generan el patrón de llenado dinámicamente sin tener en cuenta la estructura de la matriz A . Entre estos se encuentra el preconditionador ILUT [86].

Un problema que presentan los preconditionadores del tipo ILU, como se pone de manifiesto en [24], es que son inestables para cierto tipo de problemas, generalmente con matrices indefinidas. Este inconveniente puede ser minimizado en parte mediante técnicas de reordenamiento de matrices, como proponen los autores en [23, 11]. Otro inconveniente se debe a que su paralelización es difícil ya que su aplicación supone resolver sistemas triangulares que son inherentemente secuenciales.

En los últimos tiempos, en un intento de solventar estas dificultades, se ha prestado mucha atención a otro tipo de preconditionadores conocidos como preconditionadores de *inversa aproximada*. Estos preconditionadores explícitamente calculan una aproximación de la inversa de la matriz A . Por tanto, su aplicación se reduce a una multiplicación matriz-vector, operación que posee un elevado grado de paralelismo. Dentro de este grupo de preconditionadores se pueden citar el preconditionador AINV [9], FSAI [60] y SPAI [43]. En [12] se puede encontrar una revisión de todos los preconditionadores citados.

Todos los preconditionadores mencionados son preconditionadores de pro-

pósito general, es decir, de aplicación a cualquier tipo de matriz. Existen otros preconditionadores diseñados específicamente para el tipo de problema que se quiere resolver. Dentro de éstos se encuentran los métodos multinivel (sección 1.4.3), y los métodos de descomposición de dominios también conocidos como métodos de Schwarz. En [91, 68] se puede encontrar una extensa lista de referencias sobre el uso y aplicación de estos métodos. Es importante citar que tanto los métodos multinivel como los métodos de Schwarz también se utilizan para resolver sistemas de ecuaciones lineales generales, conociéndose con el nombre de métodos de Schwarz algebraicos [31, 10, 32] y métodos multinivel algebraicos [82].

1.5 Antecedentes y objetivos de la memoria

Para el estudio del comportamiento dinámico de un reactor nuclear es necesario integrar la ecuación de la difusión neutrónica dependiente del tiempo.

Para la discretización de la parte espacial se han empleado tradicionalmente diferentes métodos. En [48] se utiliza el método de los elementos finitos, en [106] y [20] métodos en diferencias finitas, mientras que se han utilizado métodos nodales en [49] y [99].

Para la integración temporal de las ecuaciones, en [38] se desarrolla un método en diferencias hacia atrás de dos pasos y un método en diferencias hacia atrás de cuatro pasos, diseñando, a su vez, una estrategia de inicialización que permitía variar el paso de integración combinando estos métodos. En [77] y [99] se utiliza un método cuasiestático adaptado a la discretización espacial utilizada.

Como se ha visto en las secciones anteriores, el hecho de utilizar métodos implícitos para la resolución de la ecuación de la difusión neutrónica implica la necesidad de resolver un sistema de ecuaciones lineales de gran tamaño, cuya matriz es no simétrica, para cada paso de integración. En [99] se de-

1.5. Antecedentes y objetivos de la memoria

sarrollan métodos que aprovechan la estructura por bloques de las matrices con muy buenos resultados. En [20] los autores resuelven estos sistemas considerando la matriz globalmente por medio de la utilización de métodos basados en subespacios de Krylov.

Recientemente en [54] y [55] los autores proponen la integración de la ecuación de la difusión neutrónica en reactores tridimensionales mediante la aplicación de técnicas multinivel. Concretamente, proponen reducir la complejidad del problema a geometrías de menor dimensión (punto y unidimensionales).

Finalmente, destacar que para comparar los diferentes métodos propuestos para la resolución numérica de la ecuación de la difusión neutrónica se utilizan fundamentalmente los problemas del reactor TWIGL (bidimensional) y el reactor de Langenbuch (tridimensional) [63].

A continuación, se exponen los principales objetivos que se persiguen. Como se ha mencionado, existen diferentes técnicas para discretizar la parte espacial de la ecuación de la difusión neutrónica. Las más habituales son los métodos en diferencias finitas centradas y los métodos de colocación nodal, que serán utilizados en la presente memoria.

Por otra parte, como se verá en el capítulo 2, la resolución numérica de la ecuación de la difusión neutrónica requiere obtener la solución de sistemas de ecuaciones lineales de gran dimensión y vacíos, por lo que es conveniente el uso de métodos iterativos para su resolución. Debido al carácter disperso, la no simetría y la clara estructura a bloques que presentan las matrices de los sistemas de ecuaciones a resolver, es interesante desarrollar métodos que exploten estas características de las matrices. Por tanto, otro de los objetivos de esta memoria es el estudio de las propiedades de convergencia de esquemas iterativos que exploten la estructura a bloques de los sistemas de ecuaciones lineales que aparecen en la resolución numérica de la ecuación de la difusión neutrónica.

Capítulo 1. Introducción

Debido al carácter no simétrico de las matrices, se pueden aplicar métodos basados en subespacios de Krylov especialmente indicados para este tipo de matrices como, por ejemplo, los métodos GMRES, BiCGSTAB y TFQMR. Por esta razón, interesa disponer de un estudio comparativo de los diferentes métodos iterativos que se pueden aplicar para resolver los sistemas de ecuaciones.

Los objetivos anteriores se cubren en los capítulos 3 y 4. En ellos se realiza un estudio teórico de las propiedades de convergencia de un método de segundo grado acelerado y los experimentos numéricos encaminados a evaluar sus prestaciones. Además se compara con los métodos basados en subespacios de Krylov antes mencionados.

En la sección 1.4.3, se han descrito brevemente las técnicas multinivel para la solución de sistemas de ecuaciones lineales. En la actualidad este tipo de técnicas son utilizadas con profusión para muy diversos tipos de problemas, empleándose también como preconditionadores para métodos basados en subespacios de Krylov. Las especiales características del método de colocación nodal, que se describirá en la sección 2.3.2, permiten el desarrollo de métodos multinivel basados en el número de polinomios de Legendre utilizado para la discretización de la ecuación de la difusión neutrónica. Éste es otro de los objetivos de la memoria que será desarrollado en el capítulo 5.

Capítulo 2

Discretización de la ecuación de la difusión neutrónica

2.1 Introducción

En este capítulo se presenta la ecuación de la difusión neutrónica dependiente del tiempo utilizada para estudiar la distribución de neutrones en el interior del núcleo de un reactor nuclear. La ecuación constituye un sistema de ecuaciones en derivadas parciales. Para su resolución numérica se han de aplicar métodos de discretización tanto para la parte espacial de la ecuación, como para la parte temporal. Es un objetivo del tema la presentación de diferentes métodos para llevar a cabo ambas tareas.

El capítulo está organizado como sigue. En la sección 2.2 se presenta la ecuación de la difusión neutrónica. En las secciones 2.3.1 y 2.3.2 se presentan dos métodos para la discretización de la parte espacial de la ecuación. Concretamente, el método de las diferencias finitas centradas y un método de colocación nodal. En la sección 2.4 se presentan diferentes métodos implícitos en diferencias hacia atrás para la discretización de la parte temporal. En la sección 2.5 se describen los problemas modelo que serán utilizados para evaluar los diferentes métodos de resolución propuestos en la presente tesis. Finalmente, en la sección 2.6 se presentan los resultados de los experimen-

2.2. Ecuación de la difusión neutrónica

tos numéricos encaminados a evaluar los dos métodos de discretización de la parte espacial de las ecuaciones presentados.

2.2 Ecuación de la difusión neutrónica

Como se mencionó en el capítulo 1, una de las magnitudes que interesa estudiar a la hora de predecir el comportamiento de un reactor nuclear es la distribución de neutrones en el núcleo del reactor como una función de la posición, la energía y el tiempo. Así pues, uno de los principales problemas de la teoría de reactores consiste en predecir esta distribución. La distribución de neutrones dentro del reactor se modeliza mediante la ecuación de la difusión neutrónica que es una aproximación de la ecuación del transporte de Boltzman [102]. Esta ecuación se obtiene teniendo en cuenta que la variación del flujo neutrónico dentro de un volumen del reactor es igual a la proporción de neutrones que entran menos la proporción de neutrones que salen del volumen. Generalmente es suficiente considerar la aproximación de dos grupos de energía que divide el espectro energético de los neutrones en dos grupos, un grupo rápido, correspondiente a los neutrones cuya energía es superior a unas unidades de electronvoltios y que denotaremos por ϕ_1 , y un grupo térmico, correspondiente a los neutrones cuya energía es menor que esta cantidad y que denotaremos por ϕ_2 . Además, se supone que no hay procesos de dispersión del grupo térmico al rápido y que no se producen neutrones en el grupo térmico. También se asume que no hay fuentes externas de neutrones. Con estas consideraciones se obtiene un sistema de ecuaciones en derivadas parciales para los grupos rápido y térmico dada por [102],

$$\begin{aligned} \frac{1}{\nu_1} \frac{\partial}{\partial t} \phi_1(r, t) - \nabla \cdot (D_1(r, t) \nabla \phi_1(r, t)) &= -(\Sigma_{a1}(r, t) + \Sigma_{12}(r, t)) \phi_1(r, t) \\ &+ (1 - \beta)(\nu \Sigma_{f1}(r, t) \phi_1(r, t) \\ &+ \nu \Sigma_{f2}(r, t) \phi_2(r, t)) \\ &+ \sum_{k=1}^{K_c} \lambda_k \mathcal{C}_k(r, t) \end{aligned} \quad (2.1)$$

Capítulo 2. Discretización de la ec. de la difusión neutrónica

$$\begin{aligned} \frac{1}{\nu_2} \frac{\partial}{\partial t} \phi_2(r, t) - \nabla \cdot (D_2(r, t) \nabla \phi_2(r, t)) &= -\Sigma_{a2}(r, t) \phi_2(r, t) \\ &+ \Sigma_{12}(r, t) \phi_1(r, t) , \end{aligned} \quad (2.2)$$

donde (r, t) indica las magnitudes espacial y temporal respectivamente, que en adelante se asumirán sin indicarse. D_g y Σ_{ag} son el coeficiente de difusión y la sección eficaz de absorción para el grupo g , respectivamente. Σ_{12} es la sección eficaz de dispersión del grupo rápido al térmico. El término $\nu\Sigma_{fg}$ es la cantidad de neutrones producidos por fisión en el grupo g . La velocidad del grupo g se denota por ν_g . \mathcal{C}_k es la concentración de precursores del tipo k . La tasa con la que un precursor de tipo k decae es $\lambda_k \mathcal{C}_k$. Además, β_k es la proporción de neutrones de fisión diferidos debidos a la transformación de un precursor de tipo k , con $\beta = \sum_{k=1}^{K_c} \beta_k$, donde K_c es el número de precursores de neutrones que se consideran.

Las ecuaciones (2.1) y (2.2) se expresan en forma matricial como sigue:

$$[\nu^{-1}] \dot{\phi} + \mathcal{L}\phi = (1 - \beta)\mathcal{M}\phi + \chi \sum_{k=1}^{K_c} \lambda_k \mathcal{C}_k , \quad (2.3)$$

donde

$$\mathcal{L} = \begin{bmatrix} -\nabla \cdot (D_1 \nabla) + \Sigma_{a1} + \Sigma_{12} & 0 \\ -\Sigma_{12} & -\nabla \cdot (D_2 \nabla) + \Sigma_{a2} \end{bmatrix} , \quad [\nu^{-1}] = \begin{bmatrix} \frac{1}{\nu_1} & 0 \\ 0 & \frac{1}{\nu_2} \end{bmatrix} ,$$

y

$$\mathcal{M} = \begin{bmatrix} \nu\Sigma_{f1} & \nu\Sigma_{f2} \\ 0 & 0 \end{bmatrix} , \quad \phi = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} , \quad \chi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} .$$

La ecuación para la concentración del grupo k -ésimo de precursores de neutrones es de la forma

$$\dot{\mathcal{C}}_k = \beta_k [\nu\Sigma_{f1} \ \nu\Sigma_{f2}] \phi - \lambda_k \mathcal{C}_k , \quad k = 1, \dots, K . \quad (2.4)$$

Las condiciones de contorno para el flujo neutrónico son $\phi|_{\Gamma} = 0$, donde Γ es la frontera del reactor. Además se asume una distribución inicial de neutrones en el reactor, $\phi(r, 0)$ que corresponde al comportamiento del reactor

2.3. Discretización espacial

en régimen estacionario, cuyo cálculo supone la solución de un problema de valores propios generalizado que ha sido estudiado por ejemplo en [34] y [100]. Para la integración numérica de las ecuaciones (2.3) y (2.4) en primer lugar se realiza una discretización de la parte espacial de las ecuaciones, seguida de una discretización temporal, como se detalla a continuación.

2.3 Discretización espacial

Para la discretización espacial de las ecuaciones (2.3) y (2.4) se pueden utilizar el método de los elementos finitos [48], métodos en diferencias finitas [106] y [20], o métodos nodales [49],[99] y [38]. A continuación se muestra el resultado de discretizar estas ecuaciones mediante el método en diferencias finitas centradas y el método de colocación nodal, métodos utilizados en los experimentos numéricos de capítulos posteriores.

2.3.1 Método en diferencias finitas centradas

Los métodos en diferencias finitas se basan en la aproximación local de las derivadas parciales que aparecen en una ecuación en derivadas parciales, como es el caso de la ecuación de la difusión neutrónica (2.3), mediante un desarrollo en serie de Taylor. Para ello, se genera una malla de puntos dividiendo el reactor en pequeñas regiones o celdas, normalmente rectangulares, y en cada punto de la malla la derivada se aproxima mediante un cociente en diferencias. El error que se comete en la discretización es tanto menor cuanto más fina es la malla generada.

Se parte de la ecuación matricial (2.3) [42]. En la figura (2.1) se muestra un ejemplo para el caso bidimensional, donde una región (reactor) rectangular de tamaño $[0, 1] \times [0, 1]$ se ha dividido mediante una malla uniforme de puntos $\{x_i, y_j : i = 0, 1, \dots, n_x + 1, j = 0, 1, \dots, n_y + 1\}$, con espaciado $h_x = 1/(n_x + 1)$ en la dirección x , y $h_y = 1/(n_y + 1)$ en la dirección y .

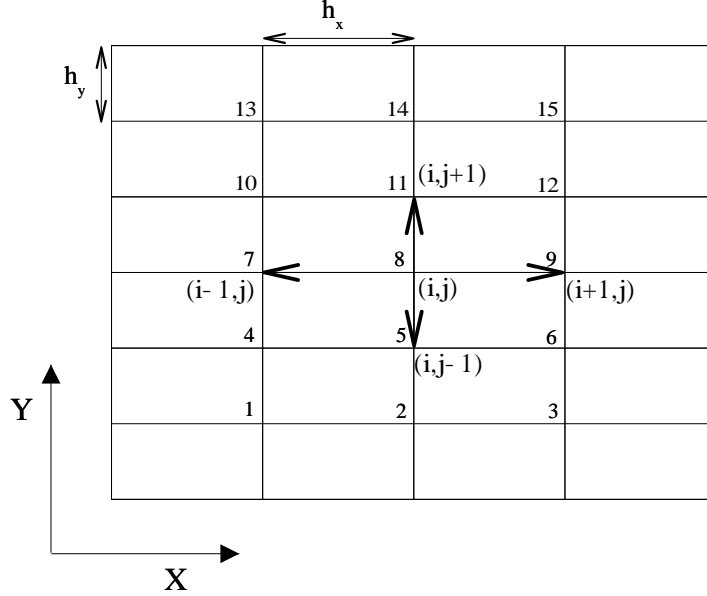


Figura 2.1: Discretización en diferencias finitas centradas. Ordenamiento natural.

Los términos de la forma

$$-\nabla \cdot (D_g \nabla \phi_g) = \frac{\partial}{\partial x} (D_g \frac{\partial \phi_g}{\partial x}) + \frac{\partial}{\partial y} (D_g \frac{\partial \phi_g}{\partial y}), \quad g = 1, 2,$$

se aproximan en el punto de la malla (i, j) mediante la ecuación en diferencias centradas dada por

$$\frac{\partial}{\partial x} (D_g \frac{\partial \phi_g}{\partial x})(i, j) \approx \frac{D_{g,i+1/2,j}(\phi_{g,i+1,j} - \phi_{g,i,j}) - D_{g,i-1/2,j}(\phi_{g,i,j} - \phi_{g,i-1,j})}{h_x^2},$$

$$\frac{\partial}{\partial y} (D_g \frac{\partial \phi_g}{\partial y})(i, j) \approx \frac{D_{g,i,j+1/2}(\phi_{g,i,j+1} - \phi_{g,i,j}) - D_{g,i,j-1/2}(\phi_{g,i,j} - \phi_{g,i,j-1})}{h_y^2},$$

donde $D_{g,i\pm 1/2,j} \equiv D_g(x_i \pm h_x/2, y_j)$, $D_{g,i,j\pm 1/2} \equiv D_g(x_i, y_j \pm h_y/2)$, y $\phi_{g,i,j}$ representa la aproximación para ϕ_g en el punto (x_i, y_j) . Esta aproximación se denomina aproximación centrada de cinco puntos. La dependencia de la derivada en un punto, con respecto a los puntos de su alrededor se representa

2.3. Discretización espacial

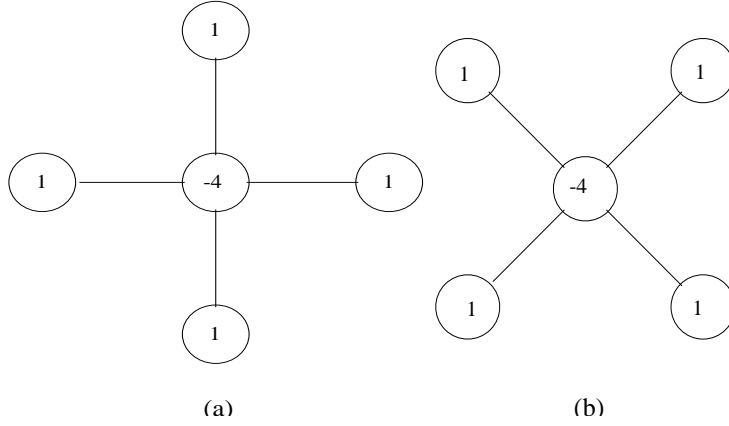


Figura 2.2: Patrones de cinco puntos: (a) patrón estándar (b) patrón diagonal.

habitualmente mediante un patrón de cinco puntos como se muestra en la figura 2.2 (a). Otro posible patrón se muestra en la figura 2.2 (b), donde se utilizan los cuatro puntos localizados en las diagonales, cambiando el tamaño de la malla. Utilizando el patrón de cinco puntos (a) en cada nodo de la malla se obtiene la ecuación

$$\begin{aligned}
 -\nabla \cdot (D_g \nabla \phi_g)(i, j) &\approx \frac{-1}{h_x^2} (D_{g,i+1/2,j}(\phi_{g,i+1,j} - \phi_{g,i,j}) - \\
 &- D_{g,i-1/2,j}(\phi_{g,i,j} - \phi_{g,i-1,j})) + \\
 &+ \frac{-1}{h_y^2} (D_{g,i,j+1/2}(\phi_{g,i,j+1} - \phi_{g,i,j}) - \\
 &- D_{g,i,j-1/2}(\phi_{g,i,j} - \phi_{g,i,j-1})) . \quad (2.5)
 \end{aligned}$$

Después de un ordenamiento natural de los nodos, que consiste en numerar los nodos de izquierda a derecha en cada línea, y progresando desde las líneas inferiores hasta las superiores (figura 2.1), se obtiene un sistema de ecuaciones diferenciales ordinarias de la forma

$$\begin{aligned}
 \begin{bmatrix} v_1^{-1} & 0 \\ 0 & v_2^{-1} \end{bmatrix} \dot{\psi} + \begin{bmatrix} L_{11} & O \\ -L_{21} & L_{22} \end{bmatrix} \psi &= (1 - \beta) \begin{bmatrix} M_{11} & M_{12} \\ O & O \end{bmatrix} \psi + \\
 &+ \begin{bmatrix} \sum_{k=1}^{K_c} \lambda_k C_k \\ O \end{bmatrix} , \quad (2.6)
 \end{aligned}$$

Capítulo 2. Discretización de la ec. de la difusión neutrónica

$$\dot{C}_k = \beta_k [M_{11} M_{12}] \psi - \lambda_k C_k, \quad (2.7)$$

donde $\psi^T = [\psi_{1n} \ \psi_{2n}]^T$, ψ_{1n} y ψ_{2n} son vectores de \mathbb{R}^n cuyas componentes corresponden a los flujos rápido y térmico respectivamente, discretizados en cada nodo de la malla. $\sum_{k=1}^{K_c} \lambda_k C_k$ es un vector de \mathbb{R}^n y corresponde a los precursores de neutrones en cada nodo. $n = n_x \times n_y$ es el número de puntos interiores en la malla. L_{ij} y M_{ij} son matrices cuadradas de dimensión $n \times n$, siendo las matrices M_{11} , M_{12} y L_{21} matrices diagonales cuyos elementos son positivos o nulos. Definiendo

$$d_{i,j} \equiv \frac{D_{g,i+1/2,j} + D_{i-1/2,j}}{h_x^2} + \frac{D_{g,i,j+1/2} + D_{g,i,j-1/2}}{h_y^2},$$

$$b_{i+1/2,j} \equiv \frac{-D_{g,i+1/2,j}}{h_x^2}, \quad c_{i+1/2,j} \equiv \frac{-D_{g,i,j+1/2}}{h_y^2},$$

las matrices L_{11} y L_{22} son matrices banda pentadiagonales, con ancho de semibanda n_x , con la siguiente estructura

$$L_{ii} = \begin{bmatrix} S_1 & T_{3/2} & & & \\ T_{3/2} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & T_{n_y-1/2} & S_{n_y} & \\ & & & & \end{bmatrix}, \quad (2.8)$$

donde

$$S_j = \begin{bmatrix} d_{1,j} & b_{3/2,j} & & & \\ b_{3/2,j} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & b_{n_x-1/2,j} & d_{n_x,j} & \\ & & & & \end{bmatrix}, \quad (2.9)$$

$$T_{j+1/2} = \begin{bmatrix} c_{1,j+1/2} & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & c_{n_x,j+1/2} \end{bmatrix}. \quad (2.10)$$

2.3. Discretización espacial

Debido a que los términos D_g , $g = 1, 2$ son mayores que cero, el siguiente teorema garantiza que cada una de las matrices L_{11} y L_{22} son simétricas y definidas positivas.

Teorema 17 [42, teorema 9.1.1] *Si $Dg \geq \alpha > 0$, entonces la matriz de coeficientes definida en (2.8), (2.9) y (2.10) es simétrica y definida positiva.*

La generalización de la discretización presentada a problemas en geometrías 3D es sencilla. En ese caso, después de aplicar un ordenamiento natural a los índices de los nodos, numerando por planos horizontales de abajo a arriba, y en cada plano horizontal como el caso 2D, se obtienen matrices L_{11} y L_{22} banda, simétricas y definidas positivas, con siete diagonales y ancho de la semibanda $n_x \times n_y$.

Es importante hacer notar que la discretización de ecuaciones elípticas de segundo orden, como es el caso del operador de Laplace que aparece en la ecuación de la difusión neutrónica (2.3), mediante métodos en diferencias finitas, conduce a matrices cuyo número de condición crece como $O(1/h^2)$ para problemas 2D ($O(1/h^3)$ para problemas 3D)(ver por ejemplo [42, corolario 9.1.2] donde se obtiene un resultado de este tipo para la ecuación de Poisson, siendo $h = h_x = h_y$ el tamaño del mallado). Como se verá en la sección 2.6, para lograr una precisión suficiente en la resolución numérica de la ecuación de la difusión neutrónica discretizando mediante el método de las diferencias finitas centradas, se necesita utilizar tamaños de malla muy pequeños. Ésta es una de las razones por las que en posteriores capítulos de la tesis se trabajará exclusivamente con el método de colocación nodal, especialmente para transitorios tridimensionales.

2.3.2 Método de colocación nodal

El método de colocación nodal desarrollado para la ecuación de la difusión neutrónica es una adaptación de los métodos de colocación clásicos

Capítulo 2. Discretización de la ec. de la difusión neutrónica

para la discretización de ecuaciones en derivadas parciales, basados en el desarrollo de la solución como combinación lineal de una base de funciones [83]. En particular, el método de colocación nodal para la ecuación de la difusión neutrónica hace uso de un desarrollo del flujo neutrónico en términos de polinomios ortonormales de Legendre. Nos restringiremos al caso tridimensional, en el que se considera al reactor dividido en nodos paralelepípedos como se muestra en la figura 2.3. Cada nodo se denotará por e . Las ecuaciones (2.3) en cada nodo se pueden escribir como

$$\begin{aligned} -\nabla \cdot (D_{1,e} \nabla \phi_{1,e} + (\Sigma_{a1,e} + \Sigma_{12,e}) \phi_{1,e}) &= S_{1,e} \\ -\nabla \cdot (D_{2,e} \nabla \phi_{2,e}) + \Sigma_{a2,e} \phi_{2,e} &= S_{2,e} \end{aligned} \quad (2.11)$$

donde

$$\begin{aligned} S_{1,e} &= (1 - \beta)(\nu \Sigma_{f1} \phi_1 + \nu \Sigma_{f2} \phi_2) + \sum_{k=1}^{K_c} \lambda_k \mathcal{C}_k - \frac{1}{\nu_1} \frac{\partial \phi_1}{\partial t}, \\ S_{2,e} &= \Sigma_{12} \phi_1 - \frac{1}{\nu_1} \frac{\partial \phi_2}{\partial t}, \end{aligned}$$

se interpretan como términos fuente de neutrones.

En esta sección se presentan las características principales del método considerando sólo un grupo de energía. La generalización al caso de dos grupos de energía es sencilla. Asumiendo que las propiedades nucleares del reactor permanecen constantes en cada nodo y, considerando una ecuación genérica, se obtiene

$$-D_{x,e} \frac{\partial^2 \phi_e}{\partial x^2} - D_{y,e} \frac{\partial^2 \phi_e}{\partial y^2} - D_{z,e} \frac{\partial^2 \phi_e}{\partial z^2} + \Sigma_{re} \phi_e = S_e. \quad (2.12)$$

Realizando el cambio de variables (ver figura 2.3)

2.3. Discretización espacial

$$\begin{aligned} u &= \frac{1}{\Delta x_e} \left[x - \frac{1}{2}(x_{m-1/2} + x_{m+1/2}) \right], \\ v &= \frac{1}{\Delta y_e} \left[y - \frac{1}{2}(y_{n-1/2} + y_{n+1/2}) \right], \\ w &= \frac{1}{\Delta z_e} \left[z - \frac{1}{2}(z_{p-1/2} + z_{p+1/2}) \right], \end{aligned} \quad (2.13)$$

la ecuación (2.12) se expresa en la forma

$$\begin{aligned} V_e S_e &= \Sigma_{re} V_e \phi_e - \frac{\Delta y_e \Delta z_e}{\Delta x_e} D_{x,e} \frac{\partial^2 \phi_e}{\partial u^2} - \\ &- \frac{\Delta x_e \Delta z_e}{\Delta y_e} D_{y,e} \frac{\partial^2 \phi_e}{\partial v^2} - \frac{\Delta x_e \Delta y_e}{\Delta z_e} D_{z,e} \frac{\partial^2 \phi_e}{\partial w^2}, \end{aligned} \quad (2.14)$$

donde $V_e = \Delta x_e \Delta y_e \Delta z_e$.

Como función de prueba para la solución en cada nodo e , se utiliza el desarrollo truncado

$$\phi_e(u, v, w) = \sum_{k_1=0}^K \sum_{k_2=0}^K \sum_{k_3=0}^K \phi_e^{k_1, k_2, k_3} P_{k_1}(u) P_{k_2}(v) P_{k_3}(w), \quad (2.15)$$

y para la fuente S_e , se toma

$$S_e(u, v, w) = \sum_{k_1=0}^K \sum_{k_2=0}^K \sum_{k_3=0}^K S_e^{k_1, k_2, k_3} P_{k_1}(u) P_{k_2}(v) P_{k_3}(w), \quad (2.16)$$

donde K es el orden del desarrollo elegido y $P_k(u)$ son los polinomios de Legendre ortonormales en el intervalo $[-1/2, 1/2]$, o lo que es lo mismo,

$$\int_{-1/2}^{1/2} P_k(u) P_l(u) du = \delta_{kl}, \quad (2.17)$$

donde δ_{kl} es la función delta de Kronecker. Estos polinomios difieren de la definición clásica de los polinomios de Legendre (véase por ejemplo [59], pág. 438-440), ya que son ortonormales y su dominio de definición es $[-1/2, 1/2]$. Ello es debido al cambio de variables (2.13) realizado para disponer de elementos cuyo volumen sea la unidad. Así pues, se utilizarán los polinomios de Legendre, $P_l(u)$, definidos por (ver [49])

$$P_0(u) = 1, \quad P_1(u) = 2\sqrt{3}u, \quad (2.18)$$

Capítulo 2. Discretización de la ec. de la difusión neutrónica

y la relación de recurrencia

$$P_{k+1}(u) = 2\sqrt{\frac{2k+3}{2k+1}} \frac{2k+1}{k+1} u P_k(u) - \sqrt{\frac{2k+3}{2k-1}} \frac{k}{k+1} P_{k-1}(u), \quad k \geq 1. \quad (2.19)$$

Para estos polinomios se satisfacen las igualdades

$$P_n(1/2) = \sqrt{2n+1}, \quad P_n(-1/2) = (-1)^n \sqrt{2n+1}, \quad (2.20)$$

y

$$P'_n(1/2) = n(n+1)\sqrt{2n+1}, \quad P'_n(-1/2) = (-1)^{n+1} n(n+1)\sqrt{2n+1}. \quad (2.21)$$

Utilizando la integración por partes, y las igualdades (2.20) y (2.21), se puede desarrollar una función genérica $f(u)$ en términos de polinomios de Legendre como $f(u) = \sum_{l=0}^{\infty} F_l P_l(u)$, obteniéndose el siguiente resultado

$$\begin{aligned} \int_{-1/2}^{1/2} du P_k(u) \frac{d^2}{du^2} f(u) &= \sqrt{2k+1} \left\{ (-1)^{k+1} \left[k(k+1)f(-1/2) + \frac{d}{du} f(-1/2) \right] \right. \\ &\quad - \left[k(k+1)f(1/2) - \frac{d}{du} f(1/2) \right] + \\ &\quad \left. + \sum_{l=0}^{K-2} [1 + (-1)^{k+l}] \sqrt{2l+1} [k(k+1) - l(l+1)] F_l \right\}. \end{aligned} \quad (2.22)$$

Introduciendo las expresiones (2.15) y (2.16) en la ecuación (2.14), multiplicando esta ecuación por la función de ponderación

$$W_{k_1, k_2, k_3}(u, v, w) = P_{k_1}(u) P_{k_2}(v) P_{k_3}(w),$$

integrando sobre todo el volumen del elemento e y utilizando la relación de ortonormalidad (2.17), se obtiene la ecuación

$$\begin{aligned} V_e S_e^{k_1, k_2, k_3} &= V_e \sum_r \phi_e^{k_1, k_2, k_3} - \Delta y_e \Delta z_e F_{e,x}^{k_1, k_2, k_3} - \\ &\quad - \Delta x_e \Delta z_e F_{e,y}^{k_1, k_2, k_3} - \Delta x_e \Delta y_e F_{e,z}^{k_1, k_2, k_3}, \end{aligned} \quad (2.23)$$

2.3. Discretización espacial

donde

$$\begin{aligned} F_{e,x}^{k_1,k_2,k_3} &= \frac{D_{x,e}}{\Delta x_e} L_{k_1} \{ \phi_{e,x}^{k_2,k_3}(u) \} , \\ F_{e,y}^{k_1,k_2,k_3} &= \frac{D_{y,e}}{\Delta y_e} L_{k_2} \{ \phi_{e,y}^{k_1,k_3}(v) \} , \\ F_{e,z}^{k_1,k_2,k_3} &= \frac{D_{z,e}}{\Delta z_e} L_{k_3} \{ \phi_{e,z}^{k_1,k_2}(w) \} . \end{aligned}$$

Las funciones $\phi_{e,x}^{k_2,k_3}(u)$, $\phi_{e,y}^{k_1,k_3}(v)$, $\phi_{e,z}^{k_1,k_2}(w)$, son, respectivamente,

$$\begin{aligned} \phi_{e,x}^{k_2,k_3}(u) &= \sum_{k=0}^K \phi_e^{k,k_2,k_3} P_k(u) , \\ \phi_{e,y}^{k_1,k_3}(v) &= \sum_{k=0}^K \phi_e^{k_1,k,k_3} P_k(v) , \\ \phi_{e,z}^{k_1,k_2}(w) &= \sum_{k=0}^K \phi_e^{k_1,k_2,k} P_k(w) , \end{aligned}$$

y el término

$$L_k(f(u)) = \int_{-1/2}^{1/2} du P_k(u) \frac{d^2}{du^2} f(u) ,$$

viene dado por la ecuación (2.22).

Ahora se imponen las condiciones de continuidad del flujo y la corriente en las seis celdas vecinas a la celda e . Para ello se numeran estos nodos de e_1 a e_6 , de la forma que se indica en la figura 2.3. Si, por ejemplo, se considera la frontera común entre los nodos e y e_1 , las condiciones de continuidad para el flujo y la corriente sobre la cara común de estos nodos, son, respectivamente,

$$\begin{aligned} \phi_{e_1}\left(\frac{1}{2}, v, w\right) &= \phi_e\left(-\frac{1}{2}, v, w\right) , \\ \frac{D_{x,e_1}}{\Delta x_{e_1}} \frac{\partial}{\partial u} \phi_{e_1}\left(\frac{1}{2}, v, w\right) &= \frac{D_{x,e}}{\Delta x_e} \frac{\partial}{\partial u} \phi_e\left(-\frac{1}{2}, v, w\right) . \end{aligned}$$

Multiplicando estas igualdades por la función de ponderación

$$W_{k_2,k_3} = P_{k_2}(v) P_{k_3}(w) ,$$

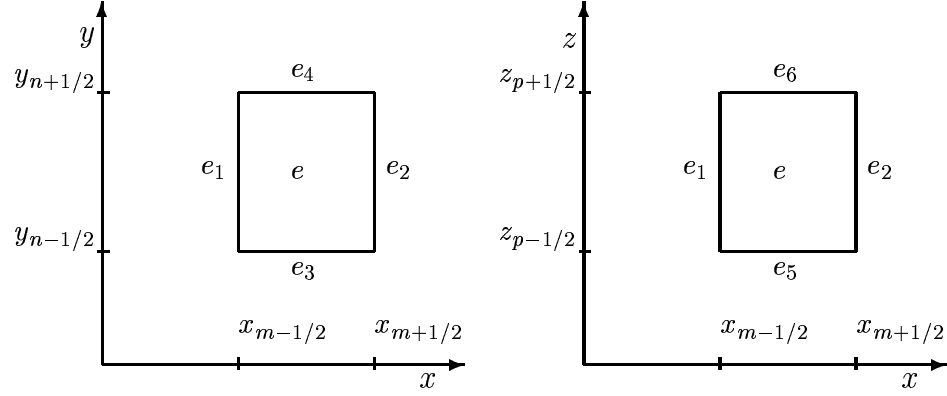


Figura 2.3: Posición de los nodos adyacentes al nodo e .

e integrando sobre la superficie común se obtiene

$$\begin{aligned} \phi_{e_1,x}^{k_2,k_3}\left(\frac{1}{2}\right) &= \phi_{e,x}^{k_2,k_3}\left(-\frac{1}{2}\right), \\ \frac{D_{x,e_1}}{\Delta x_{e_1}} \frac{d}{du} \phi_{e_1,x}^{k_2,k_3}\left(\frac{1}{2}\right) &= \frac{D_{x,e}}{\Delta x_e} \frac{d}{du} \phi_{e,x}^{k_2,k_3}\left(-\frac{1}{2}\right). \end{aligned} \quad (2.24)$$

Condiciones similares pueden obtenerse para las otras caras del elemento e . Utilizando el resultado (2.22) y las igualdades (2.20) y (2.21), se pueden obtener los resultados

$$\phi_{e,x}^{k_2,k_3}\left(-\frac{1}{2}\right) + \frac{1}{K(K+1)} \frac{d}{du} \phi_{e,x}^{k_2,k_3}\left(-\frac{1}{2}\right) = \sum_{l=0}^{K-1} (-1)^l \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{l,k_2,k_3}, \quad (2.25)$$

$$\phi_{e,x}^{k_2,k_3}\left(\frac{1}{2}\right) - \frac{1}{K(K+1)} \frac{d}{du} \phi_{e,x}^{k_2,k_3}\left(\frac{1}{2}\right) = \sum_{l=0}^{K-1} \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{l,k_2,k_3},$$

$$\phi_{e,y}^{k_1,k_3}\left(-\frac{1}{2}\right) + \frac{1}{K(K+1)} \frac{d}{dv} \phi_{e,y}^{k_1,k_3}\left(-\frac{1}{2}\right) = \sum_{l=0}^{K-1} (-1)^l \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{k_1,l,k_3},$$

$$\phi_{e,y}^{k_1,k_3}\left(\frac{1}{2}\right) - \frac{1}{K(K+1)} \frac{d}{dv} \phi_{e,y}^{k_1,k_3}\left(\frac{1}{2}\right) = \sum_{l=0}^{K-1} \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{k_1,l,k_3},$$

2.3. Discretización espacial

$$\phi_{e,z}^{k_1,k_2}(-\frac{1}{2}) + \frac{1}{K(K+1)} \frac{d}{dw} \phi_{e,z}^{k_1,k_2}(-\frac{1}{2}) = \sum_{l=0}^{K-1} (-1)^l \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{k_1,k_2,l},$$

$$\phi_{e,z}^{k_1,k_2}(\frac{1}{2}) - \frac{1}{K(K+1)} \frac{d}{dw} \phi_{e,z}^{k_1,k_2}(\frac{1}{2}) = \sum_{l=0}^{K-1} \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{k_1,k_2,l}.$$

Utilizando (2.25) y las condiciones (2.24), se puede expresar

$$\begin{aligned} \frac{d}{du} \phi_{e,x}^{k_2,k_3}(-\frac{1}{2}) &= \frac{\Delta x_e D_{x,e_1}}{\Delta x_{e_1} D_{x,e} + \Delta x_e D_{x,e_1}} \times \\ &\times \sum_{l=0}^{K-1} \sqrt{2l+1} [K(K+1) - l(l+1)] ((-1)^l \phi_e^{l,k_2,k_3} - \phi_{e_1}^{l,k_2,k_3}), \end{aligned} \quad (2.26)$$

y combinando la ecuación (2.26) y la ecuación (2.25), se obtiene

$$\begin{aligned} &\frac{D_{x,e}}{\Delta x_e} [k(k+1) \phi_{e,x}^{k_2,k_3}(-\frac{1}{2}) + \frac{d}{du} \phi_{e,x}^{k_2,k_3}(-\frac{1}{2})] = \\ &\frac{D_{x,e} k(k+1)}{\Delta x_e} \sum_{l=0}^{K-1} (-1)^l \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{l,k_2,k_3} + \\ &\frac{\Delta x_e D_{x,e_1}}{\Delta x_{e_1} D_{x,e} + \Delta x_e D_{x,e_1}} \left(1 - \frac{k(k+1)}{K(K+1)}\right) \times \\ &\times \sum_{l=0}^{K-1} \sqrt{2l+1} [K(K+1) - l(l+1)] ((-1)^l \phi_e^{l,k_2,k_3} - \phi_{e_1}^{l,k_2,k_3}). \end{aligned} \quad (2.27)$$

De forma totalmente análoga, se obtiene que

$$\begin{aligned} &\frac{D_{x,e}}{\Delta x_e} [k(k+1) \phi_{e,x}^{k_2,k_3}(\frac{1}{2}) - \frac{d}{du} \phi_{e,x}^{k_2,k_3}(\frac{1}{2})] = \\ &\frac{D_{x,e} k(k+1)}{\Delta x_e} \sum_{l=0}^{K-1} \sqrt{2l+1} \left(1 - \frac{l(l+1)}{K(K+1)}\right) \phi_e^{l,k_2,k_3} - \\ &-\frac{\Delta x_e D_{x,e_1}}{\Delta x_{e_1} D_{x,e} + \Delta x_e D_{x,e_1}} \left(1 - \frac{k(k+1)}{K(K+1)}\right) \times \\ &\times \sum_{l=0}^{K-1} \sqrt{2l+1} [K(K+1) - l(l+1)] ((-1)^l \phi_{e_2}^{l,k_2,k_3} - \phi_e^{l,k_2,k_3}). \end{aligned} \quad (2.28)$$

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Utilizando los resultados (2.27) y (2.28) se obtiene que

$$\begin{aligned}
 F_{e,x}^{k,k_2,k_3} = & -\frac{\sqrt{2k+1}}{K(K+1)} \left\{ \frac{D_{x,e}}{\Delta x_e} [K(K+1) - k(k+1)] \times \right. \\
 & \times \sum_{l=0}^{k-2} (1 + (-1)^{k+l}) \sqrt{2l+1} l(l+1) \phi_e^{l,k_2,k_3} + \\
 & + \frac{D_{x,e}}{\Delta x_e} k(k+1) \sum_{l=k}^{K-1} (1 + (-1)^{k+l}) \sqrt{2l+1} [K(K+1) - l(l+1)] \phi_e^{l,k_2,k_3} + \\
 & + \frac{1}{2} [K(K+1) - k(k+1)] \sum_{l=0}^{K-1} \sqrt{2l+1} [K(K+1) - l(l+1)] \times \\
 & \times \left[(-1)^k \frac{D_{x,e} D_{x,e_1}}{\Delta x_e D_{x,e_1} + \Delta x_{e_1} D_{x,e}} [(-1)^l \phi_e^{l,k_2,k_3} - \phi_{e_1}^{l,k_2,k_3}] - \right. \\
 & \left. - \frac{D_{x,e} D_{x,e_2}}{\Delta x_e D_{x,e_2} + \Delta x_{e_2} D_{x,e}} [(-1)^l \phi_{e_2}^{l,k_2,k_3} - \phi_e^{l,k_2,k_3}] \right] \left. \right\} . \quad (2.29)
 \end{aligned}$$

Esta expresión es invariante si se permutan los índices k y l , lo que asegura que las matrices generadas mediante este método para una sola ecuación sean simétricas.

Del mismo modo que se ha obtenido la relación (2.29) para $F_{e,x}^{k,k_2,k_3}$, se obtienen expresiones similares para $F_{e,y}^{k_1,k,k_3}$ y $F_{e,z}^{k_1,k_2,k}$, que se pueden reescribir de la forma

$$\begin{aligned}
 F_{e,x}^{k,k_2,k_3} &= \sum_{l=0}^{K-1} (A_{e,x}^{k,l;K} \phi_{e1}^{l,k_2,k_3} - B_{e,x}^{k,l;K} \phi_e^{l,k_2,k_3} + C_{e,x}^{k,l;K} \phi_{e2}^{l,k_2,k_3}) , \\
 F_{e,y}^{k_1,k,k_3} &= \sum_{l=0}^{K-1} (A_{e,y}^{k,l;K} \phi_{e3}^{k_1,l,k_3} - B_{e,y}^{k,l;K} \phi_e^{k_1,l,k_3} + C_{e,y}^{k,l;K} \phi_{e4}^{k_1,l,k_3}) , \quad (2.30) \\
 F_{e,z}^{k_1,k_2,k} &= \sum_{l=0}^{K-1} (A_{e,z}^{k,l;K} \phi_{e5}^{k_1,k_2,l} - B_{e,z}^{k,l;K} \phi_e^{k_1,k_2,l} + C_{e,z}^{k,l;K} \phi_{e6}^{k_1,k_2,l}) ,
 \end{aligned}$$

con los coeficientes

$$\begin{aligned}
 A_{e,\alpha}^{k,l;K} &= \frac{(-1)^k}{2K(K+1)} \sqrt{2k+1} \sqrt{2l+1} [K(K+1) - k(k+1)] \times \\
 &\times [K(K+1) - l(l+1)] W_{e,\alpha}^- , \quad (2.31)
 \end{aligned}$$

2.3. Discretización espacial

$$C_{e,\alpha}^{k,l;K} = \frac{(-1)^l}{2K(K+1)} \sqrt{2k+1} \sqrt{2l+1} [K(K+1) - k(k+1)] \times \quad (2.32)$$

$$\times [K(K+1) - l(l+1)] W_{e,\alpha}^+,$$

$$B_{e,\alpha}^{k,l;K} = \frac{\sqrt{2k+1} \sqrt{2l+1}}{K(K+1)} \left\{ \frac{D_{\alpha,e}}{\Delta \alpha_e} [1 + (-1)^{k+l}] [K(K+1) - k(k+1)] l(l+1) + \right.$$

$$+ \frac{1}{2} [K(K+1) - k(k+1)] [K(K+1) - l(l+1)] [(-1)^{k+l} W_{e,\alpha}^- + W_{e,\alpha}^+] \left. \right\}$$

si $l < k$,

(2.33)

y

$$B_{e,\alpha}^{k,l;K} = \frac{\sqrt{2k+1} \sqrt{2l+1}}{K(K+1)} \left\{ \frac{D_{\alpha,e}}{\Delta \alpha_e} k(k+1) [1 + (-1)^{k+l}] [K(K+1) - l(l+1)] + \right.$$

$$+ \frac{1}{2} [K(K+1) - k(k+1)] [K(K+1) - l(l+1)] [(-1)^{k+l} W_{e,\alpha}^- + W_{e,\alpha}^+] \left. \right\}$$

si $l \geq k$,

(2.34)

donde $\alpha = x, y, z$, y se han introducido los factores de acoplamiento de diferencias finitas centradas, que se definen como

$$W_{e,x}^- = W_{e_1,x}^+ = 2D_{x,e} D_{x,e_1} (\Delta x_e D_{x,e_1} + \Delta x_{e_1} D_{x,e})^{-1},$$

$$W_{e,x}^+ = W_{e_2,x}^- = 2D_{x,e} D_{x,e_2} (\Delta x_e D_{x,e_2} + \Delta x_{e_2} D_{x,e})^{-1},$$

$$W_{e,y}^- = W_{e_3,y}^+ = 2D_{y,e} D_{y,e_3} (\Delta y_e D_{y,e_3} + \Delta y_{e_3} D_{y,e})^{-1},$$

$$W_{e,y}^+ = W_{e_4,y}^- = 2D_{y,e} D_{y,e_4} (\Delta y_e D_{y,e_4} + \Delta y_{e_4} D_{y,e})^{-1},$$

$$W_{e,z}^- = W_{e_5,z}^+ = 2D_{z,e} D_{z,e_5} (\Delta z_e D_{z,e_5} + \Delta z_{e_5} D_{z,e})^{-1},$$

$$W_{e,z}^+ = W_{e_6,z}^- = 2D_{z,e} D_{z,e_6} (\Delta z_e D_{z,e_6} + \Delta z_{e_6} D_{z,e})^{-1}.$$

Estos coeficientes para los elementos que forman el contorno del reactor tienen la forma $W_{e,\alpha}^- = \frac{2D_{\alpha,e}}{\Delta \alpha_e}$, si la superficie de la izquierda de e es un contorno donde se anula el flujo, y $W_{e,\alpha}^+ = \frac{2D_{\alpha,e}}{\Delta \alpha_e}$, si la superficie de la derecha de e es un contorno de flujo cero.

Sustituyendo las expresiones (2.30) en la ecuación de conservación (2.23) obtenemos una nueva ecuación que involucra tan sólo a los coeficientes de

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Legendre del flujo neutrónico, $\phi_e^{k_1, k_2, k_3}$. Hay que hacer notar que el número de incógnitas por elemento es K^3 , a pesar de usar $(K + 1)^3$ polinomios de Legendre en la aproximación realizada. Para el caso $K = 1$, se comprueba fácilmente que el método de colocación nodal es equivalente al método de las diferencias finitas centradas [49].

En el método de colocación nodal, es posible introducir consistentemente la aproximación de *serendipita* [49]. Esta aproximación se aplica directamente al método reemplazando las ecuaciones (2.30) por

$$F_{e,x}^{k,k_2,k_3} = \sum_{l=0}^{K-1-k_2-k_3} (A_{e,x}^{k,l;K-k_2-k_3} \phi_{e1}^{l,k_2,k_3} - B_{e,x}^{k,l;K-k_2-k_3} \phi_e^{l,k_2,k_3}) \quad (2.35)$$

$$+ C_{e,x}^{k,l;K-k_2-k_3} \phi_{e2}^{l,k_2,k_3},$$

$$F_{e,y}^{k_1,k,k_3} = \sum_{l=0}^{K-1-k_1-k_3} (A_{e,y}^{k,l;K-k_1-k_3} \phi_{e3}^{k_1,l,k_3} - B_{e,y}^{k,l;K-k_1-k_3} \phi_e^{k_1,l,k_3}) \quad (2.36)$$

$$+ C_{e,y}^{k,l;K-k_1-k_3} \phi_{e4}^{k_1,l,k_3},$$

$$F_{e,z}^{k_1,k_2,k} = \sum_{l=0}^{K-1-k_1-k_2} (A_{e,z}^{k,l;K-k_1-k_2} \phi_{e5}^{k_1,k_2,l} - B_{e,z}^{k,l;K-k_1-k_2} \phi_e^{k_1,k_2,l}) \quad (2.37)$$

$$+ C_{e,z}^{k,l;K-k_1-k_2} \phi_{e6}^{k_1,k_2,l},$$

con $k_1 = 0, \dots, K-1$; $k_2 = 0, \dots, K-1-k_1$; $k_3 = 0, \dots, K-1-k_1-k_2$.

Así pues, generalizando la ecuación (2.23) para dos grupos de energía, utilizando las relaciones (2.35) y eligiendo una ordenación adecuada de los índices, a partir del sistema de ecuaciones diferenciales (2.11) se llega a un sistema de ecuaciones diferenciales ordinarias similar a (2.6) y (2.7), y por tanto con la siguiente estructura por bloques,

$$\begin{bmatrix} v_1^{-1} & 0 \\ 0 & v_2^{-1} \end{bmatrix} \dot{\psi} + \begin{bmatrix} L_{11} & O \\ -L_{21} & L_{22} \end{bmatrix} \psi = (1 - \beta) \begin{bmatrix} M_{11} & M_{12} \\ O & O \end{bmatrix} \psi + \begin{bmatrix} \sum_{k=1}^{K_c} \lambda_k C_k \\ O \end{bmatrix}, \quad (2.38)$$

$$\dot{C}_k = \beta_k [M_{11} \ M_{12}] \psi - \lambda_k C_k, \quad (2.39)$$

2.4. Discretización temporal

donde $\psi^T = [\psi_{1n} \ \psi_{2n}]^T$, ψ_{1n} y ψ_{2n} son vectores de \mathbb{R}^n cuyas componentes son los coeficientes de Legendre del flujo neutrónico rápido y térmico respectivamente, con la ordenación derivada de la elección de los índices escogida. $\sum_{k=1}^{K_c} \lambda_k C_k$ es un vector de \mathbb{R}^n y corresponde a los precursores de neutrones en cada nodo. L_{ij} y M_{ij} son matrices cuadradas de dimensión $n \times n$. Además las matrices M_{11} , M_{12} y L_{21} son matrices diagonales cuyos elementos son positivos o nulos, siendo las matrices L_{11} y L_{22} simétricas y definidas positivas [49]. Si se utiliza el método de colocación nodal con la aproximación de serendipita, la dimensión de estos vectores es $n = \frac{1}{2}K(K+1)N$ para problemas con geometría bidimensional, y $n = \frac{1}{6}K(K+1)(K+2)N$ para problemas con una geometría tridimensional, donde N es el número de nodos en que se ha discretizado el problema [49].

2.4 Discretización temporal

Como resultado de discretizar la parte espacial de las ecuaciones (2.3) y (2.4) mediante el método de colocación nodal o el método en diferencias finitas centradas se obtiene el sistema de ecuaciones diferenciales ordinarias (2.38) y (2.39), que escribimos como

$$[v^{-1}]\dot{\psi} + L\psi = (1 - \beta)M\psi + X \sum_{k=1}^K \lambda_k C_k, \quad (2.40)$$

$$\dot{C}_k = \beta_k [M_{11} \ M_{12}] \psi - \lambda_k C_k, \quad (2.41)$$

donde ψ y C_k son vectores cuyas componentes son las incógnitas correspondientes al flujo neutrónico y a los precursores de neutrones en cada celda o nodo, respectivamente. Recordemos además que L , M , y $[v^{-1}]$ son matrices con la siguiente estructura por bloques (véase la ecuación (2.38))

$$L = \begin{bmatrix} L_{11} & 0 \\ -L_{21} & L_{22} \end{bmatrix}, \quad M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & 0 \end{bmatrix}, \quad [v^{-1}] = \begin{bmatrix} v_1^{-1} & 0 \\ 0 & v_2^{-1} \end{bmatrix}, \quad X = \begin{bmatrix} I \\ O \end{bmatrix}.$$

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Los métodos de discretización espacial presentados anteriormente dan lugar a bloques L_{11} y L_{22} simétricos definidos positivos y diagonal dominantes (véase teorema 17 y [49]).

Para integrar la ecuación (2.41) desde un instante de tiempo t_n al instante de tiempo t_{n+1} , se supone que $[M_{11} M_{12}] \psi$ varía linealmente entre estos instantes de tiempo, obteniéndose por tanto una aproximación de (2.41) mediante la ecuación [38][99],

$$\begin{aligned} \dot{C}_k &= -\lambda_k C_k + \beta_k [M_{11} M_{12}]^n \psi^n + \\ &+ \frac{\beta_k}{h_t} (t - t_n) \{ [M_{11} M_{12}]^{n+1} \psi^{n+1} - [M_{11} M_{12}]^n \psi^n \}, \end{aligned} \quad (2.42)$$

siendo $h_t = t_{n+1} - t_n$ el paso de integración temporal.

Integrando la ecuación (2.42), tenemos que la solución C_k en t_{n+1} se puede expresar como

$$C_k^{n+1} = C_k^n e^{-\lambda_k h_t} + \beta_k (a_k [M_{11} M_{12}]^n \psi^n + b_k [M_{11} M_{12}]^{n+1} \psi^{n+1}), \quad (2.43)$$

donde C_k^n es el valor de C_k en t_n y los coeficientes a_k y b_k vienen dados por

$$a_k = \frac{(1 + \lambda_k h_t)(1 - e^{-\lambda_k h_t})}{\lambda_k^2 h_t} - \frac{1}{\lambda_k}, \quad b_k = \frac{\lambda_k h_t - 1 + e^{-\lambda_k h_t}}{\lambda_k^2 h_t}.$$

Para integrar la ecuación (2.40) se tiene en cuenta que esta ecuación presenta problemas de rigidez debido principalmente a que los elementos de la matriz diagonal $[v^{-1}]$ son muy pequeños, esto hace que para su integración sea conveniente recurrir a los métodos en diferencias hacia atrás¹. Pasaremos seguidamente a describir brevemente en qué consisten estos métodos [41].

Dada una ecuación diferencial ordinaria de primer orden, de la forma

$$\dot{U}(t) = f(t, U(t)),$$

un método en diferencias hacia atrás general de m pasos para la resolución de esta ecuación, consiste en una ecuación en diferencias de la forma,

$$U(t_{n+1}) + \alpha_1 U(t_n) + \alpha_2 U(t_{n-1}) + \cdots + \alpha_m U(t_{n-(m-1)}) = h_t \beta_0 f(t_{n+1}, U(t_{n+1})), \quad (2.44)$$

¹del inglés *backward*.

2.4. Discretización temporal

m	β_0	α_1	α_2	α_3	α_4	error de truncamiento
1	1	-1				$O(h_t)$
2	2/3	-4/3	1/3			$O(h_t^2)$
3	6/11	-18/11	9/11	-2/11		$O(h_t^3)$
4	12/25	-48/25	36/25	-16/25	3/25	$O(h_t^4)$

Tabla 2.1: Coeficientes de los métodos en diferencias hacia atrás.

donde h_t es el paso de integración, $\beta_0 > 0$, y $\alpha_1, \dots, \alpha_m$ se eligen de forma que se minimice el error de truncamiento. En la tabla 2.1 se muestran posibles elecciones de los parámetros del método en diferencias hacia atrás para distintos valores de m . Con esta elección de los parámetros, los métodos en diferencias hacia atrás obtenidos son estables [41]. Los métodos en diferencias hacia atrás son métodos implícitos, y su utilización para la integración de un sistema de ecuaciones diferenciales implica la necesidad de resolver un sistema de ecuaciones lineales en cada paso de integración, pero es imposible construir un método explícito que funcione bien para el tratamiento de problemas con rigidez compatible con la utilización de pasos de tiempo razonables [41].

A continuación, se desarrollan métodos en diferencias hacia atrás de uno, dos y cuatro pasos para la integración de la ecuación (2.40).

2.4.1 Método en diferencias hacia atrás de 1 paso

Discretizando la ecuación (2.40) mediante un método en diferencias hacia atrás de un paso se obtiene,

$$\frac{1}{h_t}[v^{-1}](\psi^{n+1} - \psi^n) + L^{n+1}\psi^{n+1} = (1 - \beta)M^{n+1}\psi^{n+1} + X \sum_{k=1}^K \lambda_k C_k^{n+1}. \quad (2.45)$$

Sustituyendo la ecuación (2.43) en (2.45), y considerando la estructura de las matrices L y M , podemos expresar (2.45) como el siguiente sistema

Capítulo 2. Discretización de la ec. de la difusión neutrónica

de ecuaciones

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \psi_1^{n+1} \\ \psi_2^{n+1} \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{bmatrix} \psi_1^n \\ \psi_2^n \end{bmatrix} + \sum_{k=1}^{K_c} \lambda_k e^{-\lambda_k h_t} \begin{bmatrix} C_k^n \\ 0 \end{bmatrix}, \quad (2.46)$$

donde

$$\begin{aligned} T_{11} &= \frac{1}{h_t} [v_1^{-1}] + L_{11}^{n+1} - (1 - \beta) M_{11}^{n+1} - \sum_{k=1}^{K_c} \lambda_k \beta_k b_k M_{11}^{n+1}, \\ T_{12} &= -(1 - \beta) M_{12}^{n+1} - \sum_{k=1}^{K_c} \lambda_k \beta_k b_k M_{12}^{n+1}, \\ T_{21} &= -L_{21}^{n+1}, \\ T_{22} &= \frac{1}{h_t} [v_2^{-1}] + L_{22}^{n+1}, \\ R_{11} &= \frac{1}{h_t} [v_1^{-1}] + \sum_{k=1}^{K_c} \lambda_k \beta_k a_k M_{11}^n, \\ R_{12} &= \sum_{k=1}^{K_c} \lambda_k \beta_k a_k M_{12}^n, \\ R_{22} &= \frac{1}{h_t} [v_2^{-1}]. \end{aligned} \quad (2.47)$$

Como se muestra en la tabla 2.1, el método en diferencias de orden uno tiene asociado un error de truncamiento proporcional al paso de integración h_t . Esto implica que es necesario utilizar pasos de integración muy pequeños para no cometer un error muy grande en la resolución del problema. Por ello, también es conveniente utilizar métodos en diferencias hacia atrás de orden superior para la integración de la ecuación (2.40), ya que permiten el uso de pasos de integración mayores, al ser el error de truncamiento de estos métodos proporcional a potencias del paso de integración [41]. Mientras que en el método de un paso es posible ir variando el paso de integración en cada paso de tiempo, en los métodos de orden superior el paso de integración ha de permanecer constante. Por tanto, en la práctica también se utilizan algoritmos combinados de 1, 2 y 4 pasos que permiten la utilización de pasos de integración grandes [38].

2.4. Discretización temporal

2.4.2 Método en diferencias hacia atrás de 2 pasos

Discretizando la ecuación (2.40) mediante un método en diferencias hacia atrás de dos pasos de igual forma a como se ha hecho anteriormente, se obtiene la ecuación,

$$\begin{aligned} \frac{1}{h_t}[v^{-1}] \left(\psi^{n+1} - \frac{4}{3}\psi^n + \frac{1}{3}\psi^{n-1} \right) &= -\frac{2}{3}L^{n+1}\psi^{n+1} + \\ &+ \frac{2}{3} \left((1-\beta)M^{n+1}\psi^{n+1} + X \sum_{k=1}^K \lambda_k C_k^{n+1} \right), \end{aligned}$$

Teniendo en cuenta la ecuación (2.43), de la ecuación anterior se obtiene el sistema de ecuaciones lineales siguiente:

$$\begin{aligned} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \psi^{n+1} &= \begin{bmatrix} R_{11}^1 & R_{12}^1 \\ 0 & R_{22}^1 \end{bmatrix} \psi^n + \begin{bmatrix} R_{11}^2 & 0 \\ 0 & R_{22}^2 \end{bmatrix} \psi^{n-1} + \\ &+ \frac{2}{3} \sum_{k=1}^K \lambda_k e^{-\lambda_k h_t} \begin{bmatrix} C_k^n \\ 0 \end{bmatrix}, \end{aligned} \quad (2.48)$$

donde

$$\begin{aligned} T_{11} &= \frac{1}{h_t}v_1^{-1} + \frac{2}{3}L_{11}^{n+1} - \frac{2}{3}(1-\beta)M_{11}^{n+1} - \frac{2}{3} \sum_{k=1}^K \lambda_k \beta_k b_k M_{11}^{n+1}, \\ T_{12} &= -\frac{2}{3}(1-\beta)M_{12}^{n+1} - \frac{2}{3} \sum_{k=1}^K \lambda_k \beta_k b_k M_{12}^{n+1}, \\ T_{21} &= -\frac{2}{3}L_{21}^{n+1}, \\ T_{22} &= \frac{1}{h_t}v_2^{-1} + \frac{2}{3}L_{22}^{n+1}, \\ R_{11}^1 &= \frac{4}{3} \frac{1}{h_t}v_1^{-1} + \frac{2}{3} \sum_{k=1}^K \lambda_k a_k \beta_k M_{11}^n \end{aligned} \quad (2.49)$$

$$\begin{aligned}
 R_{12}^1 &= \frac{2}{3} \sum_{k=1}^K \lambda_k a_k \beta_k M_{12}^n, \\
 R_{22}^1 &= \frac{4}{3} \frac{1}{h_t} v_2^{-1}, \\
 R_{11}^2 &= -\frac{1}{3} \frac{1}{h_t} v_1^{-1}, \\
 R_{22}^2 &= -\frac{1}{3} \frac{1}{h_t} v_2^{-1}.
 \end{aligned}$$

Para utilizar el método de orden 2 hay que resolver el sistema (2.48) para cada paso de integración. Se observa que es necesario conocer el valor de la solución en dos pasos de tiempos anteriores, es decir, ψ^0 y ψ^1 , para lo cual se puede utilizar un método en diferencias hacia atrás de 1 paso.

2.4.3 Método en diferencias hacia atrás de 4 pasos

Discretizando mediante un método en diferencias hacia atrás de orden 4 la ecuación (2.40) se obtiene la ecuación,

$$\begin{aligned}
 \frac{1}{h_t} [v^{-1}] \left(\psi^{n+1} - \frac{48}{25} \psi^n + \frac{36}{25} \psi^{n-1} - \frac{16}{25} \psi^{n-2} + \frac{3}{25} \psi^{n-3} \right) = \\
 \frac{12}{25} \left(-L^{n+1} \psi^{n+1} + (1 - \beta) M^{n+1} \psi^{n+1} + X \sum_{k=1}^K \lambda_k C_k^{n+1} \right),
 \end{aligned}$$

que, utilizando la ecuación (2.43), se puede reagrupar en un sistema de la forma

$$\begin{aligned}
 \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \psi^{n+1} &= \begin{bmatrix} R_{11}^1 & R_{12}^1 \\ 0 & R_{22}^1 \end{bmatrix} \psi^n + \begin{bmatrix} R_{11}^2 & 0 \\ 0 & R_{22}^2 \end{bmatrix} \psi^{n-1} + \\
 &+ \begin{bmatrix} R_{11}^3 & 0 \\ 0 & R_{22}^3 \end{bmatrix} \psi^{n-2} + \begin{bmatrix} R_{11}^4 & 0 \\ 0 & R_{22}^4 \end{bmatrix} \psi^{n-3} + \\
 &+ \frac{12}{25} \sum_{k=1}^K \lambda_k e^{-\lambda_k h_t} \begin{bmatrix} C_k^n \\ 0 \end{bmatrix},
 \end{aligned} \tag{2.50}$$

2.4. Discretización temporal

donde

$$\begin{aligned}
T_{11} &= \frac{1}{h_t} v_1^{-1} + \frac{12}{25} L_{11}^{n+1} - \frac{12}{25} (1 - \beta) M_{11}^{n+1} - \frac{12}{25} \sum_{k=1}^K \lambda_k \beta_k b_k M_{11}^{n+1}, \\
T_{12} &= -\frac{12}{25} (1 - \beta) M_{12}^{n+1} - \frac{12}{25} \sum_{k=1}^K \lambda_k \beta_k b_k M_{12}^{n+1}, \\
T_{21} &= -\frac{12}{25} L_{21}^{n+1}, \\
T_{22} &= \frac{1}{h_t} v_2^{-1} + \frac{12}{25} L_{22}^{n+1}, \\
R_{11}^1 &= \frac{48}{25} \frac{1}{h_t} v_1^{-1} + \frac{12}{25} \sum_{k=1}^K \lambda_k a_k \beta_k M_{11}^n, \\
R_{12}^1 &= \frac{12}{25} \sum_{k=1}^K \lambda_k a_k \beta_k M_{12}^n, \\
R_{22}^1 &= \frac{48}{25} \frac{1}{h_t} v_2^{-1}, \\
R_{11}^2 &= -\frac{36}{25} \frac{1}{h_t} v_1^{-1}, \\
R_{22}^2 &= -\frac{36}{25} \frac{1}{h_t} v_2^{-1}, \\
R_{11}^3 &= \frac{16}{25} \frac{1}{h_t} v_1^{-1}, \\
R_{22}^3 &= \frac{16}{25} \frac{1}{h_t} v_2^{-1}, \\
R_{11}^4 &= -\frac{3}{25} \frac{1}{h_t} v_1^{-1}, \\
R_{22}^4 &= -\frac{3}{25} \frac{1}{h_t} v_2^{-1}.
\end{aligned} \tag{2.51}$$

Para utilizar el método en diferencias hacia atrás de orden 4, se ha de resolver el sistema (2.50) para cada paso de integración y es necesario partir del valor de la solución en cuatro puntos, ψ^0 , ψ^1 , ψ^2 y ψ^3 . Para ello se puede utilizar un algoritmo combinado de métodos en diferencias de 1, 2, y 4 pasos [38].

Los sistemas de ecuaciones lineales (2.46), (2.48) y (2.50) se pueden es-

Capítulo 2. Discretización de la ec. de la difusión neutrónica

cribir en forma más compacta como

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \psi_1^{n+1} \\ \psi_2^{n+1} \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix} , \quad (2.52)$$

donde E_1 y E_2 dependen del método en diferencias escogido para la discretización temporal. Este sistema es no simétrico, de gran tamaño y vacío. Por tanto, para su resolución será recomendable utilizar un método iterativo. Además, se ha de remarcar que, debido a la utilización del método de colocación nodal o métodos en diferencias para la discretización de la parte espacial de las ecuaciones, los bloques T_{11} , T_{12} , T_{21} , T_{22} son simétricos, mientras que la matriz del sistema considerada globalmente no posee esta propiedad. En general los bloques T_{11} y T_{22} son simétricos y definidos positivos para pasos de integración suficientemente pequeños.

A lo largo de la presente tesis se presentan diferentes algoritmos de resolución del sistema de ecuaciones (2.52) que tratan de aprovechar las especiales características de los bloques T_{11} y T_{22} . Preferentemente y por simplicidad, sin que ello reste generalidad al estudio realizado, en los experimentos numéricos se utilizará el método en diferencias hacia atrás de 1 paso para la discretización temporal.

2.5 Problemas modelo

A continuación, se presentan tres problemas modelo que se utilizarán en los experimentos numéricos para evaluar los métodos de resolución que se presentan en los distintos capítulos. Los dos primeros corresponden a transitorios en el reactor bidimensional denominado Seed-Blanket, y el tercero es un problema tridimensional conocido como transitorio de Langenbuch.

2.5.1 Transitorios bidimensionales

Como problemas con geometría bidimensional se estudiarán dos transitorios asociados a un reactor simplificado que denominaremos reactor Seed-

2.5. Problemas modelo

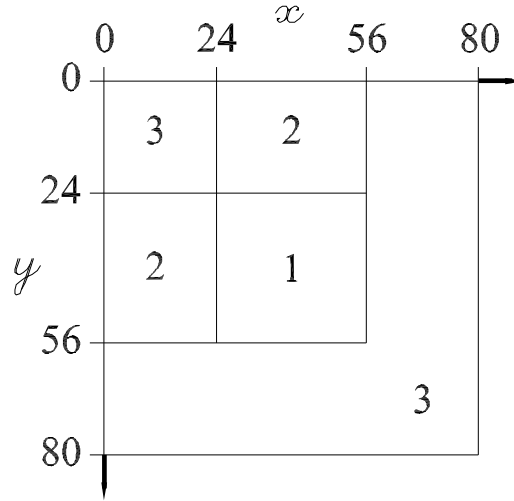


Figura 2.4: Cuadrante del reactor Seed-Blanket.

Blanket o TWIGL. Este reactor es similar al reactor utilizado en [49, 63, 92], y constituye un problema bidimensional ampliamente utilizado en la bibliografía. En este reactor se distinguen tres tipos de materiales cuya distribución en su interior se puede ver en la figura 2.4. En las tablas 2.2 y 2.3 se recogen las secciones eficaces y parámetros del grupo de precursores, respectivamente.

Cuando se aplica el método de colocación nodal para discretizar la parte espacial de la ecuación de la difusión neutrónica, el reactor se divide en 8×8 nodos de $10\text{ cm} \times 10\text{ cm}$ cada uno. Para el método en diferencias finitas centradas se escogen diferentes tamaños de malla, dependiendo de la precisión que se desea obtener con la discretización, y que se detallarán en su momento. En este reactor se estudian dos transitorios particulares. El primero es una perturbación en rampa, y el segundo es una mezcla de las perturbaciones de salto y de rampa, sin neutrones diferidos. Ambas perturbaciones se realizan sobre la sección eficaz de absorción del grupo térmico asociada al material 1 (figura 2.4).

Para el transitorio 1 se ha supuesto que \sum_{a2} responde a la función dada

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Región	Grupo	$D_g(cm)$	$\Sigma_{ag}(cm)^{-1}$	$\nu \Sigma_{fg}$	Σ_{12}
1	1	1,4	0,01	0,007	0,01
	2	0,4	0,15	0,2	
2	1	1,4	0,01	0,007	0,01
	2	0,4	0,15	0,2	
3	1	1,3	0,008	0,003	0,01
	2	0,5	0,05	0,06	

ν_1^{-1}	ν_2^{-1}
10^{-7}	10^{-5}

Tabla 2.2: Secciones eficaces de los materiales en el reactor Seed-Blanket.

β_1	λ_1
0,0064	0,08

Tabla 2.3: Parámetros de los precursores de neutrones en el reactor Seed-Blanket.

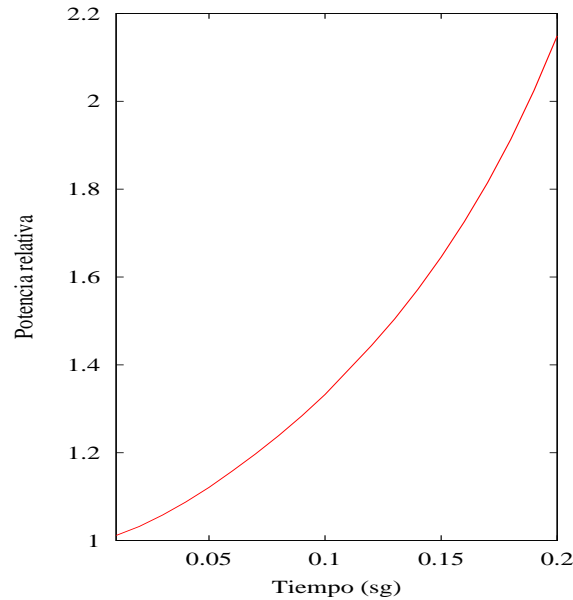


Figura 2.5: Evolución de la potencia relativa para el transitorio 1 en el reactor Seed-Blanket. Para la discretización espacial se ha utilizado el método de colocación nodal con 4 polinomios. Paso de integración de 1 milisegundo.

2.5. Problemas modelo

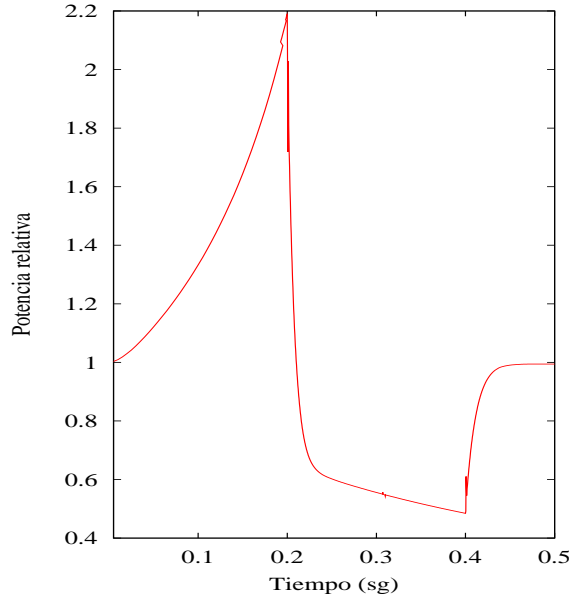


Figura 2.6: Evolución de la potencia relativa para el transitorio 2 en el reactor Seed-Blanket. Para la discretización espacial se ha utilizado el método de colocación nodal con 4 polinomios. Paso de integración de 1 milisegundo.

por

$$\sum_{a2}(t) = 0,15 - \frac{0,0035}{0,2}t, \quad 0 \leq t \leq 0,2. \quad (2.53)$$

La duración de este transitorio es de 0,2 segundos, y su curva de potencia relativa típica se puede ver en la figura 2.5 [34]. Para el transitorio 2 se ha elegido \sum_{a2} como la siguiente función dependiente del tiempo

$$\sum_{a2}(t) = \begin{cases} 0,15 - \frac{0,0035}{0,2}t & , \quad t \leq 0,2 \\ 0,15 + \frac{0,0035}{0,2}t & , \quad 0,2 < t \leq 0,4 \\ 0,15 & , \quad t > 0,4 \end{cases} \quad (2.54)$$

donde el tiempo, t , se mide en segundos. El tiempo de simulación de este transitorio es de 0,5 segundos. La evolución típica de potencia relativa para este transitorio se muestra en la figura 2.6 [34].

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Región	Grupo	D_g	$\Sigma_{ag}(cm^{-1})$	$\nu \Sigma_{fg}$	Σ_{12}
Fuel 1	1	1,423913	0,01040206	0,006477691	0,01755550
	2	0,3563060	0,08766217	0,1127328	
Fuel 2	1	1,425611	0,01099263	0,007503284	0,01717768
	2	0,3505740	0,09925634	0,1378004	
Absorbente	1	1,423913	0,01095206	0,006477691	0,01755550
	2	0,3563060	0,09146217	0,11273228	
Reflector	1	1,634227	0,002660573	0,0	0,0275963
	2	0,2640020	0,04936351	0,0	

Tabla 2.4: Secciones eficaces de los materiales en el reactor Langenbuch.

	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
β_i	0,000247	0,0013845	0,001222	0,0026455	0,000832	0,000169
$\lambda_i(s^{-1})$	0,0127	0,0317	0,115	0,311	1,4	3,87

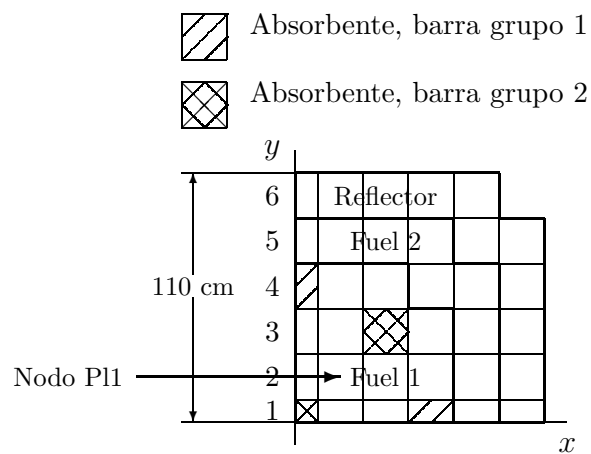
Tabla 2.5: Parámetros de los precursores de neutrones en el reactor Langenbuch.

2.5.2 Transitorio tridimensional de Langenbuch

Para comprobar los algoritmos de resolución numérica que se proponen en los siguientes capítulos en geometrías 3D, se utilizará el transitorio tridimensional de Langenbuch [63]. En la figura 2.7 se muestra la geometría del reactor modelizado. El transitorio que se estudia consiste en el deslizamiento del grupo de barras de control (1) con el absorbente desde el nivel 6 al nivel 10, y el desplazamiento del grupo de barras de control (2) desde el nivel 10 al nivel 4. El desplazamiento de las barras de control de tipo (1) desde su posición inicial a su posición final tiene lugar en el intervalo de tiempo de $t = 0$ segundos a $t = 26,6$ segundos, y las barras de control de tipo (2) se mueven desde el tiempo $t = 7,4$ segundos, a $t = 47,5$ segundos. El transitorio se sigue durante 60 segundos.

Las secciones eficaces de los materiales se muestran en la tabla 2.4. Para este transitorio se suponen 6 grupos de precursores de neutrones, cuyos parámetros se recogen en la tabla 2.5.

2.5. Problemas modelo



Sección transversal del reactor Langenbuch

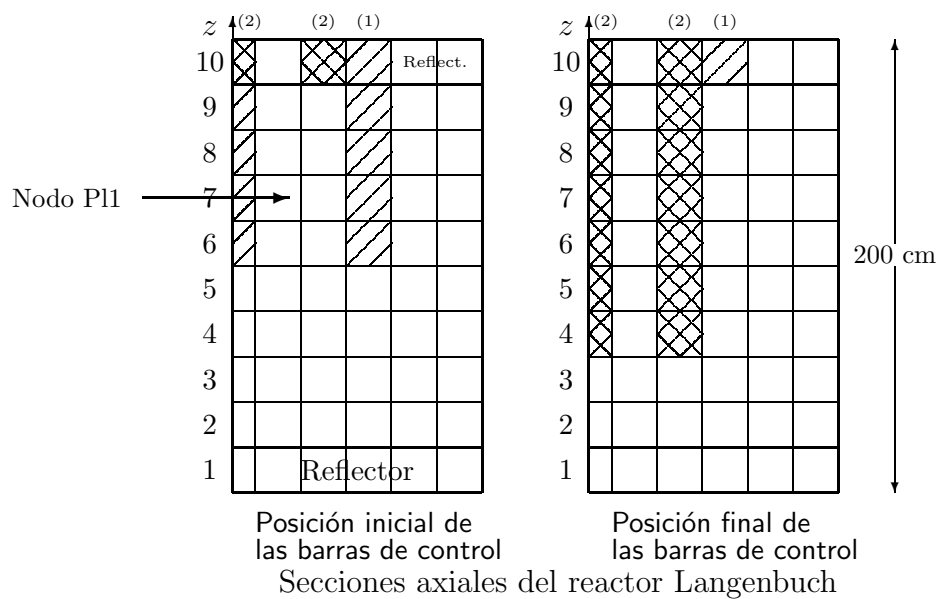


Figura 2.7: Geometría del reactor Langenbuch.

Capítulo 2. Discretización de la ec. de la difusión neutrónica

Todos los cálculos se han llevado a cabo haciendo uso de la simetría 1/4 del reactor, y se han utilizado 350 nodos para la discretización mediante el método de colocación nodal. La curva típica de potencia para este transitorio se muestra en la figura 2.8 [34]. Además, en la figura 2.7 se indica el nodo P11 que será utilizado para cuando sea necesario detallar la evolución local de la potencia.

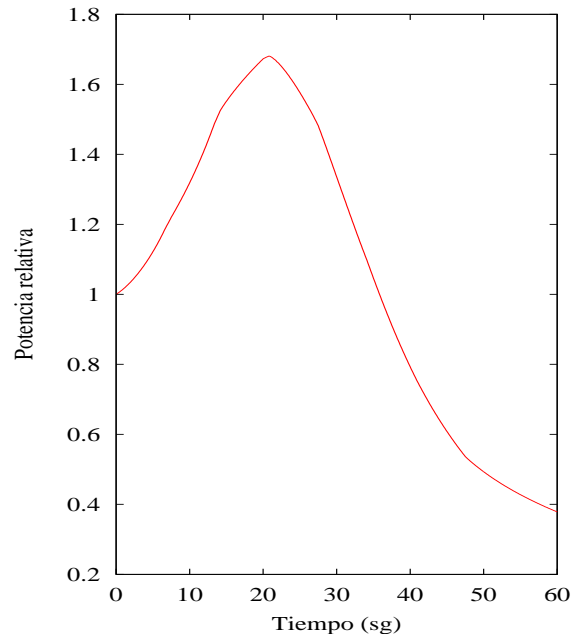


Figura 2.8: Evolución de la potencia relativa para el transitorio en el reactor Langenbuch. Obtenida con discretización espacial mediante el método de colocación nodal con 4 polinomios. Paso de integración de 125 milisegundos.

2.6 Experimentos numéricos

En esta sección se presentan los resultados de los experimentos numéricos encaminados a evaluar dos aspectos importantes. En primer lugar es interesante conocer cuál es el tamaño de la malla para la discretización en diferencias finitas centradas que proporciona una precisión en la solución final

2.6. Experimentos numéricos

próxima a la que se obtiene con el método de colocación nodal. Para ello se simulará el transitorio bidimensional 1 del TWIGL.

En segundo lugar, cuál es el efecto del número de polinomios de Legendre utilizados para la discretización mediante el método de colocación nodal en la precisión de la solución.

Para comparar las soluciones se evaluará la potencia global del reactor, relativa a la potencia del reactor en el estado estacionario, así como posibles efectos locales. El estado estacionario para los experimentos con el método de colocación nodal ha sido calculado, para ambos transitorios, mediante la resolución de la ecuación de la difusión neutrónica independiente del tiempo, que se ha estudiado en [34], [98] y [100]. Para la discretización en diferencias finitas se ha utilizado el método de la potencia.

Mientras que no se indique lo contrario, todos los códigos utilizados en la presente memoria para la obtención de los resultados de los experimentos numéricos han sido escritos en FORTRAN 77 y compilados con la opción de optimización -O3 sobre la máquina HP Exemplar S Class. Los diferentes preconditionadores y métodos de Krylov utilizados se han obtenido de las librerías ITPACK [58] y SPARSKIT [85].

2.6.1 Transitorio bidimensional

Para la resolución de los sistemas de ecuaciones en cada paso de tiempo se utiliza, en principio, el método BiCGSTAB preconditionado con ILU0 ². Como criterio de parada se ha utilizado el test de error dado por,

$$\|r_i\| \leq \|r_0\| \cdot rtol + atol ,$$

donde $r_i = b - Ax_i$ es el residuo, $rtol = 10^{-5}$ y $atol = 10^{-6}$ son la tolerancia relativa y absoluta, respectivamente. Para la discretización temporal se ha

²En posteriores capítulos se hará un estudio detallado de distintos métodos alternativos para la resolución de estos sistemas

Capítulo 2. Discretización de la ec. de la difusión neutrónica

<i>discretización espacial</i>	<i>n</i>	<i>nnz</i>
<i>Nodal</i> (2)	600	5360
<i>Nodal</i> (3)	1200	14960
<i>Nodal</i> (4)	2000	31920
<i>Diferencias</i> ($h = 4cm$)	3042	17940
<i>Diferencias</i> ($h = 3cm$)	5408	32032
<i>Diferencias</i> ($h = 2,5cm$)	7938	47124
<i>Diferencias</i> ($h = 1cm$)	50562	302100

Tabla 2.6: Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio 1 del TWIGL. n , nnz indican el tamaño y el número de elementos no nulos de las matrices a resolver en cada paso de tiempo, respectivamente.

utilizado un método en diferencias hacia atrás de un paso, con paso de tiempo de integración de 1,25 milisegundos.

En la tabla 2.6 se muestran los tamaños de las matrices y número de elementos no nulos de las diferentes discretizaciones utilizadas. Para la discretización en diferencias finitas se utiliza un tamaño de malla uniforme en las direcciones de los ejes x e y , es decir, $h_x = h_y$. Se indica en la tabla mediante h , en centímetros. Para el método de colocación nodal se ha utilizado un número de 2, 3 y 4 polinomios de Legendre. Esta discretización se denotará como *Nodal*(p), donde p indica el número de polinomios de Legendre utilizados en el desarrollo del flujo neutrónico en cada nodo.

En la figura 2.9 se muestra la curva de potencia del reactor relativa a la potencia del estado estacionario para el método de colocación nodal. Se observa como el aumento en el número de polinomios se traduce en un aumento en la precisión de la potencia pico final, es decir, en $t = 0,2$ segundos.

En la figura 2.10 se muestra la curva de potencia para el método en diferencias finitas con los diferentes tamaños de malla utilizados. Se observa como con mallas más finas la solución obtenida se aproxima a la curva correspondiente al método de colocación nodal con 4 polinomios de Legendre, siendo prácticamente similares para un tamaño de malla de 1 centímetro.

2.6. Experimentos numéricos

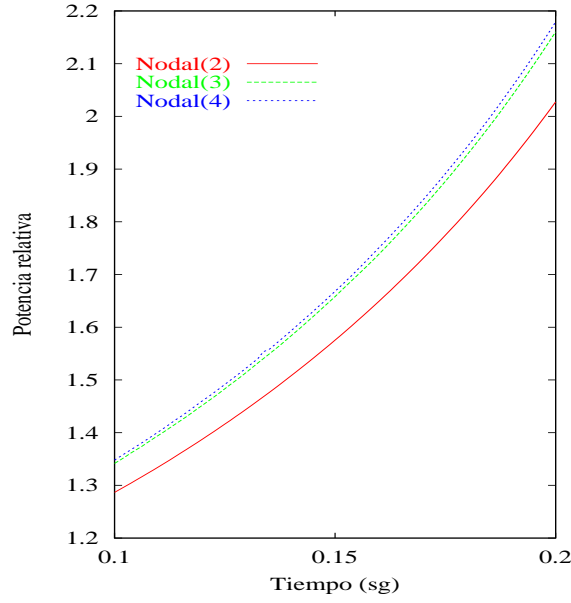


Figura 2.9: Evolución de la potencia relativa discretizando mediante el método de colocación nodal $Nodal(p)$ con $p = 2, 3, 4$.

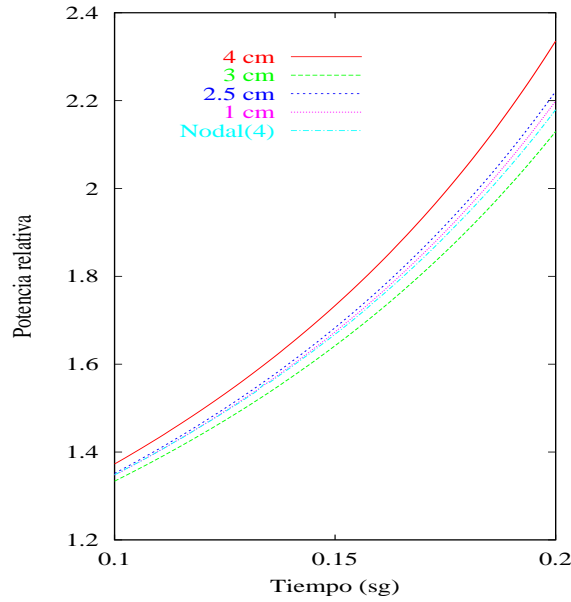


Figura 2.10: Evolución de la potencia relativa discretizando mediante el método en diferencias finitas. Se muestra también la curva para el método de colocación nodal, $Nodal(4)$.

Capítulo 2. Discretización de la ec. de la difusión neutrónica

La desventaja del método en diferencias en comparación con el método de colocación nodal es que el tamaño de las matrices a resolver y el número de elementos no nulos son muy elevados. En la tabla 2.6 se puede apreciar este hecho, por ejemplo, para tamaños de malla de 1 o 2.5 centímetros, tamaños que proporcionan las soluciones más precisas. Es importante mencionar que para un tamaño de malla de 3 centímetros se alcanza una precisión comparable a la que se obtiene con el método *Nodal(3)*, manteniendo un tamaño y número de elementos no nulos de las matrices próximos a los métodos nodales (tabla 2.6). Posteriormente, se utilizará en los experimentos numéricos este tamaño de discretización para el método en diferencias finitas centradas.

2.6.2 Transitorio de Langenbuch

De los resultados del apartado anterior se observa que para obtener una precisión comparable a la del método de colocación nodal con el método en diferencias finitas, el tamaño de discretización de la malla ha de ser pequeño, aumentando en consecuencia la dimensión y el número de elementos no nulos de las matrices a resolver en cada paso de tiempo de integración. Para el análisis del transitorio tridimensional de Langenbuch se utilizará exclusivamente el método de colocación nodal.

En esta sección se pretende evaluar el efecto del número de polinomios utilizados en el desarrollo del flujo neutrónico en la precisión de solución obtenida. Para ello se estudiará la evolución de la potencia relativa en el reactor. También se estudiarán posibles efectos locales de potencia. Concretamente, se evaluará la evolución de la potencia en el nodo P11 del reactor (figura 2.7).

Al igual que en el apartado anterior, para la resolución de los sistemas de ecuaciones en cada paso de tiempo se ha utilizado el método BiCGSTAB preconditionado con ILU0, con el mismo criterio parada. El número de polinomios de Legendre utilizados ha sido 2, 3 y 4. Para la discretización

2.6. Experimentos numéricos

<i>discretización espacial</i>	<i>n</i>	<i>nnz</i>
<i>Nodal(2)</i>	2800	31280
<i>Nodal(3)</i>	7000	106600
<i>Nodal(4)</i>	14000	270000

Tabla 2.7: Discretizaciones espaciales utilizadas para la simulación del transitorio de Langenbuch. n , nnz indican el tamaño y el número de elementos no nulos de las matrices a resolver en cada paso de tiempo, respectivamente.

temporal se ha utilizado un método en diferencias hacia atrás de un paso, con paso de tiempo de integración de 125 milisegundos.

Los tamaños de las matrices correspondientes a cada discretización espacial aparecen en la tabla 2.7.

En la figura 2.11 se muestra la curva de potencia relativa en un intervalo de tiempo del transitorio de Langenbuch próximo al pico de potencia. Se observa como el aumento en el número de polinomios se traduce en un aumento en la precisión, aunque para el transitorio de Langenbuch no existen diferencias muy significativas. En efecto, se observa como con 2 polinomios de Legendre la curva de potencia es muy próxima a las obtenidas con 3 y 4 polinomios.

Para mostrar la dependencia de la potencia local en un punto del interior del núcleo del reactor del número de polinomios de Legendre utilizados en la discretización con el método de colocación nodal, en la figura 2.12 se muestra la evolución de la potencia en el nodo P11 (ver figura 2.7). Como curva de referencia se ha tomado la curva de potencia obtenida con el código NEM, el módulo neutrónico del código ampliamente difundido TRAC/NEM [27, 6]. La figura muestra una dependencia más acusada de la evolución de la potencia local en el nodo P11 en función del número de polinomios de Legendre utilizado, que para la potencia global relativa a la potencia del estado estacionario. Claramente se obtiene mayor precisión espacial en la integración de la ecuación de la difusión neutrónica aumentando el número de polinomios. De esta manera, con 4 polinomios de Legendre se obtiene una

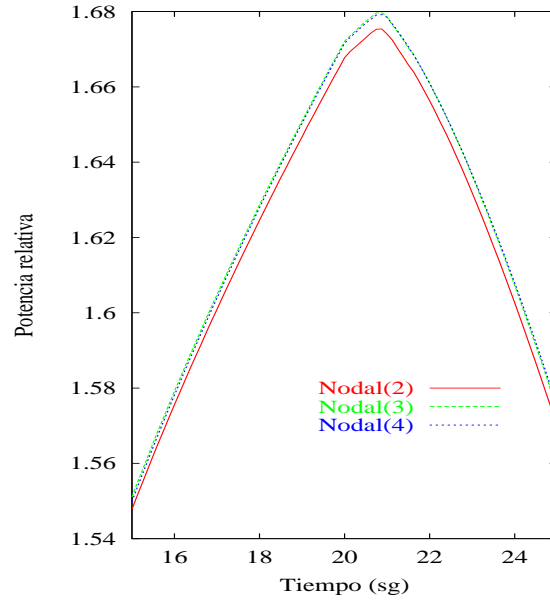


Figura 2.11: Evolución de la potencia relativa discretizando mediante el método de colocación nodal $Nodal(p)$ con $p = 2, 3, 4$.

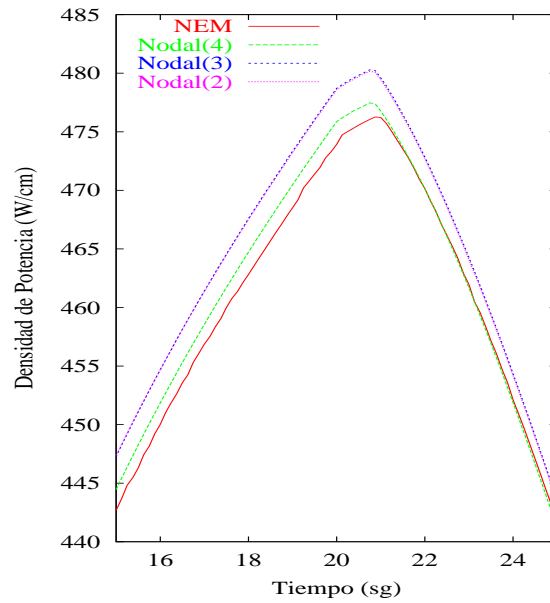


Figura 2.12: Evolución local de la potencia para el transitorio de Langenbuch en el nodo Pl1.

2.6. Experimentos numéricos

solución mucho más próxima a la curva de referencia proporcionada por el código NEM.

Por tanto, para transitorios donde los efectos locales de potencia son importantes es interesante integrar la ecuación de la difusión con suficiente precisión, es decir, discretizando con un número elevado de polinomios de Legendre si se utiliza el método de colocación nodal.

Capítulo 3

Esquema iterativo de segundo grado

3.1 Introducción

En el capítulo 2, se ha presentado la ecuación de la difusión neutrónica así como diferentes métodos para abordar su resolución numérica. La aplicación de estos métodos supone la resolución en cada paso de tiempo de un sistema de ecuaciones lineales $T\psi = e$ de la forma,

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}. \quad (3.1)$$

Este sistema de ecuaciones presenta una clara estructura por bloques siendo cada uno de los bloques T_{11} y T_{22} una matriz simétrica y definida positiva. Además, debido al gran tamaño y al carácter vacío de la matriz del sistema T en (3.1), es conveniente el empleo de esquemas iterativos que exploten al máximo las características de los bloques del sistema.

En este capítulo se estudian dos métodos iterativos que corresponden a los algoritmos 1 y 2. En el algoritmo 1, ω es un factor de relajación, similar al utilizado en el método SOR [44, 105]. A este algoritmo se le denominará *método iterativo de segundo grado A*. Para mejorar la velocidad de convergencia del método A, de forma similar a la obtención del método

3.1. Introducción

Algoritmo 1 Método iterativo por bloques de segundo grado A.

1. **Elegir** ψ_2^0
 2. **Resolver** $T_{11}\psi_1^1 = E_1 - T_{12}\psi_2^0$
 3. **Resolver** $T_{22}\psi_2^1 = E_2 - T_{21}\psi_1^1$
 4. **Para** $l = 1, 2, \dots$
 - (a) **Elegir** ω
 - (b) **Resolver** $T_{11}\psi_1^{l+1} = E_1 - T_{12}(\omega\psi_2^l + (1 - \omega)\psi_2^{l-1})$
 - (c) **Resolver** $T_{22}\psi_2^{l+1} = E_2 - T_{21}(\omega\psi_1^l + (1 - \omega)\psi_1^{l-1})$
 5. **Hasta** $\{\| \psi_1^{l+1} - \psi_1^l \| < tol \text{ y } \| \psi_2^{l+1} - \psi_2^l \| < tol\}$
-

de Gauss-Seidel a partir del método de Jacobi, si para el cálculo de ψ_2 se emplea la actualización más reciente de ψ_1 se obtiene el algoritmo 2. Al nuevo método se hará referencia en lo sucesivo como *método iterativo de segundo grado B*. El método B se ha utilizado con éxito en [99] y [38] para resolver el sistema de ecuaciones lineales (3.1). En ambos métodos se puede distinguir entre iteraciones externas, las que corresponden al bucle en el paso 4 de los algoritmos 1 y 2, e iteraciones internas, que corresponden a la solución de los sistemas de ecuaciones lineales con matrices T_{11} y T_{22} (pasos 4.(b) y 4.(c)). Debido a que cada una de estas matrices es simétrica y definida positiva es conveniente emplear el método del gradiente conjugado preconditionado. En este capítulo se estudiarán las propiedades de convergencia de los métodos iterativos de segundo grado A y B (algoritmos 1 y 2). Experimentalmente se mostrará que siempre se obtiene una convergencia más rápida del algoritmo 2, observándose en este sentido un comportamiento similar al ya conocido para los esquemas iterativos de Jacobi y Gauss-Seidel antes mencionados.

El capítulo está estructurado de la siguiente forma. En la sección 3.2 se revisan algunos conceptos y resultados básicos que se utilizarán posteriormente. En la sección 3.3 se definen los métodos iterativos de grado \hat{s} . En la

Capítulo 3. Esquema iterativo de segundo grado

Algoritmo 2 Método iterativo por bloques de segundo grado B.

1. **Elegir** ψ_2^0
 2. **Resolver** $T_{11}\psi_1^1 = E_1 - T_{12}\psi_2^0$
 3. **Resolver** $T_{22}\psi_2^1 = E_2 - T_{21}\psi_1^1$
 4. **Para** $l = 1, 2, \dots$
 - (a) **Elegir** ω
 - (b) **Resolver** $T_{11}\psi_1^{l+1} = E_1 - T_{12}(\omega\psi_2^l + (1 - \omega)\psi_2^{l-1})$
 - (c) **Resolver** $T_{22}\psi_2^{l+1} = E_2 - T_{21}(\omega\psi_1^{l+1} + (1 - \omega)\psi_1^l)$
 5. **Hasta** $\{\|\psi_1^{l+1} - \psi_1^l\| < tol \text{ y } \|\psi_2^{l+1} - \psi_2^l\| < tol\}$
-

sección 3.4 se particulariza el estudio para métodos de grado 2 o de segundo grado, y se estudian las propiedades de convergencia de estos métodos, aplicando los resultados a los métodos iterativos A y B en las secciones 3.4.1, 3.4.2 y 3.4.3. Finalmente, en la sección 3.5 se muestran los resultados de los experimentos numéricos realizados para la simulación del transitorio 1 del reactor bidimensional TWIGL y se dan algunas conclusiones.

3.2 Resultados y conceptos preliminares

En esta sección se revisan algunos conceptos y resultados básicos que serán de utilidad en el resto del capítulo. En general se trabajará con una matriz T invertible y dividida en $q \times q$ bloques de la forma

$$T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1q} \\ T_{21} & T_{22} & \cdots & T_{2q} \\ \vdots & \vdots & & \vdots \\ T_{q1} & T_{q2} & \cdots & T_{qq} \end{bmatrix}, \quad (3.2)$$

donde los bloques diagonales $T_{i,i}$, $1 \leq i \leq q$, son matrices cuadradas e invertibles de dimensión n_i , $i = 1, \dots, q$ y $\sum_{i=1}^q n_i = n$. Por tanto, la matriz

3.2. Resultados y conceptos preliminares

diagonal por bloques

$$D = \begin{bmatrix} T_{11} & O & \cdots & O \\ O & T_{22} & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & T_{qq} \end{bmatrix},$$

es invertible. Además, se puede escribir T como la suma de matrices

$$T = D - E - F,$$

con

$$E = \begin{bmatrix} O & O & \cdots & O \\ -T_{21} & O & \cdots & O \\ \vdots & \ddots & & \vdots \\ -T_{q1} & \cdots & -T_{qq-1} & O \end{bmatrix}, \quad F = \begin{bmatrix} O & -T_{12} & \cdots & T_{1q} \\ O & O & \ddots & \vdots \\ \vdots & \vdots & \ddots & -T_{q-1,q} \\ O & \cdots & O & O \end{bmatrix},$$

las partes estrictamente triangular inferior y superior a bloques de $-T$, respectivamente. Los métodos iterativos básicos descritos en el capítulo 1, sección 1.4.1, se pueden extender a matrices divididas por bloques fácilmente. Ya que D es invertible, se definen las siguientes matrices de iteración para los métodos de Jacobi, Gauss-Seidel, Gauss-Seidel acelerado (AGS) y SOR por bloques:

- Jacobi:

$$B = D^{-1}(E + F) = L + U. \quad (3.3)$$

- Gauss-Seidel:

$$\mathcal{L} = (D - E)^{-1}F = (I - L)^{-1}U, \quad (3.4)$$

- Gauss-Seidel acelerado:

$$\mathcal{L}_{AGS} = (D - \omega E)^{-1}((1 - \omega)E + F) = (I - \omega L)^{-1}((1 - \omega)L + \omega U), \quad (3.5)$$

- SOR:

$$\mathcal{L}_\omega = (D - \omega E)^{-1}((1 - \omega)D + \omega F) = (I - \omega L)^{-1}((1 - \omega)I + \omega U) , \quad (3.6)$$

donde $L = D^{-1}E$ y $U = D^{-1}F$.

Definición 18 Una matriz T de tamaño $n \times n$ se dice que es débilmente cíclica de índice p ($p > 1$) (débilmente p -cíclica) si existe una matriz de permutación π , tal que

$$\pi T \pi^T = \begin{bmatrix} O & O & \cdots & T_{1p} \\ T_{21} & O & \cdots & O \\ \vdots & \ddots & \ddots & \vdots \\ O & O & T_{pp-1} & O \end{bmatrix} ,$$

donde las matrices nulas de la diagonal son cuadradas.

Definición 19 Si la matriz de Jacobi por bloques asociada a la matriz T en (3.2) es débilmente cíclica de índice p ($p > 1$), entonces T es cíclica de índice p (p -cíclica) con respecto a la división por bloques (3.2).

Se observa que si $q = 2$ en (3.2), la matriz de Jacobi por bloques asociada es débilmente cíclica de índice 2 y, por tanto, T es 2-cíclica. Para matrices débilmente cíclicas se tiene el siguiente resultado debido a Romanovsky (1936),

Teorema 20 [97, teorema 2.4] Sea T una matriz débilmente cíclica de índice p . Entonces,

$$\det(tI - T) = t^m \prod_{i=1}^r (t^p - \sigma_i^p) ,$$

donde T tiene m valores propios iguales a cero, y rp valores propios distintos de cero, σ_i .

3.2. Resultados y conceptos preliminares

Del teorema anterior se tiene que si σ_i es un valor propio de T distinto de cero, entonces también lo son todas las raíces de la ecuación $t^p - \sigma_i^p = 0$.

Definición 21 [97, definición 4.2] Si la matriz T en (3.2) es p -cíclica, entonces se dice que T es una matriz consistentemente ordenada si todos los valores propios de la matriz

$$B(\alpha) = \alpha L + \alpha^{-(p-1)}U ,$$

que se obtiene de la matriz asociada de Jacobi $B = L + U$, son independientes de α , para $\alpha \neq 0$. En tal caso, también se dice que B es consistentemente ordenada.

El siguiente resultado relaciona los valores propios de la matriz de iteración del método SOR y el método de Jacobi.

Teorema 22 [97, teorema 4.3] Sea la matriz T como en (3.2) p -cíclica y consistentemente ordenada, con bloques diagonales invertibles. Si $\omega \neq 0$ y λ es un valor propio distinto de cero de la matriz \mathcal{L}_ω dada en (3.6), y si μ satisface la ecuación

$$(\lambda + \omega - 1)^p = \lambda^{p-1} \omega^p \mu^p , \quad (3.7)$$

entonces μ es un valor propio de la matriz del método de Jacobi B en (3.3). De la misma forma, si μ es un valor propio de B y λ satisface (3.7), entonces λ es un valor propio de \mathcal{L}_ω .

Corolario 23 [97] Sea la matriz T como en (3.2) p -cíclica y consistentemente ordenada, con bloques diagonales invertibles. Si μ es un valor propio de la matriz de Jacobi por bloques B en (3.3), entonces μ^p es un valor propio de la matriz de Gauss-Seidel \mathcal{L} en (3.4). Si λ es un valor propio distinto de cero de \mathcal{L} y $\mu^p = \lambda$, entonces μ es un valor propio de B .

3.3 Métodos iterativos de grado \hat{s}

En esta sección se presenta una forma general de expresar un método iterativo lineal de grado \hat{s} completamente consistente. Nuestro problema es encontrar la solución de un sistema de ecuaciones lineales de la forma

$$T\psi = e, \quad (3.8)$$

donde $T = [t_{i,j}]$, $i, j = 1, \dots, n$ es una matriz invertible, vacía y de gran tamaño, dividida en bloques como en (3.2). En general, un método iterativo para resolver (3.8) se puede definir como un conjunto de funciones $\phi_0(T, e)$, $\phi_1(\psi^{(0)}; T, e), \dots, \phi_n(\psi^{(0)}, \psi^{(1)}, \dots, \psi^{(n-1)}; T, e)$, donde la secuencia de vectores iterados $\psi^{(0)}, \psi^{(1)}, \dots, \psi^{(n)}$ se define de la forma

$$\begin{aligned} \psi^{(0)} &= \phi_0(T, e) \\ &\vdots \\ \psi^{(n)} &= \phi_n(\psi^{(0)}, \psi^{(1)}, \dots, \psi^{(n-1)}; T, e). \end{aligned}$$

El método iterativo se dice que es estacionario de grado \hat{s} si para algún $\hat{s} > 0$, ϕ_n es independiente de n , es decir, para todo $n \geq \hat{s}$, $\psi^{(n)}$ depende de $\psi^{(n-1)}, \psi^{(n-2)}, \dots, \psi^{(n-\hat{s})}$, pero no de $\psi^{(k)}$ para $k < n - \hat{s}$. De esta forma se puede representar un método iterativo estacionario lineal de grado \hat{s} de la forma

$$\psi^{(n)} = \sum_{i=1}^{\hat{s}} G_{\hat{s}-i} \psi^{(n-i)} + k, \quad (3.9)$$

donde $G_{\hat{s}-i}$ y k son funciones lineales de T y e . Considérese el sistema de ecuaciones lineales relacionado con el esquema anterior dado por,

$$(I - \sum_{i=1}^{\hat{s}} G_{\hat{s}-i})\psi = (I - H)\psi = k, \quad (3.10)$$

donde $H = \sum_{i=1}^{\hat{s}} G_{\hat{s}-i}$. Los siguientes teoremas relacionan el conjunto de soluciones de (3.8) con el conjunto de soluciones de (3.10).

3.3. Métodos iterativos de grado \hat{s}

Definición 24 [105, pág. 64] Sea $S(T, e)$ el conjunto de soluciones de (3.8), y sea $S(I - H, k)$ el conjunto de soluciones de (3.10). El esquema iterativo (3.9) se dice que es consistente con (3.8) si $S(T, e) \subseteq S(I - H, k)$, y completamente consistente si $S(T, e) = S(I - H, k)$.

La definición 24 implica que si un método iterativo es completamente consistente, en el caso de que converja lo hace a una solución del sistema.

Teorema 25 [105, teorema 3.2.2] Si T en (3.8) es invertible, entonces el esquema iterativo (3.9) es consistente con el sistema (3.8) si, y sólo si

$$k = (I - H)T^{-1}e.$$

Teorema 26 [105, teorema 3.2.4] Si T en (3.8) es invertible, entonces el esquema iterativo (3.9) es completamente consistente con el sistema (3.8) si, y sólo si es consistente y además $I - H$ es invertible.

Teniendo en cuenta los teoremas 25 y 26, el método iterativo de grado \hat{s} dado en (3.9) es completamente consistente con (3.8) si

$$k = (I - \sum_{i=1}^{\hat{s}} G_{\hat{s}-i})T^{-1}e,$$

y la matriz $I - \sum_{i=1}^{\hat{s}} G_{\hat{s}-i}$ es invertible, es decir, $1 \notin \sigma(\sum_{i=1}^{\hat{s}} G_{\hat{s}-i})$. Una forma de garantizar la invertibilidad de esta matriz es obtener el método de grado \hat{s} a partir de un método de primer grado convergente (definición 11, pág. 9). Dada una partición de la matriz $T = M - N$ con M invertible, se obtiene la matriz $G = M^{-1}N$ asociada a la partición. En el caso de que $\rho(G) < 1$, es decir, la partición sea convergente, el siguiente resultado nos proporciona una forma de tomar las matrices $G_{\hat{s}-i}$ para que el esquema iterativo de grado \hat{s} en (3.9) sea completamente consistente.

Capítulo 3. Esquema iterativo de segundo grado

Lema 27 Sea $T = M - N$ una partición de la matriz T del sistema (3.8), y sea la matriz $G = M^{-1}N$ tal que $\rho(G) < 1$. Sean las matrices $G_{\hat{s}-i}$ de la forma

$$G_{\hat{s}-i} = GP_{\hat{s}-i}, \quad i = 1, \dots, \hat{s}, \quad (3.11)$$

con matrices $P_{\hat{s}-i}$ tales que $\sum_{i=1}^{\hat{s}} P_{\hat{s}-i} = I$. El esquema iterativo de grado \hat{s} (3.9) es completamente consistente con (3.8) si, y sólo si

$$k = M^{-1}e.$$

Demostración. Que la matriz $(I - \sum_{i=\hat{s}-i}^{\hat{s}} G_{\hat{s}-i}) = (I - G)$ es invertible es obvio. Si el método iterativo es completamente consistente, entonces se tiene que $k = (I - G)T^{-1}e = M^{-1}e$, donde se ha hecho uso de la igualdad $(I - G) = (I - M^{-1}N) = M^{-1}(M - N) = M^{-1}T$. La condición suficiente es obvia ya que $k = M^{-1}e = (I - G)T^{-1}e$ y por el teorema 25 el esquema (3.9) es consistente con (3.8). ■

Corolario 28 Sea $T = M - N$ una partición de la matriz T del sistema (3.8), y sea la matriz $G = M^{-1}N$ tal que $\rho(G) < 1$. Si $k = M^{-1}e$ y las matrices $P_{\hat{s}-i}$ en (3.11) se toman de la forma $P_{\hat{s}-i} = \alpha_{\hat{s}-i}I$, donde $\alpha_{\hat{s}-i} \in \mathbb{R}$, $i = 1, \dots, \hat{s}$, y $\sum_{i=1}^{\hat{s}} \alpha_{\hat{s}-i} = 1$, entonces el esquema iterativo de grado \hat{s} (3.9) es completamente consistente con (3.8).

Por tanto el esquema iterativo de grado \hat{s} se puede obtener a partir de un esquema iterativo de primer grado convergente, como son los métodos definidos en el capítulo 1, sección 1.4.1. Como se verá posteriormente, el método iterativo A corresponde a un método de segundo grado obtenido a partir de la matriz de iteración del método de Jacobi por bloques aplicado al sistema de ecuaciones (3.1). Por otra parte, el método iterativo B se obtendrá a partir de la matriz de iteración del método de Gauss-Seidel acelerado (AGS).

3.4 Métodos iterativos de segundo grado

La expresión para un método iterativo estacionario de segundo grado se obtiene de la ecuación (3.9) tomando $\hat{s} = 2$ de la forma,

$$\psi^{(n+1)} = G_1 \psi^{(n)} + G_0 \psi^{(n-1)} + k. \quad (3.12)$$

El estudio de los métodos de segundo grado aparece en [40]. En [105] se realiza un estudio detallado de los mismos. En [2], los autores estudian el esquema (3.12) bajo ciertas condiciones. En particular, concluyen que si la matriz $G = G_1 + G_0 \geq O$, el esquema iterativo (3.12) no converge más rápidamente que el método de primer orden asociado a la matriz G . Además, estudian el caso particular en el que $G_1 = \omega G$ y $G_0 = (1 - \omega)I$ concluyendo que bajo ciertas suposiciones adicionales de la matriz G , (irreducible y 2-cíclica) el esquema

$$\psi^{(n+1)} = \omega G \psi^{(n)} + (1 - \omega) \psi^{(n-1)} + k', \quad (3.13)$$

obtiene una velocidad de convergencia superior al método extrapolado de primer orden dado por,

$$\psi^{(n+1)} = \omega G \psi^{(n)} + (1 - \omega) \psi^{(n)} + k'.$$

Resaltar que en [70] se propone un algoritmo paralelo basado en una versión caótica del esquema de segundo grado (3.13).

Considérese una partición de la matriz T de la forma $T = M - N$ y la matriz de iteración asociada

$$G = M^{-1}N.$$

Tomando las matrices G_0 , G_1 y k de la forma

$$G_1 = \omega G, \quad G_0 = (1 - \omega)G, \quad k = M^{-1}e, \quad (3.14)$$

Capítulo 3. Esquema iterativo de segundo grado

se obtiene el método iterativo de segundo grado dado por,

$$\psi^{(n+1)} = G(\omega\psi^{(n)} + (1 - \omega)\psi^{(n-1)}) + k. \quad (3.15)$$

Por tanto, el estudio que se realiza en esta sección asume una forma de las matrices G_0 y G_1 diferentes a las utilizadas en [2]. Además, el análisis está basado en los valores propios de la matriz de iteración G .

Asumiendo que $\rho(G) < 1$, por el corolario 28, este método es completamente consistente (basta con tomar $P_0 = (1 - \omega)I$ y $P_1 = \omega I$). Para el estudio de los métodos de segundo grado habitualmente se utiliza el método auxiliar de primer grado siguiente ([40]),

$$\begin{bmatrix} \psi^{(n)} \\ \psi^{(n+1)} \end{bmatrix} = \begin{bmatrix} 0 & I \\ G_0 & G_1 \end{bmatrix} \begin{bmatrix} \psi^{(n-1)} \\ \psi^{(n)} \end{bmatrix} + \begin{bmatrix} 0 \\ k \end{bmatrix}. \quad (3.16)$$

El método (3.16) es convergente para todo $\psi^{(0)}$ y $\psi^{(1)}$ si $\rho(\hat{G}_\omega) < 1$, donde

$$\hat{G}_\omega = \begin{bmatrix} 0 & I \\ G_0 & G_1 \end{bmatrix}, \quad (3.17)$$

es la matriz de iteración del esquema (3.16). Además, $\rho(\hat{G}_\omega) < 1$ si todos los valores propios de \hat{G}_ω son menores que uno en módulo. Denotando por λ los valores propios de \hat{G}_ω , éstos se obtienen como las raíces de la ecuación

$$\det(\lambda^2 I - \lambda G_1 - G_0) = 0. \quad (3.18)$$

Sea μ un valor propio de G . Utilizando la ecuación (3.18) y la expresión para G_0 y G_1 en (3.14) se observa que los valores propios de la matriz $\lambda\omega G + (1 - \omega)G = (\lambda\omega + (1 - \omega))G$ son de la forma λ^2 , es decir $(\lambda\omega + (1 - \omega))\mu = \lambda^2$. Por tanto, los valores propios de \hat{G}_ω están relacionados con los valores propios de G mediante la ecuación cuadrática

$$\lambda^2 - \omega\mu\lambda + (\omega - 1)\mu = 0. \quad (3.19)$$

A continuación se define el radio de las raíces de una ecuación cuadrática. Este concepto se utilizará más adelante para determinar el valor del radio espectral de \hat{G}_ω .

3.4. Métodos iterativos de segundo grado

Definición 29 [105, pág. 176] Se llama radio de las raíces de la ecuación cuadrática $x^2 - bx + c = 0$ al máximo de los módulos de sus raíces. Se denota por $\rho(b, c)$.

El siguiente lema concerniente a las raíces de la ecuación cuadrática $x^2 - bx + c = 0$ se empleará posteriormente.

Lema 30 [105, lema 6.2.1]

Si b y c son reales, las dos raíces de la ecuación cuadrática $x^2 - bx + c = 0$ son menores que uno en módulo si, y sólo si

$$|c| < 1 \text{ y } |b| < 1 + c.$$

De esta forma, si el lema 30 se cumple entonces $\rho(b, c) < 1$. Los siguientes resultados se tienen para el caso en el que todos los valores propios de G sean reales.

Teorema 31 Sea G una matriz con valores propios reales y positivos, y sea \hat{G}_ω la matriz de la ecuación (3.17). Entonces $\rho(\hat{G}_\omega) < 1$ si, y sólo si

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{\bar{\mu}}, \quad \text{y} \quad \bar{\mu} < 1,$$

donde $\bar{\mu}$ es el radio espectral de G .

Demostración. Del lema 30 se sigue que el radio de las raíces de la ecuación (3.19) es menor que 1 en módulo si, y sólo si se cumplen las siguientes condiciones para cualquier valor de μ :

1. $|c| = |(\omega - 1)\mu| < 1$, y
2. $|b| = |\omega\mu| < 1 + c$.

Capítulo 3. Esquema iterativo de segundo grado

Si $\rho(\hat{G}_\omega) < 1$, supongamos el caso $\mu = \bar{\mu}$. De la condición 1 se deducen las siguientes equivalencias,

$$|(\omega - 1)\bar{\mu}| < 1 \quad \text{si, y sólo si} \quad |\omega - 1|\bar{\mu} < 1 \quad \text{si, y sólo si} \quad |\omega - 1| < \frac{1}{\bar{\mu}}$$

$$\text{si, y sólo si} \quad 1 - \frac{1}{\bar{\mu}} < \omega < 1 + \frac{1}{\bar{\mu}} \quad .$$

De la condición 2 se sigue

$$-1 - (\omega - 1)\bar{\mu} < \omega\bar{\mu} < 1 + (\omega - 1)\bar{\mu} .$$

Entonces, de la segunda desigualdad se tiene,

$$\omega\bar{\mu} < 1 + (\omega - 1)\bar{\mu} \quad \text{si, y sólo si} \quad \bar{\mu} < 1 ,$$

y de la primera desigualdad se deduce,

$$-1 - (\omega - 1)\bar{\mu} < \omega\bar{\mu} \quad \text{si, y sólo si} \quad -2\omega\bar{\mu} < 1 - \bar{\mu} ,$$

y por tanto,

$$\omega > \frac{\bar{\mu} - 1}{2\bar{\mu}} > 1 - \frac{1}{\bar{\mu}} .$$

De esta manera, el rango de valores para ω queda acotado como sigue,

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < 1 + \frac{1}{\bar{\mu}} .$$

Para demostrar la condición suficiente, del rango de valores para ω se obtiene,

$$\frac{-(1 + \bar{\mu})}{2} < \omega - 1 < 1 ,$$

que implica,

$$-1 < \frac{-(1 + \bar{\mu})}{2} < (\omega - 1)\bar{\mu} < 1 ,$$

3.4. Métodos iterativos de segundo grado

y por tanto,

$$|c| = |(\omega - 1)\mu| \leq |(\omega - 1)\bar{\mu}| < 1 .$$

Por otro lado, para ver que $|b| < 1 + c$ consideremos dos casos:

1. Si $b > 0$, entonces

$$1 + c - |b| = 1 + c - b = 1 - \mu > 0 .$$

2. Si $b < 0$, entonces

$$1 + c - |b| = 1 + c + b = 1 - \mu + 2\omega\mu > 0 .$$

De esta forma $|c| < 1$ y $|b| < 1 + c$, y por el lema 30 se concluye que $\rho(\hat{G}_\omega) < 1$. ■

Si los valores propios de G son reales, no necesariamente positivos, el rango válido para el factor de relajación ω queda restringido como establece el siguiente resultado.

Teorema 32 *Sea G una matriz con valores propios reales, y sea \hat{G}_ω la matriz de la ecuación (3.17). Entonces $\rho(\hat{G}_\omega) < 1$ si, y sólo si*

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{2\bar{\mu}} \quad , \quad y \quad \bar{\mu} < 1 ,$$

donde $\bar{\mu}$ es el radio espectral de G .

Demostración. Se procede de la misma forma que en el teorema 31. Del lema 30 se sigue que el radio de las raíces de la ecuación (3.19) es menor que 1 en módulo si, y sólo si se cumplen las siguientes condiciones para cualquier valor de μ :

Capítulo 3. Esquema iterativo de segundo grado

1. $|c| = |(\omega - 1)\mu| < 1$, y
2. $|b| = |\omega\mu| < 1 + c$.

Supongamos que $\rho(\hat{G}_\omega) < 1$. De la condición 1 se deduce que $|(\omega - 1)\mu| < 1$ para cualquier valor de μ . En particular tomemos el valor propio μ_r , tal que $|\mu_r| = \bar{\mu}$. Se tienen las siguientes equivalencias,

$$\begin{aligned} |(\omega - 1)\mu_r| < 1 & \quad \text{si, y sólo si} \quad |\omega - 1|\bar{\mu} < 1 \\ & \quad \text{si, y sólo si} \quad |\omega - 1| < \frac{1}{\bar{\mu}} \\ & \quad \text{si, y sólo si} \quad 1 - \frac{1}{\bar{\mu}} < \omega < 1 + \frac{1}{\bar{\mu}} . \end{aligned}$$

De la condición 2 se sigue

$$-1 - (\omega - 1)\mu_r < \omega\bar{\mu} < 1 + (\omega - 1)\mu_r . \quad (3.20)$$

Entonces, de la segunda desigualdad en (3.20) se tiene,

$$\omega\bar{\mu} < 1 + (\omega - 1)\mu_r .$$

Se pueden considerar dos casos:

(a) Si $\mu_r = -\bar{\mu}$ se sigue que

$$\omega\bar{\mu} < 1 - \omega\bar{\mu} + \bar{\mu} \quad \text{si, y sólo si} \quad \omega < \frac{\bar{\mu} + 1}{2\bar{\mu}} .$$

(b) Si $\mu_r = \bar{\mu}$ se obtiene

$$\omega\bar{\mu} < 1 + \omega\bar{\mu} - \bar{\mu} \quad \text{si, y sólo si} \quad \bar{\mu} < 1 .$$

Por otro lado, de la primera desigualdad en (3.20) se deduce,

$$-1 - (\omega - 1)\mu_r < \omega\bar{\mu} .$$

3.4. Métodos iterativos de segundo grado

Nuevamente se consideran dos casos:

(a) Si $\mu_r = -\bar{\mu}$ se sigue que

$$-1 + \omega\bar{\mu} - \bar{\mu} < \omega\bar{\mu} \quad \text{si, y sólo si} \quad \bar{\mu} > -1 .$$

(b) Si $\mu_r = \bar{\mu}$ se obtiene

$$-1 - \omega\bar{\mu} + \bar{\mu} < \omega\bar{\mu} \quad \text{si, y sólo si} \quad 2\omega\bar{\mu} > \bar{\mu} - 1 \quad \text{si, y sólo si} \quad \omega > \frac{\bar{\mu} - 1}{2\bar{\mu}} .$$

De esta manera, el rango para ω queda acotado como sigue,

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{2\bar{\mu}} . \quad (3.21)$$

Para demostrar la condición suficiente, de la expresión del rango para ω y $\bar{\mu} < 1$ se obtiene,

$$\frac{\bar{\mu} - 1}{\bar{\mu}} < \frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{2\bar{\mu}} < \frac{\bar{\mu} + 1}{\bar{\mu}} .$$

Por tanto,

$$-1 < (\omega - 1)\bar{\mu} < 1 \quad \text{y} \quad |(\omega - 1)\bar{\mu}| < 1 .$$

Ya que $|\mu| \leq \bar{\mu}$ se tiene que $|c| = |(\omega - 1)\mu| < 1$.

Por otro lado, para ver que $|b| < 1 + c$, consideremos dos casos:

1. $b > 0$,

$$1 + c - |b| = 1 + c - b = 1 - \mu > 0 .$$

Capítulo 3. Esquema iterativo de segundo grado

2. $b < 0$,

$$1 + c - |b| = 1 + c + b = 1 - \mu + 2\omega\mu . \quad (3.22)$$

Para evaluar la expresión anterior, fijemos un valor cualquiera de ω en el intervalo (3.21). La ecuación (3.22) es una función lineal para valores de μ en el intervalo $[-\bar{\mu}, \bar{\mu}]$. Para $\mu = \bar{\mu}$ se tiene

$$2\omega\bar{\mu} > \bar{\mu} - 1 \quad \text{si, y sólo si} \quad 1 - \bar{\mu} + 2\omega\bar{\mu} > 0 .$$

Para $\mu = -\bar{\mu}$ se tiene de la misma forma,

$$-2\omega\bar{\mu} < -(\bar{\mu} - 1) \quad \text{si, y sólo si} \quad 1 - \bar{\mu} + 2\omega\bar{\mu} > 0 .$$

Por tanto una función lineal, que toma valores positivos en los extremos del dominio de definición, es mayor que cero en todo el dominio. De esta forma se ha comprobado que $|c| < 1$ y $|b| < 1 + c$, y por el lema 30 se concluye que $\rho(\hat{G}_\omega) < 1$. ■

El siguiente corolario establece la convergencia, asumiendo que la matriz del sistema T es 2-cíclica (definición 19, página 71), cuando la matriz de iteración G corresponde a la matriz de los métodos de Jacobi o Gauss-Seidel por bloques (ecuaciones (3.3) y (3.4)).

Corolario 33 *Sea T una matriz 2-cíclica particionada en bloques como en (3.2), con bloques diagonales invertibles T_{ii} , $1 \leq i \leq q$, y sea B la matriz de iteración asociada al método de Jacobi por bloques, de forma que los valores propios de B^2 son reales, no negativos, y además el radio espectral es $\rho(B) < 1$. Sea \mathcal{L} la matriz de iteración asociada al método de Gauss-Seidel*

3.4. Métodos iterativos de segundo grado

por bloques. Sea \hat{G}_ω la matriz de la ecuación (3.17) obtenida con $G = B$ o con $G = \mathcal{L}$, respectivamente. Entonces $\rho(\hat{G}_\omega) < 1$ si, y sólo si

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < w < \frac{\bar{\mu} + 1}{\bar{\mu}} \quad \text{con} \quad \bar{\mu} = \rho(G), \quad G = \mathcal{L}$$

o

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < w < \frac{\bar{\mu} + 1}{2\bar{\mu}} \quad \text{con} \quad \bar{\mu} = \rho(G), \quad G = B.$$

Demostración. Si T es una matriz 2-cíclica (definición 19), la matriz asociada al método iterativo de Jacobi, B , es débilmente cíclica de índice 2 (definición 18). Ya que los valores propios de B^2 son reales y no negativos, si μ_i es un valor propio de B diferente de 0 entonces $-\mu_i$ también lo es (teorema 20). Por tanto

$$-\rho(B) \leq \mu_i \leq \rho(B) < 1.$$

Por otra parte, para el método iterativo de Gauss-Seidel se tiene que los valores propios también son reales y se encuentran en el intervalo $0 \leq \mu_i \leq \rho(B)^2 < 1$ (corolario 23). Entonces, de los teoremas 32 y 31 se concluye la demostración. ■

Los teoremas 31 y 32 establecen el rango para el factor de relajación donde se garantiza que el método de segundo grado converge. El problema que queremos resolver ahora es determinar el factor de relajación, ω_b , que maximice la velocidad de convergencia del método iterativo. El siguiente resultado nos da la respuesta para el caso en el que la matriz G , a partir de la cual se obtiene el método de segundo grado, tiene valores propios reales y positivos. Es decir, proporciona el valor óptimo, ω_b , de ω que minimiza $\rho(\hat{G}_\omega)$. Además da una expresión del radio espectral $\rho(\hat{G}_{\omega_b})$.

Teorema 34 Sea G una matriz con valores propios reales y positivos de forma que $\bar{\mu} = \rho(G) < 1$, y sea \hat{G}_ω la matriz de la ecuación (3.17). Si ω_b se

Capítulo 3. Esquema iterativo de segundo grado

define como

$$\omega_b = \frac{2}{1 + \sqrt{1 - \bar{\mu}}} \quad (3.23)$$

entonces,

$$\rho(\hat{G}_{\omega_b}) = \frac{\omega_b \bar{\mu}}{2} = \sqrt{(\omega_b - 1)\bar{\mu}} \quad (3.24)$$

y si $\omega \neq \omega_b$, entonces

$$\rho(\hat{G}_\omega) > \rho(\hat{G}_{\omega_b}).$$

Además, para cualquier valor de ω en el rango $\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{\bar{\mu}}$, se tiene

$$\rho(\hat{G}_\omega) = \begin{cases} \left| \frac{\omega \bar{\mu} + \sqrt{(\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu})}}{2} \right|, & \text{si } \frac{\bar{\mu} - 1}{2\bar{\mu}} \leq \omega < \omega_b \\ \sqrt{(\omega - 1)\bar{\mu}}, & \text{si } \omega_b \leq \omega < \frac{\bar{\mu} + 1}{\bar{\mu}} \end{cases} \quad (3.25)$$

Finalmente, si $\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \omega_b$, entonces $\rho(\hat{G}_\omega)$ es una función estrictamente decreciente en ω .

Demostración. Primero daremos una prueba puramente analítica. Más adelante se dará una interpretación geométrica, que a su vez sirve como prueba alternativa.

En primer lugar hacemos notar que ω_b está determinado de forma única por las condiciones

$$\omega^2 \bar{\mu}^2 = 4(\omega_b - 1)\bar{\mu}, \quad 1 \leq \omega_b < 2.$$

Esto se debe a que ω_b es una de las raíces de la ecuación cuadrática $\omega^2 \bar{\mu}^2 = 4(\omega - 1)\bar{\mu}$, que está comprendida en el intervalo $1 \leq \omega_b < 2$. Además, el producto de las dos raíces de la ecuación anterior es $\frac{4}{\bar{\mu}}$, mayor que 4 ya que $\bar{\mu} < 1$, por lo que necesariamente la segunda raíz no puede estar en el mismo intervalo. Estudiaremos el intervalo para ω dado por

$$\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega < \frac{\bar{\mu} + 1}{\bar{\mu}},$$

3.4. Métodos iterativos de segundo grado

ya que fuera de él se tiene $\rho(\hat{G}_\omega) > 1$ (teorema 31). Definamos la función

$$\Gamma(\omega, \mu) = \left| \frac{\omega\mu + \sqrt{\omega^2\mu^2 - 4(\omega - 1)\mu}}{2} \right|.$$

$\Gamma(\omega, \mu)$ es el máximo de los módulos de los valores propios λ de \hat{G}_ω , que satisfacen la ecuación cuadrática (3.19) para un valor propio μ de G . Probemos primero el siguiente lema.

Lema 35 *Bajo las hipótesis del teorema 34, se tiene que*

$$\rho(\hat{G}_\omega) = \Gamma(\omega, \bar{\mu}) , \quad (3.26)$$

y la ecuación (3.25) se satisface.

Demostración. Primero veamos que

$$\max_{0 \leq \mu \leq \bar{\mu}} \Gamma(\omega, \mu) = \Gamma(\omega, \bar{\mu}) .$$

Si $\omega^2\bar{\mu}^2 - 4(\omega - 1)\bar{\mu} < 0$, entonces $\omega > 1$ y además $\omega^2\mu^2 - 4(\omega - 1)\mu < 0$, para $\mu \leq \bar{\mu}$. En este caso,

$$\Gamma(\omega, \mu) = \frac{\sqrt{\omega^2\mu^2 + (4(\omega - 1)\mu - \omega^2\mu^2)}}{2} = \sqrt{(\omega - 1)\mu} ,$$

que es una función creciente en μ . Si por el contrario, $\omega^2\bar{\mu}^2 - 4(\omega - 1)\bar{\mu} \geq 0$, definamos μ_c mediante la expresión,

$$\mu_c = \begin{cases} \frac{4(\omega-1)}{\omega^2}, & \omega \geq 1 \\ 0, & \omega < 1 \end{cases}$$

Claramente μ_c es una cota para los posibles valores de μ que hacen positivo el término $\omega^2\mu^2 - 4(\omega - 1)\mu$. Si $\mu \leq \mu_c$, entonces $\mu^2 \leq \frac{4(\omega - 1)}{\omega^2}\mu$, y por tanto se tiene

$$\Gamma(\omega, \mu) = \sqrt{(\omega - 1)\mu} .$$

Capítulo 3. Esquema iterativo de segundo grado

Si $\mu_c \leq \mu \leq \bar{\mu}$, entonces $\mu^2 \geq \frac{4(\omega-1)}{\omega^2}\mu$, y se tiene

$$\Gamma(\omega, \mu) = \frac{\omega\mu + \sqrt{\omega^2\mu^2 - 4(\omega-1)\mu}}{2},$$

que es una función creciente en μ . Por tanto $\Gamma(\omega, \mu) \leq \Gamma(\omega, \bar{\mu})$, para $\mu \leq \bar{\mu}$. Esto demuestra (3.26) ya que $\bar{\mu}$ es un valor propio de G y $\Gamma(\omega, \bar{\mu})$ nos da el máximo valor en módulo que puede tomar el valor propio λ de \hat{G}_ω asociado a $\bar{\mu}$.

Para ver que se cumple (3.25) dividamos el intervalo en tres tramos, esto es,

$$\frac{\bar{\mu}-1}{2\bar{\mu}} < 0 \leq \omega \leq 2 < \frac{\bar{\mu}+1}{\bar{\mu}}.$$

Consideremos el intervalo $0 < \omega < 2$. En este intervalo la función $\frac{4(\omega-1)}{\omega^2}$ es una función creciente en ω ya que

$$\frac{d}{d\omega} \left(\frac{4(\omega-1)}{\omega^2} \right) = \frac{2}{\omega^3}(2-\omega) > 0.$$

Ya que $\frac{4(\omega_b-1)}{\omega_b^2} = \bar{\mu}$, se sigue que si $\omega \geq \omega_b$, entonces $\bar{\mu} \leq \frac{4(\omega-1)}{\omega^2}$. Por tanto $\rho(\hat{G}_\omega) = \sqrt{(\omega-1)\bar{\mu}}$ para $2 > \omega \geq \omega_b$. Por otra parte, si $0 < \omega \leq \omega_b$, entonces $\bar{\mu} \geq \frac{4(\omega-1)}{\omega^2}$, y se tiene

$$\Gamma(\omega, \mu) = \frac{\omega\mu + \sqrt{\omega^2\mu^2 - 4(\omega-1)\mu}}{2}.$$

Consideremos ahora el intervalo $\frac{\bar{\mu}-1}{2\bar{\mu}} < \omega \leq 0$. La función $\omega^2\bar{\mu}^2 - 4(\omega-1)\bar{\mu}$ es positiva en los extremos del intervalo. Además es una función monótona (decreciente) en dicho intervalo, como fácilmente se obtiene al estudiar su derivada

$$\frac{d}{d\omega}(\omega^2\bar{\mu}^2 - 4(\omega-1)\bar{\mu}) = 2\omega\bar{\mu} - 4\bar{\mu}.$$

3.4. Métodos iterativos de segundo grado

Por tanto $\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu} \geq 0$ en todo el intervalo, y

$$\Gamma(\omega, \bar{\mu}) = \frac{\omega \bar{\mu} + \sqrt{\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu}}}{2}.$$

Finalmente, en el intervalo $2 \leq \omega < \frac{\bar{\mu} + 1}{\bar{\mu}}$ la función $\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu}$ es negativa como se comprueba fácilmente al observar que es negativa en los extremos, y además es una función monótona (decreciente) en dicho intervalo. Por tanto

$$\Gamma(\omega, \bar{\mu}) = \sqrt{(\omega - 1)\bar{\mu}},$$

y el lema queda demostrado. ■

Finalmente hay que comprobar que $\rho(\hat{G}_\omega)$ es estrictamente decreciente en $\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega \leq \omega_b$. En todo este intervalo, se tiene

$$\Gamma(\omega, \bar{\mu}) = \frac{\omega \bar{\mu} + \sqrt{\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu}}}{2}.$$

Además, de nuevo es fácil comprobar que la función $\omega^2 \bar{\mu}^2 - 4(\omega - 1)\bar{\mu}$ es decreciente en ese intervalo y, por tanto,

$$\Gamma(\omega, \bar{\mu}) \geq \Gamma(\omega_b, \bar{\mu}),$$

en $\frac{\bar{\mu} - 1}{2\bar{\mu}} < \omega \leq \omega_b$ completando la demostración. ■

A continuación se da una interpretación geométrica similar a la utilizada en [97, teorema 4.4]. De la ecuación (3.19), definiendo las funciones $m(\lambda) = \lambda^2$, $g(\lambda) = \omega\mu\lambda + (1 - \omega)\mu$, tenemos la igualdad $m(\lambda) = g(\lambda)$. De esta forma (3.19) puede interpretarse geométricamente como la intersección de ambas curvas, como se muestra en la figura 3.1. Para un valor determinado de μ , $g(\lambda)$ es una línea recta que pasa por el punto $(1, \mu)$, y cuya pendiente aumenta monótonamente al incrementar ω . El valor de la abcisa del punto

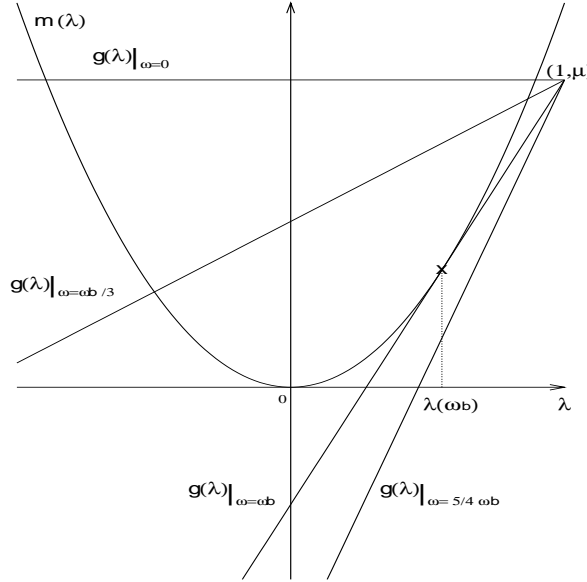


Figura 3.1: Gráfica de las ecuaciones $m(\lambda) = \lambda^2$ y $g(\lambda) = \omega\mu\lambda + (1 - \omega)\mu$, para $\mu = \bar{\mu} = 0,7995$ y diferentes valores de ω .

de intersección más alejado del origen decrece al incrementar ω hasta que $g(\lambda)$ es tangente a la curva $m(\lambda)$. El punto de tangencia se da para

$$\tilde{\omega} = \frac{2}{1 + \sqrt{1 - \mu}}, \quad (3.27)$$

es decir, cuando se tiene una raíz doble. En este punto el valor de λ es $\frac{\tilde{\omega}\mu}{2} = +\sqrt{(\tilde{\omega} - 1)\mu}$. Para $\omega > \tilde{\omega}$ el polinomio en la ecuación (3.19) tiene dos raíces complejas conjugadas de módulo $+\sqrt{(\omega - 1)\mu}$, que aumenta en módulo al incrementar ω . Por tanto, para este valor propio fijo μ de G , el valor de ω que minimiza la raíz de mayor módulo es $\tilde{\omega}$. Además, claramente el máximo valor para el módulo de λ se obtiene para $\mu = \bar{\mu} = \rho(G)$, concluyéndose

$$\min_{\omega} \rho(\hat{G}_{\omega}) = \rho(\hat{G}_{\omega_b}) = \frac{\omega_b \bar{\mu}}{2} \quad (3.28)$$

donde ω_b se define en la ecuación (3.23).

Utilizando el resultado del teorema 34, podemos establecer la velocidad de convergencia del método de segundo grado (3.15). Teniendo en cuenta la

3.4. Métodos iterativos de segundo grado

definición 15, la velocidad asintótica de convergencia para el método iterativo de segundo grado es

$$R(\hat{G}_\omega) = -\log(\rho(\hat{G}_\omega)) \quad (3.29)$$

donde \hat{G}_ω es la matriz de iteración. Si para formular el método de segundo grado, elegimos la matriz de iteración del método iterativo de Gauss-Seidel por bloques (ecuación (3.4)), se tiene el siguiente resultado.

Teorema 36 *Sea T una matriz 2-cíclica invertible con submatrices diagonales invertibles $T_{i,i}$, $1 \leq i \leq q$, y sea la matriz B correspondiente al método iterativo de Jacobi por bloques, de forma que los valores propios de B^2 son no negativos y $\rho(B) < 1$. Sea $R(\mathcal{L})$ y $R(\hat{G}_{\omega_b})$ las velocidades de convergencia de los métodos de Gauss-Seidel y segundo grado (3.15) respectivamente, donde el factor de relajación ω_b viene dado por (3.23). Si $G = \mathcal{L}$ se tiene*

$$\rho(B)R(\mathcal{L})^{1/2} \leq R(\hat{G}_{\omega_b}) \leq R(\mathcal{L})^{1/2} \left(1 + R(\mathcal{L})^{1/2}\right), \quad (3.30)$$

con la segunda desigualdad cierta si $R(\mathcal{L}) \leq 3$. Además,

$$\lim_{\rho(B) \rightarrow 1-} \left[\frac{R(\hat{G}_{\omega_b})}{R(\mathcal{L})^{1/2}} \right] = 1. \quad (3.31)$$

Demostración. Sea $\beta = -\log(\bar{\mu})$, $\bar{\mu} = \rho(B)$, donde B es la matriz de Jacobi por bloques asociada a T . Sea $R(\hat{G}_{\omega_b})$ la velocidad de convergencia del método iterativo (3.16) con el factor de extrapolación óptimo (3.23) y con G correspondiente al método iterativo de Gauss-Seidel. El radio espectral del método de Gauss-Seidel es $\bar{\mu}^2$. En estas condiciones se tienen las siguientes igualdades que usaremos posteriormente,

$$\rho(\hat{G}_{\omega_b}) = \frac{\bar{\mu}^2}{1 + \sqrt{1 - \bar{\mu}^2}} = 1 - \sqrt{1 - \bar{\mu}^2}.$$

Según se ha definido en (3.29),

$$R(\hat{G}_{\omega_b}) = -\log(\rho(\hat{G}_{\omega_b})) = -(1 - \log(\sqrt{1 - \bar{\mu}^2})).$$

Capítulo 3. Esquema iterativo de segundo grado

Aplicando el desarrollo de Taylor a la función $-\log(1-x)$ para x comprendido entre 0 y 1, se obtiene que $-\log(1-x) \geq x$. Por tanto,

$$-\log(1 - \sqrt{1 - \bar{\mu}^2}) \geq \sqrt{1 - \bar{\mu}^2} = \sqrt{1 - e^{-2\beta}} = e^{-\beta} \sqrt{e^{2\beta} - 1}.$$

Aplicando que $e^x - 1 \geq x$ para $x \geq 0$, se obtiene finalmente

$$e^{-\beta} \sqrt{e^{2\beta} - 1} \geq e^{-\beta} \sqrt{2\beta} = \bar{\mu} R(\mathcal{L})^{1/2}.$$

Por otra parte,

$$\begin{aligned} -\log(\rho(\hat{G}_{\omega_b})) &= -2\log(\bar{\mu}) + \log(1 + \sqrt{1 - \bar{\mu}^2}) \\ &= R(\mathcal{L})^{1/2} + \log(1 + \sqrt{1 - e^{-2\beta}}) \\ &\leq R(\mathcal{L})^{1/2} + \sqrt{1 - e^{-2\beta}} \\ &\leq R(\mathcal{L})^{1/2} + \sqrt{1 - (1 - 2\beta)} \\ &= R(\mathcal{L})^{1/2} + \sqrt{2\beta} \\ &= R(\mathcal{L})^{1/2}(1 + R(\mathcal{L})^{1/2}), \end{aligned}$$

donde en la segunda desigualdad se ha utilizado $e^{-y} \geq 1 - y$, para $0 < y \leq 3$.

Por tanto,

$$\bar{\mu} R(\mathcal{L})^{1/2} \leq R(\hat{G}_{\omega_b}) \leq R(\mathcal{L})^{1/2}(1 + R(\mathcal{L})^{1/2}).$$

Por último, se comprueba fácilmente que

$$\lim_{\bar{\mu} \rightarrow 1^-} \frac{R(\hat{G}_{\omega_b})}{R(\mathcal{L})^{1/2}} = 1,$$

concluyendo la demostración. ■

Ejemplo 37 *Considérese $\rho(B) = 0,9995$. En este caso se tiene las velocidades de convergencia de los métodos de Jacobi, Gauss-Seidel y sobrerelajación (SOR) son respectivamente*

$$R(B) = 0,0005, \quad R(\mathcal{L}) = 0,0010, \quad y \quad R(SOR) = 0,0628$$

3.4. Métodos iterativos de segundo grado

mientras que $R(\hat{G}_{\omega_b}) = 0,0319$ tomando $G = \mathcal{L}$. De esta manera, para velocidades de convergencia bajas de los métodos iterativos de Jacobi y Gauss-Seidel por bloques, el método iterativo de segundo grado con el parámetro de relajación óptimo ω_b proporciona una mejora sustancial, y del mismo orden que el método SOR óptimo.

En el caso en el que la matriz G tiene valores propios complejos, el estudio de la convergencia del método de segundo grado se realiza de forma diferente. Recordamos que la matriz \hat{G}_ω del esquema iterativo auxiliar (ecuación (3.16)) utilizado para estudiar el método de segundo grado (3.15) viene dada por

$$\hat{G}_\omega = \begin{bmatrix} 0 & I \\ (1-\omega)G & \omega G \end{bmatrix}. \quad (3.32)$$

Para calcular el radio espectral utilizaremos el cuadrado de la matriz que viene dado por

$$\hat{G}_\omega^2 = \begin{bmatrix} (1-\omega)G & \omega G \\ (1-\omega)\omega G^2 & (1-\omega)G + \omega^2 G^2 \end{bmatrix}. \quad (3.33)$$

Hacemos notar que dada una matriz A en $\mathbb{R}^{n \times n}$, si para algún entero $k > 0$ se tiene $\|A^k\| < 1$, entonces $\rho(A) < 1$, como se deduce del siguiente resultado.

Lema 38 [74, corolario 1.3.9] Sea $A \in \mathbb{R}^{n \times n}$. Entonces $\lim_{k \rightarrow \infty} A^k = 0$ si, y sólo si $\rho(A) < 1$. Además, $\|A^k\|$ está acotada cuando $k \rightarrow \infty$ si, y sólo si $\rho(A) < 1$.

De esta forma, si $\|\hat{G}_\omega^2\| < 1$ para alguna norma matricial, entonces el método de segundo grado (3.12) converge. El siguiente lema define una norma matricial que será utilizada en lo sucesivo. Se asume que, dada una matriz T de orden $qn \times qn$, está particionada en $q \times q$ bloques cuadrados de tamaño $n \times n$ de la forma mostrada en (3.2).

Capítulo 3. Esquema iterativo de segundo grado

Lema 39 *Sea una norma matricial consistente $\|\cdot\|_\alpha$ en el conjunto de las matrices de orden $n \times n$, y sea una matriz T de orden $qn \times qn$ particionada como en (3.2). Si se define*

$$\|T\| = \max_{1 \leq i \leq q} \sum_{j=1}^q \|T_{ij}\|_\alpha, \quad (3.34)$$

entonces, $\|\cdot\|$ es una norma matricial consistente sobre el conjunto de las matrices de orden $qn \times qn$.

Demostración. Para la demostración de este lema se comprobará que la función definida en (3.34) verifica los cinco axiomas que definen una norma matricial (ver definición 3, pág. 6).

(1) La no negatividad de la norma se deduce directamente de la no negatividad de la norma $\|\cdot\|_\alpha$.

(2) Si $\|T\| = 0$, entonces $\sum_{j=1}^q \|T_{ij}\|_\alpha = 0$ para $i = 1, \dots, q$, y por tanto $\|T_{ij}\|_\alpha = 0$ para $i, j = 1, \dots, q$. Como $\|\cdot\|_\alpha$ es una norma matricial, se tiene $T_{ij} = O$ para todo $i, j = 1, \dots, q$.

(3) $\|cT\| = \max_{1 \leq i \leq q} \sum_{j=1}^q \|cT_{ij}\|_\alpha = |c| \max_{1 \leq i \leq q} \sum_{j=1}^q \|T_{ij}\|_\alpha = |c| \cdot \|T\|$.

(4) Sean A y B dos matrices de tamaño $qn \times qn$. Entonces

$$\begin{aligned} \|A + B\| &= \max_{1 \leq i \leq q} \sum_{j=1}^q \|A_{ij} + B_{ij}\|_\alpha \leq \max_{1 \leq i \leq q} \sum_{j=1}^q (\|A_{ij}\|_\alpha + \|B_{ij}\|_\alpha) \\ &\leq \max_{1 \leq i \leq q} \sum_{j=1}^q \|A_{ij}\|_\alpha + \max_{1 \leq i \leq q} \sum_{j=1}^q \|B_{ij}\|_\alpha = \|A\| + \|B\|. \end{aligned}$$

(5) Por último, para comprobar la propiedad submultiplicativa, supongamos que el máximo se alcanza en la primera fila de bloques

$$\|B\| = \max_{1 \leq i \leq q} \sum_{j=1}^q \|B_{ij}\|_\alpha = \sum_{j=1}^q \|B_{1j}\|_\alpha.$$

3.4. Métodos iterativos de segundo grado

Entonces,

$$\begin{aligned}
\|AB\| &= \max_{1 \leq i \leq q} \sum_{j=1}^q \left\| \sum_{k=1}^q A_{ik} B_{kj} \right\|_{\alpha} \leq \max_{1 \leq i \leq q} \sum_{j=1}^q \sum_{k=1}^q \|A_{ik} B_{kj}\|_{\alpha} \\
&\leq \max_{1 \leq i \leq q} \sum_{j=1}^q \sum_{k=1}^q \|A_{ik}\|_{\alpha} \cdot \|B_{kj}\|_{\alpha} = \max_{1 \leq i \leq q} \sum_{k=1}^q \|A_{ik}\|_{\alpha} \sum_{j=1}^q \|B_{kj}\|_{\alpha} \\
&\leq \sum_{j=1}^q \|B_{1j}\|_{\alpha} \max_{1 \leq i \leq q} \sum_{k=1}^q \|A_{ik}\|_{\alpha} = \|B\| \cdot \|A\|.
\end{aligned}$$

El otro caso se razonaría de igual forma. ■

Como caso particular para $q = 2$ se tiene

$$\|T\| = \max_{1 \leq i \leq 2} \sum_{j=1}^2 \|T_{ij}\|_{\alpha}, \quad (3.35)$$

que es una norma en el conjunto de las matrices de tamaño $2n \times 2n$, y que fue utilizada en [70].

La norma (3.35) de la matriz dada en (3.33), que se denotará por $\|\hat{G}_{\omega}^2\|$, responde a la expresión,

$$\max \left\{ \|(1 - \omega)G\|_{\alpha} + \|\omega G\|_{\alpha}, \|(1 - \omega)\omega G^2\|_{\alpha} + \|(1 - \omega)G + \omega^2 G^2\|_{\alpha} \right\}. \quad (3.36)$$

Del cálculo de la norma anterior se obtiene el siguiente resultado de convergencia del método de segundo grado (ecuación (3.15)).

Teorema 40 *Sea G la matriz de iteración de un método iterativo de primer grado convergente, y sea \hat{G}_{ω} la matriz de la ecuación (3.17) para el método iterativo de segundo grado (3.15). Si*

$$-\sqrt{\frac{1 - \bar{\mu}}{2\bar{\mu}}} < \omega < \sqrt{\frac{1 + \bar{\mu}}{2\bar{\mu}}}$$

entonces $\rho(\hat{G}_{\omega}) < 1$, donde $\bar{\mu} = \rho(G)$.

Capítulo 3. Esquema iterativo de segundo grado

Demostración. Como G es la matriz de iteración de un método iterativo de primer grado convergente, entonces $\bar{\mu} < 1$. Por tanto, existe una norma matricial compatible (teorema 6, pág. 8), denotada por $\|\cdot\|_\alpha$, tal que $\|G\|_\alpha < 1$. A continuación se estudian los diferentes intervalos para el parámetro de extrapolación ω .

1. Considérese $0 < \omega < 1$. Sea β una constante real tal que $\|G\|_\alpha < \beta < 1$. Entonces se tiene para el primer término de la ecuación (3.36) la siguiente relación,

$$\|(1 - \omega)G\|_\alpha + \|\omega G\|_\alpha = (1 - \omega + \omega)\|G\|_\alpha < \beta < 1 .$$

De la misma forma, para el segundo término en (3.36) que denotaremos

$$F = \|(1 - \omega)\omega G^2\|_\alpha + \|(1 - \omega)G + \omega^2 G^2\|_\alpha ,$$

se tiene

$$\begin{aligned} F &\leq \|(1 - \omega)\omega G^2\|_\alpha + \|(1 - \omega)G\|_\alpha + \|\omega^2 G^2\|_\alpha \\ &= ((1 - \omega)\omega + \omega^2)\|G^2\|_\alpha + (1 - \omega)\|G\|_\alpha \\ &= \omega\|G^2\|_\alpha + (1 - \omega)\|G\|_\alpha \\ &\leq (\omega + 1 - \omega)\|G\|_\alpha < \beta < 1 . \end{aligned}$$

2. Considérese ahora el intervalo $1 \leq \omega < \sqrt{\frac{1 + \bar{\mu}}{2\bar{\mu}}}$. Para el primer término en (3.36) se mostrará que

$$\|(1 - \omega)G\|_\alpha + \|\omega G\|_\alpha < 1 \quad \text{si} \quad 1 \leq \omega < \frac{1 + \bar{\mu}}{2\bar{\mu}} .$$

Para ω en este intervalo, existe $\beta \in \mathbb{R}$ tal que

$$1 \leq \omega < \beta < \frac{1 + \bar{\mu}}{2\bar{\mu}} .$$

Sea $0 < \alpha < \min(1 - \bar{\mu}, \frac{1 - (2\beta - 1)\bar{\mu}}{(2\beta - 1)})$. Puesto que $\bar{\mu} < 1$, existe una norma matricial compatible tal que

$$\bar{\mu} \leq \|G\|_\alpha \leq \bar{\mu} + \alpha < 1 .$$

3.4. Métodos iterativos de segundo grado

Se tiene entonces que,

$$\|(1 - \omega)G\|_\alpha + \|\omega G\|_\alpha = (2\omega - 1)\|G\|_\alpha < (2\beta - 1)\|G\|_\alpha.$$

Además,

$$\alpha < \frac{1 - (2\beta - 1)\bar{\mu}}{(2\beta - 1)},$$

y entonces

$$(2\beta - 1) < \frac{1}{\bar{\mu} + \alpha}.$$

Se sigue finalmente que

$$(2\beta - 1)\|G\|_\alpha \leq (2\beta - 1)(\bar{\mu} + \alpha) < \frac{1}{\bar{\mu} + \alpha}(\bar{\mu} + \alpha) = 1.$$

Procediendo de igual forma para el segundo término de (3.36), existe $\beta \in \mathbb{R}$ tal que

$$1 \leq \omega < \beta < \sqrt{\frac{1 + \bar{\mu}}{2\bar{\mu}}}.$$

Sea $0 < \alpha < \min(1 - \bar{\mu}, \frac{1 - (2\beta^2 - 1)\bar{\mu}}{(2\beta^2 - 1)})$. Con un razonamiento similar al caso anterior se sigue que

$$\begin{aligned} F &\leq \|\omega(1 - \omega)G^2\|_\alpha + \|(1 - \omega)G\|_\alpha + \|\omega^2 G^2\|_\alpha \\ &= (2\omega^2 - \omega)\|G^2\|_\alpha + (\omega - 1)\|G\|_\alpha \\ &\leq (2\omega^2 - \omega + \omega - 1)\|G\|_\alpha \\ &< (2\beta^2 - 1)\|G\|_\alpha. \end{aligned}$$

Además

$$\alpha < \frac{1 - (2\beta^2 - 1)\bar{\mu}}{(2\beta^2 - 1)},$$

y por tanto

$$(2\beta^2 - 1) < \frac{1}{\bar{\mu} + \alpha}.$$

Se sigue finalmente que

$$(2\beta^2 - 1)\|G\|_\alpha \leq (2\beta^2 - 1)(\bar{\mu} + \alpha) < \frac{1}{\bar{\mu} + \alpha}(\bar{\mu} + \alpha) = 1.$$

3. Si $-\sqrt{\frac{1-\bar{\mu}}{2\bar{\mu}}} < \omega \leq 0$, procediendo de manera similar al intervalo anterior, para el primer término en (3.36) se mostrará que

$$\|(1-\omega)G\|_{\alpha} + \|\omega G\|_{\alpha} < 1 \quad \text{si} \quad -\frac{1-\bar{\mu}}{2\bar{\mu}} < \omega \leq 0.$$

Para ω en este intervalo, existe $\beta \in \mathbb{R}$ tal que

$$-\frac{1-\bar{\mu}}{2\bar{\mu}} < \beta < \omega \leq 0.$$

Sea $0 < \alpha < \min(1-\bar{\mu}, \frac{1-(1-2\beta)\bar{\mu}}{(1-2\beta)})$. Entonces se tiene que,

$$\|(1-\omega)G\|_{\alpha} + \|\omega G\|_{\alpha} = (1-2\omega)\|G\|_{\alpha} < (1-2\beta)\|G\|_{\alpha}.$$

Además, $\alpha < \frac{1-(1-2\beta)\bar{\mu}}{(1-2\beta)}$ y por tanto $(1-2\beta) < \frac{1}{\bar{\mu} + \alpha}$. Se obtiene finalmente la siguiente relación,

$$(1-2\beta)\|G\|_{\alpha} \leq (1-2\beta)(\bar{\mu} + \alpha) < \frac{1}{\bar{\mu} + \alpha}(\bar{\mu} + \alpha) = 1.$$

Para el segundo término de la ecuación (3.36) existe $\beta \in \mathbb{R}$ tal que $-\sqrt{\frac{1-\bar{\mu}}{2\bar{\mu}}} < \beta < \omega < 0$. Sea

$$0 < \alpha < \min(1-\bar{\mu}, \frac{1-(2\beta^2+1)\bar{\mu}}{(2\beta^2+1)}).$$

Se tiene que,

$$\begin{aligned} F &\leq \|\omega(1-\omega)G^2\|_{\alpha} + \|(1-\omega)G\|_{\alpha} + \|\omega^2 G^2\|_{\alpha} \\ &= (2\omega^2 - \omega)\|G^2\|_{\alpha} + (1-\omega)\|G\|_{\alpha} \\ &\leq (2\omega^2 - 2\omega + 1)\|G\|_{\alpha} < (2\omega^2 + 1)\|G\|_{\alpha} \\ &< (2\beta^2 + 1)\|G\|_{\alpha}. \end{aligned}$$

Ya que $\alpha < \frac{1-(2\beta^2+1)\bar{\mu}}{(2\beta^2+1)}$, se tiene $(2\beta^2+1) < \frac{1}{\bar{\mu} + \alpha}$, obteniéndose finalmente

$$(2\beta^2 + 1)\|G\|_{\alpha} \leq (2\beta^2 + 1)(\bar{\mu} + \alpha) < \frac{1}{\bar{\mu} + \alpha}(\bar{\mu} + \alpha) = 1.$$

3.4. Métodos iterativos de segundo grado

La demostración es completa. ■

Si se compara el resultado de este teorema con el teorema 32 se observa claramente como para valores propios complejos el intervalo válido para el factor de relajación ω se reduce, existiendo entre ambos intervalos la relación dada por la función raíz cuadrada. En las siguientes secciones se particularizará el estudio de los métodos de segundo grado a los algoritmos A y B propuestos en la sección 1.

3.4.1 Método iterativo de segundo grado A.

A continuación mostraremos que el esquema iterativo A (algoritmo 1, pág. 68) corresponde a un método de segundo grado. Para ello considérese el sistema de ecuaciones (3.8) particionado como en (3.1), es decir

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}.$$

Sea la partición de la matriz $T = M - N$ correspondiente al método iterativo de Jacobi por bloques, es decir,

$$M = \begin{bmatrix} T_{11} & O \\ O & T_{22} \end{bmatrix}, \quad N = \begin{bmatrix} O & -T_{12} \\ -T_{21} & O \end{bmatrix}.$$

Entonces, la matriz de iteración del método de Jacobi por bloques viene dada por,

$$B = \begin{bmatrix} O & -T_{11}^{-1}T_{12} \\ -T_{22}^{-1}T_{21} & O \end{bmatrix} = \begin{bmatrix} O & B_{12} \\ B_{21} & O \end{bmatrix}, \quad (3.37)$$

donde por conveniencia se han introducido las matrices $B_{12} = -T_{11}^{-1}T_{12}$ y $B_{21} = -T_{22}^{-1}T_{21}$.

Si se considera el esquema iterativo de grado 2 dado en la ecuación (3.12), con matrices $G_0 = \omega B$ y $G_1 = (1 - \omega)B$ donde ω es un factor de relajación o extrapolación, y el vector k dado por

$$k = M^{-1}e = \begin{bmatrix} T_{11}^{-1}E_1 \\ T_{22}^{-1}E_2 \end{bmatrix},$$

Capítulo 3. Esquema iterativo de segundo grado

se obtiene el siguiente sistema de ecuaciones

$$\begin{aligned} T_{11}\psi_1^{l+1} &= E_1 - T_{12}(\omega\psi_2^l + (1-\omega)\psi_2^{l-1}), \\ T_{22}\psi_2^{l+1} &= E_2 - T_{21}(\omega\psi_1^l + (1-\omega)\psi_1^{l-1}), \end{aligned}$$

que corresponde claramente al cuerpo principal del método iterativo A. De esta forma, todo el estudio anterior para un esquema iterativo de segundo grado genérico se aplica directamente al método A (algoritmo 1).

3.4.2 Método iterativo de segundo grado B.

Para el esquema iterativo B (algoritmo 2, pág. 69) también se pueden identificar ciertas matrices G_0 , G_1 y k que, sustituidas en la ecuación (3.12), nos permitan escribirlo como un esquema iterativo de segundo grado. Consideréense las matrices

$$G_0 = \begin{bmatrix} O & -(1-\omega)B_{12} \\ O & (1-\omega)\omega B_{21}B_{12} \end{bmatrix}, \quad G_1 = \begin{bmatrix} O & -\omega B_{12} \\ -(1-\omega)B_{21} & \omega^2 B_{21}B_{12} \end{bmatrix}, \quad (3.38)$$

$$k = \begin{bmatrix} T_{11}^{-1}E_1 \\ T_{22}^{-1}E_2 - \omega B_{21}T_{11}^{-1}E_1 \end{bmatrix}, \quad (3.39)$$

donde $B_{12} = T_{11}^{-1}T_{12}$, $B_{21} = T_{22}^{-1}T_{21}$ son los bloques fuera de la diagonal principal de la matriz de Jacobi por bloques de la ecuación (3.37). Al sustituir en (3.12) se obtiene el sistema de ecuaciones

$$\begin{aligned} T_{11}\psi_1^{l+1} &= E_1 - T_{12}(\omega\psi_2^l + (1-\omega)\psi_2^{l-1}), \\ T_{22}\psi_2^{l+1} &= E_2 - T_{21}(\omega\psi_1^{l+1} + (1-\omega)\psi_1^l), \end{aligned}$$

que corresponde al método B (pág. 69). Las matrices G_0 y G_1 se pueden factorizar como sigue,

$$G_0 = \begin{bmatrix} O & -B_{12} \\ -(1-\omega)B_{21} & \omega B_{21}B_{12} \end{bmatrix} \begin{bmatrix} O & O \\ O & (1-\omega)I \end{bmatrix} = H_\omega P_0, \quad (3.40)$$

$$G_1 = \begin{bmatrix} O & -B_{12} \\ -(1-\omega)B_{21} & \omega B_{21}B_{12} \end{bmatrix} \begin{bmatrix} I & O \\ O & \omega I \end{bmatrix} = H_\omega P_1, \quad (3.41)$$

3.4. Métodos iterativos de segundo grado

y de esta forma, la matriz del método ampliado equivalente (3.16) se puede escribir como

$$\hat{G}_w = \begin{bmatrix} 0 & I \\ H_\omega P_0 & H_\omega P_1 \end{bmatrix}. \quad (3.42)$$

Fácilmente se obtiene la relación

$$H_\omega P_0 + H_\omega P_1 = H_\omega (P_0 + P_1) = H_\omega.$$

Además, si se factoriza la matriz H_ω de la forma

$$H_\omega = M^{-1}N = \begin{bmatrix} T_{11}^{-1} & O \\ -\omega B_{21}T_{11}^{-1} & T_{22}^{-1} \end{bmatrix} \begin{bmatrix} O & -T_{12} \\ -(1-\omega)T_{21} & O \end{bmatrix}, \quad (3.43)$$

se observa que el término k en (3.39) responde a la expresión

$$k = M^{-1}e = \begin{bmatrix} T_{11}^{-1} & O \\ -\omega B_{21}T_{11}^{-1} & T_{22}^{-1} \end{bmatrix} \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}.$$

Por el lema 27, si además el radio espectral de la matriz H_ω es menor que 1, el método B es completamente consistente con el sistema de ecuaciones (3.1). Obsérvese que la matriz H_ω se obtiene de una partición de la matriz T del sistema de ecuaciones (3.1), $T = M_\omega - N_\omega$. En efecto, de la ecuación (3.43) se sigue que,

$$M_\omega = \begin{bmatrix} T_{11} & O \\ \omega T_{21} & T_{22} \end{bmatrix}, \quad N_\omega = \begin{bmatrix} O & -T_{12} \\ -(1-\omega)T_{21} & O \end{bmatrix}. \quad (3.44)$$

Considerando las matrices,

$$D = \begin{bmatrix} T_{11} & O \\ O & T_{22} \end{bmatrix}, \quad E = \begin{bmatrix} O & O \\ -T_{21} & O \end{bmatrix}, \quad F = \begin{bmatrix} O & -T_{12} \\ O & O \end{bmatrix},$$

se pueden escribir las matrices T y H_ω de la forma,

$$\begin{aligned} T &= (D - \omega E) - ((1 - \omega)E + F), \\ H_\omega &= (D - \omega E)^{-1}((1 - \omega)E + F). \end{aligned} \quad (3.45)$$

Claramente, la matriz de iteración H_ω corresponde al método de Gauss-Seidel acelerado (AGS) (ecuación (3.5), pág. 70), aplicado a la matriz T del sistema

Capítulo 3. Esquema iterativo de segundo grado

de ecuaciones (3.1). En la sección 3.4.3 se presentan algunos resultados de convergencia para el método AGS.

El siguiente resultado establece un rango válido para el factor de relajación ω , de manera que el método iterativo B converge a la solución del sistema. Se asume que la norma infinito de la matriz H_ω es menor que 1 (ecuación (3.43)). Al igual que en la sección anterior, se evaluará la norma del cuadrado de la matriz (3.42), utilizando para ello la norma matricial definida en el lema 39. Se tiene por tanto

$$\hat{G}_\omega^2 = \begin{bmatrix} H_\omega P_0 & H_\omega P_1 \\ H_\omega P_1 H_\omega P_0 & H_\omega P_0 + (H_\omega P_1)^2 \end{bmatrix}. \quad (3.46)$$

Teorema 41 Sea H_ω la matriz de la ecuación 3.43, de manera que $\|H_\omega\|_\infty < 1$. Si

$$2 - \frac{1}{\|H_\omega\|_\infty} < \omega < \frac{\|H_\omega\|_\infty + 1}{2\|H_\omega\|_\infty},$$

entonces $\rho(\hat{G}_\omega) < 1$.

Demostración. Si se aplica la norma matricial definida en el lema 39 a la matriz de la ecuación (3.46) se debe calcular

$$\max \{ \|HP_0\|_\infty + \|HP_1\|_\infty, \|HP_1HP_0\|_\infty + \|HP_0 + (HP_1)^2\|_\infty \}. \quad (3.47)$$

Primero se calculará la norma de las matrices P_0 y P_1 .

1. $\omega \leq 1$

$$\begin{aligned} \|P_0\|_\infty &= |(1 - \omega)| = 1 - \omega, \\ \|P_1\|_\infty &= \max\{\|I\|_\infty, \|\omega I\|_\infty\} = 1. \end{aligned}$$

2. $1 < \omega$

$$\begin{aligned} \|P_0\|_\infty &= |(1 - \omega)| = \omega - 1, \\ \|P_1\|_\infty &= \max\{\|I\|_\infty, \|\omega I\|_\infty\} = \omega. \end{aligned}$$

3.4. Métodos iterativos de segundo grado

Por tanto, se tiene

$$1. \quad 2 - \frac{1}{\|H\|_\infty} < \omega \leq 1$$

$$\|P_0\|_\infty = 1 - \omega ,$$

$$\|P_1\|_\infty = 1 .$$

$$2. \quad 1 < \omega < \frac{\|H\|_\infty + 1}{2\|H\|_\infty}$$

$$\|P_0\|_\infty = (\omega - 1) ,$$

$$\|P_1\|_\infty = \omega .$$

El siguiente paso consiste en evaluar la norma de la matriz \hat{G}_ω^2 dada en la ecuación (3.47). Procediendo de la misma forma que en la demostración del teorema 40, se tiene para cada uno de los términos de la ecuación (3.47) las relaciones siguientes:

$$1. \quad \|HP_0\|_\infty + \|HP_1\|_\infty$$

$$\text{Sea } 2 - \frac{1}{\|B\|_\infty} < \omega \leq 1, \text{ entonces}$$

$$\begin{aligned} \|HP_0\|_\infty + \|HP_1\|_\infty &\leq \|H\|_\infty(\|P_0\|_\infty + \|P_1\|_\infty) \\ &\leq \|H\|_\infty(2 - \omega) = \gamma. \end{aligned}$$

$$\text{Ya que } \omega > 2 - \frac{1}{\|H\|_\infty}, \text{ entonces se tiene}$$

$$\gamma < \|H\|_\infty \frac{1}{\|H\|_\infty} = 1.$$

$$\text{Si } 1 < \omega < \frac{\|H\|_\infty + 1}{2\|H\|_\infty}, \text{ Se tiene entonces}$$

$$\begin{aligned} \|HP_0\|_\infty + \|HP_1\|_\infty &\leq \|H\|_\infty(\|P_0\|_\infty + \|P_1\|_\infty) \\ &= \|H\|_\infty(2\omega - 1) < \|H\|_\infty \frac{1}{\|H\|_\infty} = 1 , \end{aligned}$$

$$\text{ya que } \omega < \frac{\|H\|_\infty + 1}{2\|H\|_\infty} \text{ implica } (2\omega - 1) < \frac{1}{\|H\|_\infty}.$$

Capítulo 3. Esquema iterativo de segundo grado

2. $\|HP_1HP_0\|_\infty + \|HP_1 + (HP_1)^2\|_\infty$

Para el segundo término de la ecuación (3.47), utilizando los resultados anteriores, se sigue que

$$\begin{aligned} \|HP_1HP_0\|_\infty + \|HP_1 + (HP_1)^2\|_\infty &\leq \|H\|_\infty\{\|P_1\|_\infty(\|HP_0\|_\infty + \\ &\quad + \|HP_1\|_\infty) + \|P_0\|_\infty\} \\ &< \|H\|_\infty\{\|P_1\|_\infty + \|P_0\|_\infty\} \\ &< 1. \end{aligned}$$

Finalmente, ya que la norma (3.47) es menor que 1, se tiene $\rho(\hat{G}_w) < 1$. ■

Para finalizar esta sección, obsérvese que las matrices G_0 y G_1 también se pueden factorizar de la forma,

$$G_0 = B \cdot \begin{bmatrix} 0 & -(1-\omega)\omega B_{12} \\ 0 & (1-\omega)I \end{bmatrix}, \quad (3.48)$$

$$G_1 = B \cdot \begin{bmatrix} (1-\omega)I & -\omega^2 B_{12} \\ 0 & \omega I \end{bmatrix}, \quad (3.49)$$

donde B es la matriz de Jacobi por bloques asociado a la matriz T , y que se muestra en (3.37). Se puede demostrar, de forma similar al teorema anterior, el siguiente resultado.

Teorema 42 *Sea B la matriz del método iterativo de Jacobi por bloques dividida en 2×2 bloques como en (3.37), de manera que $0.5 < \|B\|_\infty < 1$. Si*

$$2 - \frac{1}{\|B\|_\infty} < \omega < \sqrt{\frac{\|B\|_\infty + 1}{2\|B\|_\infty}},$$

entonces $\rho(\hat{G}_\omega) < 1$.

3.4. Métodos iterativos de segundo grado

3.4.3 Resultados de convergencia para el método AGS.

A continuación se revisa la bibliografía existente sobre el método de Gauss-Seidel acelerado (AGS). En la sección anterior se observó que el método iterativo B se obtiene a partir de la matriz de iteración del método AGS (ecuación (3.43) para H_ω), para el caso en el que la matriz del sistema T es una matriz dividida en 2×2 bloques (ecuación (3.1)). Por tanto, el método de segundo grado B es completamente consistente si la matriz H_ω tiene radio espectral menor que 1.

El método AGS es un caso particular del método de sobrerrelajación acelerado (AOR), ecuación (1.16) de la página 14. En efecto, tomando $\tau = 1$ en la ecuación

$$\mathcal{L}_{\omega,\tau} = (I - \omega L)^{-1}[(1 - \tau)I + (\tau - \omega)L + \tau U] ,$$

se obtiene

$$\mathcal{L}_{\omega,1} = (I - \omega L)^{-1}[(1 - \omega)L + U] , \quad (3.50)$$

que es la matriz de iteración del método AGS, \mathcal{L}_{AGS} (ecuación (3.5)). Además, en [46] se muestra que AOR coincide con el método SOR extrapolado (ESOR) cuando $\omega\tau \neq 0$. Por tanto, de la amplia y rica bibliografía sobre los métodos AOR y ESOR se pueden extraer muchos resultados de convergencia particulares para el método AGS. La bibliografía más relevante se cita a continuación.

El método clásico ESOR fue estudiado por primera vez por Sisler [90], y posteriormente ampliado por Niethammer [103]. El método AOR fue introducido por Hadjidimos [46] donde se presentan resultados de convergencia para matrices diagonal dominantes e irreducibles, y L-matrices [105, definición 2.7.1]. Algunos de los resultados que obtiene fueron hallados y ampliados independientemente por Missirlis y Evans [72] que estudian el método ESOR, y dos variantes del método de Gauss-Seidel extrapolado, una de las cuales

Capítulo 3. Esquema iterativo de segundo grado

se identifica con el método AGS. Estos autores presentan algunos resultados para M-matrices, y matrices simétricas y definidas positivas. El método AOR óptimo fue determinado, cuando la matriz del método de Jacobi asociado es consistentemente ordenada y débilmente 2-cíclica, en [90, 103, 46, 72, 3, 71]. El estudio del caso de matrices p -cíclicas y consistentemente ordenadas se realiza en [89]. Finalmente, en [47] se estudia el método AOR generalizado (GAOR), completando el estudio para matrices simétricas y definidas positivas realizado en [72].

Como se ha puesto de manifiesto, el estudio del método AGS a partir del estudio del método AOR es extenso. De cualquier modo, en esta sección se presentan algunos resultados de comparación entre el método AGS y el SOR, que no aparecen específicamente en la bibliografía citada ya que en ella se estudian métodos más generales.

Sea T una matriz invertible de tamaño $n \times n$. Se asume que la matriz T tiene la forma,

$$T = I - L - U . \quad (3.51)$$

donde $-L$ y $-U$ son las partes estrictamente triangular inferior y superior, respectivamente. Cuando todos los elementos de la diagonal son positivos, los resultados que se presentan se extienden de modo natural para matrices en la forma general

$$T = D - E - F . \quad (3.52)$$

donde D es una matriz diagonal e invertible cuyos elementos diagonales son los de T , $-E$ y $-F$ son las partes estrictamente triangular inferior y superior, respectivamente. Esta técnica es bastante habitual. Por ejemplo en [105] y [13], para M-matrices y matrices simétricas definidas positivas, el método iterativo SOR se estudia frecuentemente asumiendo que la matriz es de la forma (3.51). Aunque en principio no se asume ninguna división por bloques de la matriz, la generalización de los resultados que se presenten a matrices divididas por bloques como en (3.2) es sencilla.

3.4. Métodos iterativos de segundo grado

La matriz del método AGS (3.50) se denotará H_ω , notación introducida en el estudio del método de segundo grado, es decir,

$$H_\omega = \mathcal{L}_{\omega,1} , \quad (3.53)$$

y se obtiene para la partición de la matriz

$$T = M_\omega - N_\omega = (I - \omega L) - [(1 - \omega)L + U] . \quad (3.54)$$

Observación 43 *La matriz de iteración del método iterativo de Jacobi asociada a la matriz (3.51) es*

$$B = L + U . \quad (3.55)$$

Claramente, para $\omega = 0$, la matriz H_0 coincide con la matriz B , y para $\omega = 1$, H_1 coincide con la matriz de iteración del método de Gauss-Seidel que viene dada por,

$$\mathcal{L} = (I - L)^{-1}U . \quad (3.56)$$

Finalmente, la matriz de iteración para el método SOR se escribe como

$$\mathcal{L}_\omega = (I - \omega L)^{-1}((1 - \omega)I + \omega U) . \quad (3.57)$$

Entre las matrices de iteración del método SOR y el método AGS se tiene la relación dada por

$$\mathcal{L}_\omega = (1 - \omega)I + \omega H_\omega . \quad (3.58)$$

Se observa que el método SOR coincide con el método AGS extrapolado, con factor de extrapolación igual al de relajación.

A continuación se compara la velocidad de convergencia del método AGS en función del valor del parámetro ω . Este resultado es similar al que se tiene para el SOR [13, teorema 7.5.23].

Capítulo 3. Esquema iterativo de segundo grado

Teorema 44 *Sea T una M -matriz invertible, y sea $0 \leq \omega_1 \leq \omega_2 \leq 1$. Entonces,*

$$\rho(H_{\omega_2}) \leq \rho(H_{\omega_1}) < 1 ,$$

de forma que

$$R_{\infty}(H_{\omega_2}) \geq R_{\infty}(H_{\omega_1}) .$$

Demostración. Sin pérdida de generalidad podemos escribir $T = I - L - U$ y $H_{\omega} = M_{\omega}^{-1}N_{\omega}$ como en (3.53). De la ecuación (3.54), se observa $M_{\omega}^{-1} \geq O$, $N_{\omega} \geq O$ y, por tanto, $T = M_{\omega} - N_{\omega}$ es una partición regular (definición 11). Ya que $(1-\omega)$ es una función decreciente en ω en el intervalo $0 \leq \omega_1 \leq \omega_2 \leq 1$,

$$N_{\omega_2} \leq N_{\omega_1} ,$$

y el resultado se obtiene del teorema 13, pág. 10. ■

El siguiente resultado establece que el método también es convergente para valores del parámetro ω mayores que 1, indicando una cota máxima.

Teorema 45 *Sea T una M -matriz invertible. Entonces,*

$$\rho(H_{\omega_2}) < 1 ,$$

para todos los valores de ω en el intervalo

$$0 \leq \omega < \frac{1 + \rho(B)}{2\rho(B)} . \tag{3.59}$$

Demostración. Nótese que para M -matrices el método de Jacobi es convergente ($\rho(B) < 1$) y, por tanto, $\frac{1 + \rho(B)}{2\rho(B)} \geq 1$. La convergencia en el intervalo $0 \leq \omega < 1$ se sigue del teorema anterior, y por tanto se asumirá que $\omega \geq 1$. Considérese la matriz

$$T_{\omega} = (I - \omega L)^{-1}((\omega - 1)L + U) .$$

3.4. Métodos iterativos de segundo grado

Se tiene $T_\omega \geq O$, y $|H_\omega| \leq T_\omega$. Sea $\lambda = \rho(T_\omega)$. Por la teoría de Perron-Frobenius (teorema 8, pág. 8), λ es un valor propio de T_ω , y existe un vector no negativo $x \geq 0$ tal que $Tx = \lambda x$. De aquí,

$$((2\omega - 1) + U)x = \lambda x ,$$

y, por tanto,

$$\lambda \leq \rho((2\omega - 1) + U) .$$

Observando que $O \leq (2\omega - 1) + U \leq (2\omega - 1)(L + U) = (2\omega - 1)B$, se sigue que,

$$\lambda \leq \rho((2\omega - 1) + U) \leq (2\omega - 1)\rho(B) .$$

Si $\lambda \geq 1$, entonces

$$\omega \geq \frac{1 + \rho(B)}{2\rho(B)} .$$

Por lo tanto, si $\omega < \frac{1 + \rho(B)}{2\rho(B)}$ entonces $\lambda < 1$. Ya que $|H_\omega| \leq T_\omega$, por el teorema 9, pág. 9, se obtiene $\rho(H_\omega) \leq \lambda < 1$. ■

Observación 46 *Para el método SOR existe un resultado similar al anterior (ver [13, teorema 5.24]), pero para el intervalo*

$$0 < \omega < \frac{2}{1 + \rho(B)} . \quad (3.60)$$

Entre el intervalo (3.60) y el intervalo (3.59) del teorema 45, se puede establecer la relación,

$$0 \leq 0 < \omega \leq \frac{2}{1 + \rho(B)} < \frac{1 + \rho(B)}{2\rho(B)} .$$

Por tanto, cuando T es una M -matriz, el método iterativo con matriz H_ω converge en un intervalo más amplio que el SOR. En [72] se obtiene para el método AOR la misma cota que para el SOR, y que, particularizada para el método AGS, proporciona una cota superior más restrictiva que la obtenida en (3.59).

Capítulo 3. Esquema iterativo de segundo grado

A continuación se realiza una comparación con el método SOR del estilo del teorema de Stein y Rosenberg (teorema 16, pág. 14).

Teorema 47 Sea $T = I - L - U \in \mathbb{R}^{n \times n}$ donde $L \geq O$ y $U \geq O$ son estrictamente triangular inferior y superior, respectivamente. Entonces para $0 < \omega \leq 1$,

1. $\rho(\mathcal{L}_\omega) < 1$ si, y sólo si $\rho(H_\omega) < 1$. Además $\rho(H_\omega) \leq \rho(\mathcal{L}_\omega) < 1$.
2. $\rho(H_\omega) \geq 1$ si, y sólo si $\rho(\mathcal{L}_\omega) \geq 1$. Además $\rho(H_\omega) \geq \rho(\mathcal{L}_\omega) \geq 1$.

Demostración. Se tiene $H_\omega \geq O$ (ecuación (3.53)), y además $\mathcal{L}_\omega \geq O$, en el intervalo $0 < \omega \leq 1$. En [46, ecuación (2.7)] se muestra la relación entre los valores propios de los métodos SOR y AOR. Ya que el método AGS es una caso particular de AOR, se obtiene fácilmente la relación siguiente,

$$\lambda = (1 - \omega) + \omega\mu, \quad (3.61)$$

donde los valores propios de \mathcal{L}_ω se denotan por λ , y los de H_ω se denotan por μ .

$H_\omega \geq O$ implica que su radio espectral es un valor propio. Además, $|(1 - \omega) - \omega\mu|$ es una función creciente en $|\mu|$, y por tanto el valor máximo para el módulo de los valores propios de \mathcal{L}_ω se alcanza para $|\mu| = \bar{\mu} = \rho(H_\omega)$. Ya que $\mathcal{L}_\omega \geq O$ y por tanto su radio espectral es también un valor propio, se obtiene

$$\bar{\lambda} = \rho(\mathcal{L}_\omega) = (1 - \omega) + \omega\rho(H_\omega).$$

Si $\bar{\mu} \geq 1$, entonces

$$\bar{\lambda} = (1 - \omega) + \omega\bar{\mu} \geq 1 - \omega + \omega = 1,$$

que implica, si $\bar{\lambda} < 1$ entonces $\bar{\mu} < 1$. Además,

$$\bar{\lambda} = (1 - \omega) + \omega\bar{\mu} = \omega(\bar{\mu} - 1) + 1 \leq (\bar{\mu} - 1) + 1 = \bar{\mu},$$

3.4. Métodos iterativos de segundo grado

obteniéndose $\bar{\mu} \geq \bar{\lambda} \geq 1$.

Finalmente, si $\bar{\mu} < 1$ entonces

$$\bar{\lambda} = (1 - \omega) + \omega\bar{\mu} < 1 - \omega + \omega = 1 ,$$

que implica que si $\bar{\lambda} \geq 1$ entonces $\bar{\mu} \geq 1$. Además,

$$\bar{\lambda} = (1 - \omega) + \omega\bar{\mu} = \omega(\bar{\mu} - 1) + 1 \geq (\bar{\mu} - 1) + 1 = \bar{\mu} ,$$

obteniéndose $\bar{\mu} \leq \bar{\lambda} < 1$, que completa la demostración. ■

Corolario 48 *Sea T una M -matriz. Entonces para $0 < \omega \leq 1$,*

$$\rho(H_\omega) \leq \rho(\mathcal{L}_\omega) < 1 .$$

Demostración. Si T es una M -matriz invertible, $T^{-1} \geq O$. Se puede asumir que tiene la forma $T = I - L - U$, con $L \geq O$ y $U \geq O$. Por tanto, del teorema 45 se tiene $\rho(H_\omega) < 1$. Aplicando el teorema 47, apartado 1, se concluye la demostración. Nótese que para $0 < \omega \leq 1$, tanto H_ω como \mathcal{L}_ω se obtienen de particiones (débiles) regulares de una matriz monótona. Ya que $T^{-1}(N_{SOR} - N_\omega) = (1 - \omega)A^{-1}M_{SOR} = (1 - \omega)A^{-1}(A + N_{SOR}) \geq O$, es decir, $T^{-1}N_{SOR} \geq T^{-1}N_\omega \geq O$, con $N_{SOR} = (1 - \omega)I + \omega U$ y $N_\omega = (1 - \omega)L + U$, respectivamente, la aplicación de [79, teorema 4.18, apartado (i)] también demuestra el resultado. ■

Observación 49 *Sea T una M -matriz invertible dividida en bloques como en (3.2), y sea la matriz diagonal por bloques asociada*

$$D = \begin{bmatrix} T_{11} & O & \cdots & O \\ O & T_{22} & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \cdots & T_{qq} \end{bmatrix} .$$

Capítulo 3. Esquema iterativo de segundo grado

Por ser cada uno de los bloques diagonales menores principales de T , la matriz D es invertible, y además $D^{-1} \geq O$. Por tanto, también podemos considerar T de la forma $T = I - L - U$ donde $-L$ y $-U$ son las partes estrictamente triangular inferior y superior, respectivamente, de la matriz T . Además, $L \geq O$ y $U \geq O$, y se tiene,

$$(I - \omega L)^{-1} = I + \omega L + \omega^2 L^2 + \cdots + \omega^{q-1} L^{q-1} \geq O ,$$

donde q es el número de bloques diagonales de T . Claramente la partición del método AGS (ecuación (3.54)) para $0 \leq \omega$, es una partición (débilmente) regular de T y por tanto convergente. En consecuencia, los resultados de los teoremas 44, 45 y el corolario 48 se aplican directamente.

3.5 Experimentos numéricos

En esta sección se evalúa la eficacia de los métodos de segundo grado A y B (algoritmos 1 y 2) en la simulación del transitorio 1 del TWIGL (capítulo 2, pág. 53).

Para la discretización de la parte espacial de la ecuación de la difusión neutrónica se utilizará tanto el método de colocación nodal con 3 y 4 polinomios de Legendre, como el método en diferencias finitas con tamaño de malla uniforme $h_x = h_y = 3 \text{ cm}$. En las tablas $Nodal(p)$ hace referencia al método de colocación nodal, donde p indica el número de polinomios de Legendre utilizados en el desarrollo del flujo neutrónico en cada nodo. En la tabla 3.1 se muestran las diferentes discretizaciones espaciales utilizadas, así como el tamaño de las matrices del sistema de ecuaciones (4.42) que hay que resolver en cada paso de tiempo.

Para la discretización temporal de la ecuación de la difusión neutrónica se utilizará un método en diferencias hacia atrás de 1 paso, con paso de integración de 1,25 milisegundos. Por tanto se necesita un número total de 160 pasos para simular el transitorio completo.

3.5. Experimentos numéricos

<i>discretización espacial</i>	<i>n</i>	<i>nnz</i>
<i>Nodal(3)</i>	1200	14960
<i>Nodal(4)</i>	2000	31920
<i>Diferencias</i>	5408	32032

Tabla 3.1: Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio 1 del TWIGL. n , nnz indican el tamaño y el número de elementos no nulos de las matrices a resolver en cada paso de tiempo, respectivamente.

En cuanto a los códigos y librerías utilizadas, son válidas las consideraciones del capítulo anterior. Las pruebas se han realizado en la máquina HP Exemplar S Class. Como criterio de parada se ha utilizado el test de error,

$$\|d_i\| \leq \|d_0\| \cdot rtol, \quad (3.62)$$

con $rtol = 10^{-5}$ la tolerancia relativa, y $d_i = \psi_i - \psi_{i-1}$ el cambio previo en la solución.

Con los experimentos numéricos que se presentan se pretende comparar los métodos de segundo grado A y B. También se pretende realizar un estudio heurístico del parámetro de extrapolación ω para las matrices que aparecen en la integración de la ecuación de la difusión neutrónica, cuya forma general se muestra en la ecuación (3.1) (ver también la sección 2.4, pág. 53). Además, se estudia el mejor preconditionador para el método del gradiente conjugado que se utiliza para resolver las iteraciones internas correspondientes a la solución de los sistemas de ecuaciones con matrices T_{11} y T_{22} . Finalmente, se compararán las diferentes discretizaciones utilizadas desde el punto de vista de la precisión alcanzada en la solución final, completando el estudio realizado en el capítulo 2.

En primer lugar y tratando de responder la cuestión referente al mejor preconditionador para el método del gradiente conjugado, en la tabla 3.2 se muestran los resultados con el método de segundo grado B. Los preconditionadores utilizados han sido Jacobi, SOR simétrico (SSOR) y diferentes factorizaciones incompletas LU. Como criterio de parada para el método del

Capítulo 3. Esquema iterativo de segundo grado

<i>precondicionador</i>	<i>tiempo (sg)</i>
Sin precondicionar	57,81
Jacobi	53,24
SSOR	110,53
ILU0	76,27
ILUT(5, 10^{-2})	108,6
ILUT(5, 1)	60,13

Tabla 3.2: Simulación mediante el método iterativo B. Discretización mediante el método nodal con 4 polinomios de Legendre. Factor de extrapolación $\omega = 1, 5$.

gradiente conjugado se ha utilizado,

$$\|r_i\| \leq \|r_0\| \cdot rtol ,$$

donde $r_i = E - T\psi_i$ es el residuo, y con una tolerancia de error $rtol = 5 \times 10^{-6}$. El número máximo de iteraciones permitidas es 20 en todos los casos. Además, el preconditionador se calcula sólo una vez, en el primer paso de tiempo. Se puede observar que el preconditionador de Jacobi parece ser el más apropiado. Esta conclusión sigue la línea mostrada en [99]. De los preconditionadores restantes, los preconditionadores ILU0 y ILUT (con nivel de llenado 5 y magnitud relativa 1) obtienen resultados parecidos.

En las tablas 3.3 y 3.4 se muestran los resultados obtenidos por los métodos de segundo grado A y B en la simulación del transitorio utilizando el método nodal con 3 y 4 polinomios de Legendre. En las tablas 3.5 y 3.6 se muestran los resultados correspondientes a la discretización con método de las diferencias centradas finitas. En todas las tablas ω es el factor de extrapolación y *tiempo* es el tiempo empleado en la simulación en segundos.

Comparando los métodos de segundo grado, el método de segundo grado B muestra una convergencia más rápida que el método iterativo A para todas las discretizaciones utilizadas. Por tanto, será este el método que se utilizará con preferencia en capítulos posteriores. Se observa que el tiempo

3.5. Experimentos numéricos

<i>tiempo (sg)</i>		
ω	<i>Nodal</i> (3)	<i>Nodal</i> (4)
0	†	†
0,5	†	†
1	58,46	91,96
1,2	58,65	91,50
1,4	58,75	90,70
1,5	58,64	90,87
1,6	58,64	92,32
1,8	58,92	93,05
2	†	†

Tabla 3.3: Simulación mediante el método iterativo A. ω es el factor de extrapolación.

<i>tiempo (sg)</i>		
ω	<i>Nodal</i> (3)	<i>Nodal</i> (4)
0	†	†
0,5	†	†
1	34,46	53,93
1,2	34,57	53,78
1,4	34,42	53,60
1,5	34,41	53,24
1,6	34,93	53,45
1,8	34,50	53,96
2	†	†

Tabla 3.4: Simulación mediante el método iterativo B. ω es el factor de extrapolación.

Capítulo 3. Esquema iterativo de segundo grado

ω	tiempo (sg)
0	†
0,5	†
1	135,94
1,2	135,57
1,4	135,31
1,5	134,70
1,6	136,85
1,8	135,15
2	†

Tabla 3.5: Simulación mediante el método iterativo A. Método en diferencias con tamaño de malla $h_x = h_y = 3 \text{ cm}$. ω es el factor de extrapolación.

de simulación mediante el método B es casi la mitad del que utiliza el método A.

En cuanto a la convergencia en función del parámetro de extrapolación ω , se observa que en todos los casos, para valores de ω en el intervalo $[1, 2[$ el tiempo de simulación prácticamente permanece constante, siendo ligeramente menor para valores de ω próximos a 1,5. Para valores de ω fuera de ese intervalo no fue posible simular el transitorio, indicándose en las tablas con el símbolo †. Este hecho sugiere que la precisión en la determinación del mejor factor de extrapolación no es excesivamente importante en la aplicación del método de segundo grado a las matrices que aparecen en la integración de la ecuación de la difusión neutrónica. En cualquier caso, se tomará un valor próximo a 1.5 en las simulaciones posteriores que utilicen cualquiera de los métodos de segundo grado A o B, en consonancia con lo expuesto en [99].

Finalmente, en cuanto a la potencia de pico alcanzada al final del tiempo de simulación (en $t = 0,2$ segundos), en la tabla 3.7 se muestra una comparación para las diferentes discretizaciones utilizadas. Como potencia pico de referencia se ha tomado el valor 2,170. Esta potencia de referencia se ha obtenido discretizando mediante el método nodal con 4 polinomios de Legendre, y utilizando un paso de integración de 0,625 milisegundos. Los valores

3.5. Experimentos numéricos

ω	<i>tiempo (sg)</i>
0	†
0,5	†
1	74,57
1,2	74,85
1,4	74,98
1,5	74,79
1,6	74,89
1,8	75,11
2	†

Tabla 3.6: Simulación mediante el método iterativo B. Método en diferencias con tamaño de malla $h_x = h_y = 3 \text{ cm}$. ω es el factor de extrapolación.

<i>discretización</i>	<i>potencia</i>
<i>Nodal(3)</i>	2,160
<i>Nodal(4)</i>	2,168
<i>Diferencias</i>	2,119

Tabla 3.7: Potencia alcanzada por los métodos de segundo grado A y B en $t = 0,2$ segundos. Potencia de referencia 2,17.

de potencia pico alcanzada por los métodos A y B fueron similares. Se puede observar que mediante el método de colocación nodal se obtiene una mayor precisión en la potencia final alcanzada. Además, la utilización de este método disminuye el tiempo de simulación si se compara con la discretización mediante el método de las diferencias finitas. Por tanto, es recomendable la utilización del método de colocación nodal. También se observa que la precisión en la estimación de potencia aumenta con el número de polinomios de Legendre utilizados. Este hecho es de especial importancia si se desea estimar con suficiente precisión las variaciones locales de potencia, como se comprobará en el capítulo 5.

A modo de resumen se presentan las principales conclusiones a las que se ha llegado en este capítulo. Así se han estudiado métodos de segundo grado obtenidos a partir de la matriz de iteración de un método iterativo de primer grado y un factor de extrapolación o relajación. Se han dado diferentes re-

Capítulo 3. Esquema iterativo de segundo grado

sultados de convergencia en función de los valores propios de la matriz de iteración del método de primer grado, reales o complejos. También haciendo ciertas suposiciones se ha indicado el factor óptimo de extrapolación. Este estudio se ha particularizado a dos métodos de segundo grado propuestos para la solución de los sistemas de ecuaciones que aparecen en la integración de la ecuación de la difusión neutrónica, referenciados como métodos de segundo grado A y B. Con ellos se ha tratado de aprovechar las especiales características que presentan los sistemas de ecuaciones a resolver por bloques. El método B se diferencia del método A en que aprovecha las soluciones para el cálculo tan pronto como están disponibles. De esta forma entre ambos métodos existe una relación similar a la existente entre los métodos de Jacobi y Gauss-Seidel. Además, experimentalmente se ha comprobado que efectivamente el método B emplea aproximadamente la mitad de tiempo en la simulación del transitorio 1 del reactor TWIGL.

Con respecto a los métodos de discretización espacial utilizados, los resultados muestran que el método de colocación aumenta la precisión con el número de polinomios de Legendre utilizados. Además, tanto por precisión como por tiempo de simulación parece estar especialmente indicado para la resolución de la ecuación de la difusión neutrónica.

Finalmente, el parámetro de extrapolación $\omega = 1,5$ proporciona los mejores resultados, aunque en este sentido los resultados también sugieren que la precisión en la determinación del mejor factor de extrapolación no es excesivamente importante para el método de segundo grado.

Capítulo 4

Método variacional

4.1 Introducción

En este capítulo se estudia una técnica variacional para resolver un sistema de ecuaciones lineales de la forma

$$Ax = b \tag{4.1}$$

donde $A \in \mathbb{R}^{n \times n}$ es una matriz invertible. El objetivo final es aplicar esta técnica para acelerar la convergencia del método de segundo grado (A o B), presentado en el capítulo 3, para la solución de los sistemas de ecuaciones que se obtienen de la integración de la ecuación de la difusión neutrónica (cuya forma se muestra en las ecuaciones (2.46), (2.48) y (2.50)). Así mismo, las prestaciones del método de segundo grado acelerado se compararán con otras técnicas bien conocidas y ampliamente utilizadas para matrices no simétricas, como es el BiCGSTAB, GMRES y TFQMR, cuya descripción se vió en el capítulo de introducción (pág. 15).

El capítulo está estructurado como sigue. En la sección 4.2 se revisan algunos conceptos sobre métodos de proyección que serán empleados posteriormente. En la sección 4.3 se describe el método variacional, dándose una interpretación en términos de proyectores y algunas condiciones de convergencia en las secciones 4.3.1 y 4.3.2, respectivamente. En la sección 4.4 se

da el algoritmo para acelerar el método de segundo grado y en la sección 4.5 se muestran los resultados de los experimentos numéricos para simular el transitorio bidimensional 1 del TWIGL. Finalmente, se dan algunas conclusiones.

4.2 Preliminares

En esta sección se revisan algunos conceptos básicos sobre métodos de proyección y las condiciones de Petrov-Galerkin para obtener una solución óptima. Éstas condiciones son bien conocidas y aparecen en otras áreas del cálculo científico, como es el caso de los elementos finitos donde se utilizan habitualmente para describir los métodos de proyección en espacios de elementos finitos. Algunas referencias excelentes sobre métodos de proyección son [53], [33] y recientemente [87] y [42]. La descripción que sigue a continuación se basa en la utilizada en [87] para el estudio de diferentes métodos basados en subespacios de Krylov.

Un método de proyección extrae una solución aproximada \hat{x} para el problema (4.1) de un subespacio de \mathbb{R}^n , denominado subespacio de búsqueda y que denotaremos \mathcal{K} , imponiendo la condición de ortogonalidad del nuevo vector residuo con respecto a otro subespacio de \mathbb{R}^n , denominado subespacio de restricciones y que denotaremos \mathcal{L} . Ambos subespacios, \mathcal{L} y \mathcal{K} , son de igual dimensión m . Se pueden distinguir dos grandes grupos de métodos de proyección: *ortogonales*, cuando $\mathcal{L} = \mathcal{K}$, y *oblicuos*, cuando el subespacio \mathcal{L} es diferente del subespacio \mathcal{K} . El problema por tanto se podría formular como,

$$\begin{array}{ll} \text{Encontrar} & \hat{x} \in \mathcal{K} , \\ \text{tal que} & b - A\hat{x} \perp \mathcal{L} . \end{array} \quad (4.2)$$

Si se conoce una aproximación inicial de la solución del problema, digamos x_0 , y se desea aprovecharla para una obtención más rápida de una solución \hat{x} satisfactoria, el método consiste ahora en obtener una solución para (4.1)

Capítulo 4. Método variacional

en el espacio afín $x_0 + \mathcal{K}$, formulándose ahora el problema en los términos

$$\begin{array}{ll} \text{Encontrar} & \hat{x} \in x_0 + \mathcal{K} , \\ \text{tal que} & b - A\hat{x} \perp \mathcal{L} . \end{array} \quad (4.3)$$

De igual forma, si se escribe la nueva solución del método de proyección como sigue a continuación

$$\hat{x} = x_0 + \delta, \quad \delta \in \mathcal{K}, \quad (4.4)$$

la condición de ortogonalidad del nuevo residuo se formula como

$$b - A\hat{x} = \hat{r} = r_0 - A\delta \perp \mathcal{L}, \quad (4.5)$$

siendo $r_0 = b - Ax_0$. Nótese que $A\delta \in A\mathcal{K}$. De esta manera para minimizar la norma-2 de \hat{r} , esto es $\|\hat{r}\|_2$, en el sentido de mínimos cuadrados, se requiere que $A\delta$ sea la proyección ortogonal del residuo r_0 en el subespacio de \mathbb{R}^n , $A\mathcal{K}$. De esta forma, la elección de $\mathcal{L} = A\mathcal{K}$ nos permite optimizar la búsqueda de las nuevas soluciones en el sentido de minimización de la norma-2 del residuo. Este esquema de trabajo se conoce como condiciones de Petrov-Galerkin [87]. La mayoría de los métodos iterativos usan una sucesión de estos tipos de proyecciones. Matricialmente se pueden representar de la siguiente manera. Sean $V = [v_1, v_2, \dots, v_m]$ y $W = [w_1, w_2, \dots, w_m]$ dos matrices cuyas columnas forman una base de \mathcal{K} y \mathcal{L} , respectivamente. De la ecuación (4.4) se sigue que la solución aproximada se puede escribir de la forma

$$\hat{x} = x_0 + Vy,$$

y el residuo se expresa como

$$b - A\hat{x} = \hat{r} = r_0 - AVy.$$

Aplicando ahora la relación de ortogonalidad que expresa la ecuación (4.5) se tiene que

$$\begin{aligned} W^T \hat{r} = 0 &= W^T r_0 - W^T AVy, \\ y &= (W^T AV)^{-1} W^T r_0. \end{aligned}$$

4.3. Método variacional

Entonces las condiciones de optimización (4.4), (4.5) se escriben como,

$$\hat{x} = x_0 + V(W^T AV)^{-1}W^T r_0, \quad (4.6)$$

$$\hat{r} = r_0 - AV(W^T AV)^{-1}W^T r_0, \quad (4.7)$$

donde se ha asumido que la matriz $W^T AV$ es invertible. La matriz

$$P = AV(W^T AV)^{-1}W^T, \quad (4.8)$$

es la matriz de la proyección sobre AK ortogonalmente al subespacio \mathcal{L} . El siguiente resultado garantiza la invertibilidad de la matriz $W^T AV$ bajo ciertas condiciones para las matrices A , V y W .

Proposición 50 [87, proposición 5.1] Sean A , \mathcal{L} , y \mathcal{K} tales que satisfacen una de las dos condiciones siguientes:

- A es definida positiva y $\mathcal{L} = \mathcal{K}$, o
- A es invertible y $\mathcal{L} = AK$.

Entonces la matriz $B = W^T AV$ es invertible para cualquier base V y W de \mathcal{K} y \mathcal{L} respectivamente.

De esta manera, eligiendo $\mathcal{L} = AK$ se garantiza que el método está bien definido para cualquier matriz A invertible.

4.3 Método variacional

En esta sección se realiza el estudio de un método iterativo variacional para la solución de sistemas de ecuaciones lineales de la forma

$$Ax = b,$$

Capítulo 4. Método variacional

que responde a la ecuación siguiente [38],

$$x_{k+1} = x_k + \alpha_k r_k + \beta_k d_k, \quad (4.9)$$

donde $r_k = b - Ax_k$ y $d_k = x_k - x_{k-1}$ son el residuo y el cambio previo en la solución aproximada en la iteración k , respectivamente. Como se observa en (4.9), el método responde a una recurrencia corta de tres términos con el objetivo de realizar menos trabajo por iteración reduciendo los productos matriz-vector y vector-vector, además de utilizar menos memoria para el almacenamiento de datos. Para deducir los valores de los parámetros α y β , se busca cumplir algún criterio de optimización. En concreto en cada iteración se escogen los valores de α y β tales que minimicen la norma-2 del residuo, $\|r_k\|_2^2$. En el resto del capítulo y mientras que no se indique lo contrario, se denotara la norma-2 mediante $\|\cdot\|$. La expresión para el residuo viene dada por

$$r_{k+1} = r_k - \alpha_k Ar_k - \beta_k Ad_k. \quad (4.10)$$

Por tanto,

$$\|r_k\|^2 = (r_k - \alpha_k Ar_k - \beta_k Ad_k)^T (r_k - \alpha_k Ar_k - \beta_k Ad_k). \quad (4.11)$$

Las ecuaciones (4.9) y (4.10) definen el método variacional. Mediante un proceso de derivación estándar, se obtiene a partir de (4.11) el sistema de ecuaciones siguiente:

$$\left. \begin{aligned} \frac{\partial}{\partial \alpha_k} (\|r_k\|^2) &= 0 \\ \frac{\partial}{\partial \beta_k} (\|r_k\|^2) &= 0 \end{aligned} \right\},$$

o equivalentemente,

$$\left. \begin{aligned} -(Ar_k)^T (r_k - \alpha_k Ar_k - \beta_k Ad_k) + (r_k - \alpha_k Ar_k - \beta_k Ad_k)^T (-Ar_k) &= 0 \\ -(Ad_k)^T (r_k - \alpha_k Ar_k - \beta_k Ad_k) + (r_k - \alpha_k Ar_k - \beta_k Ad_k)^T (-Ad_k) &= 0 \end{aligned} \right\}. \quad (4.12)$$

Desarrollando el sistema (4.12) y despejando los términos independientes se obtiene,

$$\left. \begin{aligned} \alpha_k \langle Ar_k, Ar_k \rangle + \beta_k \langle Ar_k, Ad_k \rangle &= \langle Ar_k, r_k \rangle \\ \alpha_k \langle Ad_k, Ar_k \rangle + \beta_k \langle Ad_k, Ad_k \rangle &= \langle Ad_k, r_k \rangle \end{aligned} \right\}, \quad (4.13)$$

4.3. Método variacional

cuyas soluciones son,

$$\alpha_k = \frac{\langle Ar_k, r_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ad_k, r_k \rangle \langle Ar_k, Ad_k \rangle}{\langle Ar_k, Ar_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ar_k, Ad_k \rangle^2} \quad (4.14)$$

$$\beta_k = \frac{\langle Ar_k, Ar_k \rangle \langle Ad_k, r_k \rangle - \langle Ar_k, r_k \rangle \langle Ar_k, Ad_k \rangle}{\langle Ar_k, Ar_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ar_k, Ad_k \rangle^2}, \quad (4.15)$$

donde $\langle a, b \rangle = a^T b$ denota el producto escalar euclídeo. Es posible llegar a las mismas expresiones de los coeficientes α y β reformulando el problema de minimización en términos de proyectores. Además, la interpretación del esquema iterativo (4.9) como un método de proyección permitirá simplificar las expresiones de los coeficientes dadas en las ecuaciones (4.14) y (4.15).

4.3.1 Interpretación como método de proyección

El método variacional definido en (4.9) y (4.10) se puede plantear de una forma similar a las ecuaciones (4.2) y (4.3) que definen un método de proyección general, en los términos siguientes:

$$\begin{array}{ll} \text{Encontrar} & x_{k+1} = x_k + \delta_k, \\ \text{tal que} & r_{k+1} = r_k - A\delta_k \perp \mathcal{L}_k. \end{array} \quad (4.16)$$

con $\delta = \alpha_k r_k + \beta_k d_k$. Se tiene en este caso que el subespacio de búsqueda de soluciones corresponde a

$$\mathcal{K}_k = \text{Env}\{r_k, d_k\}. \quad (4.17)$$

De la misma forma, el subespacio de restricciones viene dado por

$$\mathcal{L}_k = A\mathcal{K}_k = \text{Env}\{Ar_k, Ad_k\}. \quad (4.18)$$

Por tanto se trata de un método de proyección oblicuo. Las condiciones de Petrov-Galerkin (4.16) se formulan equivalentemente como:

$$\begin{array}{ll} \text{Encontrar} & x_{k+1} = x_k + \delta_k \in x_k + \mathcal{K}_k, \\ \text{tal que} & r_{k+1} = r_k - A\delta_k \in r_k + \mathcal{L}_k \perp \mathcal{L}_k. \end{array} \quad (4.19)$$

Capítulo 4. Método variacional

Para deducir el valor de los parámetros α_k y β_k de las ecuaciones (4.9) y (4.10) se busca, por tanto, que el nuevo residuo sea ortogonal al subespacio \mathcal{L}_k . Se obtiene el siguiente sistema de ecuaciones:

$$\left. \begin{aligned} \alpha_k \langle Ar_k, Ar_k \rangle + \beta_k \langle Ad_k, Ar_k \rangle &= \langle r_k, Ar_k \rangle \\ \alpha_k \langle Ar_k, Ad_k \rangle + \beta_k \langle Ad_k, Ad_k \rangle &= \langle r_k, Ad_k \rangle \end{aligned} \right\}, \quad (4.20)$$

que coinciden con las ecuaciones (4.13) obtenidas minimizando el residuo. Por tanto las soluciones del sistema de ecuaciones (4.20) coinciden con las expresiones de los coeficientes (4.14) y (4.15).

Estos coeficientes se pueden simplificar. Nótese que

$$\begin{aligned} Ad_k &= A(x_k - x_{k-1}) = Ax_k - Ax_{k-1} \\ &= b - Ax_{k-1} - b + Ax_k = r_{k-1} - r_k \\ &= \alpha_{k-1} Ar_{k-1} + \beta_{k-1} Ad_{k-1}. \end{aligned}$$

Si se multiplican escalarmente los vectores r_k y Ad_k , se tiene,

$$\langle r_k, Ad_k \rangle = \alpha_{k-1} \langle r_k, Ar_{k-1} \rangle + \beta_{k-1} \langle r_k, Ad_{k-1} \rangle = 0, \quad (4.21)$$

donde se ha hecho uso de la ortogonalidad entre r_k y el subespacio $\mathcal{L}_{k-1} = \text{Env}\{Ar_{k-1}, Ad_{k-1}\}$. La interpretación geométrica de este resultado es ahora evidente. Debido a que

$$r_k = r_{k-1} - Ad_k, \quad (4.22)$$

se deduce que Ad_k es la proyección ortogonal del residuo r_{k-1} en el subespacio \mathcal{L}_{k-1} . De esta forma $\langle r_k, Ad_k \rangle = 0$, obteniéndose las nuevas expresiones para los coeficientes α_k y β_k dadas por

$$\begin{aligned} \alpha_k &= \frac{\langle Ar_k, r_k \rangle \langle Ad_k, Ad_k \rangle}{\langle Ar_k, Ar_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ar_k, Ad_k \rangle^2}, \\ \beta_k &= \frac{-\langle Ar_k, r_k \rangle \langle Ar_k, Ad_k \rangle}{\langle Ar_k, Ar_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ar_k, Ad_k \rangle^2}. \end{aligned} \quad (4.23)$$

A continuación se estudian las propiedades de convergencia del método variacional (4.9) y (4.10) con los coeficientes definidos en (4.23).

4.3. Método variacional

Observación 51 *Matricialmente se tienen las siguientes ecuaciones para la solución aproximada que proporciona el método iterativo y el residuo en la etapa $k + 1$:*

$$\begin{aligned} x_{k+1} &= x_k + A^{-1}Pr_k \\ r_{k+1} &= (I - P)r_k \end{aligned} \quad (4.24)$$

donde la matriz P de tamaño $k \times k$ se define como

$$P = AV \left((AV)^T AV \right)^{-1} (AV)^T, \quad (4.25)$$

siendo V la matriz de tamaño $k \times 2$ definida como

$$V = \begin{bmatrix} r_k & d_k \end{bmatrix}.$$

Denotando las matrices $V = [r_k, d_k]$ y $W = [Ar_k, Ad_k]$, se tiene por la proposición 50 de la página 122 que las ecuaciones (4.24) y (4.25), y por tanto el método de proyección, están bien definidas. En este caso la ecuación (4.8) se escribe como (4.25), que corresponde a la matriz de la proyección ortogonal en el subespacio $\mathcal{L}_k = AK_k$. De esta forma se garantiza que el vector residuo r_{k+1} (ecuación (4.24)) es el vector con norma-2 mínima en el espacio afín $r_k + AK_k$.

4.3.2 Propiedades de convergencia

De la ecuación (4.23) se deduce que los coeficientes están bien definidos si el denominador no se anula durante el proceso de iteración, es decir, si

$$\langle Ar_k, Ar_k \rangle \langle Ad_k, Ad_k \rangle - \langle Ar_k, Ad_k \rangle^2 \neq 0. \quad (4.26)$$

o lo que es lo mismo,

$$\|Ar_k\| \|Ad_k\| \neq |\langle Ar_k, Ad_k \rangle|.$$

Por la desigualdad de Cauchy-Schwarz la igualdad en (4.26) se da sólo si los vectores Ar_k y Ad_k son linealmente dependientes.

Capítulo 4. Método variacional

La pregunta que surge de forma natural es la siguiente. Cuándo los vectores Ar_k y Ad_k son linealmente dependientes?. Teniendo en cuenta el campo de valores de una matriz (definición 1, pág. 5) el siguiente resultado establece que los coeficientes α_k y β_k en (4.23) están bien definidos siempre que la solución exacta del sistema no se ha alcanzado, respondiendo así a la cuestión anterior.

Teorema 52 *Sea $\mathcal{F}(A)$ el campo de valores de A tal que $0 \notin \mathcal{F}(A)$. Los coeficientes definidos en las ecuaciones (4.23) correspondientes al método iterativo (4.9) y (4.10) están bien definidos si, y sólo si x_{k-1} no es la solución de (4.1).*

Demostración. Supongamos que $Ar_k = cAd_k$, $c \in \mathbb{R}$, ($r_k = cd_k$ por la invertibilidad de A). Bajo esta hipótesis, y debido a que (4.21) muestra que r_k es ortogonal a Ad_k , se tiene que

$$r_k \perp \mathcal{L}_k = \text{Env}\{Ar_k, Ad_k\}.$$

Entonces r_k es ortogonal también a Ar_k , es decir, $\langle r_k, Ar_k \rangle = 0$. Debido a que $0 \notin \mathcal{F}(A)$, se tiene entonces

$$\langle r_k, Ar_k \rangle = 0 \quad \text{si, y sólo si } r_k = 0,$$

y de esta manera $x_k = A^{-1}b$, es decir, se ha alcanzado la solución del sistema (4.1). Aún más, ya que $r_k = 0 = cd_k = c(r_{k-1} - r_k)$, entonces $r_{k-1} = r_k$, y la solución del sistema se ha alcanzado en la etapa $k-1$, es decir $x_{k-1} = A^{-1}b$. Por tanto, si x_{k-1} (y también x_k) no es la solución del sistema, los coeficientes están bien definidos, es decir, $Ar_k \neq cAd_k$.

Para demostrar la condición necesaria, asumamos que la ecuación (4.26) se cumple. Supóngase ahora que $x_{k-1} = A^{-1}b$. Entonces $r_{k-1} = 0$ (y $r_k = 0$), obteniéndose $Ad_k = 0$ (pues $d_k = x_k - x_{k-1} = 0$) en contra de la hipótesis. ■

4.3. Método variacional

A continuación se estudiará bajo qué condiciones el método converge a la solución del sistema, para cualquier valor de r_0 y d_0 . Antes de abordar esta cuestión se reescribirán las ecuaciones del esquema iterativo.

En (4.21) se comprobó que $r_k \perp Ad_k$. Debido a que el nuevo residuo r_{k+1} se calcula a partir del residuo anterior r_k restándole la proyección ortogonal en $\mathcal{L}_k = \text{Env}\{Ar_k, Ad_k\}$ (ecuación (4.24)), se puede realizar la misma operación restándole su proyección ortogonal en una dirección previamente ortogonalizada con respecto al vector Ad_k , es decir, en la dirección $A\bar{r}_k$ donde

$$\bar{r}_k = r_k - b_k d_k, \quad (4.27)$$

donde

$$b_k = \frac{\langle Ar_k, Ad_k \rangle}{\langle Ad_k, Ad_k \rangle}.$$

Ahora, los vectores solución y residuo responden a las expresiones

$$x_{k+1} = x_k + a_k A\bar{r}_k, \quad (4.28)$$

donde

$$a_k = \frac{\langle r_k, A\bar{r}_k \rangle}{\langle A\bar{r}_k, A\bar{r}_k \rangle},$$

y

$$r_{k+1} = r_k - a_k A\bar{r}_k. \quad (4.29)$$

De las ecuaciones (4.27), (4.28) y (4.29) se tiene que

$$\langle r_k, A\bar{r}_k \rangle = \langle r_k, Ar_k \rangle, \quad (4.30)$$

$$\langle A\bar{r}_k, A\bar{r}_k \rangle = \langle Ar_k, Ar_k \rangle - \frac{\langle Ar_k, Ad_k \rangle^2}{\langle Ad_k, Ad_k \rangle}. \quad (4.31)$$

Se puede comprobar fácilmente que r_{k+1} en (4.29) es equivalente a la ecuación (4.10), con los coeficientes α_k, β_k definidos en (4.23).

Capítulo 4. Método variacional

Observación 53 *De la ecuación (4.31) se cumple,*

$$\|A\bar{r}_k\|^2 \leq \|Ar_k\|^2.$$

En el caso en que $0 \notin \mathcal{F}(A)$, el teorema 52 muestra que si la solución del sistema de ecuaciones no se ha alcanzado, los coeficientes (4.23) están bien definidos, y de esta manera la norma-2 del residuo se reduce en cada iteración como establece el siguiente resultado.

Teorema 54 *La norma-2 del residuo en el método iterativo (4.9) y (4.10) con coeficientes (4.23) decrece monótonamente para cualquier vector inicial r_0 y d_0 , $r_0 \neq d_0$, si, y solo si $0 \notin \mathcal{F}(A)$.*

Demostración. La prueba es similar a la del teorema 2.2.1 de [42]. Supongamos que los coeficientes están bien definidos, es decir, se cumple la ecuación (4.26). Si no fuera el caso, la solución se habría alcanzado (teorema 52). Si $0 \in \mathcal{F}(A)$ y r_k es un vector que satisface $\langle r_k, A\bar{r}_k \rangle = \langle r_k, Ar_k \rangle \equiv r_k^T Ar_k = 0$ (véase la ecuación (4.30)), entonces $\|r_{k+1}\| = \|r_k\|$ (ecuación (4.29)), en contra de la hipótesis. Para la condición suficiente, si $0 \notin \mathcal{F}(A)$ entonces $\langle r_k, A\bar{r}_k \rangle$ no puede ser 0 y $\|r_{k+1}\| < \|r_k\|$. ■

Observación 55 *Las condiciones del teorema 54 se satisfacen cuando A es una matriz definida positiva.*

El teorema 54 establece que el residuo en el método iterativo (4.9) se reduce en cada iteración. El siguiente resultado muestra además que la reducción que se produce está acotada por un factor mínimo fijo.

Teorema 56 *El método iterativo (4.9) y (4.10) con coeficientes (4.23) converge a la solución del sistema (4.1), $A^{-1}b$, para cualquier vector inicial r_0*

4.3. Método variacional

y $d_0, r_0 \neq d_0$, si, y solo si $0 \notin \mathcal{F}(A)$. En este caso, la norma-2 del residuo satisface

$$\|r_{k+1}\| \leq \sqrt{1 - \frac{d^2}{\|A\|^2}} \|r_k\|,$$

para todo k , donde d es la distancia del origen al campo de valores de la matriz A .

Demostración. El teorema 54 muestra que el residuo decrece monótonamente. Por tanto, para demostrar que el método iterativo converge a la solución del sistema basta ahora con demostrar que además lo hace reduciéndose en cada iteración al menos por un cantidad fija. Ya que el campo de valores de una matriz es un conjunto cerrado [52], si $0 \notin \mathcal{F}(A)$ entonces existe un entero positivo d , la distancia del origen al campo de valores $\mathcal{F}(A)$, de forma que $|\frac{y^T A^T y}{y^T y}| \geq d$ para todos los vectores $y \neq 0$. De la ecuación (4.29), calculando el producto escalar del vector r_{k+1} consigo mismo, y usando las ecuaciones (4.30) y (4.31), y la observación 53 se obtiene,

$$\begin{aligned} \langle r_{k+1}, r_{k+1} \rangle &= \langle r_k, r_k \rangle - \frac{\langle r_k, A\bar{r}_k \rangle^2}{\langle A\bar{r}_k, A\bar{r}_k \rangle} = \langle r_k, r_k \rangle - \frac{|\langle r_k, A\bar{r}_k \rangle|^2}{\langle A\bar{r}_k, A\bar{r}_k \rangle} \\ &= \langle r_k, r_k \rangle - \frac{|\langle r_k, Ar_k \rangle|^2}{|r_k^T r_k|^2} \frac{\|r_k\|^4}{\|A\bar{r}_k\|^2} \\ &= \|r_k\|^2 \left(1 - \left| \frac{r_k^T A^T r_k}{r_k^T r_k} \right|^2 \cdot \frac{\|r_k\|^2}{\|A\bar{r}_k\|^2} \right) \\ &\leq \|r_k\|^2 \left(1 - \left| \frac{r_k^T A^T r_k}{r_k^T r_k} \right|^2 \cdot \frac{\|r_k\|^2}{\|Ar_k\|^2} \right) \\ &\leq \|r_k\|^2 \left(1 - d^2 \cdot \frac{\|r_k\|^2}{\|Ar_k\|^2} \right) \\ &\leq \|r_k\|^2 \left(1 - \frac{d^2}{\|A\|^2} \right). \end{aligned}$$

Hacemos notar que $d \leq \nu(A) \leq \|A\|$, donde $\nu(A)$ es el radio numérico de A

Capítulo 4. Método variacional

(definición 1, pág. 5). Entonces,

$$\|r_{k+1}\| \leq \sqrt{1 - \frac{d^2}{\|A^2\|}} \|r_k\|.$$

■

Observación 57 *Un resultado similar se puede encontrar en [42] para el método ORTHOMIN(1). Del teorema 56 se tiene,*

$$\langle r_{k+1}, r_{k+1} \rangle \leq \|r_k\|^2 \left(1 - \left| \frac{r_k^T A^T r_k}{r_k^T r_k} \right|^2 \cdot \frac{\|r_k\|^2}{\|Ar_k\|^2} \right) = \langle \hat{r}_{k+1}, \hat{r}_{k+1} \rangle ,$$

donde \hat{r}_{k+1} es el residuo que produce el método ORTHOMIN(1) (ver [42], ecuación (2.8)).

Cuando la matriz del sistema A es simétrica la norma-2 del residuo, $\|r_{k+1}\|$, es minimizada sobre el espacio afín de dimensión $(k+1)$

$$r_0 + \text{Env}\{A\bar{r}_0, \dots, A\bar{r}_k\} ,$$

como establece el siguiente resultado.

Teorema 58 *Sea A simétrica, $0 \notin \mathcal{F}(A)$, y considérese el método variacional (4.4). Si la solución del sistema $A^{-1}b$ no se ha alcanzado en el paso k , entonces*

$$\langle r_{k+1}, A\bar{r}_j \rangle = \langle A\bar{r}_{k+1}, A\bar{r}_j \rangle = 0 , \quad \forall j \leq k$$

Además, de todos los vectores del espacio afín

$$r_0 + \text{Env}\{A\bar{r}_0, A\bar{r}_1, \dots, A\bar{r}_k\} , \tag{4.32}$$

r_{k+1} es el vector con norma-2 mínima. También se tiene que la solución se obtiene como máximo en la etapa n , es decir, $r_n = 0$, donde n es la dimensión de la matriz A .

4.3. Método variacional

Demostración. Por el teorema 54, los coeficientes del método iterativo están bien definidos hasta la etapa $j \leq k - 1$. La demostración se hará por inducción sobre k . Por construcción (ecuaciones (4.27), (4.28) y (4.29)) se tiene

$$\langle r_1, A\bar{r}_0 \rangle = \langle A\bar{r}_1, Ad_1 \rangle = \langle A\bar{r}_1, A\bar{r}_0 \rangle = 0,$$

y en general,

$$\begin{aligned} A\bar{r}_k &= Ar_k - \frac{\langle Ar_k, Ad_k \rangle}{\langle Ad_k, Ad_k \rangle} Ad_k = Ar_k - \frac{\langle Ar_k, a_{k-1} A\bar{r}_{k-1} \rangle}{a_{k-1}^2 \langle A\bar{r}_{k-1}, A\bar{r}_{k-1} \rangle} a_{k-1} A\bar{r}_{k-1} \\ &= Ar_k - \frac{\langle Ar_k, A\bar{r}_{k-1} \rangle}{\langle A\bar{r}_{k-1}, A\bar{r}_{k-1} \rangle} A\bar{r}_{k-1}, \end{aligned}$$

y por tanto $A\bar{r}_k \perp A\bar{r}_{k-1}$. Supongamos que $\langle r_k, A\bar{r}_j \rangle = \langle A\bar{r}_k, A\bar{r}_j \rangle = 0$, para todo $j \leq k - 1$. Los coeficientes en la etapa $k + 1$ se eligen de forma que $\langle r_{k+1}, A\bar{r}_k \rangle = \langle A\bar{r}_{k+1}, A\bar{r}_k \rangle = 0$. Para $j \leq k - 1$ se tiene,

$$\langle r_{k+1}, A\bar{r}_j \rangle = \langle r_k - a_k A\bar{r}_k, A\bar{r}_j \rangle = 0 \quad (4.33)$$

por la hipótesis de inducción. Además,

$$\begin{aligned} \langle A\bar{r}_{k+1}, A\bar{r}_j \rangle &= \langle Ar_{k+1} - \frac{\langle Ar_{k+1}, A\bar{r}_k \rangle}{\langle A\bar{r}_k, A\bar{r}_k \rangle} A\bar{r}_k, A\bar{r}_j \rangle \\ &= \langle Ar_{k+1}, A\bar{r}_j \rangle = \langle Ar_{k+1}, a_j^{-1}(r_j - r_{j+1}) \rangle \\ &= \langle Ar_{k+1}, a_j^{-1}(\bar{r}_j + b_j d_j - \bar{r}_{j+1} - b_{j+1} d_{j+1}) \rangle \\ &= a_j^{-1} \langle r_{k+1}, A(\bar{r}_j + b_j d_j - \bar{r}_{j+1} - b_{j+1} d_{j+1}) \rangle = 0, \end{aligned}$$

donde en la última igualdad se ha aplicado (4.33). Nótese que por la ecuación (4.30), y las hipótesis del teorema, $a_j \neq 0$. Además se comprueba fácilmente que, de la ecuación (4.29), se tiene

$$r_{k+1} \in r_0 + \text{Env}\{A\bar{r}_0, \dots, A\bar{r}_k\}.$$

Ya que r_{k+1} es ortogonal al subespacio $\text{Env}\{A\bar{r}_0, \dots, A\bar{r}_k\}$, se sigue que r_{k+1} es el vector con norma mínima en (4.32). Por tanto, si la solución exacta no se ha alcanzado antes de la etapa n , entonces $r_n = 0$. ■

Observación 59 *Si se toma $d_0 = 0$, $\bar{r}_0 = r_0$, se puede comprobar fácilmente que*

$$\text{Env}\{A\bar{r}_0, A\bar{r}_1, \dots, A\bar{r}_k\} = \text{Env}\{Ar_0, A^2r_0, \dots, A^{k+1}r_0\},$$

y en este caso r_{k+1} es el vector de norma mínima en el espacio afín

$$r_0 + \text{Env}\{Ar_0, A^2r_0, \dots, A^{k+1}r_0\}.$$

De esta manera, el método variacional es equivalente al método del residuo mínimo MINRES [78].

4.3.3 Implementación práctica

En esta sección se propone una implementación del método variacional (4.9) y (4.10). En el teorema 52 de la sección anterior se estableció que los coeficientes α_k y β_k (ecuación (4.23)), están bien definidos, es decir, su denominador no se anula, siempre que la solución del sistema no se haya alcanzado. El teorema es válido bajo la hipótesis de que el 0 no pertenece al campo de valores de la matriz A , $\mathcal{F}(A)$. A continuación se propone un algoritmo más estable y válido para matrices invertibles en general, de forma que los coeficientes α_k y β_k siempre se puedan calcular.

Como se expuso en la sección 4.3.1 (pág. 124), las condiciones de optimización de Petrov-Galerkin para el método variacional (4.9) se expresan de la siguiente manera (véase ecuación (4.19))

$$\begin{array}{ll} \text{Encontrar} & x_{k+1} \in x_k + \text{Env}\{r_k, d_k\}, \\ \text{tal que} & r_{k+1} (\in r_k + \text{Env}\{Ar_k, Ad_k\}) \perp \text{Env}\{Ar_k, Ad_k\}. \end{array} \quad (4.34)$$

Las condiciones (4.34) garantizan que r_{k+1} es el vector en el espacio afín $r_k + \text{Env}\{Ar_k, Ad_k\}$ con norma-2 mínima. Por tanto, el método variacional en la etapa $k + 1$ encuentra una solución de la forma

$$x_{k+1} = x_k + [r_k, d_k]y_k$$

4.3. Método variacional

con

$$y_k = \begin{bmatrix} \alpha_k \\ \beta_k \end{bmatrix} .$$

Los coeficientes en y_k se escogen de manera que se minimice el residuo en la etapa $k + 1$, es decir,

$$y_k = \arg \min_y \|r_k - A[r_k, d_k]y\| . \quad (4.35)$$

donde $\arg \min_y \|\cdot\|$ es el valor de y que minimiza la norma $\|\cdot\|$. Debido a que la matriz A es invertible, el problema de minimos cuadrados (4.35) siempre tiene solución única. Para resolver la ecuación (4.35) se construye una base ortonormal del subespacio vectorial de \mathbb{R}^n ,

$$\text{Env}\{r_k, Ar_k, Ad_k\} , \quad (4.36)$$

que denotaremos como $\mathcal{B} = \{v_1, v_2, v_3\}$. Los vectores de la base \mathcal{B} se calculan como sigue a continuación:

$$\begin{aligned} \beta v_1 &= r_k , \\ h_{21}v_2 &= (Ar_k - h_{11}v_1) , \\ h_{32}v_3 &= (Ad_k - h_{12}v_1 - h_{22}v_2) , \end{aligned} \quad (4.37)$$

donde los coeficientes β, h_{ij} ($i = 1, 2, 3, j = 1, 2$) vienen dados por

$$\begin{aligned} \beta &= \|r_k\| , \\ h_{11} &= \langle Ar_k, v_1 \rangle, \quad h_{21} = \|(Ar_k - h_{11}v_1)\| , \\ h_{12} &= \langle Ad_k, v_1 \rangle, \quad h_{22} = \langle Ad_k, v_2 \rangle, \quad h_{32} = \|(Ad_k - h_{12}v_1 - h_{22}v_2)\| . \end{aligned} \quad (4.38)$$

Las ecuaciones (4.37) y (4.38) corresponden al algoritmo de Gram-Schmidt para construir una base ortogonal del subespacio (4.36) [39, pág. 218]. Matricialmente estas ecuaciones se escriben de la forma

$$A[r_k, d_k] = Q_k H_{3,2} , \quad (4.39)$$

donde las columnas de la matriz Q_k son los vectores de la base ortonormal \mathcal{B} , $Q_k = [v_1, v_2, v_3]$, y la matriz $H_{3,2}$ es la matriz de Hessenberg triangular superior de dimensión 3×2 dada por,

$$H_{3,2} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ 0 & h_{32} \end{bmatrix} . \quad (4.40)$$

Capítulo 4. Método variacional

Retomando la ecuación (4.35) para el vector de incógnitas y_k , y utilizando la relación (4.39), se obtiene el problema equivalente de mínimos cuadrados siguiente:

$$\begin{aligned} y_k &= \arg \min_y \|r_k - A[r_k, d_k]y\| \\ &= \arg \min_y \|r_k - Q_k H_{3,2}y\| \\ &= \arg \min_y \|Q_k(\beta e_1 - H_{3,2}y)\| \\ &= \arg \min_y \|\beta e_1 - H_{3,2}y\| \end{aligned} \tag{4.41}$$

donde $e_1 = (1, 0, 0)^T$. Hemos obtenido por tanto un problema de mínimos cuadrados de dimensión 3×2 , frente al problema de dimensión $n \times 2$ original (ecuación (4.35)), que se puede resolver fácilmente mediante un método directo. Además, la ecuación (4.21) (pág. 125) muestra que $r_k \perp Ad_k$, y en consecuencia el coeficiente h_{12} en (4.40) vale 0, reduciéndose así la complejidad del sistema (4.41). Una posible implementación del método variacional se muestra en el algoritmo 3.

Algoritmo 3 Implementación práctica del método variacional (4.9) y (4.10).

1. **Elegir** x_0 y d_0 . Calcular $r_0 := b - Ax_0$, $\beta := \|r_0\|$ y $v_1 := r_0/\beta$.
 2. **Para** $k = 1, 2, \dots$,
 - (a) **Calcular** $w_1 := Ar_k$ y $w_2 := Ad_k$
 - (b) **Para** $j = 1, 2$
 - i. **Para** $i = 1, \dots, j$
 - ii. $h_{ij} := \langle w_j, v_i \rangle$
 - iii. $w_j := w_j - h_{ij}v_j$
 - (c) $h_{j+1,j} := \|w_j\|$
 - (d) $v_{j+1} := w_j/h_{j+1,j}$
 3. **Calcular** y_k que minimiza $\|\beta e_1 - H_{3,2}y\|$
 4. **Calcular** $x_{k+1} := x_k + [r_k, d_k]y_k$, $r_{k+1} := r_k - Q_k H_{3,2}y_k$ y $d_{k+1} = x_{k+1} - x_k$
-

4.4 Método de segundo grado acelerado

En esta sección se presenta el algoritmo utilizado en los experimentos numéricos para acelerar el método de segundo grado mediante la técnica variacional descrita en la sección 4.3.

El método de segundo grado utilizado preferentemente corresponde al método de segundo grado B (algoritmo 2, pág. 69), analizado en el capítulo 3. Este método se utiliza para resolver los sistemas de ecuaciones lineales que surgen de la integración de la ecuación neutrónica que se escriben de la forma,

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} \psi_1^{n+1} \\ \psi_2^{n+1} \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}, \quad (4.42)$$

Se comprobó que obtenía mejores resultados que el método de segundo grado A (algoritmo 1, pág. 68).

Para acelerar este método iterativo se utiliza la técnica variacional descrita en la sección 4.3. La aceleración se realiza de la forma siguiente: cada r iteraciones del método de segundo grado con parámetro de extrapolación ω , se realizan q iteraciones del método variacional descrito (algoritmo 3). Al método acelerado se denominará $ASD(\omega, r, q)$, y se muestra en el algoritmo 4.

Observación 60 *La aceleración del método de segundo grado A (algoritmo 1, pág. 68) mediante el método variacional se realiza de la misma forma, obteniéndose un algoritmo similar al algoritmo 4 pero con el paso 4.(c) de la forma,*

$$T_{22}\psi_2^{(l+1)} = E_2 - T_{21}(\omega\psi_1^{(l)} + (1 - \omega)\psi_1^{(l-1)}).$$

El método de segundo grado A acelerado se denotará mediante $ASD^(\omega, r, q)$. En el capítulo 3 se comprobó experimentalmente que el método de segundo grado B empleaba menos tiempo de simulación que el método A, hecho*

Capítulo 4. Método variacional

Algoritmo 4 Método de segundo grado acelerado, $ASD(\omega, r, q)$.

1. **Elegir** $\psi_2^{(0)}$.
 2. **Resolver** $T_{11}\psi_1^{(1)} = E_1 - T_{12}\psi_2^{(0)}$.
 3. **Resolver** $T_{22}\psi_2^{(1)} = E_2 - T_{21}\psi_1^{(1)}$.
 4. **Para** $l = 1, 2, \dots, r$
 - (a) **Elegir** ω .
 - (b) **Resolver** $T_{11}\psi_1^{(l+1)} = E_1 - T_{12}(\omega\psi_2^{(l)} + (1 - \omega)\psi_2^{(l-1)})$.
 - (c) **Resolver** $T_{22}\psi_2^{(l+1)} = E_2 - T_{21}(\omega\psi_1^{(l+1)} + (1 - \omega)\psi_1^{(l)})$.
 - (d) **Si** se cumple criterio de parada, FIN.
 5. **Acelerar** con q iteraciones del algoritmo 3, tomando $x_0 = \psi^{(r+1)}$, $r_0 = b - A\psi^{(r+1)}$ y $d_0 = \psi^{(r+1)} - \psi^{(r)}$.
 6. **Volver** al punto 4.
-

que se repite para sus versiones aceleradas como muestran los experimentos numéricos que se presentan en la próxima sección.

4.5 Experimentos numéricos

Para evaluar la eficiencia de los métodos de segundo grado acelerados descritos en las secciones anteriores se ha analizado el transitorio 1 correspondiente al reactor bidimensional TWIGL (curva 2.5, pág. 55) y el transitorio tridimensional de Langenbuch (curva 2.8, pág. 58).

En cuanto a los códigos y librerías utilizadas, son válidas las consideraciones de los capítulos anteriores. Las pruebas se han realizado en la máquina HP Exemplar S Class.

Como criterio de parada se han utilizado dos test de error diferentes. El primero de ellos, referenciado como *Test 1*, comprueba el valor relativo del

4.5. Experimentos numéricos

residuo en cada iteración y responde a la ecuación,

$$\|r_i\| \leq \|r_0\| \cdot rtol + atol , \quad (4.43)$$

donde $rtol$ y $atol$ son la tolerancia relativa y absoluta. El segundo criterio es similar al anterior pero utiliza el cambio previo en la solución según la ecuación,

$$\|d_i\| \leq \|d_0\| \cdot rtol + atol , \quad (4.44)$$

y será referenciado como *Test 2*.

4.5.1 Transitorio bidimensional

En esta sección se presentan los experimentos numéricos correspondientes a la simulación del transitorio 1 del reactor bidimensional TWIGL (figura 2.4, pág. 53) mediante el método de segundo grado B acelerado $ASD(\omega, r, q)$ detallado en el algoritmo 4 (sección 4.4).

El reactor ha sido discretizado utilizando el método de las diferencias finitas con tamaño de malla uniforme $h_x = h_y = 3 \text{ cm}$, y también mediante el método de colocación nodal con nodos cuadrados de tamaño 8 cm . Esta discretización se denotará como $Nodal(p)$, donde p indica el número de polinomios de Legendre utilizados en el desarrollo del flujo neutrónico en cada nodo. Se han utilizado los valores $p = 3$ y $p = 4$. En la tabla 3.1 (pág. 112) se mostraron el tamaño y el número de elementos no nulos de las matrices del sistema de ecuaciones (4.42) que hay que resolver en cada paso de tiempo para cada una de estas discretizaciones.

Para la discretización temporal de la ecuación de la difusión neutrónica se utilizará un método en diferencias hacia atrás de 1 paso, con paso de integración de 1,25 milisegundos. Por tanto se necesita un número total de 160 pasos para simular el transitorio completo. Como potencia pico de referencia se eligió el valor 2.17, calculada con una discretización espacial

Capítulo 4. Método variacional

Nodal(4) y un paso de integración $h = 6.25 \times 10^{-4}$ segundos. La curva de evolución del transitorio se puede ver en la figura 2.5 (pág. 55).

El método $ASD(\omega, r, q)$ se comparará con el método de segundo grado A acelerado, que será denotado como $ASD^*(\omega, r, q)$. Para resolver las iteraciones internas en los métodos acelerados $ASD(\omega, r, q)$ y $ASD^*(\omega, r, q)$, es decir, los sistemas de ecuaciones correspondientes a las matrices T_{11} y T_{22} (pasos 4.(b) y 4.(c) en el algoritmo 4), se utilizará el método del gradiente conjugado preconditionado con el método de Jacobi (JCG). Los experimentos realizados en el capítulo 3 mostraron que era el método más apropiado para resolver estos sistemas. Además, esos mismos experimentos mostraron que $\omega = 1,5$ era el mejor valor del factor de extrapolación, aunque ligeras variaciones del parámetro en el intervalo $[1, 2[$ no degradaban significativamente la convergencia.

También se ha realizado un estudio comparativo con otros métodos basados en subespacios de Krylov (ver la sección 1.4.2, del capítulo 1). En concreto con los métodos GMRES(k), TFQMR y BiCGSTAB preconditionados. En el caso del GMRES(k), la dimensión máxima del subespacio de Krylov será $k=10$ o $k=20$. Como preconditionadores se utilizarán las factorizaciones incompletas ILU0 y ILUT (sección 1.4.4). Es importante resaltar que estos preconditionadores se calculan para los bloques diagonales T_{11} y T_{22} de la matriz de coeficientes sólo una vez en el primer paso del transitorio, ahorrando tiempo y memoria en el paso de preconditionado. Además, se ha observado que el número de iteraciones es similar al caso del preconditionador calculado sobre la matriz T completa. De esta forma el tiempo de simulación es menor manteniéndose la velocidad de convergencia.

En las tablas de resultados (*rtol*, *atol*) indica la tolerancia relativa y absoluta utilizada para los test de error (4.43) y (4.44). En el caso de los métodos $ASD(\omega, r, q)$ y $ASD^*(\omega, r, q)$ método hace referencia al método utilizado en las iteraciones internas. La columna *iter.* indica el número medio de itera-

4.5. Experimentos numéricos

r	q	$tiempo\ (sg)$
1	1	35,85
2	1	34,82
2	2	43,01
3	1	30,50
3	2	43,14
3	3	32,51
4	1	30,93
4	2	44,03
4	3	32,80
4	4	45,18
5	1	29,20

Tabla 4.1: Tiempo de simulación del método $ASD(1.5, r, q)$ para diferentes valores de r y q . Discretización utilizada: método nodal con 4 polinomios.

ciones¹ empleado para resolver el sistema de ecuaciones (4.42) en cada paso de tiempo. El tiempo total de simulación en segundos y la potencia (pico) alcanzada en $t = 0,2$ segundos se indican en las columnas *tiempo* y *potencia*, respectivamente. Para los métodos BiCGSTAB, GMRES(10) y TFQMR, la columna *precond.* indica el preconditionador utilizado. Para el caso del preconditionador ILUT se indica entre paréntesis el nivel y la tolerancia de llenado.

En primer lugar se estudian dos aspectos importantes: la mejor combinación de los parámetros r y q para el método $ASD(1.5, r, q)$, y la efectividad del método variacional para acelerar el método de segundo grado B. En la tabla 4.1 se puede observar que el método $ASD(1.5, r, q)$ con $r = 5$ y $q = 1$, es decir, una iteración de aceleración por cada cinco del método de segundo grado (algoritmo 4), obtiene los mejores resultados en cuanto a tiempo de simulación. Por ello, en el resto de experimentos se empleará el método $ASD(1.5, 5, 1)$. Además, si se compara con el método de segundo grado B sin acelerar cuyos resultados se muestran en la tabla 3.4 (pág. 114), se observa que el método $ASD(1.5, 5, 1)$ tarda aproximadamente la mitad que el

¹Iteraciones externas para los métodos $ASD(\omega, r, q)$ y $ASD^*(\omega, r, q)$.

Capítulo 4. Método variacional

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
JCG(20, 5×10^{-6})	$10^{-5}/0$	6	12, 9	2, 160
JCG(100, 5×10^{-6})	$10^{-5}/0$	6	13, 0	2, 160
JCG(20, 10^{-5})	$10^{-5}/0$	6	12, 3	2, 112
JCG(20, 10^{-5})	$10^{-7}/0$	6	12, 6	2, 123

Tabla 4.2: Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización $Nodal(3)$. Criterio de parada utilizado $Test\ 2$ (ecuación (4.44)).

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
JCG(20, 5×10^{-6})	$10^{-5}/0$	6	29, 2	2, 168
JCG(100, 5×10^{-6})	$10^{-5}/0$	6	30, 5	2, 168
JCG(20, 10^{-5})	$10^{-5}/0$	6	26, 2	2, 097
JCG(20, 10^{-5})	$10^{-7}/0$	6	26, 2	2, 097

Tabla 4.3: Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización $Nodal(4)$. Criterio de parada utilizado $Test\ 2$ (ecuación (4.44)).

método B sin acelerar para simular el transitorio completo. Este hecho pone de manifiesto que el método variacional es muy efectivo en la aceleración del método de segundo grado B.

También es importante estudiar el efecto del grado de precisión con el que se resuelven las iteraciones internas. En las tablas 4.2, 4.3 y 4.4 se observa que para el método $ASD(1.5, 5, 1)$ el mejor resultado se obtiene resolviendo mediante el JCG con un máximo de iteraciones de 20. Para este método, se ha utilizado como criterio de parada el $Test\ 1$ con $rtol = 5 \times 10^{-6}$ y $atol = 0.0$.

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
JCG(20, 5×10^{-6})	$10^{-5}/0$	6	31, 4	2, 119
JCG(100, 5×10^{-6})	$10^{-5}/0$	6	31, 6	2, 119
JCG(20, 10^{-5})	$10^{-5}/0$	5	27, 6	2, 110
JCG(20, 10^{-5})	$10^{-7}/0$	6	29, 0	2, 110

Tabla 4.4: Resultados de simulación del transitorio 1 del TWIGL con el método $ASD(1.5, 5, 1)$ para la discretización en diferencias finitas. Criterio de parada utilizado $Test\ 2$ (ecuación (4.44)).

4.5. Experimentos numéricos

<i>discretización</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
<i>Nodal</i> (3)	24	41, 0	2, 160
<i>Nodal</i> (4)	24	80.1	2, 168
<i>Diferencias</i>	22	85, 6	2, 119

Tabla 4.5: Resultados de simulación del transitorio 1 del TWIGL con el método $ASD^*(1.5, 5, 1)$ para diferentes discretizaciones. Criterio de parada utilizado *Test 2* con $rtol = 10^{-5}$, $atol = 0$ (ecuación (4.44)). Método JCG(20, 5×10^{-6}) para las iteraciones internas.

Se indica como JCG(20, $5 \cdot 10^{-6}$). Es importante resolver las iteraciones internas con suficiente precisión para alcanzar un nivel de potencia satisfactorio, como se pone de manifiesto en la tabla 4.2 para el caso JCG(20, 10^{-5}). En este ejemplo, incluso resolviendo las iteraciones externas con una tolerancia del error de 10^{-7} , la potencia pico alcanzada es tan sólo de 2.123. Esto se explica por la utilización de la solución en un paso de tiempo como solución inicial en el siguiente. Si el transitorio no es muy abrupto la diferencia en las soluciones en dos pasos de tiempo consecutivos no es elevada. Por tanto, la solución inicial puede satisfacer el criterio de error en pocas iteraciones, obteniéndose una solución para el nuevo sistema de ecuaciones prácticamente idéntica a la del paso de tiempo anterior. Conclusiones similares se deducen para las discretizaciones *Nodal*(4) y diferencias finitas, tablas 4.3 y 4.4, respectivamente.

Comparando con el método $ASD^*(1.5, 5, 1)$ (método de segundo grado A acelerado) cuyos resultados se muestran en la tabla 4.5, se observa claramente que los tiempos obtenidos por el método $ASD(1.5, 5, 1)$ son sensiblemente inferiores. Este hecho está en consonancia con el comportamiento observado para los métodos de segundo grado A y B (capítulo 3).

Las tablas 4.6 a 4.14 muestran los resultados para los métodos BiCGSTAB, GMRES(k) y TFQMR.

Para el método BiCGSTAB se observa que los preconditionadores ILUT e ILU0 obtienen resultados muy parecidos para las dos discretizaciones nodales,

Capítulo 4. Método variacional

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	56	22, 3	2, 160
ILU0	<i>Test 2</i>	$10^{-5}/0$	41	18, 7	2, 160
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	27	14, 2	2, 159
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	27	14, 1	2, 159
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	37	12, 0	2, 160
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	37	12, 1	2, 161
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	32	14, 8	2, 160
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	32	15, 02	2, 160

Tabla 4.6: Resultados del método BiCGSTAB para la discretización *Nodal*(3).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	77	74, 7	2, 167
ILU0	<i>Test 2</i>	$10^{-5}/0$	60	59, 2	2, 167
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	39	43, 3	2, 167
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	38	43, 5	2, 167
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	37	40, 3	2, 168
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	39	43, 5	2, 167
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	53	46, 9	2, 169
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	39	42, 9	2, 167

Tabla 4.7: Resultados del método BiCGSTAB para la discretización *Nodal*(4).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	58	157, 13	2, 128
ILU0	<i>Test 2</i>	$10^{-5}/0$	39	109, 3	2, 128
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	30	86, 8	2, 128
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	27	78.42	2, 128
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	96	174	2, 128
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	89	163	2, 128
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	21	94, 7	2, 128
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	19	87, 73	2, 128

Tabla 4.8: Resultados del método BiCGSTAB para la discretización en diferencias finitas.

4.5. Experimentos numéricos

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
ILU0	<i>Test 1</i>	$10^{-5}/0$	118	55	2, 160
ILU0	<i>Test 2</i>	$10^{-5}/0$	77	37, 77	2, 160
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	33	20, 51	2, 150
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	33	20, 30	2, 150
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	82	23, 98	2, 150
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	85	24, 29	2, 150
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	60	28, 87	2, 152
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	60	28, 64	2, 152

Tabla 4.9: Resultados del método GMRES(10) para la discretización Nodal(3).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
ILU0	<i>Test 1</i>	$10^{-5}/0$	119	151	2, 167
ILU0	<i>Test 2</i>	$10^{-5}/0$	75	99, 3	2, 167
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	38	56, 15	2, 167
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	38	56	2, 167
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	78	60, 71	2, 165
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	79	61, 3	2, 167
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	75	80, 6	2, 167
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	75	80, 7	2, 167

Tabla 4.10: Resultados del método GMRES(20) para la discretización Nodal(4).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	<i>potencia</i>
ILU0	<i>Test 1</i>	$10^{-5}/0$	58	156	2, 128
ILU0	<i>Test 2</i>	$10^{-5}/0$	39	108	2, 128
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	30	85, 9	2, 128
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	27	77, 9	2, 128
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	97	173, 9	2, 128
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	89	162, 2	2, 128
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	21	94	2, 128
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	19	86, 7	2, 128

Tabla 4.11: Resultados del método GMRES(10) para la discretización en diferencias finitas.

Capítulo 4. Método variacional

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	63	33	2,160
ILU0	<i>Test 2</i>	$10^{-5}/0$	49	27,2	2,160
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	37	21,79	2,160
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	37	21,8	2,160
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	71	22,6	2,159
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	71	22,5	2,159
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	53	26,9	2,159
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	52	26,7	2,159

Tabla 4.12: Resultados del método TFQMR para la discretización *Nodal*(3).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	85	104	2,167
ILU0	<i>Test 2</i>	$10^{-5}/0$	72	90,9	2,167
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	51	67,3	2,167
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	51	67	2,167
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	102	66,2	2,167
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	102	66,6	2,167
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	84	81,9	2,167
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	84	82	2,167

Tabla 4.13: Resultados del método TFQMR para la discretización *Nodal*(4).

<i>precond.</i>	<i>Test</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>	potencia
ILU0	<i>Test 1</i>	$10^{-5}/0$	58	156	2,128
ILU0	<i>Test 2</i>	$10^{-5}/0$	39	107	2,128
ILU0	<i>Test 1</i>	$10^{-5}/10^{-6}$	30	85,8	2,128
ILU0	<i>Test 2</i>	$10^{-5}/10^{-6}$	27	77,1	2,128
ILUT(5,1)	<i>Test 1</i>	$10^{-5}/10^{-6}$	96	175	2,128
ILUT(5,1)	<i>Test 2</i>	$10^{-5}/10^{-6}$	95	171	2,128
ILUT(5,10 ⁻²)	<i>Test 1</i>	$10^{-5}/10^{-6}$	21	93,6	2,122
ILUT(5,10 ⁻²)	<i>Test 2</i>	$10^{-5}/10^{-6}$	19	87,1	2,128

Tabla 4.14: Resultados del método TFQMR para la discretización en diferencias finitas.

4.5. Experimentos numéricos

utilizando como criterio de parada el test de error *Test 1*. Para la discretización en diferencias es el preconditionador ILU0 con criterio de parada *Test 2* el que mejor resultados proporciona. La tolerancia de error absoluta (*atol*) se ha de tomar ligeramente inferior a la relativa, pero nunca 0 ya que en este caso el tiempo empleado en la simulación es excesivo.

Comparando el método BiCGSTAB con el método $ASD(1.5, 5, 1)$ se observa que para la discretización *Nodal*(3) los resultados en cuanto a tiempo de simulación y precisión en la solución final son prácticamente similares. Sin embargo, para las discretizaciones *Nodal*(4) y diferencias finitas el método $ASD(1.5, 5, 1)$ se muestra claramente superior, empleando aproximadamente la mitad de tiempo en la simulación del transitorio. Es importante resaltar como para este método el número de iteraciones externas permanece aproximadamente constante para las tres discretizaciones empleadas.

Para los métodos GMRES(k) y TFQMR el mejor preconditionador es ILU0. Hay que destacar que con el método GMRES(k) fue necesario utilizar una base del subespacio de Krylov con 20 términos (k=20) para poder simular el transitorio cuando la discretización espacial utilizada era *Nodal*(4) (tabla 4.10). Las conclusiones relativas al criterio de parada son similares a las del BiCGSTAB.

Por otro lado, comparando los tres métodos basados en subespacios de Krylov claramente el método que mejores resultados proporciona en cuanto a tiempo de simulación y precisión en la solución final es el método BiCGSTAB. Esta observación sigue la línea marcada en [20] donde los autores llegan a conclusiones similares.

Finalmente, hay que hacer notar que el método ASD^* presenta unas prestaciones similares a los métodos GMRES y TFQMR para la discretización en diferencias finitas.

4.5.2 Transitorio de Langenbuch

En esta sección se presentan los experimentos numéricos correspondientes a la simulación del transitorio del reactor tridimensional Langenbuch (figura 2.7, pág. 58) mediante el método de segundo grado B acelerado $ASD(1.5, 5, 1)$ detallado en el algoritmo 4 (sección 4.4) y estudiado numéricamente para el transitorio bidimensional del reactor TWIGL en la sección anterior, donde se observó que los valores $\omega = 1.5$, $r = 5$ y $q = 1$ proporcionaban los mejores resultados.

Atendiendo a los resultados de la sección anterior y del capítulo 2, que mostraban una mayor precisión de la solución mediante la discretización con el método de colocación nodal, los experimentos que se presentan a continuación se han realizado únicamente para este método. En la tabla 4.15 se muestran las diferentes discretizaciones espaciales utilizadas, así como el tamaño de las matrices del sistema de ecuaciones (4.42) que hay que resolver en cada paso de tiempo.

Para la discretización temporal de la ecuación de la difusión neutrónica se utilizará un método en diferencias hacia atrás de 1 paso, con paso de integración de 125 milisegundos. Por tanto se necesita un número total de 480 pasos para simular el transitorio completo.

Además, también se ha realizado un estudio comparativo con los métodos basados en subespacios de Krylov utilizados en la sección anterior, es decir, los métodos GMRES(k), TFQMR y BiCGSTAB preconditionados con ILU0. Al igual que antes, el preconditionador se calcula para los bloques diagonales T_{11} y T_{22} de la matriz de coeficientes sólo una vez en el primer paso del transitorio.

Como criterio de parada se han utilizado el *Test 2* para el método de segundo grado y el *Test 1* para los métodos de Krylov (ecuaciones (4.44) y (4.43)). El significado de las diferentes entradas de las tablas de resultados es el mismo que en la sección anterior. En éstas, no se hace referencia a

4.5. Experimentos numéricos

<i>discretización espacial</i>	<i>n</i>	<i>nnz</i>
<i>Nodal</i> (2)	2800	31280
<i>Nodal</i> (3)	7000	106600
<i>Nodal</i> (4)	14000	270000

Tabla 4.15: Tamaño y número de elementos no nulos de las matrices para las discretizaciones espaciales utilizadas para la simulación del transitorio de Langenbuch. n , nnz indican el tamaño y el número de elementos no nulos de las matrices a resolver en cada paso de tiempo, respectivamente.

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>
<i>ASD</i> (1.5, 5, 1)	$10^{-5}/0$	16	468
BiCGSTAB	$10^{-5}/10^{-6}$	19	565
GMRES(10)	$10^{-5}/10^{-6}$	19	576
TFQMR	$10^{-5}/10^{-6}$	40	632

Tabla 4.16: Resultados de simulación del transitorio de Langenbuch para la discretización *Nodal*(2).

la potencia alcanzada ya que los resultados mostrados corresponden a la obtención de la solución final con la misma precisión (próxima a la curva de referencia dada en la figura 2.8, pág. 59).

En la tabla 4.16 se observa que para la discretización con el método *Nodal*(2), el método *ASD*(1.5, 5, 1) emplea menos tiempo para simular el transitorio completo. Para un número de 3 polinomios (tabla 4.17) es el método BiCGSTAB el que proporciona mejores resultados seguido del GMRES, siendo el tiempo de simulación empleado por el método *ASD*(1.5, 5, 1) ligeramente superior, aunque todavía inferior al tiempo empleado por el método TFQMR. Para 4 polinomios, claramente el método BiCGSTAB obtiene los mejores resultados en tiempo de simulación, seguido del método de segundo grado acelerado.

Por otro lado, de entre los métodos basados en subespacios de Krylov es el método BiCGSTAB el que menos tiempo emplea en la simulación del transitorio completo, al igual que ocurría para el transitorio bidimensional del TWIGL.

Capítulo 4. Método variacional

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>
<i>ASD</i> (1.5, 5, 1)	$10^{-5}/0$	16	1462
BiCGSTAB	$10^{-5}/10^{-6}$	22	1200
GMRES(10)	$10^{-5}/10^{-6}$	23	1369
TFQMR	$10^{-5}/10^{-6}$	41	1805

Tabla 4.17: Resultados de simulación del transitorio de Langenbuch para la discretización *Nodal*(3).

<i>método</i>	<i>rtol/atol</i>	<i>iter.</i>	<i>tiempo (sg)</i>
<i>ASD</i> (1.5, 5, 1)	$10^{-5}/0$	16	5160
BiCGSTAB	$10^{-5}/10^{-6}$	36	4620
GMRES(20)	$10^{-5}/10^{-6}$	41	5090
TFQMR	$10^{-5}/10^{-6}$	62	5578

Tabla 4.18: Resultados de simulación del transitorio de Langenbuch para la discretización *Nodal*(4).

Finalmente, es importante resaltar que habitualmente la simulación de los transitorios, cuando se discretizan las ecuaciones mediante el método de colocación nodal, se realiza con un número pequeño de polinomios, típicamente 2 ó 3. Solamente cuando se necesita estudiar con precisión los efectos locales de la variación del flujo neutrónico en el interior del reactor se utiliza un número elevado de polinomios de Legendre.

En el capítulo 5 se presentará un método multinivel para resolver los sistemas de ecuaciones cuando se utiliza un número elevado de polinomios en la discretización mediante el método de colocación nodal.

Recapitulando lo anteriormente expuesto, se han realizado experimentos numéricos para evaluar el comportamiento del método de segundo grado acelerado, denotado como $ASD(\omega, r, q)$. Se ha determinado que los valores $\omega = 1.5$, $r = 5$ y $q = 1$ proporcionaban el mejor comportamiento del método. Así mismo, se ha comparado con tres métodos basados en subespacios de Krylov. Concretamente se ha comparado con los métodos BiCGSTAB, GMRES(k) y TFQMR preconditionados. El estudio comparativo permite concluir que el método $ASD(\omega, r, q)$ presenta la mejor relación *tiempo de sim-*

4.5. Experimentos numéricos

ulación / precisión de la solución para los transitorios bidimensionales del TWIGL. Para el transitorio tridimensional del reactor Langenbuch esto continúa siendo cierto para un número de polinomios de Legendre de 2. Para un número de polinomios mayor el método BiCGSTAB obtiene los mejores resultados. También es importante resaltar que de los tres métodos basados en subespacios de Krylov, es el método BiCGSTAB el más indicado para el tipo de problemas estudiado.

Por tanto, se puede concluir que para transitorios bidimensionales, y transitorios tridimensionales discretizados con un número de polinomios pequeño², el método $ASD(\omega, r, q)$ es competitivo frente a los métodos BiCGSTAB, GMRES(k) y TFQMR. Destacar también que la modelización bidimensional (o incluso unidimensional) es utilizada para la simulación de transitorios en reactores 3D. Recientemente en [54, 55] los autores proponen la integración de la ecuación de la difusión neutrónica en reactores 3D reduciendo la complejidad del problema a geometrías de menor dimensión (punto y unidimensionales). El paso a geometrías tridimensionales se realiza mediante un mecanismo adaptativo únicamente cuando se detectan fuertes variaciones en el flujo neutrónico.

²Se utiliza un número elevado de polinomios cuando se necesita simular con suficiente precisión los efectos de variación rápida y efectos locales del flujo neutrónico en el interior del reactor.

Capítulo 5

Métodos multinivel

5.1 Introducción

El análisis de transitorios donde aparecen grandes gradientes del flujo neutrónico requiere el uso de discretizaciones de la parte espacial de la ecuación de la difusión neutrónica suficientemente precisas. Cuando se utiliza el método de colocación nodal, la precisión se mejora aumentando el número de polinomios de Legendre utilizados para la expansión del flujo neutrónico en el interior de cada celda del reactor (capítulo 2). Ya que el tamaño de los sistemas de ecuaciones lineales a resolver aumenta considerablemente con el número de polinomios utilizados en la discretización, el coste por iteración también aumenta.

En este capítulo se presenta un método multinivel basado en el número de polinomios de Legendre utilizado en el método de colocación nodal. A partir de la solución obtenida discretizando con un número de polinomios relativamente pequeño, se pretende obtener una aproximación inicial de la solución lo suficientemente buena que permita resolver en un número menor de iteraciones el sistema de ecuaciones de mayor tamaño, es decir, aquel que se obtiene discretizando con el número máximo de polinomios. Por tanto, los algoritmos que se proponen son del tipo iteración anidada, brevemente descritos en el capítulo 1 (pág. 19).

5.2. Métodos multinivel

A diferencia de los métodos multinivel habituales, donde las diferentes mallas se obtienen cambiando el tamaño de los nodos (tamaño de la triangularización cuando se utilizan técnicas de elementos finitos, o el paso de discretización para el método en diferencias finitas), la estrategia que se presenta no está basada en esta metodología. Manteniendo un tamaño de nodo fijo, lo que caracteriza a cada malla o nivel es el número de polinomios utilizado en la discretización. Por tanto, cada nivel estará asociado a un número diferente de polinomios de Legendre.

El capítulo se estructura como sigue. En la sección 5.2 se describen los componentes básicos para la formulación de los métodos multinivel propuestos (operadores de restricción e interpolación), y dos métodos multinivel diferentes, denominados multinivel algebraico y geométrico. En la sección 5.3 se presentan los resultados de los experimentos numéricos para los transitorios bidimensionales del reactor TWIGL y el tridimensional de Langenbuch. Al final se revisan las conclusiones más destacadas.

5.2 Métodos multinivel

Considérese el sistema de ecuaciones lineales que se obtiene de la discretización de la ecuación de la difusión neutrónica y que habitualmente se denota por,

$$T\psi = E, \quad (5.1)$$

donde T es una matriz de tamaño $n \times n$ (ecuación (2.52), pág. 53). La matriz T se obtiene discretizando dichas ecuaciones en el dominio del reactor, que se denotará por Ω , mediante un número de polinomios de Legendre determinado, K . Este número K , se elige lo suficientemente grande para obtener una solución con un grado de precisión aceptable.

En los capítulos 3 y 4 se han presentado diferentes métodos iterativos

Capítulo 5. Métodos multinivel

para obtener la solución de (5.1), y que responden a la forma general,

$$\psi^{nueva} = G(T, \psi^{anteriores}, E). \quad (5.2)$$

Una forma de mejorar la convergencia de estos métodos es mediante la obtención de una solución inicial próxima a la solución del sistema (5.1), pero con un coste tan pequeño como sea posible. Esta idea es la base de los métodos de iteración anidada (pág. 19).

A continuación se detallan los componentes básicos que permitirán definir diferentes métodos multinivel para resolver el sistema de ecuaciones lineales (5.1) [15, 68]:

1. Un conjunto de m mallas o niveles Ω^k , $k = 1, \dots, m$ de tal forma que $\Omega^1 \subset \Omega^2 \subset \dots \subset \Omega^m$. Por similitud con la terminología habitual utilizada en los métodos multinivel, Ω^k denota el dominio discretizado. El nivel k corresponde a la discretización de la parte espacial de la ecuación de la difusión neutrónica con un número de polinomios de Legendre diferente e igual a $P(k)$. La entrada k -ésima del vector $P \in \mathbb{N}^m$, contiene el número de polinomios utilizados en el nivel k , con $P(m) = K$, es decir, el número máximo de polinomios de Legendre utilizados.
2. Asociado a cada nivel hay un sistema de ecuaciones lineales que hay que resolver de la forma

$$T^{(k)}\psi^{(k)} = E^{(k)}, \quad (5.3)$$

y cuya solución, obtenida mediante algún método de relajación, se utiliza como solución inicial para el siguiente nivel (más fino). Además, $T^{(k)} \in \mathbb{R}^{N \cdot n_k \times N \cdot n_k}$, y $\psi^{(k)}, E^{(k)}$ son vectores de $\mathbb{R}^{N \cdot n_k}$, donde N es el número de nodos en los que se divide el reactor, y n_k es el número de coeficientes del desarrollo del flujo neutrónico en un nodo cualquiera. Como se vió en el capítulo 2 (pág. 45) para problemas bidimensionales,

$$n_k = \frac{1}{2}P(k)(P(k) + 1),$$

5.2. Métodos multinivel

mientras que para problemas tridimensionales,

$$n_k = \frac{1}{6}P(k)(P(k) + 1)(P(k) + 2) .$$

Ω^k también se utilizará para representar el espacio vectorial al que pertenecen los vectores y matrices definidos en el nivel k , es decir, $\Omega^k \equiv \mathbb{R}^{N \cdot n_k}$, con $\Omega^m \equiv \mathbb{R}^n$. De esta forma, $T^{(k)}$ se puede interpretar como la versión (restricción) en la malla Ω^k de la matriz de coeficientes del sistema (5.1), $T = T^{(m)}$.

3. Para utilizar el vector $\psi^{(k)}$ como solución inicial para el problema asociado con una malla más fina $\Omega^{k'}$, $k' > k$, es necesario definir un operador de *interpolación* o *prolongación*, $\mathcal{I}_k^{k'} : \Omega^k \rightarrow \Omega^{k'}$. Entonces,

$$\psi^{(k')} = \mathcal{I}_k^{k'} \psi^{(k)} .$$

De igual forma, un vector $\psi^{(k')}$ en el nivel $\Omega^{k'}$ se representa en una malla más gruesa Ω^k , $k < k'$, mediante el operador de *restricción* $\mathcal{I}_{k'}^k : \Omega^{k'} \rightarrow \Omega^k$, de la forma

$$\psi^{(k)} = \mathcal{I}_{k'}^k \psi^{(k')} .$$

Como ya se ha mencionado, los algoritmos multinivel que se presentan para el método de colocación nodal se basan en la generación de diferentes niveles o mallas caracterizados por el número de polinomios de Legendre utilizados en la discretización. En cada nivel, el número de nodos en los que se divide el reactor permanece fijo, y lo que cambia es el número de polinomios utilizado para la expansión del flujo neutrónico en su interior.

Si se quiere resolver la ecuación de la difusión neutrónica con K polinomios de Legendre, para cada paso de tiempo, denotado t_{n+1} , se realiza una iteración anidada tomando como solución inicial el vector de coeficientes del paso de tiempo anterior para el nivel más fino Ω^m , es decir, $\psi^{(m)}(t_n)$. El algoritmo 5 muestra la forma general del método multinivel.

Algoritmo 5 Algoritmo de la iteración anidada general.

1. **Elegir** m niveles tales que, $\Omega^1 \subset \Omega^2 \subset \dots \subset \Omega^m$.
 2. **Elegir** $P \in \mathbb{N}^m$, el número de polinomios de Legendre en cada nivel.
 3. **Restringir al nivel más grueso:** $\psi_0^{(1)}(t_{n+1}) = \mathcal{I}_m^1 \psi^{(m)}(t_n)$.
 4. **Resolver** $T^{(1)} \psi^{(1)}(t_{n+1}) = E^{(1)}$ tomando $\psi_0^{(1)}(t_{n+1})$ como solución inicial.
 5. **Para** $k = 1, \dots, m$
 - (a) **Interpolar al siguiente nivel:** $\psi_0^{(k+1)}(t_{n+1}) = \mathcal{I}_k^{k+1} \psi^{(k)}(t_{n+1})$.
 - (b) **Resolver** $T^{(k+1)} \psi^{(k+1)}(t_{n+1}) = E^{(k+1)}$ tomando $\psi_0^{(k+1)}(t_{n+1})$ como solución inicial.
 6. **Fin.**
-

Para completar el algoritmo se necesita definir los operadores de interpolación y restricción. Además, también se ha de definir como se genera la matriz $T^{(k)}$ en el nivel k . Se proponen dos formas diferentes de generar estas matrices que dan lugar a su vez a dos algoritmos distintos: el algoritmo *multinivel algebraico* y el algoritmo *multinivel geométrico*.

5.2.1 Operadores de interpolación y restricción

Supongamos que se desea determinar el flujo neutrónico utilizando ℓ polinomios de Legendre en el paso de tiempo t_{n+1} . Sea i el nivel correspondiente al número de polinomios ℓ , es decir, $P(i) = \ell$. Sea e un nodo del reactor (ver figura 2.3, pág. 41). Si se utiliza el método de colocación nodal con la aproximación serendipita, el flujo neutrónico en el nodo responde al desarrollo dado por (véase ecuación 2.35, pág. 45),

$$\phi_e^\ell(u, v, w) = \sum_{k_1=0}^{\ell-1} \sum_{k_2=0}^{\ell-1-k_1} \sum_{k_3=0}^{\ell-1-k_1-k_2} \phi_e^{\ell; k_1, k_2, k_3} P_{k_1}(u) P_{k_2}(v) P_{k_3}(w). \quad (5.4)$$

5.2. Métodos multinivel

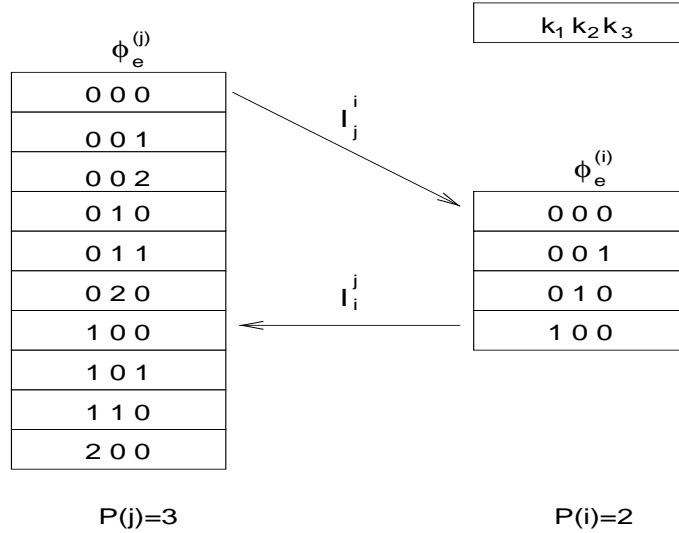


Figura 5.1: Acción de los operadores prolongación (\mathcal{I}_j^i) y restricción (\mathcal{I}_i^j) sobre los coeficientes en el nodo e . El número de polinomios de Legendre utilizado es $P(j) = 3$ y $P(i) = 2$ para los niveles Ω^j y Ω^i , respectivamente.

El vector de incógnitas, $\psi^{(i)} \in \Omega^i$, en la ecuación (5.3) estará formado por los coeficientes $\phi_e^{\ell'; k_1, k_2, k_3}$ correspondientes a todos los nodos del reactor adecuadamente ordenados. La interpolación del vector $\psi^{(i)}$ desde el nivel Ω^i al nivel Ω^j , con $j > i$ ($P(j) = \ell' > P(i) = \ell$), se define para cada nodo como,

$$\phi_e^{\ell'; k_1, k_2, k_3}(t_{n+1}) = \begin{cases} \phi_e^{\ell; k_1, k_2, k_3}(t_{n+1}) & \text{si } k_1 + k_2 + k_3 < \ell - 1, \\ \phi_e^{\ell'; k_1, k_2, k_3}(t_n) & \text{si } k_1 + k_2 + k_3 \geq \ell - 1, \end{cases} \quad (5.5)$$

con $k_1 = 0, \dots, \ell' - 1$, $k_2 = 0, \dots, \ell' - 1 - k_1$ y $k_3 = 0, \dots, \ell' - 1 - k_1 - k_2$. Por tanto, para aquellos coeficientes del flujo neutrónico para los que no se dispone de una actualización proveniente del nivel inferior en el paso de tiempo t_{n+1} , se utilizan los del paso de tiempo anterior y en el mismo nivel.

De la misma forma, la restricción del vector $\psi^{(j)}$ definido en el nivel $\Omega^{(j)}$ al nivel $\Omega^{(i)}$, $i < j$, se define como,

$$\phi_e^{\ell; k_1, k_2, k_3}(t_{n+1}) = \phi_e^{\ell'; k_1, k_2, k_3}(t_{n+1}), \quad (5.6)$$

con $k_1 = 0, \dots, \ell - 1$, $k_2 = 0, \dots, \ell - 1 - k_1$ y $k_3 = 0, \dots, \ell - 1 - k_1 - k_2$.

Capítulo 5. Métodos multinivel

Ambos operadores responden a la operación de inyección que simplemente copia coeficientes en las posiciones adecuadas sin ningún tipo de promediado. Su elección se justifica de forma intuitiva por el hecho de que cualquier función de la forma

$$P_{k_1}(u)P_{k_2}(v)P_{k_3}(w) ,$$

que forme parte de la expansión (5.4) en dos niveles diferentes, $\ell \neq \ell'$, debe tener el mismo coeficiente, es decir, $\phi_e^{\ell; k_1, k_2, k_3} = \phi_e^{\ell'; k_1, k_2, k_3}$. Además, si se considera el espacio generado por el conjunto de funciones,

$$\begin{aligned} \mathcal{P}^\ell = \{ & P_{k_1}(u)P_{k_2}(v)P_{k_3}(w) : k_1 = [0, \ell - 1], \\ & k_2 = [0, \ell - 1 - k_1], \\ & k_3 = [0, \ell - 1 - k_1 - k_2] \} , \end{aligned}$$

la operación de restricción dada en (5.6) simplemente es la restricción de la expansión del flujo neutrónico ϕ_e^ℓ (ecuación(5.4)) al subespacio generado por el conjunto de funciones $\mathcal{P}^{\ell'} \subset \mathcal{P}^\ell$, con $\ell' < \ell$, y que corresponde a su expansión con ℓ' polinomios. De igual forma se justificaría la operación de prolongación.

Para el ejemplo de la figura 5.1, el operador de restricción de los coeficientes correspondientes al nodo e en el nivel Ω^j sobre el nivel Ω^i se expresa matricialmente como,

$$\mathcal{I}_j^i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} .$$

Formalmente, esta matriz puede escribirse como

$$\mathcal{I}_j^i = [I_i | O] \pi_j , \tag{5.7}$$

donde I_i es la identidad en \mathbb{R}^{n_i} y π_j es una matriz de permutación en \mathbb{R}^{n_j} ¹. La operación de restricción (5.6) se formula de forma equivalente como,

$$\phi_e^{\ell; k_1, k_2, k_3}(t_{n+1}) = \mathcal{I}_j^i \phi_e^{\ell'; k_1, k_2, k_3}(t_{n+1}) . \tag{5.8}$$

¹Ver, por ejemplo, [31] donde se introduce esta nomenclatura.

5.2. Métodos multinivel

Obsérvese que el operador traspuesto del anterior, es decir,

$$\mathcal{I}_i^j = (\mathcal{I}_j^i)^T, \quad (5.9)$$

inyecta los coeficientes desde el nivel Ω^i en las posiciones adecuadas del vector de coeficientes del nivel Ω^j .

Si se define el operador $\mathcal{I}_j^{-i} : \mathbb{R}^{n_j} \rightarrow \mathbb{R}^{n_j - n_i}$ como,

$$\mathcal{I}_j^{-i} = [O|I_{-i}] \pi_j, \quad (5.10)$$

donde I_{-i} es la identidad en $\mathbb{R}^{n_j - n_i}$, la operación de prolongación definida por la ecuación (5.5) se puede expresar como,

$$\phi_e^{\ell'; k_1, k_2, k_3}(t_{n+1}) = (\mathcal{I}_j^{-i})^T \mathcal{I}_j^{-i} \phi_e^{\ell'; k_1, k_2, k_3}(t_n) + \mathcal{I}_i^j \phi_e^{\ell'; k_1, k_2, k_3}(t_{n+1}). \quad (5.11)$$

La matriz $(\mathcal{I}_j^{-i})^T \mathcal{I}_j^{-i}$ es una matriz diagonal de tamaño $n_j \times n_j$, con unos en la diagonal en las filas donde $(\mathcal{I}_j^{-i})^T$ tiene elementos distintos de cero. El efecto que produce sobre el vector $\phi_e^{\ell'; k_1, k_2, k_3}(t_n)$ es anular aquellos coeficientes que serán interpolados desde el nivel inferior.

Observación 61 *Los operadores de prolongación y restricción definidos por las ecuaciones (5.7-5.11), son iguales para todos los nodos. Por tanto, los operadores de prolongación y restricción completos equivalentes (los que actúan sobre los vectores de coeficientes completos, $\psi^{(j)}$ y $\psi^{(i)}$) quedan perfectamente definidos. Por ejemplo, sea $\psi^{(j)}$ un vector en $\Omega^j \equiv \mathbb{R}^{N \cdot n_j}$, donde N es el número de nodos en los que está dividido el reactor, y n_j es el número de coeficientes del flujo neutrónico correspondientes a un nodo cualquiera e en el nivel j . De la misma forma, $\psi^{(i)}$ es un vector en $\Omega^i \equiv \mathbb{R}^{N \cdot n_i}$, con n_i el número de coeficientes correspondientes a un nodo en el nivel i . El operador de restricción $\mathcal{I}_j^i : \mathbb{R}^{N \cdot n_j} \rightarrow \mathbb{R}^{N \cdot n_i}$ tal que,*

$$\psi^{(i)} = \mathcal{I}_j^i \psi^{(j)},$$

es una matriz diagonal por bloques de tamaño $N \cdot n_i \times N \cdot n_j$, con N bloques diagonales rectangulares de la forma,

$$\mathcal{I}_j^i = \begin{bmatrix} [I_i|O] \pi_j & & & \\ & \ddots & & \\ & & [I_i|O] \pi_j & \\ & & & \ddots \\ & & & & [I_i|O] \pi_j \end{bmatrix},$$

con la matriz $[I_i|O] \pi_j$ como se indicó en la ecuación (5.7). De forma similar se construiría el operador completo equivalente a \mathcal{I}_j^{-i} , ecuación (5.10), que permite definir la operación de prolongación (5.11). Mientras que no haya peligro de confusión, se utilizará indistintamente la misma notación para los operadores de restricción y prolongación restringidos a un nodo, y sus operadores completos equivalentes.

5.2.2 Multinivel algebraico

Las matrices $T^{(k+1)}$ que aparecen en el algoritmo 5, pueden generarse utilizando el operador de restricción dado en (5.7) de la forma,

$$T^{(k+1)} = \mathcal{I}_m^{k+1} T^{(m)} \mathcal{I}_{k+1}^m, \quad (5.12)$$

es decir, la restricción de la matriz de coeficientes $T = T^{(m)}$ del sistema de ecuaciones (5.1) al espacio columna de la matriz $(\mathcal{I}_m^{k+1})^T$. Esta operación consiste en tomar de la matriz T aquellos elementos que aparecen en el nivel Ω^{k+1} . De la misma forma, el término independiente $E^{(k)}$ se obtiene de la restricción del término independiente $E^{(m)} = E$ al nivel $\Omega^{(k+1)}$. Con la elección de $T^{(k+1)}$ en la forma (5.12) se obtiene el algoritmo denominado multinivel algebraico o AML.

El método AML se muestra en el algoritmo 6, con los operadores de interpolación y restricción definidos en la sección 5.2.1.

5.2. Métodos multinivel

Algoritmo 6 Algoritmo de la iteración anidada AML.

1. **Elegir** m niveles tales que, $\Omega^1 \subset \Omega^2 \subset \dots \subset \Omega^m$.
2. **Elegir** $P \in \mathbb{N}^m$, el número de polinomios de Legendre en cada nivel.
3. **Restringir al nivel más grueso:** $\psi_0^{(1)}(t_{n+1}) = \mathcal{I}_m^1 \psi^{(m)}(t_n)$.
4. **Resolver** $T^{(1)} \psi^{(1)}(t_{n+1}) = E^{(1)}$ tomando $\psi_0^{(1)}(t_{n+1})$ como solución inicial, $T^{(1)} = \mathcal{I}_m^1 \mathcal{T}^{(m)} \mathcal{I}_1^m$ y $E^{(1)} = \mathcal{I}_m^1 E^{(m)}$.
5. **Para** $k = 1, \dots, m$

(a) **Interpolar al siguiente nivel:**

$$\psi_0^{(k+1)}(t_{n+1}) = (\mathcal{I}_j^{-i})^T \mathcal{I}_j^{-i} \psi^{(k+1)}(t_n) + \mathcal{I}_k^{k+1} \psi^{(k)}(t_{n+1}).$$

- (b) **Resolver** $T^{(k+1)} \psi^{(k+1)}(t_{n+1}) = E^{(k+1)}$ tomando $\psi_0^{(k+1)}(t_{n+1})$ como solución inicial, $T^{(k+1)} = \mathcal{I}_m^{k+1} \mathcal{T}^{(m)} \mathcal{I}_{k+1}^m$ y $E^{(k+1)} = \mathcal{I}_m^{k+1} E^{(m)}$.

6. **Fin.**

5.2.3 Multinivel geométrico

El algoritmo geométrico surge de la observación de que los elementos de la matriz $T^{(k+1)}$ en (5.12), difieren de los que aparecen en la matriz de coeficientes del sistema de ecuaciones

$$(T\psi = E)^{(k+1)} , \quad (5.13)$$

generado para un número de polinomios $P(k+1)$ explícitamente. Este hecho se puede apreciar al examinar la ecuación (2.29) (pág. 43), que muestra la relación entre los coeficientes del flujo neutrónico, y por tanto, la forma de las entradas de la matriz $T^{(k+1)}$. Se observa, que el número de polinomios con los que se hace la discretización se utiliza para obtener el valor de dichos elementos. Lo mismo ocurre con el término independiente $E^{(k+1)}$.

De esta manera, la segunda estrategia multinivel se basa en la generación de la matriz de coeficientes y el término independiente en cada nivel, teniendo en cuenta el número de polinomios. A este método se le denominará multinivel geométrico o GML y se recoge en el algoritmo 7.

La notación que se muestra en la ecuación (5.13) se usará para indicar que el sistema de ecuaciones ha sido generado a partir del número de polinomios, en contraposición a (5.12) que se obtiene de una restricción.

5.3 Experimentos numéricos

Para evaluar la eficiencia de los métodos AML y GML descritos en las secciones 5.2.2 y 5.2.3, respectivamente, se han analizado los dos transitorios correspondientes al reactor bidimensional TWIGL (curvas 2.5 y 2.6) y el transitorio tridimensional de Langenbuch (curva 2.8, pág. 58), descritos en la sección 2.5.1 (pág. 53).

En cuanto a los códigos y librerías utilizadas, son válidas las consideraciones de los capítulos anteriores. Las pruebas se han realizado en la máquina

5.3. Experimentos numéricos

Algoritmo 7 Algoritmo de la iteración anidada GML.

1. **Elegir** m niveles tales que, $\Omega^1 \subset \Omega^2 \subset \dots \subset \Omega^m$.
2. **Elegir** $P \in \mathbb{N}^m$, el número de polinomios de Legendre en cada nivel.
3. **Restringir al nivel más grueso:** $\psi_0^{(1)}(t_{n+1}) = \mathcal{I}_m^1 \psi^{(m)}(t_n)$.
4. **Generar** $(T\psi = E)^{(1)}$ para $P(1)$ polinomios de Legendre.
5. **Resolver** $T^{(1)}\psi^{(1)}(t_{n+1}) = E^{(1)}$ tomando $\psi_0^{(1)}(t_{n+1})$ como solución inicial.
6. **Para** $k = 1, \dots, m$

(a) **Interpolar al siguiente nivel:**

$$\psi_0^{(k+1)}(t_{n+1}) = (\mathcal{I}_j^i)^T \mathcal{I}_j^i \psi^{(k+1)}(t_n) + \mathcal{I}_k^{k+1} \psi^{(k)}(t_{n+1}).$$

- (b) **Generar** $(T\psi = E)^{(k+1)}$ para $P(k+1)$ polinomios de Legendre.
- (c) **Resolver** $T^{(k+1)}\psi^{(k+1)}(t_{n+1}) = E^{(k+1)}$ tomando $\psi_0^{(k+1)}(t_{n+1})$ como solución inicial.

7. **Fin.**

Capítulo 5. Métodos multinivel

h_t (sg)	tiempo (sg)	método de relajación
0,001	177	$ASD(1.5, 5, 1)$
0,005	114	
0,01	92	
0,001	293	BiCGSTAB
0,005	79	
0,01	44	

Tabla 5.1: Tiempo CPU (en segundos) utilizado por NOBACK1 para simular el transitorio 1.

HP Exemplar S Class.

Al igual que en capítulos anteriores, en todos los experimentos se utiliza una fórmula de discretización temporal de un paso (capítulo 2, pág. 46).

5.3.1 Transitorios bidimensionales

En las tablas 5.1 a 5.6 se muestra el tiempo de CPU utilizado para simular los transitorios 1 y 2 del TWIGL con diferentes pasos de tiempo. Las tablas 5.1 y 5.2 muestran el tiempo de simulación utilizado por el algoritmo 7, sin hacer uso del algoritmo multinivel², es decir, para una sola malla correspondiente al número máximo de 5 polinomios de Legendre. Este algoritmo es referenciado como NOBACK1 [38].

Con 5 polinomios de Legendre, las matrices de coeficientes de los sistemas de ecuaciones a resolver en cada paso de tiempo son de dimensión 3000 con un total de 58480 elementos no nulos.

Para resolver los sistemas de ecuaciones lineales se utiliza el método de segundo grado acelerado, $ASD(1.5, 5, 1)$, descrito en los capítulos 3 y 4, y el método BiCGSTAB preconditionado con ILU0. Ambos métodos se muestran como los más apropiados para el reactor bidimensional TWIGL (véase los experimentos numéricos en el capítulo 4).

²Hay que hacer notar que los algoritmos 6 y 7 coinciden cuando se aplican con una sola malla, es decir, sin la técnica multinivel.

5.3. Experimentos numéricos

h_t (sg)	tiempo (sg)	método de relajación
0,001	521	ASD(1.5, 5, 1)
0,005	252	
0,01	184	
0,001	555	BiCGSTAB
0,005	155	
0,01	94	

Tabla 5.2: Tiempo CPU (en segundos) utilizado por NOBACK1 para simular el transitorio 2.

Las tablas 5.3 y 5.4 muestran el tiempo utilizado por los algoritmos multinivel geométrico (GML) y algebraico (AML), para el transitorio 1. De la misma forma, las tablas 5.5 y 5.6 muestran el tiempo para el segundo transitorio.

En todas las tablas, h_t corresponde al paso de tiempo utilizado en segundos, *tiempo* es el tiempo de CPU utilizado en la simulación del transitorio completo. La columna *método de relajación* indica el método utilizado para resolver el sistema de ecuaciones lineales en cada nivel y en cada paso de tiempo. El número de mallas máximo utilizado es $m = 4$, con un número de polinomios de Legendre para cada nivel $P(k) = \{2, 3, 4, 5\}$, $k = 1, 2, 3, 4$. Para los algoritmos multinivel AML y GML, *niveles* indica las mallas utilizadas. Por ejemplo, $k = \{2, 4\}$ indicaría el uso de los niveles Ω^2 y Ω^4 correspondientes a un número de polinomios $P(2) = 3$ y $P(4) = 5$, respectivamente.

Se observa que el tiempo utilizado por el algoritmo GML para los dos transitorios siempre es menor que el tiempo de simulación con NOBACK1. Este comportamiento se repite para los tres pasos de tiempo considerados. Para pasos de tiempo pequeños, 0.001 y 0.005 segundos, ya que NOBACK1 obtiene la solución del sistema de ecuaciones lineales con pocas iteraciones, el uso de la estrategia GML con ciclos multinivel largos produce resultados peores (el tiempo de simulación aumenta), como se observa en las tablas 5.3 y 5.5 para el ciclo $k = \{1, 2, 3, 4\}$. Esto se debe a que en la malla más fina

Capítulo 5. Métodos multinivel

h_t (sg)	tiempo (sg)	niveles Ω^k	método de relajación
0,001	236	$k = \{1, 2, 3, 4\}$	$ASD(1.5, 5, 1)$
0,001	150	$k = \{2, 4\}$	
0,005	61	$k = \{1, 2, 3, 4\}$	
0,005	50	$k = \{2, 4\}$	
0,01	31	$k = \{1, 2, 3, 4\}$	
0,01	45	$k = \{2, 4\}$	
0,001	298	$k = \{1, 2, 3, 4\}$	BiCGSTAB
0,001	228	$k = \{2, 4\}$	
0,005	75	$k = \{1, 2, 3, 4\}$	
0,005	55	$k = \{2, 4\}$	
0,01	36	$k = \{1, 2, 3, 4\}$	
0,01	49	$k = \{2, 4\}$	

Tabla 5.3: Tiempo de CPU (en segundos) utilizado por GML para simular el transitorio 1. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$.

h_t (sg)	tiempo (sg)	niveles Ω^k	método de relajación
0,001	353	$k = \{1, 2, 3, 4\}$	$ASD(1.5, 5, 1)$
0,001	312	$k = \{2, 4\}$	
0,005	95	$k = \{1, 2, 3, 4\}$	
0,005	70	$k = \{2, 4\}$	
0,01	35	$k = \{1, 2, 3, 4\}$	
0,01	45		
0,001	398	$k = \{1, 2, 3, 4\}$	BiCGSTAB
0,001	350	$k = \{2, 4\}$	
0,005	107	$k = \{1, 2, 3, 4\}$	
0,005	87	$k = \{2, 4\}$	
0,01	51	$k = \{1, 2, 3, 4\}$	
0,01	64	$k = \{2, 4\}$	

Tabla 5.4: Tiempo de CPU (en segundos) utilizado por AML para simular el transitorio 1. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$.

5.3. Experimentos numéricos

h_t (sg)	tiempo (sg)	niveles Ω^k	método de relajación
0,001	674	$k = \{1, 2, 3, 4\}$	ASD(1.5, 5, 1)
0,001	387	$k = \{2, 4\}$	
0,005	151	$k = \{1, 2, 3, 4\}$	
0,005	108	$k = \{2, 4\}$	
0,01	76	$k = \{1, 2, 3, 4\}$	
0,01	80	$k = \{2, 4\}$	
0,001	713	$k = \{1, 2, 3, 4\}$	BiCGSTAB
0,001	561	$k = \{2, 4\}$	
0,005	152	$k = \{1, 2, 3, 4\}$	
0,005	154	$k = \{2, 4\}$	
0,01	77	$k = \{1, 2, 3, 4\}$	
0,01	86	$k = \{2, 4\}$	

Tabla 5.5: Tiempo de CPU (en segundos) utilizado por GML para simular el transitorio 2. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$.

h_t (sg)	tiempo (sg)	niveles Ω^k	método de relajación
0,001	920	$k = \{1, 2, 3, 4\}$	ASD(1.5, 5, 1)
0,001	795	$k = \{2, 4\}$	
0,005	224	$k = \{1, 2, 3, 4\}$	
0,005	189	$k = \{2, 4\}$	
0,01	111	$k = \{1, 2, 3, 4\}$	
0,01	122	$k = \{2, 4\}$	
0,001	925	$k = \{1, 2, 3, 4\}$	BiCGSTAB
0,001	834	$k = \{2, 4\}$	
0,005	227	$k = \{1, 2, 3, 4\}$	
0,005	208	$k = \{2, 4\}$	
0,01	127	$k = \{1, 2, 3, 4\}$	
0,01	142	$k = \{2, 4\}$	

Tabla 5.6: Tiempo de CPU (en segundos) utilizado por AML para simular el transitorio 2. Número de polinomios de Legendre, $P(k) = \{2, 3, 4, 5\}$.

(nivel correspondiente a 5 polinomios de Legendre) sólo se puede obtener una pequeña reducción en el número de iteraciones, y de esta forma, el trabajo extra del algoritmo multinivel GML no puede ser compensado. En estos casos es necesario utilizar ciclos más cortos, en particular el ciclo $k = \{2, 4\}$, para mejorar los tiempos de NOBACK1 si el paso de integración es corto. Los ciclos largos son adecuados para pasos de integración grandes.

Con respecto al algoritmo multinivel algebraico, en las tablas 5.4 y 5.6 se observa que los tiempos de simulación casi siempre son ligeramente más elevados a los que obtiene NOBACK1. Sólo cuando se utiliza como método de relajación el algoritmo de segundo grado acelerado, el algoritmo AML es capaz de obtener mejores tiempos. Además, como se ha mencionado para el caso GML, utilizando ciclos multinivel cortos para pasos de tiempo pequeños se reduce el tiempo de simulación.

Por otro lado, y completando el estudio realizado en el capítulo 4, si se comparan los métodos de relajación se observa que el método de segundo grado acelerado, $ASD(1.5, 5, 1)$, resulta más conveniente para los algoritmos multinivel que el método BiCGSTAB.

Se puede concluir que el algoritmo multinivel GML combinado con el método de segundo grado acelerado como método de relajación para resolver los sistemas de ecuaciones lineales en cada nivel, produce los mejores resultados para los dos transitorios del TWIGL, independientemente del paso de tiempo utilizado para integrar la ecuación de la difusión neutrónica.

5.3.2 Transitorio de Langenbuch

Los resultados de simulación de los problemas bidimensionales de la sección anterior mostraban un mejor comportamiento del algoritmo GML frente al algoritmo AML. Por ello se ha elegido la variante geométrica para los experimentos numéricos con el transitorio de Langenbuch.

El número de mallas máximo utilizado es $m = 3$, con un número de

5.3. Experimentos numéricos

h_t (sg)	tiempo (sg)	niveles Ω^k	método de relajación
0, 125	6188	$k = \{1, 2, 3\}$	$ASD(1.5, 5, 1)$
	4556	$k = \{1, 3\}$	
0, 125	3000	$k = \{1, 2, 3\}$	BiCGSTAB
	3128	$k = \{1, 3\}$	

Tabla 5.7: Tiempo de CPU (en segundos) utilizado por el algoritmo GML para simular el transitorio de Langenbuch. Número de polinomios de Legendre, $P(k) = \{2, 3, 4\}$.

h_t (sg)	$NOBACK1_{[BiCGSTAB]}$ (sg)	$NOBACK1_{[ASD]}$ (sg)	NEM (sg)
0, 125	4620	5160	64800

Tabla 5.8: Tiempo de CPU en segundos utilizado por los códigos GML, NOBACK1 y NEM para simular el transitorio de Langenbuch. Entre corchetes se indica el método de relajación utilizado para NOBACK1.

polinomios de Legendre para cada nivel $P(k) = \{2, 3, 4\}$, $k = 1, 2, 3$. Los métodos de relajación utilizados han sido el método $ASD(1.5, 5, 1)$ y el método BiCGSTAB preconditionado con ILU0. Además, se ha comparado con el método NOBACK1 y con el código NEM [27, 6].

Como en la sección anterior, NOBACK1 hace referencia al algoritmo que no utiliza la técnica multinivel, es decir, el algoritmo 7 para una sola malla correspondiente al número máximo de 3 polinomios de Legendre. Los métodos de relajación utilizados con NOBACK1 han sido el método BiCGSTAB y el método $ASD(1.5, 5, 1)$.

En la tabla 5.7 se observan los resultados para el método GML con los métodos de relajación $ASD(1.5, 5, 1)$ y BiCGSTAB. Se han utilizado dos ciclos multinivel, los correspondientes a $k = 1, 2, 3$ y $k = 1, 3$. Se observa que el algoritmo GML relajado con el método BiCGSTAB obtiene mejores resultados que con el método de segundo grado acelerado. Además, el ciclo $k = 1, 2, 3$ obtiene resultados ligeramente mejores que el ciclo más corto $k = 1, 3$.

Por otro lado, la tabla 5.8 muestra el tiempo de CPU utilizado para sim-

Capítulo 5. Métodos multinivel

ular el transitorio por los códigos NOBACK1 y NEM. Se observa claramente que el algoritmo multinivel GML obtiene el mejor tiempo de simulación, especialmente si se compara con el código NEM.

Resumiendo, mediante la simulación de los dos transitorios correspondientes al reactor TWIGL se han comparado los dos algoritmos multinivel AML y GML con el método denominado NOBACK1, observándose que el algoritmo GML obtiene tiempos de simulación considerablemente mejores.

Además, también se ha simulado el transitorio de Langenbuch mediante el algoritmo GML y se ha comparado con el código NEM, código de producción muy difundido, obteniendo unos resultados excelentes.

Para los transitorios bidimensionales, el método GML relajado con el método de segundo grado acelerado $ASD(1.5, 5, 1)$, obtiene mejores tiempos que con el método BiCGSTAB. Este comportamiento se invierte para el transitorio tridimensional de Langenbuch.

Finalmente, es importante resaltar que la estrategia multinivel presentada proporciona una herramienta básica para implementar métodos de integración capaces de cambiar la precisión de la solución de un paso de tiempo al siguiente. De esta manera, la estrategia multinivel posibilita la implementación de un control para la precisión espacial similar al control existente en muchos códigos de producción para la integración temporal, y que básicamente consiste en el cambio del paso de tiempo de integración.

Conclusiones y líneas futuras

Las conclusiones principales que se pueden extraer de los estudios presentados en esta memoria son las siguientes.

Se han aplicado diferentes técnicas de discretización para la parte espacial y temporal de la ecuación de la difusión neutrónica dependiente del tiempo. Concretamente, para la parte espacial se ha utilizado el método de las diferencias finitas centradas y un método de colocación nodal que se basa en el desarrollo del flujo neutrónico en términos de polinomios de Legendre. La parte temporal se ha discretizado mediante métodos implícitos en diferencias hacia atrás.

Los experimentos numéricos realizados con los transitorios bidimensionales correspondientes al reactor TWIGL y el transitorio tridimensional de Langenbuch han mostrado la conveniencia de utilizar el método de colocación nodal para la discretización de la parte espacial. Ésto es debido a que con el método en diferencias finitas se han de resolver sistemas de ecuaciones con matrices de mayor tamaño y con un elevado número de elementos no nulos para obtener precisiones similares a las obtenidas mediante el método de colocación nodal.

Con respecto a la resolución de los sistemas de ecuaciones lineales en cada paso de tiempo de integración, se han estudiado métodos de segundo grado obtenidos a partir de la matriz de iteración de un método iterativo de primer grado y un factor de extrapolación o relajación. Se han dado diferentes resultados de convergencia en función de los valores propios de la

Conclusiones y líneas futuras

matriz de iteración del método de primer grado, reales o complejos, indicando, bajo ciertas suposiciones, el factor óptimo de extrapolación. Este estudio se ha particularizado a dos métodos de segundo grado para la solución de los sistemas de ecuaciones que aparecen en la integración de la ecuación de la difusión neutrónica, referenciados como métodos de segundo grado A y B. Con ellos se ha tratado de aprovechar las especiales características que presentan los sistemas de ecuaciones a resolver por bloques. El método B se diferencia del método A en que aprovecha las soluciones para el cálculo tan pronto como están disponibles. De esta forma entre ambos métodos existe una relación similar a la existente entre los métodos de Jacobi y Gauss-Seidel. Además, experimentalmente se ha comprobado que efectivamente el método B emplea aproximadamente la mitad de tiempo en la simulación del transitorio bidimensional del reactor TWIGL.

Para acelerar la convergencia del método de segundo grado B se ha presentado un método variacional. El método minimiza la norma del residuo proyectando sobre un subespacio bidimensional. Además, se han dado ciertas condiciones para su convergencia. De los experimentos numéricos para evaluar el comportamiento del método de segundo grado acelerado (denotado como $ASD(\omega, r, q)$) realizados, se ha determinado que los valores $\omega = 1.5$, $r = 5$ y $q = 1$ proporcionaban el mejor comportamiento del método. Se ha observado que el método $ASD(1.5, 5, 1)$ tarda aproximadamente la mitad que el método de segundo B sin acelerar. Este hecho pone de manifiesto que el método variacional es muy efectivo en la aceleración del método de segundo grado B.

Así mismo, se ha comparado con los métodos BiCGSTAB, GMRES(k) y TFQMR preconditionados. El estudio comparativo permite concluir que el método $ASD(\omega, r, q)$ presenta la mejor relación *tiempo de simulación / precisión de la solución* para los transitorios bidimensionales del TWIGL. Para el transitorio tridimensional del reactor Langenbuch esto continúa siendo cierto para un número de polinomios de Legendre de 2. Para un número

Conclusiones y líneas futuras

de polinomios mayor el método BiCGSTAB obtiene los mejores resultados. También es importante resaltar que de los tres métodos basados en subespacios de Krylov, es el método BiCGSTAB el más indicado para el tipo de problemas estudiado. Por tanto se puede concluir que para transitorios bidimensionales, y transitorios tridimensionales discretizados con un número de polinomios pequeño, el método $ASD(\omega, r, q)$ es competitivo frente a los métodos BiCGSTAB, GMRES(k) y TFQMR.

Finalmente, se han presentado dos algoritmos multinivel para la integración de la ecuación de la difusión neutrónica dependiente del tiempo, denominados GML y AML. Estos métodos se basan en el método de colocación nodal para la discretización de la parte espacial de la ecuación, de manera que los diferentes niveles se caracterizan por el número de polinomios de Legendre utilizados para la expansión del flujo neutrónico. La diferencia entre ambos radica en la forma de generar las matrices de coeficientes en cada nivel.

Mediante la simulación de los transitorios bidimensionales del reactor TWIGL se ha observado que el algoritmo multinivel GML combinado con el método de segundo grado acelerado $ASD(1.5, 5, 1)$ como método de relajación para resolver los sistemas de ecuaciones lineales en cada nivel, produce los mejores resultados independientemente del paso de tiempo utilizado para integrar la ecuación de la difusión neutrónica.

Para el transitorio tridimensional de Langenbuch, se han comparado los dos algoritmos multinivel con otros métodos. En particular el método denominado NOBACK1 y el código NEM, código de producción muy difundido. Se ha observado como el algoritmo GML relajado con el método BiCGSTAB obtiene tiempos de simulación considerablemente mejores. Incluso el método $ASD(1.5, 5, 1)$ sin estrategia multinivel resulta ser muy competitivo frente al código NEM.

Es importante resaltar que la estrategia multinivel presentada proporciona una herramienta básica para implementar métodos de integración ca-

Conclusiones y líneas futuras

paces de cambiar la precisión de la solución de un paso de tiempo al siguiente. De esta manera, la estrategia multinivel posibilita la implementación de un control para la precisión espacial similar al control existente en muchos códigos de producción para la integración temporal.

Por todo lo dicho anteriormente, para integrar un transitorio tridimensional donde se produce una variación brusca de la potencia, como es el caso del problema de Langenbuch, se recomienda utilizar un método optimizado del tipo GML relajado con el método BiCGSTAB. Para el análisis de sistemas más pequeños o con una variación de la potencia más pequeña se puede abordar de forma efectiva con un método GML relajado con el método $ASD(1.5, 5, 1)$.

Mencionar que parte de los resultados presentados en esta memoria aparecen recogidos en [18], [19], [36], [37] y [38].

A continuación se esquematizan algunas líneas de trabajo de interés para su desarrollo futuro. Uno de los objetivos fundamentales que se persigue es la utilización de algoritmos numéricos paralelos para la resolución de los sistemas de ecuaciones lineales que se obtienen en cada paso de integración. En este sentido se pueden identificar dos vías de actuación.

La primera sería considerar el sistema de ecuaciones a resolver en su totalidad y aplicar alguna técnica general para su resolución en paralelo. Concretamente, dado que el método BiCGSTAB ha demostrado ser una opción interesante especialmente para problemas con geometrías 3D, se buscaría paralelizar su aplicación en este tipo de problemas. Como se mencionó en el capítulo de introducción, la eficiencia de los métodos basados en subespacios de Krylov depende de su preconditionado. Por tanto, para la paralelización del método BiCGSTAB se ha de buscar alguna técnica de preconditionamiento con un alto nivel de paralelismo. Es importante recordar que en los experimentos numéricos presentados los preconditionadores que mejores resultados proporcionaban eran del tipo ILU, de difícil paralelización pues su aplicación requiere la solución de sistemas triangulares. Por tanto, no son en principio

una opción válida.

De entre las diferentes técnicas de preconditionamiento existentes, los *precondicionadores basados en multiparticiones* y *precondicionadores polinomiales* permiten la obtención de preconditionadores paralelos. En [73, 16, 17, 22] se pueden encontrar ejemplos de preconditionadores de este tipo susceptibles de ser utilizados para conseguir el fin que se persigue.

Otro tipo de preconditionamiento que está despertando un creciente interés son los denominados *precondicionadores de inversa aproximada* (brevemente descritos en la sección 1.4.4 del capítulo 1). De entre ellos destaca por sus prestaciones el preconditionador AINV. En [7, 8] se propone la paralelización de este preconditionador mediante técnicas de particionado de grafos, aplicándose con éxito a la solución de diferentes problemas incluidos problemas de difusión. Por tanto sería interesante su aplicación a las matrices que se obtienen en la discretización de la ecuación de la difusión neutrónica.

El otro camino para la obtención de paralelismo surge de las especiales características de los sistemas de ecuaciones lineales que hay que resolver. Es por tanto un paralelismo dependiente de la aplicación, en contraposición a las líneas de actuación propuestas anteriormente. Después de reordenar las incógnitas, el sistema (2.52) (pág. 53) se puede reescribir como

$$\begin{bmatrix} T_{12} & T_{11} \\ T_{22} & T_{21} \end{bmatrix} \begin{bmatrix} \psi_2 \\ \psi_1 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \end{bmatrix}.$$

Debido a que el bloque T_{21} es invertible (el bloque T_{12} no tiene necesariamente esta propiedad), mediante un proceso de eliminación gaussiana se obtiene el sistema equivalente,

$$\begin{bmatrix} S & 0 \\ T_{22} & T_{21} \end{bmatrix} \begin{bmatrix} \psi_2 \\ \psi_1 \end{bmatrix} = \begin{bmatrix} E_1 - T_{11}T_{21}^{-1}E_2 \\ E_2 \end{bmatrix},$$

donde

$$S = T_{12} - T_{11}T_{21}^{-1}T_{22},$$

Conclusiones y líneas futuras

es el complemento de Schur del bloque T_{21} . El cálculo de las incógnitas en ψ_2 se puede llevar a cabo independientemente aplicando un método basado en subespacios de Krylov. Si se distribuyen adecuadamente las matrices entre los procesadores este proceso posee un elevado grado de paralelismo. Además, una vez calculadas las incógnitas en ψ_2 , ψ_1 se calcula directamente en paralelo como $\psi_1 = T_{21}^{-1}(E_2 - T_{22}\psi_2)$. En [35] los resultados presentados muestran que esta metodología es una opción de paralelización prometedora.

Bibliografía

- [1] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.
- [2] G. Avdelas, J. de Pillis, A. Hadjidimos, and M. Neumann. A guide to the acceleration of iterative methods whose iteration matrix is non-negative and convergent. *Siam J. Matrix Anal. Appl.*, 9(3):329–342, 1988.
- [3] G. Avdelas and A. Hadjidimos. Optimum accelerated overrelaxation method in a special case. *Mathematics of Computation*, 36(153):183–187, 1981.
- [4] O. Axelsson. A generalized SSOR method. *BIT*, 12:443–467, 1972.
- [5] O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, New York, 1994.
- [6] T. M. Beam, K.N. Ivanov, A. J. Baratta, and H. Finnemann. Kinetics model upgrade in the Penn state coupled TRAC/NEM codes. *Annals of Nuclear Energy*, 26:1205–1219, 1999.
- [7] M. Benzi, J. Marín, and M. Tuma. Parallel preconditioning with factorized sparse approximate inverses. En *Proc. of the Ninth SIAM Conference on Parallel Processing for Scientific Computing 1999*, San Antonio, TX USA, march 1999. SIAM.

Bibliografía

- [8] M. Benzi, J. Marín, and M. Tuma. A two-level parallel preconditioner based on sparse approximate inverses. En David R. Kincaid and A. C. Elster, editores, *Iterative Methods in Scientific Computation IV*, volume 5 of *IMACS Series in Computational and Applied Mathematics*, pages 167–178. IMACS, New Brunswick, NJ, 1999.
- [9] M. Benzi, C. D. Meyer, and M. Tuma. A sparse approximate inverse preconditioner for the conjugate gradient method. *SIAM Journal on Scientific Computing*, 17:1135–1149, 1996.
- [10] M. Benzi, R. Nabben, and D. B. Szyld. Algebraic theory of multiplicative Schwarz methods. Technical Report 00–2–10, Department of Mathematics, Temple University, Philadelphia, 2000.
- [11] M. Benzi, D. B. Szyld, and A. van Duin. Orderings for incomplete factorization preconditioning of nonsymmetric problems. *Journal on Scientific Computing*, 20:1652–1670, 1999.
- [12] M. Benzi and M. Tuma. A comparative study of sparse approximate inverse preconditioners. *Applied Numerical Mathematics*, 30(2–3):305–340, 1999.
- [13] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*, volume 9 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1994. Originally published by Academic Press, New York, 1979.
- [14] C. Brezinski. *Projection Methods for Systems of Equations*. North-Holland, Amsterdam, 1997.
- [15] W. L. Briggs. *A Multigrid Tutorial*. SIAM, Philadelphia, 1987.
- [16] R. Bru, C. Corral, A. Martínez, and J. Mas. Multisplitting preconditioners based on incomplete Choleski factorizations. *SIAM J. Mat. Anal.*, 16(4):1210–1222, 1995.

Bibliografía

- [17] R. Bru, C. Corral, and J. Mas. A preconditioned conjugate gradient method on a distributed memory multiprocessor. *Applied Mathematics Letters*, 8(3):49–53, 1995.
- [18] R. Bru, D. Ginestar, , J. Marín, G. Verdú, and V. Vidal. A second degree method for the neutron diffusion equation. En F. J. Cobos, J. R. Gómez, and F. Mateos, editores, *Encuentro de Análisis Matricial y Aplicaciones, EAMA-97*, pages 87–93, Sevilla, septiembre 1997. Universidad de Sevilla.
- [19] R. Bru, D. Ginestar, T. Manteuffel, J. Marín, and G. Verdú. Iterative schemes for the neutron diffusion equation. En *Proc. of the VIII Copper Mountain Conference on Iterative Methods*, volume II, Copper Mountain, Colorado USA, march 1998. The University of Colorado. In cooperation with SIAM Group on Numerical Linear Algebra.
- [20] G. S. Chen, J. M Christenson, and D. Y. Yang. Application of the preconditioned transpose-free quasi-minimal residual method for two-group reactor kinetics. *Annals of Nuclear Energy*, 24(5):339–35, 1996.
- [21] P. Concus, G. H. Golub, and G. Meurant. Block preconditioning for the conjugate gradient method. *SIAM J. Sci. Statist. Comput.*, 6:220–252, 1985.
- [22] C. Corral, I. Giménez, J. Marín, and J. Mas. Parallel m-step preconditioners for the conjugate gradient method. *Parallel Computing*, 25:265–281, 1999.
- [23] I. S. Duff and G. A. Meurant. The effect of ordering on preconditioned conjugate gradients. *BIT*, 29:635–657, 1989.
- [24] H. C. Elman. A stability analysis of incomplete LU factorizations. *Mathematics of Computations*, 47:191–217, 1986.

Bibliografía

- [25] L. Elsner. Comparisons of weak regular splittings and multisplitting methods. *Numerische Mathematik*, 56:283–289, 1989.
- [26] V. Faber and T. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.*, 21:352–361, 1984.
- [27] H. Finnmann, F. Bennewitz, and M. R. Wagner. Interface current techniques for multidimensional reactor calculations. *Atomkernenergie (ATKE)*, 30:123–128, 1977.
- [28] R. Fletcher. Conjugate gradient methods for indefinite systems. En G. A. Watson, editor, *Proceedings of the Dundee Biennial Conference on Numerical Analysis 1974*, pages 73–89, New York, 1975. Springer-Verlag.
- [29] R. W. Freund. A transpose-free quasi-minimal residual algorithm for non-hermitian linear systems. *SIAM J. Sci. Comput.*, 14:470–482, 1993.
- [30] R. W. Freund and N. M. Nachtigal. QMR: A quasi-minimal residual method for non-hermitian linear systems. *Numer. Math.*, 60:315–339, 1991.
- [31] A. Frommer and D. Szyld. Weighted max norms, splittings, and overlapping additive Schwarz iterations. *Numerische Mathematik*, 83:259–278, 1999.
- [32] A. Frommer and D. B. Szyld. An algebraic convergence theory for restricted additive Schwarz methods using weighted max norms. Technical Report 00–3–23, Department of Mathematics, Temple University, Philadelphia, 2000.
- [33] N. Gastinel. *Analyse Numerique Lineaire*. Hermann, Paris, 1966.

Bibliografía

- [34] D. Ginestar. *Integración de la Ecuación de la Difusión Neutrónica en Geometrías Multidimensionales. Aplicación a Reactores Nucleares. Cálculo de los Modos Lambda*. Tesis doctoral, Universidad Politécnica de Valencia, 1995.
- [35] D. Ginestar, J. Marín, A. Martínez, and G. Verdú. A parallel Schur complement method for the neutron diffusion equation. En *Proc. of the 16th IMACS World Congress 2000*, Laussana, August 2000. IMACS.
- [36] D. Ginestar, J. Marín, and G. Verdú. Multilevel methods to solve the neutron diffusion equation. sometido para su publicación.
- [37] D. Ginestar, J. Marín, and G. Verdú. A multilevel method for the neutron diffusion equation. En José M. Aragonés, editor, *Proc. of Mathematics and Computation, Reactor Physics and Environmental Analysis in Nuclear Applications*, pages 401–410, Madrid, September 1999. Senda Editorial.
- [38] D. Ginestar, G. Verdú, V. Vidal, R. Bru, J. Marín, and J.L. Muñoz-Cobo. High order backward discretization of the neutron diffusion equation. *Annals of Nuclear Energy*, 25(1–3):47–64, 1998.
- [39] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore and London, 2 edition, 1989.
- [40] G. H. Golub and R.S. Varga. Chebyshev semi-iterative methods, successive over-relaxation iterative methods, and second-order Richardson iterative methods. *Numerische Mathematik*, 3:147–168, 1961.
- [41] S. Grainville. *The Numerical Solution of Ordinary and Partial Differential Equations*. Academic Press, New York and London, 1988.
- [42] A. Greenbaum. *Iterative Methods for Solving Liner Systems*, volume 17 of *Frontiers in applied mathematics*. SIAM, Philadelphia, 1997.

Bibliografía

- [43] M. Grote and T. Huckle. Parallel preconditioning with sparse approximate inverses. *SIAM Journal on Scientific Computing*, 18:838–853, 1997.
- [44] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Springer-Verlag, Berlin, New York, 1994.
- [45] W. Hackbusch and U. Trottenberg. *Multigrid Methods*. Springer-Verlag, Berlin, New York, 1982.
- [46] A. Hadjidimos. Accelerated overrelaxation method. *Mathematics of Computation*, 32(141):149–157, January 1978.
- [47] A. Hadjidimos, A. Psimarni, and A. Yeyios. On the convergence of some generalized iterative methods. *Linear algebra and its applications*, 75:117–132, 1986.
- [48] A. Hébert. Application of the Hermite method for finite element reactor calculations. *Nuclear Science and Engineering*, 91:34–58, 1985.
- [49] A. Hébert. Development of the nodal collocation method for solving the neutron diffusion equation. *Annals of Nuclear Energy*, 14(10):527–541, 1987.
- [50] A. F. Henry. *Nuclear Reactor Analysis*. The M.I.T Press, 1975.
- [51] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards*, 49:409–435, 1952.
- [52] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, London, UK, 1985.
- [53] A. S. Householder. *Theory of Matrices in Numerical Analysis*. Blaisdell Pub. Co., Johnson, CO, 1964.

Bibliografía

- [54] C. J. Jackson, D. G. Cacuci, and H. B. Finnemann. Dimensionally adaptive neutron kinetics for multidimensional reactor safety transients i: new features of RELAP5/PANBOX. *Nuclear Science and Engineering*, 131:143–163, 1999.
- [55] C. J. Jackson, D. G. Cacuci, and H. B. Finnemann. Dimensionally adaptive neutron kinetics for multidimensional reactor safety transients ii: dimensionally adaptive switching algorithms. *Nuclear Science and Engineering*, 131:164–186, 1999.
- [56] D. Jespersen. Multigrid methods for partial differential equations. En *Studies in numerical analysis*, volume 24 of *Studies in Mathematics*. Mathematical Association of America, 1984.
- [57] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*, volume 16 of *Frontiers in Applied Mathematics*. SIAM, Philadelphia, 1995.
- [58] D. R. Kincaid, J. R. Respass, and D. M. Young. ITPACK 2C: Large sparse linear systems by adaptive accelerated iterative methods. Technical report, University of Texas at Austin, Austin, TX, 1980.
- [59] A.N. Kolmogórov and S. V. Fomín. *Elementos de la Teoría de Funciones y del Análisis Funcional*. MIR, 1972.
- [60] L. Yu. Kolotilina and A. Yu. Yeremin. Factorized sparse approximate inverse preconditionings I. Theory. *SIAM J. Matrix Anal. Applic.*, 14:45–58, 1993.
- [61] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Stand.*, 45:255–282, 1950.
- [62] C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Stand.*, 49:33–53, 1952.

Bibliografía

- [63] S. Langenbuch, W. Maurer, and W. Werner. Coarse-mesh flux-expansion method for the analysis of space-time effects in large light water reactor cores. *Nuclear Science and Engineering*, 63:437–456, 1977.
- [64] T. A. Manteuffel. An incomplete factorization technique for positive definite linear systems. *Math. Comp.*, 34:473–497, 1980.
- [65] J. March-Leuba. *Dynamic Behavior of BWR*. Tesis doctoral, University of Tennessee, Knoxville, 1984.
- [66] J. March-Leuba and E. D. Blakeman. A mechanism for out-of-phase instabilities in boiling water reactors. *Nuclear Science and Engineering*, pages 107–173, 1991.
- [67] I. Marek and D. B. Szyld. Comparison theorems for weak splittings of bounded operators. *Numerische Mathematik*, 58:389–397, 1990.
- [68] Stephen F. McCormick, editor. *Multigrid Methods*. SIAM, Philadelphia, 1987.
- [69] J. A. Meijerink and H. A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comp.*, 31:148–162, 1977.
- [70] M. V. Migallón. *Modelos Iterativos Caóticos Síncronos y Asíncronos para la Resolución de Sistemas Lineales*. Tesis doctoral, Departamento de Tecnología Informática y Computación, Universidad de Alicante, 1993.
- [71] N. M. Missirlis. Convergence theory of extrapolated iterative methods for a certain class of non-symmetric linear systems. *Numerische Mathematik*, 45:447–458, 1984.

Bibliografía

- [72] N. M. Missirlis and D. J. Evans. On the convergence of some generalized preconditioned iterative methods. *SIAM Journal on Numerical Analysis*, 18(4):591–596, 1979.
- [73] D. P. O’Leary and R. E. White. Multisplittings of matrices and parallel solution of linear systems. *SIAM J. Alg. Discr. Meth.*, 6:630–640, 1985.
- [74] J. M. Ortega. *Numerical Analysis, A Second Course*. Academic Press, New York, 1972. Reprinted by SIAM, Philadelphia, 1990.
- [75] J. M. Ortega. *Introduction to Parallel and Vector Solution of Linear Systems*. Plenum Press, New York, 1988.
- [76] J. M. Ortega and W. Rheinbold. Monotone iterations for nonlinear equations with applications to Gauss-Seidel methods. *SIAM Journal on Numerical Analysis*, 4:171–190, 1967.
- [77] K. O. Ott. Accuracy of the quasistatic treatment of spatial reactor kinetics. *Nuclear Science and Engineering*, 36:402–411, 1969.
- [78] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 11:197–209, 1974.
- [79] C. Perea. *Teoremas de convergencia y comparación para particiones y multiparticiones*. Tesis doctoral, Universidad de Alicante, 1997.
- [80] B. Pershagen. *Light Water Reactor Safety*. Pergamon Press, Oxford, 1989.
- [81] R. Nabben. A note on comparison theorems for splittings and multisplittings of hermitian positive definite matrices. *Linear Algebra and its Applications*, 233:67–80, 1996.
- [82] J. W. Ruge and K. Stüben. Algebraic multigrid. En S. F. McCormick, editor, *Multigrid Methods*, chapter 4. SIAM, Philadelphia, 1987.

Bibliografía

- [83] R.D. Russell. A comparison of collocation and finite differences for two point boundary problems. *SIAM Journal on Numerical Analysis*, 14(1):19–39, 1977.
- [84] Y. Saad. Preconditioning techniques for indefinite and nonsymmetric linear systems. *Journal of Computational and Applied Mathematics*, 24:155–169, 1988.
- [85] Y. Saad. SPARSKIT: A basic tool kit for sparse matrix computations. Technical Report 90–20, Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffet Field, CA, 1990.
- [86] Y. Saad. ILUT: a dual threshold incomplete ILU factorization. *Numerical Linear Algebra with Applications*, 1:387–402, 1994.
- [87] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publishing Company, Boston, 1996.
- [88] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statis. Comput.*, 7:856–869, 1986.
- [89] Y. G. Saridakis. Generalized consistent orderings and the accelerated overrelaxation method. *BIT*, 26:369–376, 1986.
- [90] M. Sisler. Über die optimierung eines zweiparametrigen iterationsverfahren. *Apl. Mat.*, 20:126–142, 1975.
- [91] B. Smith, P. Bjorstad, and W. Gropp. *Domain decomposition: parallel multilevel methods for elliptic partial differential equations*. Cambridge University Press, New York, 1996.
- [92] J-W. Song and J-K. Kim. An efficient nodal method for transient calculations in light water reactors. *Nuclear Technology*, 103:157–167, 1993.

Bibliografía

- [93] P. Sonneveld. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 10:36–52, 1989.
- [94] G. Sordas and R. Varga. Comparisons of regular splittings of matrices. *Numerische Mathematik*, 44:23–35, 1984.
- [95] H. A. van der Vorst. Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 12(6):631–644, 1992.
- [96] R. S. Varga. Boundary problems in differential equations. En R. E. Lander, editor, *Factorization and normalized iterative methods*. The University of Wisconsin Press, Madison, 1960.
- [97] R. S. Varga. *Matrix Iterative Analysis*. Prentice–Hall, Englewood Cliffs, NJ, 1962.
- [98] G. Verdú, D. Ginestar, V. Vidal, and J.L. Muñoz-Cobo. 3D λ –modes of the neutron diffusion equation. *Annals of Nuclear Energy*, 21(7):405–421, 1994.
- [99] G. Verdú, D. Ginestar, V. Vidal, and J.L. Muñoz-Cobo. A consistent multidimensional nodal method for transient calculations. *Annals of Nuclear Energy*, 22(6):395–410, 1995.
- [100] V. Vidal. *Métodos Numéricos para la Obtención de los Modos Lambda de un Reactor Nuclear. Técnicas de Aceleración y Paralelización*. Tesis doctoral, Universidad Politécnica de Valencia, 1997.
- [101] H. F. Walker. Implementation of the GMRES method using Householder transformations. *SIAM J. Sci. Statist. Comput.*, 9:152–163, 1988.
- [102] J. R. Weston and M. Stacey. *Space–Time Nuclear Reactor Kinetics*. Academic Press, 1970.

Bibliográfia

- [103] W.Niethammer. On different splittings and the associated iteration methods. *SIAM Journal on Numerical Analysis*, 16(2):186–200, 1979.
- [104] Z. I. Woźnicki. Nonnegative splitting theory. *Japan Journal of Industrial and Applied Mathematics*, 11:289–342, 1994.
- [105] D. M. Young. *Iterative Solution of Linear Systems*. Academic Press, New York and London, 1971.
- [106] V. G. Zimin and H.Ninokata. Acceleration of the outer iterations of the space-dependent neutron kinetics equations solution. *Annals of Nuclear Energy*, 23(17):1407–1424, 1996.