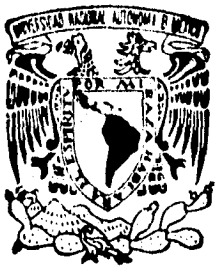


UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

DIVISION DE ESTUDIOS DE POSGRADO

FENOMENOS DE SUPERCONVERGENCIA EN METODOS NODALES PARA PROBLEMAS UNIDIMENSIONALES DE VALORES INICIALES CON APLICACIONES A PROBLEMAS DE TRANSPORTE DE PARTICULAS

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE

MAESTRIA EN CIENCIAS

MATEMATICAS

P R E S E N T A

BLANCA EVELIA FLORES SOTO

DIRECTOR DE TESIS: DR. JEAN-PIERRE HENNART BOUDET

MEXICO, D. F. 1996

A MIS PADRES, por haberme dado la vida y enseñarme que la mejor herencia es la educación.

A MI MADRE, por soportar mi ausencia.

A MI PADRE, por estar con nosotros en todo momento.

A MIS HERMANOS:

SALVADOR, que crecimos juntos descubriendo y abriendo camino a los demás.

ANA, mi única hermana.

TOMAS, el más chico, esperando que aprendas de todos nosotros solo lo mejor.

A MI TIA LUZ, que siempre ha estado con nosotros y que siempre estará, así como a MI ABUELA que la queremos tanto.

A MIS AMIGOS, que no enlisto por temor a olvidarme de alguno, gracias por ser como son.

A TI ... , por ser y estar conmigo hoy, esperando que sea siempre.

AGRADECIMIENTOS

Deseo expresar mi más sincero agradecimiento al Dr. J. P. Hennart y a M. C. Edmundo del valle por su confianza y apoyo, así como también por el empeño y tiempo dedicados para la realización de esta tesis.

Al Dr. Jesús López Estrada por su orientación y apoyo durante mis estudios de Maestría.

Al M. C. José A. Gómez O. por aceptar ser mi tutor de Maestría.

Al rector de la universidad de Sonora, M. C. Jorge Luis Ibarra Mendivil, por haberme apoyado para la obtención del grado.

Al Dr. Fernando Avila Murillo por su apoyo, confianza y consejos durante mucho tiempo.

A Eloisa y Ezequiel por apoyarme y recibirme en su casa.

A los miembros del jurado por todas sus sujerencias. .

A la Universidad de Sonora por la oportunidad y el apoyo brindado para hacer realidad esta tesis.

Al CONACYT por haberme otorgado una beca para realizar mis estudios de Maestría.

Y por último a todas aquellas personas e instituciones que de manera directa o indirecta hicieron posible esto.

Contenido

Indice de Tablas

Indice de Figuras

RESUMEN

En la mayoría de los casos en los que se utiliza una ecuación diferencial o integro-diferencial para describir algún fenómeno físico, encontrar su solución exacta es muy complicado razón por la cual se simplifica el problema y/o se resuelve por métodos numéricos tal es el caso de la ecuación de transporte de partículas, en especial la ecuación de transporte de neutrones en un reactor nuclear.

Como los que los métodos nodales surgieron a raíz de problemas en Ingeniería Nuclear serán utilizados en este trabajo.

Los métodos nodales se basan en aproximar la solución de dicha ecuación por una combinación lineal de *funciones base y momentos*. En el Capítulo 2 se definen los momentos y se obtienen las funciones base para el problema de valores iniciales que trabajamos (ecuación de transporte de partículas).

Para la solución de este problema se conocen dos tipos de métodos (del tipo nodal), que son: el *método de momentos continuos* y el de *momentos discontinuos*, [11], [12] y [13]; en el Capítulo 3 se hace una descripción general de estos métodos y se presentan algunos ejemplos.

Además, se sabe que para estos dos métodos se tienen resultados de superconvergencia. En un Apéndice A se reportan los teoremas con los que se aseguran estos resultados y en la primera sección del Capítulo 4, se resumen de los resultados numéricos que confirman lo anterior.

Utilizamos estos resultados de superconvergencia para elevar el orden de convergencia continuo (CCO), el cual es el único (de los presentados en [13]) que no exhibe superconvergencia; en la segunda parte del Capítulo 4 se aclara el significado del término *superconvergencia*, y se muestra la forma de elevar el CCO.

De esta forma, diremos que la finalidad de este trabajo es el de *obtener órdenes de convergencia más altos que los conocidos en cualquier punto del dominio* del problema (en este caso: de valores iniciales).

Para lograrlo implementaremos una técnica que llamaremos *postprocesamiento*, que describiremos y explicaremos con detalle en la segunda sección del Capítulo 4.

CAPITULO 1

INTRODUCCION

Muchos modelos que describen un sin-número de procesos físicos se expresan en forma de ecuaciones diferenciales, tal es el caso de los reactores nucleares, en los que por ejemplo se desea conocer la población de neutrones en cierta región del reactor (densidad de neutrones) que contiene una mezcla arbitraria, pero conocida, de materiales. En los modelos más sencillos, estas ecuaciones son ordinarias, es decir la función incógnita depende de solo una variable independiente y por regla general no son lineales.

En la industria nuclear resulta de gran importancia el disponer de herramientas de cálculo que permitan la predicción del comportamiento del núcleo del reactor nuclear, es decir una vez que conocida la densidad de neutrones en el reactor puede emplearse para predecir la forma en la que se llevan a cabo las reacciones nucleares (fisión, dispersión, etc.) ya sea en estado estacionario o transitorio. Estas herramientas resultan ser programas de computadora en los que los diseñadores de los reactores nucleares pueden estudiar diferentes opciones de operación o tipos de combustible que permitan cubrir los requerimientos de producción de energía respetando las normas de seguridad.

Las características del núcleo de un reactor cambian a medida que se va consumiendo el combustible. El estudio de estos cambios se efectúa analizando el comportamiento del reactor en estado estacionario a diferentes tiempos de "vida" de éste. Esto permite desarrollar estrategias que aseguren el quemado eficiente y seguro del combustible.

El problema central de la física de reactores puede establecerse, en forma simple, como el cálculo en cualquier tiempo de las características de la población de neutrones en cierta región del espacio que contiene una mezcla arbitraria, pero conocida de materiales. En particular deseamos conocer el número de neutrones en cualquier volumen, que viajan en cierta dirección.

El determinar las características de la población de neutrones se simplifica por el hecho de que, en muchos casos de interés, esta población es tan grande que los neutrones pueden ser tratados como un fluido. Así, no tenemos que enfrentar el problema de seguir con detalle la vida de cada neutrón.

Cuando el "fluido" de neutrones se encuentra en un medio material los neutrones interactúan con éste, y es describiendo esta interacción en forma matemática que obtenemos una ecuación integro-diferencial, llamada ecuación de transporte (de partículas), que si pudiéramos resolverla, podríamos diseñar y predecir el comportamiento de los reactores nucleares en forma muy precisa.

Más adelante, en este mismo capítulo, describiremos sin mucho detalle la forma de obtener la ecuación de transporte de partículas. El estado estacionario esta ecuación es en lo que nos centraremos en nuestra atención a lo largo de este trabajo.

Una vez que se tiene el modelo matemático (ecuación integro-diferencial) es necesario resolverla, hasta aquí no hemos mencionamos la forma de resolver la ecuación de transporte de partículas, existe la posibilidad de resolverla de una forma analítica para casos sencillos y bajo ciertas suposiciones.

Otra forma de resolverla es utilizando métodos numéricos, los cuales han sido desarrollados después de la segunda guerra mundial y en paralelo a la aparición y desarrollo de las computadoras digitales cada vez más poderosas.

El primer método que apareció para resolver de manera aproximada problemas en los que aparecieran derivadas fue el de diferencias finitas. Más

tarde aparecieron los métodos de residuos pesados clásico, así como los métodos de elementos finitos; siguiéndoles los métodos de elementos finitos nodales.

A continuación describiremos brevemente en qué consisten los métodos que mencionamos con anterioridad.

Aproximar la solución u, que en nuestro caso sería la solución de la ecuación de transporte y que nos indica la densidad de neutrones en cierta dirección, con el método de *diferencias finitas* consiste en representar dicha solución por un número finito de valores en puntos particulares del dominio donde deseamos conocerla. La discretización por el método de diferencias finitas de cualquier problema implica dos etapas distintas: la primera es la discretización del dominio, Ω. Es decir, reemplazar a Ω por un número finito de puntos x_i, donde queremos aproximar la solución. La segunda etapa, es la discretización de la ecuación; que consiste en expresar las diferentes derivadas que aparecen en la ecuación, en un punto dado por el valor de u en este punto y algunos otros puntos vecinos cercanos, por ejemplo para aproximar la primera derivada se puede hacer de la forma $\frac{du}{dx} = \frac{u_{i+1}-u_i}{h}$, donde $h = x_{i+1} - x_i$, $u_{i+1} = u(x_i)$ y $u_i = u(x)$, con los x_i's los puntos de discretización del dominio; de forma similar se puede hacer para aproximar derivadas de orden más alto.

El otro método utilizado para resolver este tipo de problemas es el método de *residuos pesados clásico* que utiliza la idea de balance. Por "balance", queremos decir que la ecuación original se satisface "en promedio" sobre cada celda que forma la discretización del dominio, es decir, la integral sobre cada celda de la ecuación es cero. La idea básica consiste en buscar una aproximación a la solución mediante una expresión analítica de la forma general:

$$u_h(x) = \sum_{i=1}^{N} U_i \cdot u_i(x) \tag{1.1}$$

donde las incógnitas son los coeficientes U_i y las $u_i(x)$ son funciones dadas linealmente independientes, globales en el sentido de que se extienden a todo el dominio donde se desea aproximar la solución, debido a ésto corremos el

riesgo de obtener sistemas mal condicionados ya que este método nos lleva a un sistema de la forma $AU = F$, donde $A = [a_{ij}]$ es una matriz $N \times N$, U y F vectores N-dimensionales, en el cual $a_{ij} \neq 0$ lo que hace que A sea llena y sin ninguna estructura en especial. Pero aún así los métodos de *residuos pesados* tienen una gran ventaja respecto a los de *diferencias finitas*: la solución aproximada puede estimarse en cualquier punto del dominio y no solamente en los puntos de la malla como en el método de diferencias finitas. La desventaja está en la necesidad de calcular una serie de integrales que por lo general, no son triviales.

Hasta aquí podemos mencionar cuales son las principales diferencias entre el método de *diferencias finitas* y el de *residuos pesados:*

1. En el método de *diferencias finitas* la aproximación es discreta contrario al método de *residuos pesados*, en el que podemos aproximar la solución en cualquier punto.

2. En *diferencias finitas* las incógnitas básicas son los valores de la aproximación en los puntos discretos de la malla, en cambio, en el de *residuos pesados* estas incógnitas son los coeficientes U_i de (1.1), que no tienen un sentido en particular.

3. Determinar el sistema de ecuaciones algebraicas que se satisface para *diferencias finitas* es trivial, siempre y cuando se tenga una geometría regular, mientras que en *residuos pesados* puede ser más o menos complicado dependiendo del método particular que se utilice (subdominios, colocación, Galerkin).

4. El sistema obtenido con *diferencias finitas* es hueco y estructurado, pero en *residuos pesados* (en su forma clásica) la matriz del sistema puede llegar a ser llena y en muchos casos bastante mal condicionada, dependiendo del problema.

Conociendo las ventajas y desventajas de los métodos de *diferencias finitas* así como de *residuos pesados* mencionaremos en qué consiste el *método de elementos finitos.*

El método de *elementos finitos* [18], es un método de discretización muy general que combina dos factores: primeramente (como en residuos pesados) la ecuación original no es considerada tal cual, sino que se multiplica por algún peso y se integra llegando a nuevas formulaciones (para problemas de valores a la frontera se dicen formulaciones débiles). Lo segundo es una selección muy particular de las *funciones base* en la aproximación, la cual se escribe en forma semejante a (1.1) y que al mismo tiempo aprovecha las ventajas del método de diferencias finitas y el de residuos pesados. En su forma más simple es un proceso de construir subespacios, los cuales son llamados espacios de elemento finito. Su construcción se caracteriza por:

- una discretización del dominio de interés en dominios elementales.
- La aproximación a la solución es calculable en cualquier punto del dominio, ya que la solución aproximada u_h de u se construye en forma similar a (1.1) donde las u_i (funciones base) se encuentran en un espacio polinomial o casi-polinomial.
- Las funciones base u_i tienen un soporte (la cerradura del dominio donde son diferentes de cero) tan pequeño como sea posible.

Por otra parte, los *métodos de elementos finitos nodales* son métodos considerados intermedios entre los métodos de elementos finitos y los métodos de diferencias finitas. Los métodos nodales fueron introducidos a finales de la década de los 70's en cálculos de reactores nucleares; estos métodos poseen características que se consideran muy favorables heredadas de los métodos de diferencias finitas y de elementos finitos. Del método de diferencias finitas posee la característica de producir sistemas de ecuaciones cuyas matrices son relativamente huecas lo que facilita su manipulación. Al igual que en el método de elementos finitos los métodos nodales aproximan la función de interés por una combinación lineal de unas funciones, normalmente polinomiales llamadas *funciones base* y *momentos*, dando por resultado poder calcular la función en cualquier punto dentro del dominio de interés.

Como pudimos observar del resumen de los métodos numéricos existentes y en relación al problema que trabajaremos (transporte de partículas en un

reactor nuclear), los métodos idóneos para atacarlo son los nodales, razón por la cual serán utilizados. Además, en el Capítulo 2 se encontrarán las funciones base y se definirán los momentos, con los cuales expresaremos la solución aproximada a la ecuación de transporte que manejaremos en este trabajo.

Con anterioridad mencionamos que nuestro principal problema será el concerniente a la ecuación de transporte de partículas (que es un problema de valores iniciales), a continuación y de una forma breve mencionaremos los principales puntos para obtenerla, en el caso de una dimensión, en estado estacionario e isotrópico.

1.1 PROBLEMA DE VALORES INICIALES

Como se dijo anteriormente, uno de los principales problemas en la teoría de reactores nucleares es la distribución de los neutrones en el reactor nuclear. Esta distribución se determina por el proceso de transporte neutrónico (ecuación de transporte). Para obtener dicha ecuación (en estado estacionario e isotrópico) es necesario efectuar un balance de neutrones en el reactor.

En nuestro caso, el unidimensional, se obtiene la ecuación de transporte de partículas sobre un dominio $\overline{\Omega} = [a, b]$ considerando:

1. rapidez con la que se pierden neutrones debido a:

   - que son removidos, que se representa por

     $$q_t u(x, \mu),$$

     donde q_t es la sección transversal total, es decir la cantidad de neutrones (en promedio) que interactúan por absorción o dispersión por unidad de longitud y u es el flujo angular de neutrones, es decir, la densidad de neutrones que se mueven en cierta dirección y μ la dirección angular en la que se desplazan los neutrones
   - fugas, matemáticamente

     $$\mu \frac{\partial u}{\partial x},$$

reactor nuclear), los métodos idóneos para atacarlo son los nodales, razón por la cual serán utilizados. Además, en el Capítulo 2 se encontrarán las funciones base y se definirán los momentos, con los cuales expresaremos la solución aproximada a la ecuación de transporte que manejaremos en este trabajo.

Con anterioridad mencionamos que nuestro principal problema será el concerniente a la ecuación de transporte de partículas (que es un problema de valores iniciales), a continuación y de una forma breve mencionaremos los principales puntos para obtenerla, en el caso de una dimensión, en estado estacionario e isotrópico.

1.1 PROBLEMA DE VALORES INICIALES

Como se dijo anteriormente, uno de los principales problemas en la teoría de reactores nucleares es la distribución de los neutrones en el reactor nuclear. Esta distribución se determina por el proceso de transporte neutrónico (ecuación de transporte). Para obtener dicha ecuación (en estado estacionario e isotrópico) es necesario efectuar un balance de neutrones en el reactor.

En nuestro caso, el unidimensional, se obtiene la ecuación de transporte de partículas sobre un dominio $\overline{\Omega} = [a, b]$ considerando:

1. rapidez con la que se pierden neutrones debido a:

   - que son removidos, que se representa por

     $$q_t u(x, \mu),$$

     donde q_t es la sección transversal total, es decir la cantidad de neutrones (en promedio) que interactúan por absorción o dispersión por unidad de longitud y u es el flujo angular de neutrones, es decir, la densidad de neutrones que se mueven en cierta dirección y μ la dirección angular en la que se desplazan los neutrones
   - fugas, matemáticamente

     $$\mu \frac{\partial u}{\partial x},$$

2. rapidez con la que se ganan neutrones por:

- dispersión, es decir

$$\frac{q_s}{2}\int_{-1}^{1} u(x,\mu')\,d\mu',$$

donde q_s es la sección transversal de dispersión, que representa la cantidad de neutrones que se dispersan (en promedio) por unidad de longitud.

- una fuente independiente, que llamaremos Q.

Entonces, en estado estacionario la rapidez de ganancia de neutrones es igual a la rapidez de pérdidas, de tal forma que:

$$\mu\frac{\partial u}{\partial x} + q_t u = \frac{q_s}{2}\int_{-1}^{1} u(x,\mu')\,d\mu' + Q, \tag{1.2}$$

donde Q depende (en general) de x y u, el flujo angular, debe ser una función real positiva y continua que cumple con las condiciones

$$u(a,\mu) = 0 \quad \text{para} \quad \mu > 0, \quad u(b,\mu) = 0 \quad \text{para} \quad \mu < 0,$$

por lo tanto, el espacio al que pertenecerá u será

$$U = \left\{u \mid u \in C^0; \quad u(a,\mu) = 0 \text{ si } \mu > 0, \text{ o } u(b,\mu) = 0 \text{ si } \mu < 0\right\}. \tag{1.3}$$

Como podemos observar la ecuación (1.2) es monoenergética e independiente del tiempo, la independencia en t (tiempo) le da el carácter de autónoma.

En nuestro caso unidimensional, se puede obtener una aproximación llamada S_N, la cual consiste en discretizar la variable dirección angular μ' en un conjunto de N direcciones y aproximar las integrales de la ecuación de transporte (1.2) con una cuadratura numérica, siendo la más común la de Gauss-Legendre en la que (por conveniencia) se hace que las raíces de los

polinomios correspondan a las direcciones angulares seleccionadas en la discretización de esta variable.

Discretizar μ, consiste en elegir un conjunto discreto de valores de μ, digamos μ_n; $n = 1, \ldots N$, de esta manera la integral que aparece en (1.2) se aproxima de la forma

$$\int_{-1}^{1} u(x, \mu')\, d\mu' \cong \sum_{m=1}^{N} \omega_m u_m.$$

donde ω_m son los pesos de cuadratura (de Gauss-Legendre) y u_m es u restringida a los puntos de cuadratura μ_m, es decir $u_m = u\,|_{\mu_m} = u(x, \mu_m)$, con $\mu_m \in (-1, 1) \,/\, \{0\}$, excluyendo el cero para evitar una ecuación degenerada.

Como u es restringida a μ_m entonces, es necesario tomar el resto de los términos que aparecen en (1.2) bajo la misma restricción, obteniendo

$$\mu_i \frac{du_i}{dx} + q_t u_i = q_s \sum_{m=1}^{N} \omega_m u_m + Q_i, \qquad i = 1, 2, \ldots N \tag{1.4}$$

un sistema de ecuaciones diferenciales ordinarias, discretizada para ciertos valores de μ bien definidos, entonces las condiciones iniciales solo deben cumplirse para esos valores de μ_i lo cual hace que estas condiciones sean más débiles que la que se tenía con anterioridad (antes de efectuar la discretización) y se expresan como

$$u_i(a) = 0 \quad \text{para} \quad \mu_i > 0, \quad u_i(b) = 0 \quad \text{para} \quad \mu_i < 0, \tag{1.5}$$

donde a y b son los extremos de cada subintervalo en el que se discretizó μ.

Conviene observar que el sistema (1.4) no es totalmente monodireccional pues en su lado derecho aparecen flujos angulares con direcciones μ_m, $m = 1, 2, \ldots N$, (u_m). Resulta natural, aplicar un procesó iterativo que nos permita reducir nuestro sistema (1.4) a uno totalmente desacoplado (es decir, un sistema de ecuaciones monodireccionales). En efecto, dadas u_i^0, $i = 1, 2, \ldots N$, se calculan las soluciones u_i^{n+1}, $i = 1, 2, \ldots N$, del sistema

$$\mu_i \frac{du_i^{n+1}}{dx} + q_t u_i^{n+1} = f_i^n, \quad i = 1, \ldots N, \tag{1.6}$$

donde

$$f_i^n = \frac{q_s}{2} \sum_{m=1}^{N} \omega_m u_m^n + Q_{n_i} \quad i = 1, \ldots N.$$

Así hemos llegado al sistema de ecuaciones (1.6), con condiciones iniciales (1.5), pero podemos observar que, en general tenemos una ecuación de la forma

$$\mu \frac{du}{dx} + qu = f, \tag{1.7}$$

en cada una de las celdas que discretizan a μ, es decir, resolver (1.6) significa resolver N ecuaciones monodireccionales (una sola dirección angular, μ) de transporte monoenergéticas de la forma (1.7), con una función fuente f que depende de u, donde una solución aproximada u_i se determina de izquierda a derecha sobre intervalos sucesivos para $\mu > 0$, y de derecha a izquierda para $\mu < 0$.

Entonces, como resolver (1.4) es equivalente a resolver varias ecuaciones de la forma (1.7), por lo que en adelante nos referiremos a (1.7) como el problema a tratar junto con las condicones iniciales (1.5).

CAPITULO 2

FUNCIONES BASE

Como mencionamos en el capítulo anterior, en los métodos nodales la solución aproximada, que llamaremos u_h de u, la solución de (1.7), se expresa como una combinación lineal de unas funciones llamadas base, así como de momentos.

Cabe aclarar que la selección de las *funciones base* no es única, por lo que, en este capítulo, nuestro problema será estudiar como escoger dichas *funciones base* para el problema de valores iniciales (1.7) con condiciones iniciales (1.5). Las *funciones base* podrían seleccionarse de una manera global, es decir, definidas en todo el intervalo $[a, b]$, pero se corre el riesgo de que la matriz del sistema resulte mal condicionada y llena. Debido a lo anterior haremos una selección particular de las *funciones base* $u_i(x)$.

Para realizar lo anterior y como las funciones base se escogen por partes, primeramente discretizaremos el dominio $\overline{\Omega} = [a, b]$, reemplazándolo por un número finito de dominios elementales (subintervalos) Ω_e, $e = 1, \ldots, E$ tales que

$$\bigcup_{e=1}^{E} \Omega_e = \Omega \quad \text{y} \quad \Omega_e \cap \Omega_f = \emptyset \quad \text{si } e \neq f.$$

En nuestro caso (unidimensional) es equivalente a introducir una malla o partición sobre Ω constituida por E intervalos no necesariamente del mismo tamaño:

$$a \equiv x_1 < x_2 < \ldots < x_{E-1} < x_E \equiv b,$$

de tal forma que el elemento Ω_e es el intervalo (x_e, x_{e+1}) de tamaño $h_e = x_{e+1} - x_e$. Llamaremos h al $\max_e h_e$, de esta manera h nos indica lo grueso o fino de la partición.

En algunas ocasiones nos referiremos a Ω_e como el intervalo (x_l, x_r).

La forma de construir las N_e *funciones base*, que denotaremos por u_i, es un punto importante. Estas se escogen por partes, es decir, restringidas al elemento Ω_e van ha ser funciones sencillas para su manejo en la computadora, usualmente polinomios (no siempre), estas funciones sencillas pertenecen localmente (sobre Ω_e) a un cierto espacio funcional S_e, por ejemplo los polinomios de grado menor o igual a k, de dimensión $\dim S_e$ definido completamente por un conjunto de funcionales lineales que llamaremos grados de libertad (en el sentido de que son los parámetros que definen a u_h) D_e locales, de cardinalidad $\text{card } D_e = \dim S_e$; de hecho los grados de libertad son funcionales lineales $L_i^e : S_e \to R$. En la mayoría de los casos estas funcionales son valores de u_h en puntos particulares de Ω_e y posiblemente de algunas de sus derivadas.

En los métodos nodales también encontramos parámetros del tipo momentos.

Además tenemos que unir estas funciones base por pedazos para que las $u_i(x)$ pertenezcan globalmente a U ($S_h \subset U$ donde U es el espacio de funciones dado por (1.3)), aunado a esto deben (las u_i) tener un soporte pequeño, es decir, que sean diferentes de cero en un número muy pequeño de elementos, de tal forma que las *funciones base* sean casi ortogonales y nos lleven a matrices huecas, con pocos elementos diferentes de cero y concentrados alrededor de la diagonal principal.

Resumiendo los principales puntos en la construcción de S_h ($u_h \in S_h$) tenemos:

1. Discretización de Ω en elementos finitos Ω_e, $e = 1, \ldots, E$, de diámetro máximo h.

2. Las *funciones base* $u_i(x)$ restringidas al elemento Ω_e para toda e, deben ser *sencillas*.

3. Las $u_i(x) \in U$ globalmente.

4. Reducir el soporte de $u_i(x)$ a un número pequeño de elementos.

Para que el punto 2) tenga sentido, S_e debe estar definido por un conjunto D_e de *grados de libertad* y se necesita que $\dim S_e = \text{card } D_e = N_e$, esto para que la determinación u_h, formada con una combinación lineal de funciones base y momentos (más adelante se dará su forma explícita), sea única, es decir, se requiere que D_e sea $S_e - unisolvente$, en otras palabras: dados N_e reales α_i existe una única función $u \in S_e$ tal que

$$L_i^e(u) = \alpha_i, \ i = 1, \ldots, N_e, \quad L_i^e \in D_e.$$

Una forma de probarlo es exhibiendo un conjunto N_e de *funciones base* u_i, $i = 1, \ldots N_e$, las cuales satisfacen

$$L_i^e(u_j) = \delta_{ij} = \begin{cases} 1 & \text{si} \quad i = j \\ 0 & \text{si} \quad i \neq j \end{cases} ; \quad i = 1, \ldots, N_e, \tag{2.1}$$

cosa que haremos en este capítulo.

Con relación al punto 3), como las $u_i(x)$ están definidas localmente, es decir sobre Ω_e, en un cierto espacio funcional S_e definido por un conjunto D_e de grados de libertad locales, con card $D_e = \dim S_e$, entonces tenemos que unir estas funciones definidas por pedazos. Al unirlas estamos haciendo que $u_i \in S_h \subset U$ globalmente y así los N_e grados de libertad por elemento se transforman en N grados de libertad globales, a menos de que no se le pida ninguna clase de continuidad a $u_i(x)$, si éste es el caso $N < \sum_{e=1}^{E} N_e$, es decir, algunos grados de libertad globales serán compartidos localmente.

En general, D_e y S_e son los mismos para cada uno de los elementos, así que simplemente los llamaremos D y S, respectivamente. Con esto, resulta muy conveniente desarrollar las *funciones base* sobre un intervalo de referencia

$\tilde{\Omega} = [-1, 1]$ y regresar al intervalo real por medio de una transformación lineal invertible, $[x_l, x_r] \xrightarrow{T} [-1, 1]$, por ejemplo $T \equiv \frac{2x - x_l - x_r}{h} \equiv \zeta$.

Hasta ahora, las *funciones base* están definidas localmente por un conjunto D de grados de libertad y por un espacio funcional S, además definiremos los grados de libertad como *momentos*, de la función sobre el elemento.

Recordando lo que se ha dicho hasta ahora, en este capítulo y el anterior, nuestro problema de valores iniciales (1.4) con condiciones iniciales (1.5) restringido a un solo elemento $\Omega_e = [x_l, x_r]$ nos queda

$$\mu \frac{du(x)}{dx} + qu(x) = f, \tag{2.2}$$

con condición inicial

$$u(a) = 0 \quad \text{si} \quad \mu > 0,$$

o

$$u(b) = 0 \quad \text{si} \quad \mu > 0,$$

donde a y b son los extremos del subintervalo correspondiente al valor μ, que podemos rescribirla como:

$$\frac{du}{dx} - \frac{1}{\lambda} u = \tilde{f}, \tag{2.3}$$

donde $1/\lambda = -hq/2\mu$ y $\tilde{f} = fh/2\mu$, con $\mu \neq 0$, h es la longitud del intervalo $[x_l, x_r]$ y la derivada de u es con respecto a $\xi = (2x - x_r - x_l)/h$, $\xi \in [-1, 1]$.

2.1 POLINOMIOS Y MOMENTOS DE LEGENDRE

Como mencionamos anteriormente, las *funciones base* deben ser *sencillas* y podemos tomarlas como polinomios o casipolinomios. Aún así, nuestro problema sigue siendo construir dichas *funciones base*, para lo cual utilizaremos los *polinomios de Legendre*.

El *polinomio de Legendre* cumple con:

$$P_k(+1) = 1 \qquad y \qquad P_k(-1) = (-1)^k, \tag{2.4}$$

además

$$\int_{-1}^{1} P_k(x)P_l(x)dx = \delta_{kl}N_k, \tag{2.5}$$

y

$$P_k(-x) = (-1)^k P_k(x), \tag{2.6}$$

donde δ_{kl} es la delta de Kronecker y

$$N_k = \int_{-1}^{1} P_k^2(x)dx = \frac{2}{2k+1}. \tag{2.7}$$

En el caso de un intervalo cualesquiera $[x_l, x_r]$ el polinomio de Legendre de orden k correspondiente es:

$$p_k(x) = P_k(\tfrac{2x-x_l-x_r}{x_r-x_l}).$$

Hasta donde hemos mencionado, las *funciones base* están definidas localmente por un conjunto D de grados de libertad y por un espacio funcional S. En particular definiremos los grados de libertad como *momentos* de la función sobre el elemento que estarán definidos sobre el intervalo de referencia $\hat{\Omega} = [-1, 1]$ como:

$$m_l(u) = u(-1) = U_l, \tag{2.8}$$

$$m_r(u) = u(1) = U_r, \tag{2.9}$$

$$m_c^i(u) = \frac{1}{N_i} \int_{-1}^{1} P_i(x)u(x) = U_c^i \quad \text{para} \quad i = 0, \ldots, k, \tag{2.10}$$

donde los subíndices l, r, y c indican "left", "right" y "cell", respectivamente, y $P_i(x)$ es el polinomio normalizado de Legendre de grado i. Las expresiones (2.8) y (2.9) son *momentos*. Gracias a la normalización adoptada, las expresiones (2.8), (2.9) y (2.10) son válidas sobre cualquier celda $\Omega_e = (x_l, x_r)$ y no sólo sobre la celda $\hat{\Omega}$, por ejemplo, sobre Ω_c tenemos:

$$m_c^i(u) = \frac{\int_{x_l}^{x_r} p_i(x)u(x)dx}{\int_{x_l}^{x_r} (p_i(x))^2 dx} \qquad \text{para} \qquad i = 0, \ldots, k,$$

por lo que, dada cualquier celda Ω_e, ir y regresar de Ω_e a la celda de referencia $\hat{\Omega}$ es muy fácil. En efecto, todas las *funciones base* que serán introducidas son válidas para $\hat{\Omega}$ en términos de los P_k.

Como mencionamos anteriormente, el conjunto de grados de libertad está formado por los *momentos* de $u_h(x)$ sobre el elemento, dichos momentos son las formas lineales definidas en (2.8), (2.9) y (2.10).

En lo sucesivo $u(x)$ será aproximada sobre cada celda $[x_l, x_r]$ o, equivalentemente sobre $\hat{\Omega}$ por una función u_h perteneciente a alguno de los siguientes espacios funcionales

$$\mathcal{P}_k = \left\langle \left\{ 1, x, \ldots, x^k \right\} \right\rangle \tag{2.11}$$

$$\mathcal{E}_k = \left\langle \left\{ 1, x, \ldots, x^{k-1}, \exp(\frac{1}{\lambda}x) \right\} \right\rangle \tag{2.12}$$

donde λ es algún real negativo constante, siempre y cuando μ y q sean considerados constantes por pedazos es decir, sobre Ω_e o, en su defecto, sobre $\hat{\Omega}$.

Y $\exp\left(\frac{1}{\lambda}x\right)$ surge de manera natural al tomar (2.3) en su forma homogénea (i.e. $f = 0$).

Este comportamiento continuo por pedazos de $u_h \sim u$, justifica el hecho de haber hablado de un *formalismo de elemento finito nodal generalizado*, donde *generalizado* significa que el comportamiento por pedazos no se restringe a ser *polinomial* ($\mathcal{P}_k$), sino que puede ser *analítico* ($\mathcal{E}_k$). Por *polinomial* se entiende que escogeremos funciones base polinomiales, es decir estarán expresadas en términos de polinomios. Por *analítico* entenderemos que usaremos *funciones base* no polinomiales, podríamos hablar de *funciones base* casipolinomiales. *Nodal* significará que los parámetros básicos que describen a u_h sobre cada celda serán los *momentos:* en los extremos de cada celda y los de celda.

Ya sea que usemos $\mathcal{P}_k$ o $\mathcal{E}_k$, vamos a suponer que $u(x)$ es aproximada por $u_h(x)$ en términos de $k+1$ grados de libertad y para determinar los elementos de D que definen a $u_h(x)$ en cada intervalo se puede decir que en el caso de valores iniciales tenemos dos técnicas (del tipo residuos pesados) llamados *método de momentos continuos* y *método de momentos discontinuos*.

En los métodos de *momentos continuos*, u_h es una aproximación continua de $u(x)$, es decir, u_h es continua de celda a celda esto es $u_h(x_l) = u_l$ en cada celda se toma igual a $u_h(x_r) = u_r$ de la celda anterior, por lo tanto requiere de los valores de u_h en x_l y x_r. Entonces en este caso el conjunto D de grados de libertad está dado por

$$D = \left\{m_l(u_h), m_r(u_h), m_c^i(u_h), \quad \text{para} \quad i = 0, \ldots, k-2\right\}, \tag{2.13}$$

en el espacio

$$S_h = \{u \mid u \in C^0,\ u(a) = 0 \text{ si } \mu > 0,\ u(b) = 0 \text{ si } \mu > 0\},$$

donde a y b son los extremos del subintervalo correspondiente al valor μ.

Y para los métodos de *momentos discontinuos* supondremos que u_h no es continua de celda a celda: es decir, que $u_h(x_r) = u_r$ en el extremo derecho

de cada celda no necesariamente será igual a $u_h(x_l+0)$ de la celda siguiente. De esta forma tenemos que

$$D' = \left\{ m_r(u_h), m_c^i(u_h), \quad \text{para} \quad i = 0, \ldots, k-1 \right\}, \tag{2.14}$$

con el espacio

$$S_h = \{ u \mid u \in C^{-1},\ u(a) = 0 \text{ si } \mu > 0,\ u(b) = 0 \text{ si } \mu > 0 \}.$$

donde, de nuevo, a y b son los extremos del subintervalo correspondiente al valor μ.

Además, tanto en (2.13) como en (2.14) card D =card $D' = k+1$.

La determinación de las *funciones base* se puede hacer sobre la celda de referencia $\hat{\Omega}$, y una vez conocida la representación de u_h sobre $\hat{\Omega}$ es simple obtener u_h sobre la celda real Ω_e por medio de la transformación afín inversa: $x = [(x_r - x_l)\xi + x_r + x_l]/2$.

En resumen: sobre la celda de referencia $\hat{\Omega}$, u_h está definida por un espacio S y un conjunto de grados de libertad D (funcionales lineales L_i actuando sobre S) dados por

$$S = \mathcal{P}_k, \text{ o } \mathcal{E}_k,$$

y D por (2.13) en el caso de momentos continuos y (2.14) en el de momentos discontinuos. Con dim$S$ =card D =card $D' = k+1$.

2.2 METODOS DE MOMENTOS CONTINUOS

Con $\mathcal{P}_k$, o $\mathcal{E}_k$, supondremos que sobre cada celda $\Omega_e = (x_l, x_r)$, $u(x)$ es aproximada en $\mathcal{P}_k$ o $\mathcal{E}_k$ en términos de $k+1$ grados de libertad los cuales, de acuerdo con (2.13) y la definición de momentos, serán

$$\begin{aligned} U_r &= u_h(x_r) = m_r(u_h), \\ U_l &= u_h(x_l) = m_l(u_h), \\ U_c^i &= m_c^i(u_h(x)) \quad \text{para} \quad i = 0, \ldots, k-2. \end{aligned}$$

Llamaremos u_l, u_r y u_c^i a las *funciones base* correspondientes a $m_l(u_h) = U_l$, $m_r(u_h) = U_r$ y $m_c^i(u_h) = U_c^i$. De esta forma una función nodal general en $\mathcal{P}_k$, o $\mathcal{E}_k$ se puede expresar como sigue

$$u_h(x) = U_l u_l(x) + U_r u_r(x) + \sum_{i=0}^{k-2} U_c^i u_c^i(x), \qquad x \in [x_l, x_r]. \tag{2.15}$$

Así nuestro problema se reduce, primeramente a encontrar las *funciones base* $u_l(x)$, $u_r(x)$ y $u_c^i(x)$. Una manera de hallar estas funciones base es expresandolas en términos de los polinomios o casi-polinomios de Legendre, según sea el caso *polinomial* o *analítico*.

2.2.1 El caso Polinomial ($\mathcal{P}_k$)

En este caso, deseamos encontrar las expresiones de las *funciones base* $u_l(x)$, $u_r(x)$ y $u_c^i(x)$, asociadas a la frontera izquierda, derecha y celda respectivamente, en términos de los polinomios normalizados de Legendre P_k [1] (en el sentido de (2.1)). Antes de proseguir recordemos que es muy conveniente trabajar directamente con los polinomios de Legendre P_l, debido a que se puede aprovechar su ortogonalidad, por lo que pensaremos en el espacio de polinomios de grado a lo más k, $\mathcal{P}_k$ generado por los primeros k polinomios de Legendre, es decir

$$\mathcal{P}_k = \langle \{P_0, \ldots, P_k\} \rangle. \tag{2.16}$$

Hecho lo anterior procederemos a encontrar las expresiones para las *funciones base*.

Primeramente lo haremos para el extremo izquierdo, la cual tiene la forma general

[1] Entonces u_h también esta en terminos de ellos, que han sido utilizados por tradición en problemas de transporte por varios investigadores desde la guerra por ejemplo en métodos como el de armonicos esféricos

$$u_l(x) = \sum_{j=0}^{k} \alpha_j P_j(x), \tag{2.17}$$

y debe cumplir con

$$\begin{aligned} m_l(u_l(x)) &= 1 \\ m_r(u_l(x)) &= 0 \\ m_c^i(u_l(x)) &= 0 \quad i = 0, \ldots, k-2, \end{aligned} \tag{2.18}$$

que son las condiciones de cardinalidad de los polinomios de interpolación de Legendre extendidas al caso de momentos y de acuerdo con (2.10) tenemos que

$$\int_{-1}^{1} P_i(x) u_l(x) dx = 0 \quad i = 0, \ldots, k-2,$$

de aquí observamos que $u_l(x)$ es ortogonal a $P_i(x)$ para $i = 0, \ldots, k-2$, y de acuerdo con (2.17) podemos expresar a $u_l(x)$ como una combinación lineal de $P_{k-1}(x)$ y $P_k(x)$, es decir

$$u_l(x) = \alpha_{k-1} P_{k-1}(x) + \alpha_k P_k(x), \tag{2.19}$$

y relacionando (2.8) y (2.9) con (2.17) tenemos que

$$u_l(1) = 1 \qquad y \qquad u_l(-1) = 0, \tag{2.20}$$

entonces evaluando (2.19) en 1 y -1, verificando que se cumpla (2.20) y utilizando (2.4) obtenemos un sistema de dos ecuaciones con dos incógnitas (α_{k-1} y α_k), resolviéndolo obtenemos

$$\alpha_{k-1} = \frac{1}{2}(-1)^{k-1} \qquad y \qquad \alpha_k = -\frac{1}{2}(-1)^{k-1},$$

de esta forma tenemos que

$$u_l(x) = \frac{1}{2}(-1)^{k-1} [P_{k-1}(x) - P_k(x)]. \tag{2.21}$$

Procediendo de una manera análoga, se obtiene para el extremo derecho

$$u_r(x) = \frac{1}{2}\left[P_{k-1}(x) + P_k(x)\right]. \tag{2.22}$$

Por último, para las *funciones base* asociadas a los momentos en la celda tenemos que su forma general es:

$$u_c^i(x) = \sum_{j=0}^{k} \gamma_j P_j(x), \tag{2.23}$$

con las restricciones de que

$$\begin{gathered} m_l(u_c^i(x)) = 0, \\ m_r(u_c^i(x)) = 0, \\ m_c^j(u_c^i(x)) = \delta_{ij} \quad j = 0, \ldots, k-2, \end{gathered}$$

de la última restricción tenemos que $u_c^i(x)$ es ortogonal a P_j para $j \neq i$ entonces, podemos expresarla como:

$$u_c^i(x) = \gamma_i P_i(x) + \sum_{j=1}^{2} \eta_j P_{k+j-2}(x),$$

evaluando en 1 y -1, y de acuerdo con (2.8) y (2.9) obtenemos un sistema de ecuaciones, que al resolver obtenemos valores para γ_i y η_j para $j = 1, 2$ que nos llevan a:

$$u_c^i(x) = P_i(x) - P_{k-2+m(i)}(x) \qquad i = 0, \ldots, k-2, \tag{2.24}$$

donde $m(i) = 1$ o 2, tal que $k - 2 + m(i)$ e i tengan la misma paridad.

Podemos decir entonces que obtuvimos las expresiones que deseábamos encontrar, así (2.21), (2.22) y (2.24) forman el conjunto de *funciones base* para el caso del *método polinomial continuo*.

2.2.2 El caso Analítico ($\mathcal{E}_k$)

Similarmente, como en el caso polinomial, $\mathcal{E}_k$ se puede elegir (tomar) como el espacio generado por los primeros $k-1$ polinomios de Legendre y un término más que nos complete la base, que es la solución al problema homogéneo (2.3) $\exp(\frac{1}{\lambda}x)$, que lo hace que no sea polinomial en su totalidad, es decir:

$$\mathcal{E}_k = \left\langle \left\{ P_o, \ldots, P_{k-1}, \exp(\frac{1}{\lambda}x) \right\} \right\rangle, \tag{2.25}$$

o bien por

$$\mathcal{E}_k = \left\langle \left\{ P_o, \ldots, P_{k-1}, \exp_k(\frac{1}{\lambda}x) \right\} \right\rangle, \tag{2.26}$$

esto último con el fin de obtener una base ortogonal, donde $\exp_k(\alpha x)$ está definido como $\exp(\alpha x)$ menos sus componentes $P_o, \ldots, P_{k-1}$, es decir

$$\exp_k(\alpha x) = \exp(\alpha x) - \sum_{i=0}^{k-1} m_c^i [\exp(\alpha x)] P_i(x), \tag{2.27}$$

y

$$\exp_k(-\alpha x) = \exp(-\alpha x) - \sum_{i=0}^{k-2} (-1)^i m_c^i [\exp(\alpha x)] P_i(x), \tag{2.28}$$

con

$$m_c^i [\exp(\alpha x)] = \frac{1}{N_i} \int_{-1}^{1} P_i(x) \exp(\alpha x) dx. \tag{2.29}$$

Además tenemos que

$$\cosh_k(\tfrac{1}{\lambda}x) = \frac{1}{2}\left[\exp_k(\tfrac{1}{\lambda}x) + \exp_k(-\tfrac{1}{\lambda}x)\right],$$
$$\operatorname{senh}_k(\tfrac{1}{\lambda}x) = \frac{1}{2}\left[\exp_k(\tfrac{1}{\lambda}x) - \exp_k(-\tfrac{1}{\lambda}x)\right]. \tag{2.30}$$

Como en el caso polinomial, P_{k-1} y P_k juegan un papel muy especial, ya que ellos definen a $u_l(x)$ y a $u_r(x)$, por lo que es conveniente reemplazar a $\exp_k(\alpha x)$ por

$$Q_k(x) = a_k P_{k-1}(x) + b_k \exp_k(\frac{1}{\lambda}x), \tag{2.31}$$

y como debe satisfacer (2.37), obtenemos un sistema de ecuaciones para a_k y b_k, al resolverlo llegamos a que

$$a_k = \frac{\exp_k(-\frac{1}{\lambda}) - (-1)^k \exp_k(\frac{1}{\lambda})}{\exp_k(-\frac{1}{\lambda}) + (-1)^k \exp_k(\frac{1}{\lambda})}, \tag{2.32}$$

$$b_k = \frac{2(-1)^k}{\exp_k(-\frac{1}{\lambda}) + (-1)^k \exp_k(\frac{1}{\lambda})}. \tag{2.33}$$

Con estos valores podemos llegar a una representación general de los Q_i's, la que al usar (2.30) se obtiene

$$Q_i(x) = P_i(x) \qquad i = 0, \ldots, k-2, \tag{2.34}$$

$$Q_{k-1}(x) = \begin{cases} \cosh_k(\frac{1}{\lambda}x), & k \text{ impar} \\ \operatorname{senh}_k(\frac{1}{\lambda}x), & k \text{ par} \end{cases}, \tag{2.35}$$

$$Q_k(x) = \begin{cases} \operatorname{senh}_k(\frac{1}{\lambda}x), & k \text{ impar} \\ \cosh_k(\frac{1}{\lambda}x), & k \text{ par} \end{cases}, \tag{2.36}$$

los cuales satisfacen:

$$Q_k(1) = 1 \qquad y \qquad Q_k(-1) = (-1)^k. \tag{2.37}$$

Finalmente podemos definir a $\mathcal{E}_k$ como

$$\mathcal{E}_k = \langle \{P_o, \ldots, P_{k-1} \equiv Q_{k-1}, Q_k\} \rangle .$$

Ahora, procedemos a encontrar la representación de las *funciones base* $u_l(x)$, $u_r(x)$ y $u_c^i(x)$, para este caso (analítico continuo). Primero lo haremos para el extremo izquierdo, es decir, $u_l(x)$. Tenemos que su forma general es

$$u_l(x) = \sum_{j=0}^{k-2} \alpha_j P_{k-2}(x) + \sum_{j=1}^{2} \alpha_{k+j-2} Q_{k+j-2}(x),$$

donde $P_{k-1} \equiv Q_{k-1}$. De igual forma como en el caso polinomial se debe de cumplir (2.18) obteniendo de nuevo que $u_l(x)$ es ortogonal a $P_i(x)$, $i = 0, \ldots, k-2$, por lo tanto podemos expresar a $u_l(x)$ como una combinación lineal de $Q_{k-1}(x)$ y $Q_k(x)$, es decir

$$u_l(x) = \alpha_{k-1} Q_{k-1}(x) + \alpha_k Q_k(x),$$

que de acuerdo con (2.18), (2.8) y (2.9), también se tiene que cumplir (2.20), de esta forma llegamos a un sistema muy similar al caso polinomial, que resolviendo obtenemos finalmente que

$$u_l(x) = \frac{1}{2}(-1)^{k-1} [Q_{k-1}(x) - Q_k(x)]. \tag{2.38}$$

Como se observó, el procedimiento es completamente análogo al caso polinomial, entonces procediendo de igual forma para el extremo derecho y para los momentos de celda obtenemos

$$u_r(x) = \frac{1}{2} [Q_{k-1}(x) + Q_k(x)], \tag{2.39}$$

$$u_c^i(x) = Q_i(x) - Q_{k-2+m(i)}(x) \qquad i = 0, \ldots, k-2, \tag{2.40}$$

donde $m(i) = 1$ o 2, tal que $k - 2 + m(i)$ e i tengan la misma paridad. Y $Q_i(x) = P_i(x)$ para $i = 0, \ldots k-1$.

2.2.3 Momentos de Indice k y k − 1

En ambos casos *polinomial* $(\mathcal{P}_k)$ o *analítico* $(\mathcal{E}_k)$, los grados de libertad de u_h son los momentos de celda $U_c^i = m_c^i(u_h)$, $i = 0, \ldots, k-2$ más $U_l = m_l(u_h)$ y $U_r = m_r(u_h)$, entonces es posible calcular dos momentos más de (2.24) y (2.40), además de que dichas expresiones las utilizaremos más adelante. Por lo tanto necesitamos calcular $m_c^{k-1}(u_h)$ y $m_c^k(u_h)$. De acuerdo con la expresión de u_h (2.15) también será necesario calcular $m_c^{k-1}(u_l)$, $m_c^{k-1}(u_r)$, $m_c^k(u_l)$, $m_c^k(u_r)$, así como $m_c^{k-1}(u_c^i)$, y $m_c^k(u_c^i)$ para $i = 0, \ldots k-2$. Si recordamos la forma que toma $u_l(x)$, $u_r(x)$ y $u_c^i(x)$ para $i = 0, \ldots k-2$, tanto para el caso polinomial como para el analítico, lo primero que hay que calcular es $m_c^k(Q_k)$, $m_c^k(P_k)$, $m_c^{k-1}(Q_k)$, $m_c^{k-1}(P_k)$, $m_c^k(Q_{k-1})$, y $m_c^{k-1}(Q_{k-1})$, pero también recordemos que $Q_k(x)$ se reduce a $P_k(x)$ tomando a $\left(\frac{1}{\lambda}\right)^i = i!$, con ésto sólo calcularemos los momentos para el caso analítico, pudiendo obtener de ellos el caso polinomial fácilmente.

Primeramente encontraremos $m_c^{k-1}(Q_{k-1})$. Aplicando a (2.35) el momento de orden $k-1$ tenemos:

$$m_c^{k-1}(Q_{k-1}) = \frac{1}{2}\begin{cases} m_c^{k-1}\left[\exp_k(\frac{1}{\lambda}x)\right] + m_c^{k-1}\left[\exp_k(-\frac{1}{\lambda}x)\right], & k \text{ impar} \\ m_c^{k-1}\left[\exp_k(\frac{1}{\lambda}x)\right] - m_c^{k-1}\left[\exp_k(-\frac{1}{\lambda}x)\right], & k \text{ par} \end{cases}, \tag{2.41}$$

como podemos observar tenemos que calcular

$$m_c^{k-1}\left[\exp_k(\tfrac{1}{\lambda}x)\right] \qquad \text{y} \qquad m_c^{k-1}\left[\exp_k(-\tfrac{1}{\lambda}x)\right],$$

para esto utilizamos las expresiones (2.28) y (2.29) y después de algunas operaciones obtenemos

$$m_c^{k-1}\left[\exp_k(\frac{1}{\lambda}x)\right] = \frac{1}{\lambda^{k-1}(k-1)!}, \tag{2.42}$$

$$m_c^{k-1}\left[\exp_k(-\frac{1}{\lambda}x)\right] = \frac{(-1)^{k-1}}{\lambda^{k-1}(k-1)!}, \tag{2.43}$$

tomando estos valores y sustituyéndolos en (2.41) tenemos que

$$m_c^{k-1}(Q_{k-1}) = \frac{1}{\lambda^{k-1}(k-1)!}, \tag{2.44}$$

haciendo lo mismo para $Q_k(x)$, aplicamos el momento de orden $k-1$ a (2.36) y utilizando (2.42) y (2.43) llegamos a:

$$m_c^{k-1}(Q_k) = 0. \tag{2.45}$$

Por otra parte aplicando el momento de orden k a (2.35) y (2.36) necesitamos los valores de $m_c^k\left[\exp_k(\frac{1}{\lambda}x)\right]$ y $m_c^k\left[\exp_k(-\frac{1}{\lambda}x)\right]$, que podemos calcularlos de una manera similar a como lo hicimos con los momentos de orden $k-1$, obtenemos

$$m_c^k\left[\exp_k(\frac{1}{\lambda}x)\right] = \frac{1}{\lambda^k k!}, \tag{2.46}$$

$$m_c^k\left[\exp_k(-\frac{1}{\lambda}x)\right] = \frac{(-1)^k}{\lambda^k k!}, \tag{2.47}$$

entonces

$$m_c^k(Q_k) = \frac{1}{\lambda^k k!}, \tag{2.48}$$

$$m_c^k(Q_{k-1}) = 0. \tag{2.49}$$

Una vez teniendo (2.44), (2.45), (2.48) y (2.49) podemos calcular los momentos de orden k y $k-1$ de las *funciones base* $u_l(x)$, $u_r(x)$ y $u_c^i(x)$ para $i = 0, \ldots k-2$, así tenemos que

$$m_c^{k-1}(u_l) = \frac{(-1)^{k-1}}{2\lambda^{k-1}(k-1)!}, \tag{2.50}$$

$$m_c^k(u_l) = \frac{(-1)^k}{2\lambda^k k!}, \tag{2.51}$$

$$m_c^{k-1}(u_r) = \frac{1}{2\lambda^{k-1}(k-1)!}, \tag{2.52}$$

$$m_c^{k}(u_r) = \frac{1}{2\lambda^{k}k!}. \tag{2.53}$$

Sólo nos hace falta $m_c^{k-1}(u_c^i)$ y $m_c^k(u_c^i)$ para $i = 0, \ldots, k-2$, de acuerdo con (2.40) necesitamos $m_c^n(Q_i) = m_c^n(P_i)$ para $i = 0, \ldots, k-2$, y $n = k-1, k$, ésto último gracias a (2.34), de aquí que

$$m_c^n(Q_i) = 0 \qquad i = 0, \ldots, k-2, \quad n = k-1, k, \tag{2.54}$$

además tenemos

$$m_c^n(Q_{k-2+m(i)}) = \begin{cases} m_c^n(Q_{k-1}) & \text{si } m(i) = 1 \\ \\ m_c^n(Q_k) & \text{si } m(i) = 2 \end{cases} \quad \text{para } n = k-1, k\,, \tag{2.55}$$

con (2.54) y (2.55) obtenemos

$$m_c^{k-1}(u_c^i) = \frac{-\delta_{1,m(i)}^{k-1}}{\lambda(k-1)!} \qquad i = 0, \ldots, k-2, \tag{2.56}$$

$$m_c^{k}(u_c^i) = \frac{-\delta_{2,m(i)}^{k}}{\lambda k!} \qquad i = 0, \ldots, k-2. \tag{2.57}$$

2.3 METODOS DE MOMENTOS DISCONTINUOS

En este método supondremos que sobre cada celda $\Omega_e = [x_l, x_r]$, u es aproximada por u_h en un espacio de funciones, definida en términos de $k+1$ grados de libertad, los cuales serán $u_h(x_r) = U_r$ así como los momentos de celda de $u_h(x)$ hasta de orden $k-1$, es decir, $U_c^i = m_c^i(u_h(x))$ para $i = 0, \ldots k-1$. Nótese que en contraste con el método de *momentos continuos* $u_h(x_l) = U_l$ no está tomado como un parámetro y entonces, en principio $u_h(x)$ es discontinua en los extremos de las celdas; de una forma más precisa,

$u_h(x_l+0) \neq u_h(x_l-0)$, donde $u_h(x_l-0)$ es U_r para la celda anterior. Como una consecuencia, tenemos disponible un grado más de libertad que en el caso continuo.

De acuerdo con lo anterior, de la misma forma que en el caso de momentos continuos la determinación de las *funciones base* puede hacerse sobre la celda de referencia $\hat{\Omega}$. Donde $u_h(x)$ está definida por un espacio S y un conjunto de grados de libertad D' dados por:

$$\begin{aligned} S &= \mathcal{P}_k \quad o \quad \mathcal{E}_k, \\ D' &= \{U_r, U_c^i, \quad i = 0, \ldots k-1\}, \end{aligned} \tag{2.58}$$

con $\dim S = \text{card } D' = k+1$, donde $\mathcal{P}_k$ es de la forma (2.16) y $\mathcal{E}_k$ de la forma (2.25).

En este caso en lugar de (2.15) tenemos que la función nodal general está dada por

$$u_h(x) = U_r u_r(x) + \sum_{i=0}^{k-1} U_c^i u_c^i(x). \tag{2.59}$$

2.3.1 El caso Polinomial ($\mathcal{P}_k$)

Ahora, sólo deseamos encontrar las expresiones de las *funciones base* $u_r(x)$ y $u_c^i(x) \quad i = 0, \ldots, k-1$, asociadas a la frontera derecha y celda respectivamente. En primer lugar analizaremos para la frontera derecha, donde $u_r(x)$ tiene la forma

$$u_r(x) = \sum_{j=0}^{k} \beta_j P_j(x), \tag{2.60}$$

y debe cumplir con

$$\begin{aligned} m_r(u_r(x)) &= 1, \\ m_c^i(u_r(x)) &= 0 \quad i = 0, \ldots, k-1, \end{aligned} \tag{2.61}$$

y de acuerdo con (2.10) tenemos que

$$\int_{-1}^{1} P_i(x) u_r(x) dx = 0 \quad i = 0, \ldots, k-1,$$

entonces podemos concluir que $u_r(x)$ es ortogonal a $P_k(x)$, por lo tanto

$$u_r(x) = \alpha P_k(x),$$

donde $u_r(x)$ debe cumplir con (2.61), así

$$u_r(1) = 1,$$

por lo tanto

$$\alpha P_k(1) = \alpha = 1,$$

de esta manera

$$u_r(x) = P_k(x). \tag{2.62}$$

Por otro lado, la forma general de $u_c^i(x)$ para $i = 0, \ldots, k-1$ es la misma que (2.23) con

$$\begin{aligned} m_r(u_c^i(x)) &= 0, \\ m_c^i(u_c^i(x)) &= 1 \quad i = 0, \ldots, k-1, \end{aligned}$$

de aquí tenemos que $u_c^i(x)$ es ortogonal a $P_j(x)$ si $i \neq j$, entonces podemos expresarla como

$$u_c^i(x) = \gamma_i P_i(x) + \gamma_k P_k(x),$$

evaluando esta expresión en $x = 1$ y de acuerdo con (2.9) obtenemos una ecuación y resolviéndola llegamos a

$$u_c^i(x) = P_i(x) - P_k(x) \qquad i = 0, \ldots, k-1. \tag{2.63}$$

2.3.2 El caso analítico ($\mathcal{E}_k$)

Con el mismo conjunto D' de grados de libertad como en (2.16) pero con $\mathcal{E}_k$ en lugar de $\mathcal{P}_k$, podemos checar de una manera análoga a los casos anteriores que

$$\begin{aligned} u_r(x) &= Q_k(x), \\ u_c^i(x) &= P_i(x) - Q_k(x) \qquad i = 0, \ldots, k-1, \end{aligned} \tag{2.64}$$

aquí $Q_k(x)$ es proporcional a $\exp_k(\frac{1}{\lambda}x)$, el cual fue definido en (2.27) como $\exp(\frac{1}{\lambda}x)$ menos sus k primeras componentes. La normalización escogida es tal que $Q_k(1) = 1$, es decir, $Q_k(x) = \exp_k(\frac{1}{\lambda}x)/\exp(\frac{1}{\lambda})$.

Como los grados de libertad básicos de $u_h(x)$ son los momentos de celda $u_c^i(x)$, $i = 0, \ldots, k-1$, más $u_r(x)$, podemos calcular un momento más de $u_h(x)$, ya sea en el caso polinomial (2.62) y (2.63) o analítico (2.64). Aunque sólo es necesario hacerlo para el caso analítico ya que $Q_k(x)$ se reduce a $P_k(x)$ si $b_k = 1$ y $a_k = 0$, con ésto tenemos que para encontrar $m_c^k(u_r)$ y $m_c^k(u_c^i)$ $i = 0, \ldots, k-1$, necesitamos los valores de $m_c^k(Q_k)$ y $m_c^k(P_i)$ $i = 0, \ldots, k-1$.

Tenemos que

$$m_c^k(P_i) = 0 \quad i = 0, \ldots, k-1,$$

faltándonos solamente $m_c^k(Q_k)$, para ésto recordemos que $Q_k(x)$ tiene la expresión (2.31), de aquí que tengamos que calcular $m_c^k(P_{k-1})$ y $m_c^k(Q_k)$ que de acuerdo con (2.49) y (2.46) tenemos

$$m_c^k(P_{k-1}) = 0 \qquad \text{y} \qquad m_c^k(\exp_k(\tfrac{1}{\lambda}x) = 1,$$

así que

$$\begin{aligned} m_c^k(u_r) &= a_k, \\ m_c^k(u_c^i) &= -a_k. \end{aligned} \tag{2.65}$$

CAPITULO 3

DESCRIPCION GENERAL

Para nuestro problema (*valores iniciales*) como se mencionó en el capítulo anterior se tienen dos métodos, el de *momentos continuos* y el de *momentos discontinuos* de los cuales analizaremos y presentaremos ejemplos en este capítulo, tomados de [11] y [12], así como un resumen de los resultados de convergencia que se conocen de los dos métodos que analizamos (nodales continuos y discontinuos) que aparecen en [13]. Los teoremas que demuestran estos resultados de convergencia se presentan en el Apéndice A.

Como el problema original (1.4) se puede ver como varias ecuaciones de la forma (1.7), entonces lo que haremos es una descripción general del método de elementos finitos para este problema y se presentan ejemplos para el mismo, (1.7) con condiciones iniciales (1.5).

3.1 METODOS DE MOMENTOS CONTINUOS

En el capítulo anterior hablamos de escoger un espacio polinomial $\mathcal{P}_k$ o analítico $\mathcal{E}_k$ en el cual se encuentran las funciones base que nos definen a la solución aproximada u_h, como u_h es de la forma (2.15) se puede decir que $u_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ dependiendo de donde se encuentren las funciones base.

Entonces, con $u_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ en el método de momentos continuos, es decir estamos pidiendo continuidad de la solución aproximada de celda a celda, tenemos que uno de los $k+1$ parámetros (que definen a u_h, card$D = k+1$)

es fijo y en términos de él se definen los demás, por lo que solo determinaremos k. Para esto lo que haremos es tomar los momentos de Legendre (de orden 0 a $k-1$) del residual

$$R(u_h) = \mu(x)Du_h(x) + q(x)u_h(x) - f(x) \equiv Lu_h - f, \tag{3.1}$$

y hasta entonces procurar que se cumpla que $R(u_h) = 0$ lo que nos asegura que la solución aproximada u_h converge a u, donde L es el operador

$$L(u_h) = \mu(x)Du_h(x) + q(x)u_h(x).$$

Entonces, calcularemos $R(u_h)$ en términos de las U_i y se tiene que cumplir $R(u_h) = 0$ al multiplicarlo por unas funciones de prueba (primer factor del método de elementos finitos), que en este caso serán los p_i (polinomios de Legendre), es decir obtendremos los momentos de Legendre de $R(u_h)$, de esta forma tenemos que

$$\int_{x_l}^{x_r} (Lu_h - f)\, p_i(x)\, dx = 0, \qquad i = 0, \ldots, k-1, \tag{3.2}$$

y sustituyendo (3.1) en (3.2) obtenemos

$$\int_{x_l}^{x_r} \mu(x)Du_h(x)p_i(x)dx + \int_{x_l}^{x_r} q(x)u_h(x)p_i(x)\, dx = \int_{x_l}^{x_r} f(x)p_i(x)dx, \tag{3.3}$$

calculamos la primera integral y regresando a la celda de referencia $\widehat{\Omega}$ encontramos que

$$\frac{2\mu}{hN_i}\left\{\left[U_r - (-1)^i U_l\right] - \int_{-1}^{1} u_h(x)DP_i(x)dx\right\} + qU_c^i = m_c^i(f) \qquad i = 0, \ldots, k-1. \tag{3.4}$$

donde U_l, U_r y U_c^i son los momentos asociados a los extremos izquierdo, derecho y de celda respectivamente, a (3.4) se le llama la ecuación de momentos continuos.

A continuación escribiremos en forma explícita estas ecuaciones para $i = 0, 1, 2$, y posteriormente las utilizaremos para obtener los ejemplos en los casos $k = 1, 2, 3$, obteniendo U_r, U_c^i para $i = 0, 1, 2$ en términos de U_l (los necesarios dependiendo de k) ésto significa que U_l será el parámetro fijo del que habíamos hablado. Una vez que se tienen es fácil sustituirlos en (2.15) para obtener la solución aproximada u_h para los diferentes ejemplos ($K = 1, 2, 3$).

Para $i = 0$, con $P_0 = 1$, $DP_0 = 0$ y $N_0 = 2$, obtenemos:

$$\frac{\mu}{h}(U_r - U_l) + qU_c^0 = m_c^0(f), \tag{3.5}$$

mientras que para $i = 1$, con $DP_1 = P_0$ y $N_1 = \frac{2}{3}$, (3.4) nos queda

$$\frac{3\mu}{h}(U_r + U_l) - \frac{6\mu}{h}U_c^0 + qU_c^1 = m_c^1(f), \tag{3.6}$$

finalmente, para $i = 2$, con $DP_2 = 3P_1$ y $N_2 = \frac{2}{5}$ tenemos

$$\frac{5\mu}{h}(U_r - U_l) - \frac{10\mu}{h}U_c^1 + qU_c^2 = m_c^2(f). \tag{3.7}$$

3.1.1 EJEMPLOS

Para obtener los ejemplos desarrollaremos la u_h de la forma (2.15) hasta los términos necesarios $(k+1)$ y utilizando (3.5), (3.6) y (3.7) según sea el caso, expresaremos U_r y U_c^i para $i = 0, 1, 2$ en términos de U_l, construyendo con ellos u_h. Como u_h se expresa en términos de u_i (funciones base) y U_i (momentos) y las u_i dependen del espacio en el que se tomen, entonces se ejemplificaran para los dos casos, primeramente para el polinomial y posteriormente para el analítico, además presentamos una sección en la que se mencionan los órdenes de convergencia que se conocen para los métodos de momentos continuos.

CASO POLINOMIAL

Recordemos que tenemos dos casos, uno donde el espacio de funciones base son del tipo polinomial ($\mathcal{P}_k$) y otro en el que son del tipo casipolinomial ($\mathcal{E}_k$), en el capítulo anterior se detalló la forma de tomar dichos espacios. Primeramente presentaremos ejemplos en el caso polinomial ($\mathcal{P}_k$), es decir, $u_h \in \mathcal{P}_k$ y consideraremos los casos particulares correspondientes a $k = 1, 2,$ y 3.

En el caso $k = 1$, (2.15) nos queda

$$u_h = u_l(x)U_l + u_r(x)U_r, \tag{3.8}$$

y gracias a (2.50) y (2.52) tenemos que:

$$U_c^0 = \frac{1}{2}(U_l + U_r). \tag{3.9}$$

Sustituyendo (3.9) en (3.5) nos lleva directamente a

$$(\mu + \frac{qh}{2})U_r = (\mu - \frac{qh}{2})U_l + hm_c^0(f). \tag{3.10}$$

Introduciendo $\epsilon = qh/\mu$ podemos obtener U_r en términos de U_l resolviendo

$$U_r = \frac{(2-\epsilon)U_l + \frac{2\epsilon}{q}m_c^0(f)}{2+\epsilon}. \tag{3.11}$$

En el caso $k = 2$, utilizamos de nuevo la expresión (2.15), (2.50) y (2.52), además de la (2.56), con estas tres últimas encontramos una expresión para $m_c^1(u_h)$; que queda:

$$U_c^1 = \frac{1}{2}(U_r - U_l). \tag{3.12}$$

Insertando este valor en (3.6) llegamos a un sistema de dos ecuaciones para U_r conociendo U_l de la forma (con $\epsilon = qh/\mu$)

$$U_r = \frac{(12 - 6\epsilon + \epsilon^2)U_l + \frac{12\epsilon}{q} m_c^0(f) + \frac{2\epsilon^2}{q} m_c^1(f)}{12 + 6\epsilon + \epsilon^2}, \tag{3.13}$$

con

$$U_c^0 = \frac{12U_l + (6\epsilon + \epsilon^2)\frac{m_c^0(f)}{q} - \frac{2\epsilon}{q} m_c^1(f)}{12 + 6\epsilon + \epsilon^2}. \tag{3.14}$$

Por último para $k = 3$, de manera similar encontramos que:

$$\begin{aligned} U_r = \; & (120 + 60\epsilon + 12\epsilon^2 + \epsilon^3)^{-1} [(120 - 60\epsilon + 12\epsilon^2 - \epsilon^3)U_l \\ & + (120\epsilon + 2\epsilon^3)m_c^0(f)/q + 20\epsilon^2 m_c^1(f)/q + 2\epsilon m_c^2(f)/q], \end{aligned}$$

$$\begin{aligned} U_c^0 = \; & (120 + 60\epsilon + 12\epsilon^2 + \epsilon^3)^{-1} [(120 + 2\epsilon^2)U_l \\ & + (60\epsilon + 10\epsilon^2 + \epsilon^3)m_c^0(f)/q - 20\epsilon^2 m_c^1(f)/q - 2\epsilon m_c^2(f)/q], \end{aligned}$$

con

$$U_c^1 = \frac{-60\epsilon U_l + \frac{60\epsilon}{q} m_c^0(f) - \frac{(12\epsilon^2 + \epsilon^3)}{q} m_c^1(f) - \frac{(12\epsilon + 6\epsilon^2)}{q} m_c^2(f)}{120 + 60\epsilon + 12\epsilon^2 + \epsilon^3}.$$

CASO ANALITICO

Aquí, $u_h \in \mathcal{E}_k$, en lugar de pertenecer a $\mathcal{P}_k$ y también consideraremos tres casos, $k = 1, 2, 3$.

Primeramente para $k = 1$, la ecuación (3.5) la seguimos teniendo, pero u_h está dada por la ecuación (3.8). Aquí, u_l y u_r están dadas por las ecuaciones (2.39) y (2.40), de tal forma que:

$$U_c^0 = \frac{1 - a}{2} U_l + \frac{1 + a}{2} U_r, \tag{3.15}$$

donde $a \equiv a_1$ dada por las ecuación (2.32), con (3.15), (3.5) llegamos a

$$U_r = \frac{[2-\epsilon(1-a)]\,U_l + \frac{2\epsilon}{q} m_c^0(f)}{2+\epsilon(1+a)},$$

como

$$\frac{2-\epsilon(1-a)}{2+\epsilon(1+a)} = \exp(-\epsilon),$$

mientras que

$$\frac{2\epsilon}{2+\epsilon(1+a)} = 1-\exp(-\epsilon),$$

de tal forma que

$$U_r = \exp(-\epsilon)U_l + \frac{1-\exp(-\epsilon)}{q} m_c^0(f).$$

En el caso $k = 2$, debemos usar las ecuaciones (3.5) y (3.6) junto con u_h, mientras que

$$U_c^1 = \frac{(1+a)\,U_r - (1-a)\,U_l}{2} - aU_c^0,$$

donde a, ahora es a_2 dada por (2.32). Resolviendo para U_r y U_c^0, finalmente obtenemos

$$\begin{aligned} U_r = \; & \exp(-\epsilon)U_l + [1-\exp(-\epsilon)]\frac{m_c^0(f)}{q} \\ & +2\left[1-(1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\frac{m_c^1(f)}{q}, \end{aligned}$$

$$\begin{aligned} U_c^0 = \; & \left(\frac{1-\exp(-\epsilon)}{\epsilon}\right)U_l \\ & +\left(1-\left[\frac{1-\exp(-\epsilon)}{\epsilon}\right]\right)\frac{m_c^0(f)}{q} \\ & -2\left[1-(1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\frac{m_c^1(f)}{\epsilon q}. \end{aligned}$$

Finalmente en el caso $k = 3$, también debemos considerar la ecuación (3.7), además

$$U_c^2 = \frac{(1-a)U_l + (1+a)U_r}{2} - U_c^0 - aU_c^1,$$

donde $a \equiv a_3$ con a_3 dada por (2.32). Resolviendo para U_r, U_c^0 y U_c^1 tenemos que

$$\begin{aligned} U_r = \; & \exp(-\epsilon)U_l + 1 - \exp(-\epsilon)\frac{m_c^0(f)}{q} \\ & + 2\left[1-\exp(-\epsilon)\right]\left[1 - \frac{2+\epsilon}{2\epsilon}\right]\frac{m_c^1(f)}{q} \\ & + \left\{1 - \exp(-\epsilon) - 12\left[1 - (1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\right\}\frac{m_c^2(f)}{\epsilon q}, \end{aligned}$$

$$\begin{aligned} U_c^0 = \; & \frac{1-\exp(-\epsilon)}{\epsilon}U_l + \left(1 - \frac{1-\exp(-\epsilon)}{\epsilon}\right)\frac{m_c^0(f)}{q} \\ & - 2\left[1 - (1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\frac{m_c^1(f)}{\epsilon q} \\ & - \left\{1 - \exp(-\epsilon) - 12(1-\exp(-\epsilon))\left[1 - \left(\frac{2+\epsilon}{2\epsilon}\right)\right]\right\}\frac{m_c^2(f)}{\epsilon q}, \end{aligned}$$

y

$$\begin{aligned} U_c^1 = \; & -\left[3\left(\frac{1-\exp(-\epsilon)}{\epsilon}\right) - 6\left(\frac{1-\exp(-\epsilon)}{\epsilon^2}\right)\right]U_l \\ & + \left\{6\left[1 - (1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\right\}\frac{m_c^0(f)}{\epsilon q} \\ & + \left\{1 - 6\left(\frac{2+\epsilon}{\epsilon^2}\right)\left[1 - (1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\right\}\frac{m_c^1(f)}{q} \\ & - \left\{3(2+\epsilon)\left[1 - \exp(-\epsilon) - \frac{12}{\epsilon} - 6\left(1 - \exp(-\epsilon)\left(\frac{2+\epsilon}{\epsilon^2}\right)\right)\right]\right\}\frac{m_c^2(f)}{\epsilon^2 q}. \end{aligned}$$

PROPIEDADES DE CONVERGENCIA

Los ejemplos presentados en la sección anterior, cumplen o tienen ciertas propiedades de convergencia, ya sea para la función o para los momentos las cuales podemos resumirlas en las siguientes:

- en los puntos de la malla que discretiza al dominio se tiene un orden de convergencia, llamado discreto, de $O\left(h^{2k}\right)$,donde h es el tamano máximo de los intervalos usados sobre $[a, b]$.
- El orden de convergencia en los momentos locales (en cada celda) es de $O\left(h^{2k-i}\right)$, para $i \leq k$,
- y en cualquier punto del dominio tenemos un orden de convergencia (llamado orden de convergencia continuo) de $O\left(h^{k+1}\right)$.

Aquí, el primer y tercer punto son resultados de *superconvergencia*, observable para $k > 1$.

El enunciado del teorema, así como su demostración, se encuentra en el Apéndice A, o bien en [11].

El resultado del segundo punto, para los momentos es en el sentido de que son del tipo *local*, es decir, están dados para cada subintervalo $[x_l, x_r]$. Para obtener los correspondientes momentos globales, tenemos que pensar en un intervalo fijo, por ejemplo $[x_m, x_n]$, y llamamos H a su longitud (fija), entonces el $l-$ésimo momento de u sobre $[x_m, x_n]$ está dado por

$$m_c^l(u) = \frac{2l+1}{H} \int_{x_m}^{x_n} p_l\left[\frac{2x - (x_m + x_n)}{H}\right] u(x)\, dx. \tag{3.16}$$

Por lo tanto, el momento aproximado correspondiente, se obtiene reemplazando u por u_h en (3.16). Con u_h, la integral de x_m a x_n se puede expresar como una suma sobre cada celda $[x_j, x_{j+1}]$, que nos lleva a

$$m_c^l(u_h) = \frac{2l+1}{H} \sum_{j=m}^{n-1} \frac{h_j}{2} \int_{-1}^{1} p_l\left[\zeta; x_m, x_n; x_j, x_{j+1}\right] u_h(\zeta)\, dx,$$

donde $\zeta = \frac{2x-(x_j+x_{j+1})}{h_j}$ con $h_j = x_{j+1} - x_j$, mientras que

$$p_l\left[\zeta; x_m, x_n; x_j, x_{j+1}\right] = p_l\left\{\left[\zeta h_j + (x_j + x_{j+1}) - (x_m + x_n)\right]/H\right\}. \tag{3.17}$$

Los polinomios resultantes sobre cada $[x_j, x_{j+1}]$ son una combinación lineal de los polinomios normalizados de Legendre de grado 0 a grado l sobre ese intervalo y todos los momentos locales de orden 0 a l aparecen sobre cada subintervalo. Entonces, de (3.17), en la expresión de los momentos globales, el i–ésimo momento local será multiplicado por un factor $(h_j)^i/l!$ de orden $O(h^i)$, donde $h \equiv \max_j h_j$ (o $h = h_j$ en el caso de una partición uniforme) de $[x_m, x_n]$. En otras palabras, el decrecimiento en el exponente de los momentos locales en un orden de i, expresado en el Teorema 3 (véase Apéndice A) (A1), es compensado por el factor h^i para el momento global de orden i. Esto se puede resumir en:

- para los momentos globales se tiene que

$$\left|m_c^i(u) - m_c^i(u_h)\right| = O\left(h^{2k}\right), \qquad \text{para } i \in N \tag{3.18}$$

Cabe aclarar que los resultados de estos dos teoremas fueron tomados de [11], y que nos servirán para obtener nuevos resultados.

3.2 METODOS DE MOMENTOS DISCONTINUOS

En este caso, recordemos que u_h no es considerada continua de celda a celda, es decir que $u_h(x_r - 0) = u_r$ en el extremo derecho de cada celda no necesariamente es igual a $u_h(x_l + 0)$ de la siguiente celda. Así que con $u_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ en el método de momentos discontinuos, en cada celda hay $k+1$ parámetros libres (uno más que en el caso continuo, que determinan a u_h), que determinaremos considerando los momentos de Legendre (de orden 0 a k) del residual (3.1) y expresando que son cero (como en el caso continuo, con el fin de asegurar convergencia de u_h a u) tenemos:

$$\begin{aligned} &\lim_{\epsilon \to 0} \int_{x_l - \epsilon}^{x_r - \epsilon} (Lu_h - f) p_i(x) dx = \int_{x_l}^{x_r} (Lu_h - f) p_i(x) dx \\ &+ p(x) p_i(x) \left[u_h(x_l + 0) - u_h(x_l - 0)\right] = 0 \qquad i = 0, \ldots, k, \end{aligned} \tag{3.19}$$

donde u_h en la frontera izquierda ($x = x_l$) tiene un salto finito. Regresando a la celda de referencia $\hat{\Omega}$ y después de integrar por partes el término $Lu_h p_i(x)$ las ecuaciones nos quedan

$$\frac{2\mu}{hN_i}\{(-1)^i[u_h(x_l+0)-U_l]+[U_r-(-1)^i u_h(x_l+0)] \\ -\int_{-1}^{1} u_h(x)DP_i(x)dx\Big\}+qU_c^i = m_c^i(f),$$

para $i = 0,\ldots,k$, o equivalentemente

$$\frac{2\mu}{hN_i}\left\{\left[U_r-(-1)^i U_l\right]-\int_{-1}^{1} u_h(x)DP_i(x)dx\right\}+qU_c^i = m_c^i(f) \qquad i=0,\ldots,k, \tag{3.20}$$

en la cual $q(x)$ y $\mu(x)$ han sido consideradas constantes sobre $\hat{\Omega}$ y por simplicidad hemos tomado a $u_h(x_l-0)$ como U_r. Y tal y como se encuentra (3.20), estas ecuaciones son estrictamente equivalentes a (3.4) en el caso continuo, la única diferencia está en la definición de u_h en cada intervalo.

Las formas explícitas de (3.20) son exactamente las mismas que para el caso continuo, es decir tenemos de nuevo las expresiones (3.5), (3.6) y (3.7), pero con dos diferencias significativas: primero, en este caso los índices son para $i = 0,1,2$, mientras que en el caso continuo eran para $i = 1,2,3$. Y la otra diferencia está en que ahora U_l no es un parámetro que aparece en la representación explícita de u_h en la celda considerada, como en el caso continuo.

En sección siguiente presentaremos los ejemplos para $k = 0,1,2$ en los casos en los que $u_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ así como un resumen de los resultados de convergencia para el caso discontinuo.

3.2.1 EJEMPLOS

CASO POLINOMIAL

Aquí, con u_h en P_k, consideraremos los casos particulares correspondientes a $k = 0,1$ y 2.

En el caso $k = 0$, (2.64) queda

$$u_h = U_r u_r(x), \tag{3.21}$$

con

$$U_c^0 = m_c^0(u_h) = U_r,$$

de tal forma que de (3.5) obtenemos

$$U_r = \frac{U_l + \frac{\epsilon}{q} m_c^0(f)}{1+\epsilon},$$

donde $\epsilon = qh/\mu$.

En el caso $k = 1$, (2.64) nos queda

$$u_h = U_r u_r(x) + U_c^0 u_c^0(x), \tag{3.22}$$

con

$$U_c^1 = m_c^1(u_h) = U_r - U_c^0,$$

de tal forma que seguimos teniendo (3.5) y (3.6) queda de la forma

$$\frac{3\mu}{h}(U_r + U_l) - \frac{6\mu}{h} U_c^0 + q\left(U_r - U_c^0\right) = m_c^1(f).$$

Resolviendo las dos ecuaciones anteriores para U_r y U_c^0 obtenemos

$$U_r = \frac{(6+2\epsilon)U_l + (6\epsilon+\epsilon^2)\frac{m_c^0(f)}{q} + \epsilon^2 \frac{m_c^1(f)}{q}}{6+4\epsilon+\epsilon^2},$$

y

$$U_c^0 = \frac{(6+\epsilon)U_l + (3\epsilon+\epsilon^2)\frac{m_c^0(f)}{q} - \epsilon \frac{m_c^1(f)}{q}}{6+4\epsilon+\epsilon^2}.$$

En el caso $k = 2$, (2.64) toma la forma

$$u_h = U_r u_r(x) + U_c^0 u_c^0(x) + U_c^1 u_c^1(x), \tag{3.23}$$

con

$$U_c^2 = m_c^2(u_h) = U_r - U_c^0 - U_c^1,$$

y se sigue cumpliendo la ecuación (3.5), y las ecuaciones (3.6) y (3.7) quedan de la forma

$$\frac{3\mu}{h}(U_r + U_l) - \frac{6\mu}{h}U_c^0 + qU_c^1 = m_c^1(f),$$

y

$$\frac{5\mu}{h}(U_r - U_l) - \frac{10\mu}{h}U_c^1 + q\left(U_r - U_c^0 - U_c^1\right) = m_c^2(f),$$

de tal forma que resolviéndolas para U_r ,U_c^0 y U_c^1 llegamos a

$$U_r = (60 + 36\epsilon + 9\epsilon^2 + \epsilon^3)^{-1}\left((60 - 24\epsilon + 3\epsilon^2)U_l + (60\epsilon + 6\epsilon^2 + \epsilon^3)\frac{m_c^0(f)}{q} + (10\epsilon^2 + \epsilon^3)\frac{m_c^1(f)}{q} + \epsilon^3\frac{m_c^2(f)}{q}\right),$$

y

$$U_c^0 = (60 + 36\epsilon + 9\epsilon^2 + \epsilon^3)^{-1}\left((60 + 6\epsilon + \epsilon^2)U_l + (30\epsilon + 8\epsilon^2 + \epsilon^3)\frac{m_c^0(f)}{q} - (10\epsilon + \epsilon^2)\frac{m_c^1(f)}{q} - \epsilon^2\frac{m_c^2(f)}{q}\right),$$

$$U_c^1 = (60 + 36\epsilon + 9\epsilon^2 + \epsilon^3)^{-1}\left(-(30\epsilon + 3\epsilon^2)U_l + (30\epsilon + 3\epsilon^2)\frac{m_c^0(f)}{q} + (6\epsilon^2 + \epsilon^3)\frac{m_c^1(f)}{q} - (6\epsilon + 3\epsilon^2)\frac{m_c^2(f)}{q}\right).$$

CASO ANALITICO

En este caso $u_h \in \mathcal{E}_k$, y consideraremos los siguientes ejemplos.

En el caso $k = 0$, la ecuación (3.5) la seguimos teniendo con u_h dada por (3.21). Aquí, por lo tanto u_r está dada por

$$U_r = \frac{\exp(x/\lambda)}{\exp(1/\lambda)},$$

tal que

$$U_c^0 = \frac{\lambda U_r \operatorname{senh}(1/\lambda)}{\exp(1/\lambda)},$$

sustituyendo esta expresión en (3.5) y resolviendo para U_r llegamos a

$$U_r = \frac{U_l + \frac{\epsilon}{q} m_c^0(f)}{1 - \epsilon\eta},$$

con $\eta = \dfrac{\lambda \operatorname{senh}(1/\lambda)}{\exp(1/\lambda)}$. Con $\lambda = -\epsilon/2$, $1/(1+\epsilon\eta) = \exp(-\epsilon)$, la ecuación anterior nos queda:

$$U_r = \exp(-\epsilon) U_l + \epsilon \exp(-\epsilon) \frac{m_c^0(f)}{q}.$$

En el caso $k = 1$, trabajando de manera similar y después de manipulaciones algebraicas obtenemos:

$$\begin{aligned} U_r = \; & \exp(-\epsilon) U_l + [1 - \exp(-\epsilon)] \frac{m_c^0(f)}{q} \\ & + [1 - \exp[-\epsilon] - \epsilon \exp(-\epsilon)] \frac{m_c^1(f)}{3q}, \end{aligned}$$

y

$$\begin{aligned} U_c^0 = \; & \frac{1 - \exp(-\epsilon)}{\epsilon} U_l + \left(1 - \frac{1 - \exp(-\epsilon)}{\epsilon}\right) \frac{m_c^0(f)}{q} \\ & - [1 - \exp(-\epsilon) - \epsilon \exp(-\epsilon)] \frac{m_c^1(f)}{3\epsilon q}. \end{aligned}$$

Finalmente para $k = 2$ tenemos que:

$$\begin{aligned} U_r = \; & \exp(-\epsilon) U_l + [1 - \exp(-\epsilon)] \frac{m_c^0(f)}{q} \\ & + 2 \left[1 - (1 - \exp(-\epsilon)) \left(\frac{2+\epsilon}{2\epsilon}\right)\right] \frac{m_c^1(f)}{q} \\ & + 2 \left(3 + 2\epsilon + \frac{\epsilon^2}{2}\right) (\exp(-\epsilon) + \epsilon - 3) \frac{m_c^2(f)}{5\epsilon q}. \end{aligned}$$

con

$$U_c^0 = \frac{1-\exp(-\epsilon)}{\epsilon}U_l + \left(1 - \frac{1-\exp(-\epsilon)}{\epsilon}\right)\frac{m_c^0(f)}{q}$$
$$-2\left[1-(1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\frac{m_c^1(f)}{\epsilon q}$$
$$+2\left(3+2\epsilon+\frac{\epsilon^2}{2}\right)(\exp(-\epsilon)+\epsilon-3)\frac{m_c^2(f)}{5\epsilon^2 q},$$

$$U_c^1 = \frac{6(1-\exp(-\epsilon))-3\epsilon(1-\exp(-\epsilon))}{\epsilon^2}U_l$$
$$+6\left[1-(1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\frac{m_c^0(f)}{\epsilon q}$$
$$+\left(1-\left[1-(1-\exp(-\epsilon))\left(\frac{2+\epsilon}{2\epsilon}\right)\right]\right)\left(\frac{12+6\epsilon}{\epsilon^2}\right)\frac{m_c^1(f)}{q}$$
$$-6(2+\epsilon)\left(\left[3+2\epsilon+\frac{\epsilon^2}{2}\right]\exp(-\epsilon)+\epsilon-3\right)\frac{m_c^1(f)}{5\epsilon^3 q}.$$

PROPIEDADES DE CONVERGENCIA

Al igual que para el método de momentos continuos, en el método de momentos discontinuos también podemos agrupar la propiedades de convergencia en dos teoremas, el primero se encuentra en su totalidad en el Apéndice A que en resumen dice:

- en los puntos discretos del dominio se tiene un orden de convergencia de $O\left(h^{2k+1}\right)$, donde h es el tamaño máximo de los intervalos usados sobre$[a,b]$.
- Para los momentos locales se tiene que el orden es de $O\left(h^{2k+1-i}\right)$, para $i \leq k$,
- finalmente, el orden de convergencia continuo (en cualquier punto del dominio diferente a los de discretización) es $O\left(h^{k+1}\right)$.

en este caso el primer y tercer resultado es llamado de *superconvergencia*, observable para $k > 0$.

A igual que el caso continuo la demostración se encuentra en el Apéndice A, y también en [12] junto con el orden de convergencia para los momentos globales, en el cual se señala que:

- $$\left|m_c^i(u) - m_c^i(u_h)\right| = O\left(h^{2k+1}\right), \qquad \text{para } i \in N \tag{3.24}$$

CAPITULO 4

RESULTADOS DE CONVERGENCIA

Con el capítulo 2 (Funciones Base) y Capítulo 3 se completan los dos elementos básicos de un método de elementos finitos, como las U_i son momentos entonces estamos hablando de elementos finitos nodales. En este capítulo se utilizaran las expresiones de cada uno de los esquemas, presentados como ejemplos en el capítulo anterior para encontrar así, la solución aproximada de un problema modelo y una vez que se tiene, en una primera parte (sección 4.1) se confirmará desde un punto de vista computacional los resultados teóricos referentes a los órdenes de convergencia en los puntos de la malla, en cualquier otro punto del dominio y momentos globales. Cabe hacer la aclaración que todos estos resultados son de [13]. Además se presentará una nueva técnica, que llamamos *postprocesamiento*, que nos permitirá elevar el orden de convergencia en cualquier punto del dominio del problema de algunos de los esquemas que se presentaron en el Capítulo 3. Los programas que se utilizaron se anexan en disco flexible.

4.1 DESCRIPCION DEL PROBLEMA MODELO

El problema que usaremos como modelo es referente a transporte de partículas, para esto recordemos que las ecuaciones ordinarias para transporte de partículas en geometría plana consta de un sistema de ecuaciones diferenciales

ordinarias de la forma

$$\mu_i \frac{du_i}{dx} + q u_i = f_i, \quad i = 1, \ldots, N, \quad \forall \quad x \in (a, b), \tag{4.1}$$

sujeta a las condiciones iniciales

$$\begin{aligned} u_i(a) &= 0, \quad \text{para} \quad \mu_i > 0, \\ u_i(b) &= 0, \quad \text{para} \quad \mu_i < 0, \end{aligned}$$

donde a y b son los extremos del intervalo que discretiza a μ, es decir son los extremos del intervalo para una dirección μ_i. La u (flujo angular) correspondientes a μ positiva (negativa respectivamente), se aproxima por una u_h que se determina de izquierda a derecha (derecha-izqnierda resp.) sobre cada celda $[x_l, x_r]$. La fuente f_i es de la forma

$$f_i(x) = \sum_{j=1}^{N} \omega_j u_j(x) + q_i(x),$$

donde los ω_j son los pesos de cuadratura asociados con los puntos de cuadratura $\mu_i \in (-1, 1) / \{0\}$, los cuales son los de Gauss-Legendre y u_j es la u restringida a la a μ_i los direcciones angulares en las que se discretizó μ, como en el primer capítulo.

Para aproximar la solución de (4.1) se pueden usar técnicas de iteración estandar. De éstas se obtiene

$$\mu_i D u_i^{n+1} + q u_i^{n+1} = f_i^n, \quad n = 0, 1, 2, \ldots,$$

la fuente puede estimarse iterativamente de los flujos angulares anteriores.

Para obtener los órdenes de convergencia de cada uno de los esquemas numéricos descritos en el capítulo anterior, consideraremos un problema modelo de la forma (4.1) tomado de [14] que consiste en obtener la solución analítica de la aproximación S_2 de la ecuación, esto significa, de acuerdo con el Capítulo 1, que discretizamos el dominio en dos direcciones angulares; además se considera una fuente externa y condiciones de frontera nulas

en ambos lados del dominio $[a,b]$, el cual está definido como $[-1,1]$. Esta aproximación S_2 es de la forma:

$$\mu_n \frac{du_n}{dx} + q_t u_n = q_s \sum_{m=1}^{2} w_m u_m + Q_n; \qquad n = 1,2, \tag{4.2}$$

donde $\mu_1 = -\mu_2 = 1/\sqrt{3}$, estos son los puntos de cuadratura de Gauss-Legendre para dos puntos. $Q_1 = Q_2 = 1$, significa que tenemos una fuente que nos manda un neutrón por unidad de volumen. Además, $w_1 = w_2 = 1$, los pesos de cuadratura de Gauss-Legendre para dos puntos.

La sección transversal total q_t, (L^{-1}) tiene un valor de 2 (en promedio, cada unidad de longitud 2 neutrones interactúan por absorción y/o dispersión) y la sección transversal de dispersión, q_s, de 1 (en promedio, cada unidad de longitud 1 neutrón es absorbido).

Todo lo anterior significa que tenemos un fenómeno físico de transporte de partículas que se puede modelar con una ecuación de la forma (1.2), en el dominio $[-1,1]$, discretizamos respecto a μ con dos puntos, de Gauss-Legendre, obteniendo una ecuación de la forma (1.4) en cada subintervalo de esta discretización (respecto a μ), que se pueden agrupar en (4.2), de esta forma como $\mu_1 > 0$ se tiene que $u(a) = 0$, y $\mu_2 < 0$, $u(b) = 0$. Entonces resolver (4.2) significa encontrar dos soluciones, una para cada valor de μ, las cuales se obtienen (numéricamente) barriendo el dominio de izquierda a derecha para μ_1 y de derecha a izquierda para μ_2.

En (4.2), $u_1(x)$ y $u_2(x)$ son los flujos angulares (analíticos) de neutrones en las direcciones μ_1 y μ_2 respectivamente, dados por

$$u_1(x) = 1 - \frac{e^{\sqrt{6}x} + \left(3+2\sqrt{2}\right) e^{-\sqrt{6}x}}{e^{-\sqrt{6}} + \left(3+2\sqrt{2}\right) e^{\sqrt{6}}}, \tag{4.3}$$

y

$$u_2(x) = 1 - \frac{e^{-\sqrt{6}x} + \left(3+2\sqrt{2}\right) e^{\sqrt{6}x}}{e^{-\sqrt{6}} + \left(3+2\sqrt{2}\right) e^{\sqrt{6}}}. \tag{4.4}$$

4.1.1 PRUEBAS DE CONVERGENCIA PARA EL PROBLEMA MODELO

Para obtener los órdenes de convergencia discretos y continuos para los diferentes esquemas numéricos, se hicieron varios experimentos con el problema modelo previamente descrito.

Primeramente, para los esquemas *continuos*, polinomiales (CMP) y analíticos (CMA). Se denotará con DCO (Discrete Convergence Order) y CCO (Continuous Convergence Order) a los *órdenes de convergencia discretos y continuos* correspondiente a u_h, la aproximación del flujo angular u para una malla de tamaño h.

Para obtener DCO, primeramente se calcularon los *errores discretos* de la forma

$$e_h = \max_i |u(x_i) - u_h(x_i)|. \tag{4.5}$$

para diferentes particiones del dominio. Las particiones se hicieron dividiendo el dominio del problema modelo $[-1,1]$, en $2,4,8,16,32$ y 64 subintervalos de igual longitud. Después, se calculó $u(x_i)$, el valor exacto de la solución en los puntos de la malla, así como la solución aproximada $u_h(x_i)$, también en los puntos de la malla. Esta última es de la forma (2.15), de tal manera que encontrando el máximo de su diferencia en cada punto de la malla encontramos los errores que necesitamos en cada una de las particiones y sustituyéndolos en

$$\mathrm{DCO} = \frac{\ln\left(e_h/e_{h/2}\right)}{\ln 2},$$

se obtiene una estimación del *orden de convergencia discreto* (DCO).

En el caso de CCO, al igual que para DCO, tenemos que calcular *errores* de la misma forma que en (4.5) con la diferencia de que en este caso se trata de un orden de convergencia continuo, es decir, tenemos que encontrar errores en cualquier punto del dominio y no solo en los extremos de cada subinetervalo en los que fue dividido. Por tal razón, los errores se calcularon evaluando la solución exacta u (que dependiendo de en que parte del dominio se encuentre

el en el que se evalúa, esta dada por (4.3) o (4.4)) y la aproximada u_h en 129 puntos uniformemente distribuidos sobre el dominio $[-1,1]$. La u_h en estos puntos se obtuvo utilizando la solución aproximada que se construye con (2.15) (combinación lineal de momentos y funciones base), que es un polinomio de grado k ($k = 2,3$) en cada subintervalo, tomando en cuenta los valores de u_h en los extremos de cada subintervalo de la malla, así como los $k-1$ momentos locales (en el sentido en que son calculados sobre cada subintervalo que conforma la partición del dominio) necesarios.

Para representar los órdenes de convergencia discretos correspondientes a los *momentos globales*: cero, primero, segundo y tercero, de u_h, se utilizaron los números M0, M1, M2 y M3, calculados de la forma:

$$\mathrm{M}l = \frac{\ln\left(m_c^l(e_h)/m_c^l\left(e_{h/2}\right)\right)}{\ln 2}, \qquad l = 0,1,2,3,$$

donde

$$e_h = u(x) - u_h(x).$$

son los errores, y

$$m_c^l(e_h) = \frac{2l+1}{I} \sum_{j=\frac{I}{2}+1}^{I} \int_{-1}^{1} p_l\left[2\zeta/I + x_j + x_{j+1} - 1\right] e_h(\zeta)\, d\zeta,$$

la cual se obtiene de (3.16); donde I es el número de subintervalos en los cuales se dividió el dominio. De la misma forma que antes, se divide el dominio en $2, 4, 8, 16, 32$ y 64 subintervalos. Además, lo de *global* es en el sentido de que son calculados sobre un conjunto de celdas vecinas y no sobre cada celda como en los casos anteriores, en particular, los resultados que se presentan están hechos de tal forma que el conjunto de celdas (o subintervalos) vecinas que se tomó para calcular los momentos globales formara el intervalo $(0,1)$, independientemente en cuantas celdas o subintervalos haya sido dividido el dominio. Los resultados que se presentan en este capítulo son para los casos $k = 2,3$; ya que estos son los que presentan resultados de *superconvergencia* y son los que utilizaremos para cálculos posteriores y se resumen en las Tablas 4.1,4.2,4.3 y 4.4.

No. de Interv.	DCO	CCO	M0	M1	M2	M3
2-4	3.749	2.171	4.312	4.528	2.905	2.993
4-8	4.093	2.608	4.094	4.105	3.837	3.322
8-16	4.013	2.823	4.024	4.025	3.963	3.862
16-32	3.997	2.919	4.006	4.006	3.991	3.967
32-64	4.002	8.300	4.002	4.001	3.998	3.992
Teórico	4.0	3.0	4.0	4.0	4.0	4.0

Tabla 4.1: Ordenes de convergencia para el esquema CMP2

No. de Interv.	DCO	CCO	M0	M1	M2	M3
2-4	5.646	3.188	6.244	6.236	6.240	3.899
4-8	6.062	3.571	6.063	6.062	6.062	5.751
8-16	6.005	3.776	6.016	6.016	6.016	5.944
16-32	5.995	3.889	6.003	6.004	6.004	5.987
32-64	5.996	3.945	5.947	5.989	6.000	5.999
Teórico	6.0	4.0	6.0	6.0	6.0	6.0

Tabla 4.2: Ordenes de convergencia para el esquema CMP3

No. de Interv.	DCO	CCO	M0	M1	M2	M3
2-4	3.400	2.104	2.551	3.335	3.038	3.512
4-8	3.798	2.592	3.670	3.791	3.198	2.569
8-16	3.945	2.820	3.916	3.944	3.825	3.720
16-32	3.986	2.926	3.979	3.986	3.957	3.934
32-64	3.997	7.620	3.995	3.997	3.990	3.983
Teórico	4.0	3.0	4.0	4.0	4.0	4.0

Tabla 4.3: Ordenes de convergencia para el esquema CMA2

No. de Interv.	DCO	CCO	M0	M1	M2	M3
2-4	5.419	3.125	5.528	5.191	5.082	2.763
4-8	5.813	3.534	5.839	5.767	5.746	5.408
8-16	5.937	3.764	5.956	5.939	5.935	5.862
16-32	5.984	3.885	5.989	5.985	5.983	5.966
32-64	5.986	3.943	6.020	5.994	5.996	5.992
Teórico	6.0	4.0	6.0	6.0	6.0	6.0

Tabla 4.4: Ordenes de convergencia para el esquema CMA3

Las estimaciones numéricas para CCO en el esquema CMP3 es una muy buena aproximación con el resultado teórico predecido, pero en esquema CMP2 hay un punto importante; si observamos las Tablas 4.1 y 4.3, en el renglón 16-32, la estimación numérica para CCO es muy cercana al valor teórico pero en el renglón 32-64, CCO incrementa en más de dos veces al resultado teórico predecido. En particular, este hecho se debe a que el punto medio también exhibe superconvergencia, ya que evaluamos el error en una malla con 128 subintervalos de igual longitud, y los puntos en los que se hace la evaluación son puntos de la malla o puntos medios de los cálculos hechos para 64 subintervalos.

Para los esquemas *discontinuos*: polinomial (DMP) y analítico (DMA) se calcularon dos órdenes de convergencia discretos (DCO), ya que como el mismo nombre lo dice, las aproximaciones no necesariamente son continuas en los extremos de cada celda, así que tenemos DCO(r) y DCO(l) asociados a $u_h(x_i - 0)$ y a $u_h(x_i + 0)$ respectivamente, es decir, corresponden a los valores derecho e izquierdo de u_h en cada celda.

La forma de calcular DCO(r) y DCO(l) es similar a la de DCO, pero con

$$e_h = \max_i |u(x_{i+1}) - u_h(x_{i+1} - 0)|, \tag{4.6}$$

para DCO(r) y

$$e_h' = \max_i |u(x_i) - u_h(x_i + 0)|, \tag{4.7}$$

No. de Interv.	DCO(r)	DCO(l)	CCO	M0	M1	M2	M3
2-4	2.432	1.183	1.183	5.634	2.281	5.049	-0.242
4-8	2.882	1.503	1.503	2.917	2.657	2.180	1.315
8-16	2.898	1.724	1.724	2.117	2.837	2.132	2.563
16-32	2.934	1.855	1.855	2.741	2.920	2.718	2.842
32-64	2.969	1.926	1.926	2.899	2.961	2.883	2.935
Teórico	3.0	2.0	2.0	3.0	3.0	3.0	3.0

Tabla 4.5: Ordenes de convergencia para el esquema DMP1

para DCO(l). Entonces

$$\mathrm{DCO}(r) = \frac{\ln\left(e_h/e_{h/2}\right)}{\ln 2},$$

y

$$\mathrm{DCO}(l) = \frac{\ln\left(e'_h/e'_{h/2}\right)}{\ln 2}.$$

Como podemos observar, la única diferencia con respecto a los métodos de momentos continuos es que en lugar de calcular un error en cada punto de la malla, calculamos dos errores en los mismos puntos, ya que tanto u como u_h no necesariamente son continuas, así que tenemos que evaluar u y u_h en los puntos x_{i+1} (x_i) y $x_{i+1} - 0$ ($x_i + 0$) para poder encontrar e_h y calcular DCO(r) (DCO(l)).

En el caso del orden de convergencia continuo CCO, se hace exactamente igual, y solamente se toma en cuenta la discontinuidad al momento de construir los la solución aproximada, que es un polinomio de grado k de la forma (2.59), es decir, hay un brinco de polinomio a polinomio en cada punto de la malla. Por último los errores en los momentos globales (Ml, $l = 0, 1, 2, 3$) se calculan de la misma forma que en los esquemas continuos, sin dejar de considerar la discontinuidad de u.

Con esto tenemos que se obtuvieron las tablas 4.5, 4.6, 4.7 y 4.8. (presentaremos los resultados para los casos $k = 1, 2$)

No. de Interv.	DCO(r)	DCO(l)	CCO	M0	M1	M2	M3
2-4	4.363	2.120	2.120	4.641	3.888	3.092	2.348
4-8	4.918	2.512	2.512	4.789	4.616	4.518	4.219
8-16	4.923	2.743	2.743	4.900	4.854	4.831	4.764
16-32	4.958	2.868	2.868	4.953	4.940	4.933	4.914
32-64	4.980	2.933	2.933	4.977	4.973	4.971	4.965
Teórico	5.0	3.0	3.0	5.0	5.0	5.0	5.0

Tabla 4.6: Ordenes de convergencia para el esquema DMP2

No. de Interv.	DCO(r)	DCO(l)	CCO	M0	M1	M2	M3
2-4	2.409	1.242	1.242	2.501	2.419	3.448	2.354
4-8	2.876	1.654	1.654	2.875	2.876	1.989	-0.058
8-16	2.997	1.829	1.829	2.980	2.996	2.786	2.547
16-32	3.015	1.915	1.915	3.002	3.013	2.938	2.867
32-64	3.011	1.957	1.957	3.004	3.011	2.979	2.952
Teórico	3.0	2.0	2.0	3.0	3.0	3.0	3.0

Tabla 4.7: Ordenes de convergencia para el esquema DMA1

No. de Interv.	DCO(r)	DCO(l)	CCO	M0	M1	M2	M3
2-4	4.418	2.385	2.385	4.128	4.402	4.431	4.769
4-8	4.847	2.654	2.654	4.817	4.845	4.849	4.161
8-16	4.972	2.821	2.821	4.983	4.972	4.972	4.812
16-32	4.999	2.909	2.909	5.012	5.000	4.998	4.946
32-64	5.003	2.953	2.953	5.011	5.003	5.003	4.982
Teórico	5.0	3.0	3.0	5.0	5.0	5.0	5.0

Tabla 4.8: Ordenes de convergencia para el esquema DMA2

4.2 POSTPROCESAMIENTO

Después de haber expuesto la teoría en el Capítulo 3 y los resultados numéricos existentes en la sección anterior, nos damos cuenta de que el *orden de convergencia continuo* (CCO), en cualquiera de los casos, es menor que el orden de convergencia discreto y que el orden de convergencia en los momentos; por esta razón trataremos de mejorarlo, es decir, aumentarlo.

En el capítulo anterior mencionamos que en los Teoremas 3 y 5 (que aparecen en el Apéndice A) se obtienen resultados de *superconvergencia*, entendiendo por *superconvergencia* el hecho de obtener ordenes de convergencia más altos de lo que razonablemente se espera, es decir, debido a que se tiene una interpolación se espera una convergencia con una unidad más que el grado de dicha interpolación, por lo que si obtenemos un orden de convergencia que lo sobrepase, entonces se dice que es *superconvergente*. En los esquemas *continuos* y *discontinuos* (polinomial y analítico) la solución aproximada es un polinomio de k, de la forma (2.15) y (2.59) respectivamente, que de acuerdo con los resultados básicos de la teoría de interpolación se espera un orden de convergencia $k+1$. Como podemos observar, en los esquemas continuos, para el problema modelo y los casos que se consideraron ($k = 2, 3$), del Teorema 3 (Apéndice A) y las tablas 4.1, 4.2, 4.3 y 4.4, el orden de convergencia discreto superó lo que se esperaba y en el caso del *orden de convergencia continuo* coincidió con lo esperado, es decir:

	DCO	CCO
k	(Teórico y Numérico)	(Teórico y Numérico)
2	4	3
3	6	4

En los métodos de momentos discontinuos, se observa el mismo fenómeno, pero solamente en DCO(r) obtenemos superconvergencia y no así para el orden de convergencia discreto en el extremo izquierdo de cada celda, DCO(l). En resumen

	DCO(r)	CCO, DCO(l)
k	(Teórico y Numérico)	(Teórico y Numérico)
1	3	2
2	5	3

Además de la superconvergencia en DCO se tiene en los momentos M0, M1, M2 y M3, y se presentan en las Tablas

- 4.1 y 4.2 para los esquemas polinomiales continuos (CMP2 y CMP3 respectivamente),
- 4.3 y 4.4 para los esquemas analíticos continuos (CMA2 y CMA3 respectivamente);
- 4.5 y 4.6 para los casos polinomiales discontinuos (DMP1 y DMP2 respectivamente),
- 4.7 y 4.8 para los casos analíticos discontinuos (DMA1 y DMA2 respectivamente).

Del resumen hecho con anterioridad, podemos concluir que en ambos métodos (nodal continuo y discontinuo), no obtenemos superconvergencia en el *orden de convergencia continuo* (CCO), es decir, sólo llegamos al orden de convergencia natural de la interpolación, y debido a ésto, nuestro objetivo será tratar de igualar el CCO con DCO.

Aunque el proceso para obtener un CCO más alto, en los esquemas continuos y discontinuos es muy similar, lo trataremos por separado para evitar confusiones.

4.2.1 ESQUEMAS CONTINUOS

Para obtener un CCO (*orden de convergencia continuo*), lo haremos aprovechando las *superconvergencias* (A.1) y (3.18) (de DCO y Ml, con $l = 0, \ldots 3$), que se refieren a que tenemos superconvergencia en los puntos de la malla así como en los momentos. La pregunta que nos hacemos entonces es, ¿De qué manera podemos obtener un CCO más alto? La respuesta es la siguiente: utilizaremos los valores de u_h en los puntos de la malla y los momentos globales necesarios, con ellos *postprocesaremos* la u_h obtenida en una $\tilde{u}_h$ que exhiba un orden de convergencia más alto, de esta forma obtendremos órdenes de convergencia continuos (postprocesados, CCOP) de la forma: $O(h^{2k})$,

para $k = 2,3$ (se excluye el caso $k = 1$ ya que solo es observable para $k > 1$). Otra cosa que hay qué aclarar es en qué consiste o que significa *postprocesar*. En secciones anteriores se mencionó la forma de obtener OOC (*orden de convergencia continuo*) y mencionamos que se evaluó la solución aproximada en 129 puntos uniformemente distribuidos en el dominio, esta solución aproximada se construyo fue de la forma

$$u_h = U_l u_l(x) + U_r u_r(x) + \sum_{i=0}^{k-2} U_c^i u_c^i(x), \tag{4.8}$$

donde $u_l(x)$, $u_r(x)$ y $u_c^i(x)$ son las funciones base dadas por (2.21), (2.22) y (2.24) en el caso polinomial y (2.38), (2.39) y (2.40) en el caso analítico respectivamente y U_l, U_r, U_c^i, $i = 0, \ldots k-2$ son los momentos asociados al extremos derecho, izquierdo y celda respectivamente, recordemos que es una u_h de esta forma en cada celda. Entonces, el *postprocesamiento* consiste en construir un nuevo polinomio de interpolación de grado más alto que (4.8) y de manera global, es decir, no debe construirse tomando en cuenta los parámetros (valor a la derecha, izquierda y momentos de u) de cada celda, sino que deben ser respecto a más de una celda (2,3,... celdas vecinas). Este nuevo polinomio de interpolación puede tomarse de la forma

$$\tilde{u}_h(x) = U_l u_l(x) + U_r u_r(x) + \sum_{i=0}^{2k-3} U_c^i u_c^i(x), \tag{4.9}$$

donde U_c^i son momentos globales, en el sentido de que se toman respecto a más de una celda (2,3,... celdas vecinas). Como podemos observar, el polinomio (4.9) es de grado $2k-1$, cosa que nos beneficia ya que por resultados básicos de la teoría de interpolación se espera un orden de convergencia $2k$ (para $k = 2, 3$ esperaríamos un orden de 4 y 6), siendo ésta la cota a la que deseamos llegar. Cabe aclarar que las expresiones para los momentos son diferentes a las que se presentaron en el Capítulo 2, para ellos encontramos expresiones explícitas que se presentarán más adelante (Apéndice B). Podemos, entonces resumir todo lo anterior en el siguiente Teorema:

Teorema 1 *La solución $\tilde{u}_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ de la ecuación de momentos continuos (3.4), tiene la siguiente propiedad de convergencia: para cualquier punto $x \in [a,b]$ se cumple que para $k \geq 2$*

$$|u(x) - \tilde{u}_h(x)| = O\left(h^{2k}\right).$$

Demostración. Recordemos que $\tilde{u}_h(x)$ tiene la forma (4.9), así que

$$\begin{aligned}\epsilon(x) &= u(x) - \tilde{u}_h(x) \\ &= \left[u(x_l) - \tilde{U}_l\right] u_l(x) + \left[u(x_r) - \tilde{U}_r\right] u_r(x) \\ &+ \sum_{i=0}^{2k-3} \left[m_r^i(u) - m_c^i(\tilde{u}_h)\right] u_c^i(x) + R(x),\end{aligned}$$

suponiendo que $h\lambda \sim O(1)$ si $\mathcal{E}_k$ se usa para evitar usar un exponencial degenerado ($\exp\mu$ para $-\mu$ muy grande), entonces existen constantes c_l, c_r y c_c^i tal que

$$|u_l(x)| \leq c_l, \quad |u_r(x)| \leq c_r, \quad \left|u_c^i(x)\right| \leq c_c^i \quad \forall\, i,$$

entonces

$$\begin{aligned}\epsilon(x) &= c_l \left|u(x_l) - \tilde{U}_l\right| + c \left|_r u(x_r) - \tilde{U}_r\right| \\ &+ \sum_{i=0}^{2k-3} c_c^i \left|m_c^i(u) - m_c^i(\tilde{u}_h)\right| + |R(x)|,\end{aligned}$$

y como x no es un punto malla, usando (A.1) y (3.18) así como los resultados de interpolación: $|R(x)| \propto O\left(h^{2k}\right)$ obtenemos

$$\epsilon(x) = O\left(h^{2k}\right).$$

RESULTADOS NUMERICOS

Para obtener los resultados numéricos que corroboren el Teorema 1 se hicieron varias pruebas numéricas, las cuales consistieron en lo siguiente. Primeramente, es claro que para encontrar órdenes de convergencia se comparan errores obtenidos, éstos tienen que ser uno local comparado con uno global; local en el sentido de que la interpolación se hace sobre cada celda y en el global la interpolación se hace sobre varias celdas vecinas (ver Figuras 4.1 y 4.2). En cualquier caso habría que asegurarse si tenemos que tomar *postprocesamiento*, es claro que en el global sí, pero para los errores locales se tiene que verificar.

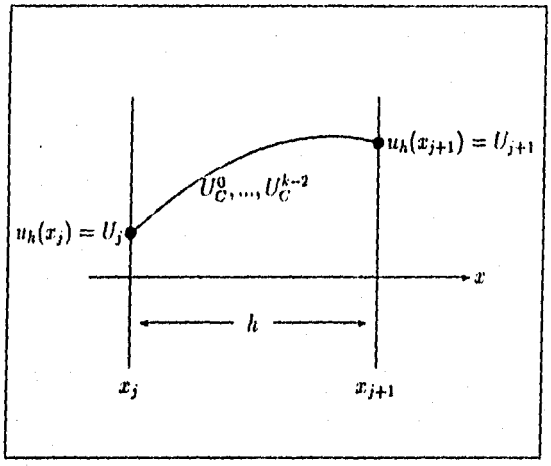


Figura 4.1: Interpolación Local (Continua) de grado k, por ejemplo $k = 2$

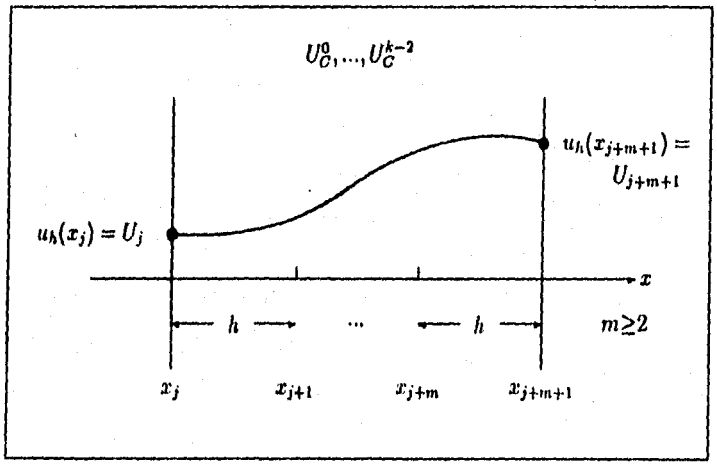


Figura 4.2: Interpolación Global (Continua) de grado k, por ejemplo $k = 3$

ERRORES LOCALES Para calcular los errores locales se puede hacer de dos formas, una con y otra sin postprocese, entendiendo que *sin postprocesar* significa utilizar la solución aproximada original (2.15), por ejemplo

- Para CMP2 ($k = 2$)

1. Sin postprocesar, y de acuerdo con (2.15) el polinomio de interpolación es de grado 2, las funciones base son las siguientes

$$\begin{aligned} u_l &= -\tfrac{1}{2}(P_1 - P_2) \\ u_r &= +\tfrac{1}{2}(P_1 + P_2) \\ u_c^0 &= P_0 - P_2, \end{aligned}$$

sustituyéndolas en (2.15) el polinomio nos queda de la forma

$$u_h = -\frac{1}{2}(P_1 - P_2)U_l + \frac{1}{2}(P_1 + P_2)U_r + (P_0 - P_2)U_c^0.$$

2. Con *postprocesamiento*, de acuerdo con (4.9) el polinomio es de grado 3 y las funciones base quedan de la forma

$$\begin{aligned} u_l &= \tfrac{1}{2}(P_2 - P_1) \\ u_r &= \tfrac{1}{2}(P_2 + P_3) \\ u_c^0 &= P_0 - P_2 \\ u_c^1 &= P_1 - P_3, \end{aligned}$$

además de que

$$U_c^1 = \frac{1}{2}(u_r - u_l)$$

así que, después de sustituirlas en (4.9) y factorizar, el nuevo polinomio de interpolación nos queda:

$$\tilde{u}_h = -\frac{1}{2}(P_1 - P_2)U_l + \frac{1}{2}(P_1 + P_2)U_r + (P_0 - P_2)U_c^0.$$

Con ésto verificamos que en el caso de CMP2 ($k = 2$), tener una aproximación local sin postprocesar y postprocesada, es lo mismo, claro que cuando

postprocesamos localmente, el momento uno U_c^1 se aproxima o construye con las funciones base u_r y u_l ya que en este caso no se tiene.

Para los demás casos, se hace de manera muy similar, y están desarrollados en el Apéndice (C).

Entonces, resumiendo tenemos que para los esquemas continuos no es necesario hacer un *postprocesamiento* para encontrar los errores locales ya que éstos son exactamente iguales a los que se obtienen sin postprocesar (originales), así que tomamos los errores locales originales que son de la forma (4.5) con u_h de la forma (2.15) en cada celda.

ERRORES GLOBALES. Para encontrar los errores globales, recordemos que global significa tomar los momentos y funciones base respecto a más de una celda. Así que para encontrar los errores globales (postprocesados) construimos un polinomio de interpolación de la forma (4.9) construido de manera global (respecto a más de una celda), en particular se construyó un polinomio cada dos celdas vecinas, tomándose también los momentos y funciones base respecto a cada dos celdas vecinas. Entonces, recordemos que los momentos son diferentes a los presentados en el Capítulo 2. Las expresiones explícitas para estos momentos se encuentran en el Apéndice B, los cuales se calcularon de la forma

$$U_c^i = \frac{\int_{x_{2k-1}}^{x_{2k+1}} p_i\left(\frac{2x-x_{2k-1}-x_{2k+1}}{2h}\right) u_h(x)\,dx}{\int_{x_{2k-1}}^{x_{2k+1}} p_i^2\left(\frac{2x-x_{2k-1}-x_{2k+1}}{2h}\right) dx}. \tag{4.10}$$

En (4.10), las integrales están tomadas en los intervalos (x_{2k-1}, x_{2k+1}) para $k = 1, 2, \ldots$ lo que significa que agrupa a dos celdas, lo que le da el sentido de global. De nuevo, los errores se calculan de la forma (4.5), pero, en este caso u_h es reemplazada por $\tilde{u}_h$, el polinomio de interpolación (4.9) global (respecto a dos celdas vecinas).

Así que, una vez teniendo los errores locales y globales (postprocesados), se calcula el orden de convergencia continuo postprocesado (que llamaremos CCOP) de la siguiente forma

$$\text{CCOP} = \frac{\ln\left(e_h/\tilde{e}_{h/2}\right)}{\ln 2}, \tag{4.11}$$

donde e_h tiene la forma (4.5) y

$$\tilde{e}_h = \max_x |u(x) - \tilde{u}_h(x)|,$$

aquí, hay que aclarar de que forma se hace la comparación de errores, por ejemplo si el dominio del problema modelo $(-1,1)$ se divide en 4 subintervalos, para encontrar e_h (error local) se construyen 4 polinomios de la forma (2.15), uno en cada subintervalo o celda, tomando como parámetros el valor de u_h en los extremos de cada celda y los momentos locales necesarios (ver Figura 4.3) y en el caso de $\tilde{e}_h$, se construyen solo 2 polinomios cada dos celdas vecinas (ver Figura 4.4) de la forma (4.9). En general, si tenemos n celdas para encontrar los errores locales se construyen n polinomios de grado k de la forma (4.8) uno en cada celda, tomando como parámetros los valores de u_h en los extremos de cada celda más los momentos locales U_c^i, $i = 0, \ldots k-2$; y para los errores globales se construyen $n/2$ polinomios de grado $2k-1$ de la forma (4.9) cada dos celdas con los parámetros: valor de u_h en el extremo izquierdo de la primer celda y extremo derecho de la segunda celda, así como los momentos globales U_c^i $i = 0, \ldots 2k-3$.

Una vez que tenemos e_h y $\tilde{e}_h$ (errores local y global respectivamente) se comparan de tal forma que ambos estén tomados sobre el mismo subintervalo, es decir, si se calcula e_h para n celdas de la forma(x_j, x_{j+1}) el correspondiente error global es aquel en el que cada dos celdas vecinas de la forma (x_{2k-1}, x_{2k+1}) coincidan con la celda (x_j, x_{j+1}), es decir, $x_j = x_{2k-1}$ y $x_{j+1} = x_{2k+1}$. En otras palabras, si se toma e_h en n celdas, tenemos que tomar el error global para $n/2$ celdas, que es $\tilde{e}_{h/2}$, gráficamente se puede ver en la Figuras 4.5, 4.6 o bien en la Figura 4.7, es decir, se toma e_h y $\tilde{e}_h$ en un

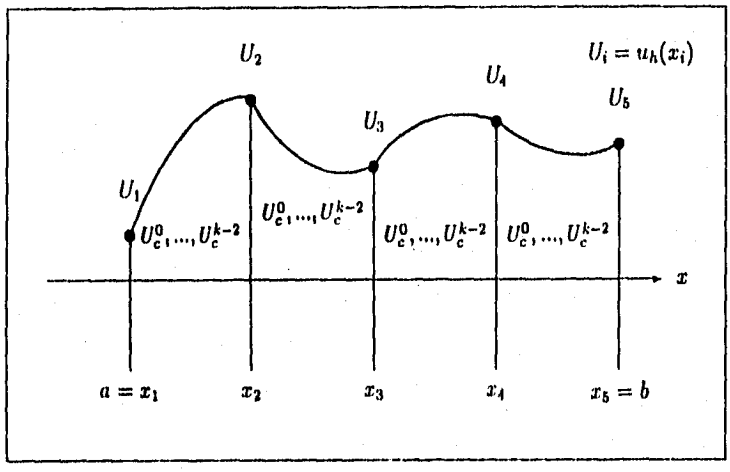


Figura 4.3: Interpolaciones Locales (Continuas sobre cada celda) de grado k, por ejemplo $k = 2$

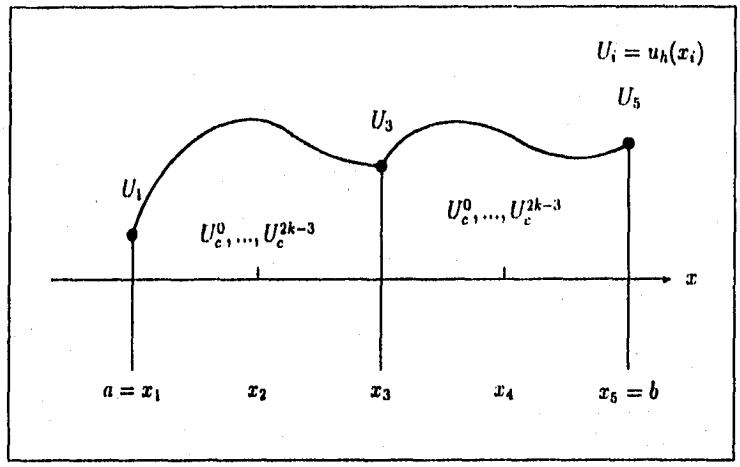


Figura 4.4: Interpolaciones Globales (Continuas cada dos celdas) de grado k, por ejemplo $k = 3$

	CMP2		CMP3	
n	e_h	$\tilde{e}_h$	e_h	$\tilde{e}_h$
2	3.9032×10^{-2}	4.2446×10^{-2}	5.8181×10^{-3}	2.3508×10^{-3}
4	8.6688×10^{-3}	6.6065×10^{-3}	6.3864×10^{-4}	1.7526×10^{-4}
8	1.4222×10^{-3}	8.8728×10^{-4}	5.3749×10^{-5}	1.2331×10^{-5}
16	2.0101×10^{-4}	1.0982×10^{-4}	3.9245×10^{-6}	8.9030×10^{-7}
32	2.6572×10^{-5}	1.3207×10^{-5}	2.6498×10^{-7}	6.0433×10^{-8}
64	8.4325×10^{-8}	1.6177×10^{-6}	1.7210×10^{-8}	4.1253×10^{-9}

Tabla 4.9: Errores sin y con postprocesamiento global para CMP2 y CMP3

	CMA2		CMA3	
n	e_h	$\tilde{e}_h$	e_h	$\tilde{e}_h$
2	1.5576×10^{-2}	8.3176×10^{-3}	2.2042×10^{-3}	1.5890×10^{-3}
4	3.6248×10^{-3}	2.7460×10^{-3}	2.5262×10^{-4}	1.3425×10^{-4}
8	6.0140×10^{-4}	4.0622×10^{-4}	2.1798×10^{-5}	8.1099×10^{-6}
16	8.5153×10^{-5}	5.0224×10^{-5}	1.6047×10^{-6}	4.2868×10^{-7}
32	1.1202×10^{-5}	5.6623×10^{-6}	1.0862×10^{-7}	2.5249×10^{-8}
64	5.6917×10^{-8}	7.1490×10^{-7}	7.0612×10^{-9}	1.7005×10^{-9}

Tabla 4.10: Errores sin y con postprocesamiento global para CMA2 y CMA3

subintervalo de la misma longitud, dividiendo el dominio de prueba $[-1,1]$ en n subintervalos, encontramos el error e_h y se compara con $\tilde{e}_h$ calculado para $n/2$ intervalos ($\tilde{e}_{h/2}$). Los errores que se obtuvieron se presentan en las Tablas 4.9 y 4.10.

Sustituyendo estos errores en (4.11) se calcularon los Ordenes de Convergencia Continuos (postprocesados, que llamaremos CCOP), los cuales están en la Tabla 4.11.

Con esta tabla ilustramos numéricamente los resultados del Teorema 1.

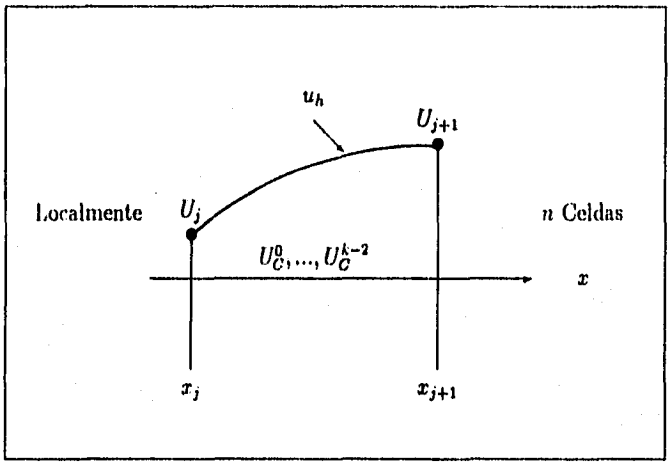


Figura 4.5: Polinomio de Interpolación Local (Continuo), por ejemplo $k = 2$

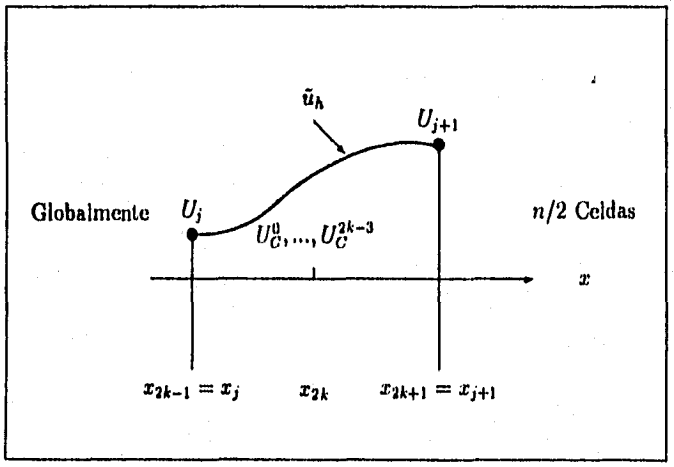
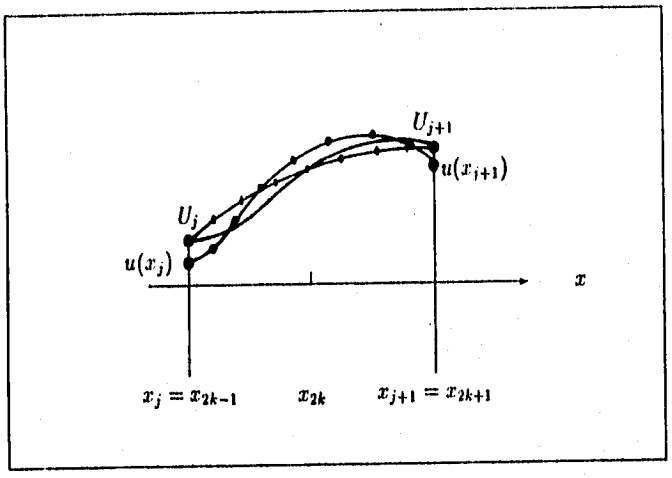


Figura 4.6: Polinomio de Interpolación (Global Continuo) Postprocesado, por ejemplo $k = 3$



u_h —•—•— Interpolación Local

$\tilde{u}_h$ —— Interpolación Global

u —•—•— Solución Exacta

Figura 4.7: Comparación de Interpolaciones y Solución Exacta ($k = 3$)

CCOP				
	CMP		CMA	
Intervalos	CMP2	CMP3	CMA2	CMA3
2-4	2.563	5.053	2.504	4.037
4-8	3.288	5.695	3.158	4.961
8-16	3.695	5.916	3.582	5.668
16-32	3.928	6.021	3.910	5.990
32-64	4.038	6.005	3.970	5.997
Teórico	4	6	4	6

Tabla 4.11: Ordenes de Convergencia Continuos (postprocesado)

4.2.2 ESQUEMAS DISCONTINUOS

Para los esquemas discontinuos obtener un *orden de convergencia continuo* (CCO) más alto, se hace aprovechando las *superconvergencias* (B.1) (Apéndice B) y (3.24) para $k = 1, 2$. De manera análoga que el caso anterior (continuo) para obtener un CCO más alto utilizaremos el valor de u_h en el extremo derecho de las celdas, ya que gracias a DCO(r) tenemos superconvergencia en ellos, así como los momentos globales necesarios (M0, M1 para $k = 2$ y M0, M1, M2, y M3 para $k = 3$), con ellos *postprocesaremos* la u_h obtenida en una $\tilde{u}_h$ que exhiba un orden de convergencia más alto, de esta forma obtendremos órdenes de convergencia continuos de la forma $O\left(h^{2k+1}\right)$, para $k = 1, 2$. Cuando hablamos de la forma de obtener CCO mencionamos que se evaluó la solución aproximada (un polinomio de grado k en cada celda) de la forma

$$u_h = U_r u_r(x) + \sum_{i=0}^{k-1} U_c^i u_c^i(x), \tag{4.12}$$

donde $u_r(x)$ y $u_c^i(x)$ son las funciones base dadas por (2.62) y (2.63) respectivamente en el caso polinomial y por (2.64) en el caso analítico respecto a cada celda.

En este caso, la prueba de *postprocesamiento* que hicimos fue del tipo discontinuo, es decir, se utilizaron polinomios los cuales no necesariamente eran continuos de celda a celda, esto significa que $u_h(x_r)$ en el extremo derecho de cada no necesariamente es igual a $u_h(x_l + 0)$ de la siguiente celda.

A diferencia del los esquemas continuos, el polinomio con el que se interpoló para el *postprocesamiento* (en forma global) fue de la forma

$$\tilde{u}_h(x) = U_r u_r(x) + \sum_{i=0}^{2k-1} U_c^i u_c^i(x), \tag{4.13}$$

que es de grado $2k$ y con momentos globales U_c^i, en el sentido de que se toman sobre más de una celda, de esta forma, teniendo un polinomio de interpolación de grado $2k$ esperamos (por resultados básicos de interpolación) un orden de convergencia de $2k+1$. Si llegamos a demostrar esta hipótesis, entonces superaríamos el *orden de convergencia continuo* (conocido) $k+1$. Al igual que para los métodos CMP y CMA las expresiones para U_c^i son diferentes a las que se presentaron en el Capítulo 2, y también se encontraron expresiones en forma explícita que se presentaran más adelante (Apéndice B). Resumiendo tenemos:

Teorema 2 *La solución $\tilde{u}_h \in \mathcal{P}_k$ o $\mathcal{E}_k$ de la ecuación de momentos discontinuos (3.20), tiene la siguiente propiedad de convergencia: para cualquier punto $x \in [a,b]$*

$$|u(x) - \tilde{u}_h(x)| = O\left(h^{2k+1}\right),$$

donde $\tilde{u}_h$ es de la forma (4.13).

Demostración. Recordemos que $\tilde{u}_h(x)$ tiene la forma (4.13), así que

$$\begin{aligned}\epsilon(x) &= u(x) - \tilde{u}_h(x)\\ &= \left[u(x_r) - \tilde{U}_r\right] u_r(x)\\ &+ \textstyle\sum_{i=0}^{2k-1} \left[m_c^i(u) - m_c^i(\tilde{u}_h)\right] u_c^i(x) + R(x),\end{aligned}$$

suponiendo que $h\lambda \sim O(1)$ si $\mathcal{E}_k$ se usa para evitar usar un exponencial degenerado, entonces existen constantes c_r y c_c^i tal que

$$|u_r(x)| \le c_r, \quad \left|u_c^i(x)\right| \le c_c^i \quad \forall\, i,$$

entonces

$$\begin{aligned}\epsilon(x) &= c_r \left|u(x_r) - \tilde{U}_r\right|\\ &+ \textstyle\sum_{i=0}^{2k-1} c_c^i \left|m_c^i(u) - m_c^i(\tilde{u}_h)\right| + |R(x)|,\end{aligned}$$

y como x no es un punto malla, usando (B.1) y (3.24) así como el resultado de interpolación $|R(x)| = O\left(h^{2k+1}\right)$ obtenemos

$$e(x) = O\left(h^{2k+1}\right).$$

RESULTADOS NUMERICOS. Para ejemplificar el Teorema 2 se hicieron varias pruebas numéricas, al igual que para los esquemas continuos las pruebas consistieron en comparar errores locales contra globales, en el mismo sentido que se menciono antes siempre y cuando se respete la discontinuidad de estos esquemas (ver Figuras 4.8 y 4.9); de la misma forma que para los CMP y CMA, tenemos que verificar si tomamos *postprocesamiento* para calcular los dos tipos de errores (locales y globales), es claro que en el global sí, pero en el local se tiene que corroborar.

ERRORES LOCALES. Para calcular los errores locales, tenemos dos formas de hacerlo, sin y con *postprocesamiento*, en la primera se construye un polinomio de la forma (4.12) (ver Figura 4.10) y en la segunda de la forma (4.13) en cada celda (no necesariamente continuo de celda a celda), como es de esperarse al igual que en los métodos continuos obtuvimos los mismos errores y esto se debe a que de cualquiera de las dos formas es el mismo polinomio, por ejemplo:

- Para DMP1

1. Sin postprocesar, el polinomio es de grado 1, las funciones base son las siguientes

$$\begin{aligned} u_r &= P_1 \\ u_c^0 &= P_0 - P_1, \end{aligned}$$

y el polinomio de interpolación, de acuerdo a (2.59) nos queda de la forma

$$u_h = P_1 U_r + (P_0 - P_1) U_c^0.$$

2. Con *postprocesamiento*, el polinomio es de grado 2 y las funciones base quedan de la forma

$$\begin{aligned} u_r &= P_2 \\ u_c^0 &= P_0 - P_2 \\ u_c^1 &= P_1 - P_2, \end{aligned}$$

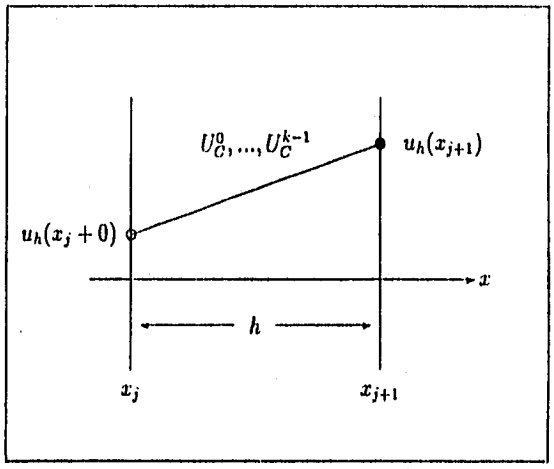


Figura 4.8: Interpolación Local de grado k (Discontinua), por ejemplo $k = 1$

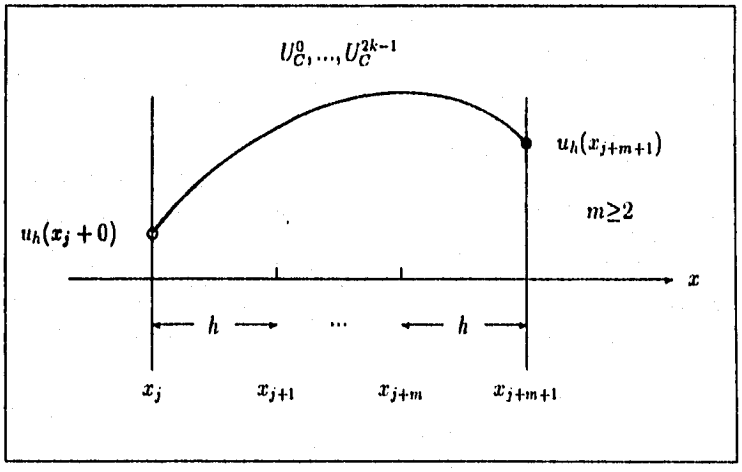


Figura 4.9: Interpolación Global de grado k (Discontinua), por ejemplo $k = 2$

con

$$U_c^1 = u_r - u_c^0$$

así, que el nuevo polinomio de interpolación, después de factorizar nos queda:

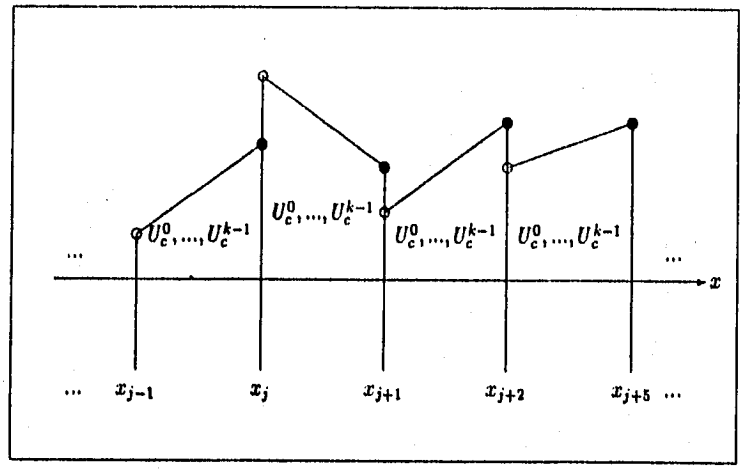
$$\tilde{u}_h = P_1 U_r + (P_0 - P_1) U_c^0.$$

Con esto podemos asegurar que en el caso de DMP1, tener una aproximación local sin postprocesar y una postprocesada, es exactamente lo mismo, siempre y cuando al postprocesar localmente el momento uno U_c^1 se aproxime con el valor de u_h en el extremo derecho de cada celda u_r y la función base u_c^0, cosa que tiene que hacerse ya que no se conoce, debido a esto los errores son exactamente los mismos.

Para los demás casos, se hace de manera muy similar, y están desarrollados en el Apéndice C. Por tal motivo, los errores locales serán los que se tiene originalmente.

ERRORES GLOBALES. Recordemos de nuevo que global significa tomar los momentos y funciones base respecto a más de una celda, así que seguiremos tomándolas respecto a dos celdas adyacentes o vecinas. Además los momentos son diferentes a los presentados en el Capítulo 2: Las expresiones para estos momentos se encuentran en el Apéndice B, los cuales se calcularon de la forma (4.10). Entonces, una vez que se tienen los errores locales y globales postprocesados, se calcula CCOP con (4.11). Al igual que para los métodos continuos, para calcular los errores globales construimos un polinomio de la forma (4.13) cada dos celdas vecinas (ver Figura 4.11). La comparación de errores se hace tomando e_h y $\tilde{e}_h$ en un subintervalo de la misma longitud, es decir si se divide el dominio de prueba $[-1, 1]$ en n subintervalos, se toma el error e_h y se compara con $\tilde{e}_h$ calculado para $n/2$ intervalos, llamado $\tilde{e}_{h/2}$ (ver Figuras 4.12 y 4.13).

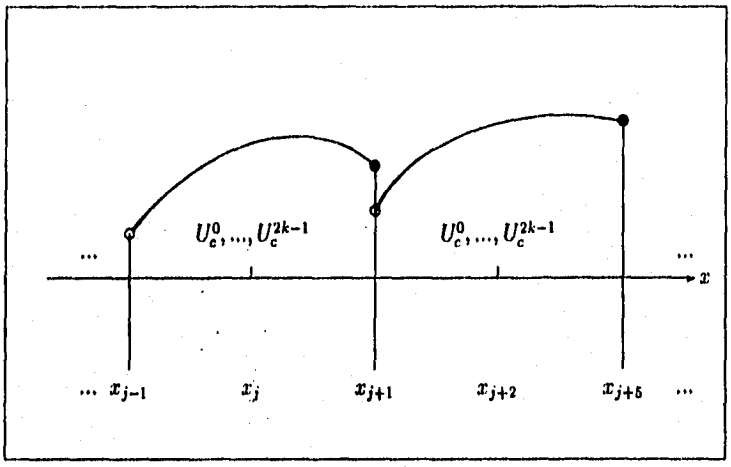
Los errores obtenidos, en estos casos ($k = 1, 2$) se presentan en las Tablas 4.12 y 4.13.



$\circ\, U_i$

$\bullet\, u_h(x_j + 0)$

Figura 4.10: Interpolaciones Locales de grado k (Discontinuas sobre cada celda), por ejemplo $k = 1$



- o U_i
- • $u_h(x_j + 0)$

Figura 4.11: Interpolaciones Globales de grado k (Discontinuas sobre cada dos celdas, por ejemplo $k = 2$

	DMP1		DMP2	
n	e_h	$\tilde{e}_h$	e_h	$\tilde{e}_h$
2	0.27075	0.25261	6.5832×10^{-2}	4.1117×10^{-2}
4	0.11927	0.10169	1.5135×10^{-2}	5.8561×10^{-3}
8	4.2103×10^{-2}	30660×10^{-2}	2.6547×10^{-3}	6.9977×10^{-4}
16	1.2742×10^{-2}	8.0462×10^{-3}	3.9671×10^{-4}	7.7762×10^{-5}
32	3.5221×10^{-3}	2.0102×10^{-3}	5.4345×10^{-5}	8.7422×10^{-6}
64	9.2694×10^{-4}	4.9753×10^{-4}	7.1156×10^{-6}	1.0175×10^{-6}

Tabla 4.12: Errores sin y con postprocesamiento global (con interpolación discontinua) para DMP1 y DMP2

	DMA1		DMA2	
n	e_h	$\tilde{e}_h$	e_h	$\tilde{e}_h$
2	0.15668	0.21921	4.0771×10^{-2}	1.0205×10^{-2}
4	6.3729×10^{-2}	6.2558×10^{-2}	7.8049×10^{-3}	3.0716×10^{-3}
8	2.0254×10^{-2}	1.5969×10^{-2}	1.2396×10^{-3}	3.4314×10^{-4}
16	5.6984×10^{-3}	3.7586×10^{-3}	1.7550×10^{-4}	3.5713×10^{-5}
32	1.5113×10^{-3}	8.8268×10^{-4}	2.3373×10^{-5}	3.8496×10^{-6}
64	3.8921×10^{-4}	2.1139×10^{-4}	3.0167×10^{-6}	4.3723×10^{-7}

Tabla 4.13: Errores sin y con postprocesamiento global (con interpolación discontinua) para DMA1 y DMA2

CCOP				
	DMP		DMA	
Intervalos	DMP1	DMP2	DMA1	DMA2
2-4	1.413	3.491	1.325	3.730
4-8	1.960	4.435	1.997	4.508
8-16	2.388	5.093	2.430	5.117
16-32	2.664	5.504	2.691	5.511
32-64	2.824	5.739	2.838	5.740
Teórico	3	5	3	5

Tabla 4.14: Ordenes de Convergencia Continuos (con postprocesamiento discontinuo)

Con estos errores, se calcularon los Ordenes de Convergencia Continuos postprocesados CCOP, los cuales están en la Tabla 4.14.

Como podemos observar en la Tabla 4.14, los resultados teóricos no coinciden con los numéricos en los casos DMP2 y DMA2, sino que estos son más altos de lo que se esperaba teóricamente, sin contradecir el Teorema 2, aunque teóricamente solo podemos asegura un CCOP de 5. Para $k = 1$ tenemos una muy buena similitud entre el resultado teórico y numérico.

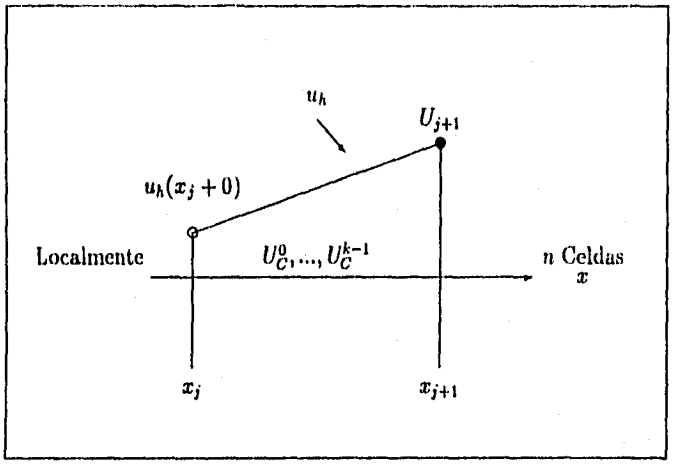


Figura 4.12: Polinomio de Interpolación Local (Discontinuo), por ejemplo $k=1$

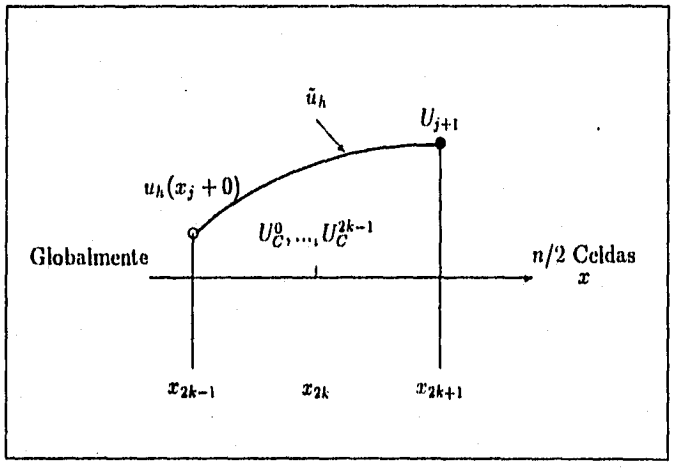


Figura 4.13: Polinomio de Interpolación Global (Postprocesado Discontinuo), por ejemplo $k=2$

CAPITULO 5

CONCLUSIONES

En este trabajo se han revisado dos tipos de métodos nodales de aproximación, *el continuo* y *el discontinuo*. En ambos métodos se analizaron varios ejemplos para los casos *polinomial* ($\mathcal{P}_k$) y *analítico* ($\mathcal{E}_k$). Estos fueron los siguientes:

- Para el método continuo, en ambos casos se analizaron los ejemplos para $k = 2, 3$
- y para el discontinuo fue para $k = 1, 2$

la razón de analizarlos fue para verificar el *orden de convergencia* de estos métodos. Al hacerlo encontramos que tienen propiedades de *superconvergencia* en el o los extremos de cada celda que divide al dominio (DCO) y en los momentos, notando que en todos los casos CCO (*orden de convergencia continuo*) es menor que DCO (*orden de convergencia discreto*).

Con estas observaciones, lo que se hizo fue mejorar CCO, es decir, obtener un CCO más alto, concretamente se igualó a DCO. Para ello se aprovecharon las superconvergencias en uno o los dos extremos de cada celda (DCO y DCO(r)) así como la de algunos *momentos*, esto se hizo construyendo nuevos polinomios de interpolación (grado más alto que las soluciones aproximadas por métodos nodales), a lo que llamamos *postprocesamiento*. Los nuevos polinomios difieren de los que se utilizaron originalmente para calcular CCO [11] y [12], en:

- se construyen de manera global, es decir, en el sentido de que son respecto a más de una celda (aquí se construyeron cada 2 celdas), tomando como parámetros los valores de u_h en uno o los extremos (dependiendo del esquema) de éstas dos celdas así como sus momentos (globales respecto a las dos celdas).

En el Capítulo 4 se detalló la forma en que fueron tomados los nuevos polinomios de interpolación, así como de qué forma se hizo la comparación de errores. Además se presentan 2 teoremas nuevos (Teoremas 1 y 2) que son la principal aportación de este trabajo, agregando que se hicieron varias pruebas numéricas que ilustran los resultados teóricos.

En los casos DMP2 y DMA2, cuando se postprocesaron numéricamente se encontró un orden de convergencia más alto del que se esperaba, aunque teóricamente no se puede demostrar.

De todo lo anterior podemos concluir que efectivamente después de haber postprocesado la u_h obtenida en una $\tilde{u}_h$, los órdenes de convergencia continuos (CCO) se pueden mejorar, es decir, la $\tilde{u}_h$ puede alcanzar un orden de convergencia más alto que el de u_h en cualquier punto del dominio, de tal forma que se iguale al que se tiene en los puntos de la malla (DCO), que ya es superconvergente. En las pruebas que se realizaron, tenemos que para los esquemas continuos (polinomial y analítico) y para los discontinuos (polinomial y analítico) en el caso $k = 1$, se igualó el CCOP (orden de convergencia continuo con postprocesamiento) con DCO (orden de convergencia discreto) tanto teórica como analíticamente. En los casos restantes DMP2 y DMA2 el resultado numérico superó al teórico.

En estos últimos casos (DMP2 y DMA2) no se puede obtener teóricamente un CCOP mayor a 5; el análisis de este fenómeno puede ser un tema para un trabajo futuro, así como verificar si en el caso del problema de valores a la frontera

$$-\frac{d}{dx}\left(\mu\frac{du}{dx}\right) + qu = f, \qquad \forall \quad x \in (a,b),$$

sujeta a las condiciones de frontera

$$u(a) = \frac{du}{dx}(b) = 0.$$

se pueden obtener comportamientos del mismo tipo, es decir obtener ordenes de convergencia continuos más altos, para los cuales pensamos que si existen fenómenos muy similares. Además creemos que se pueden extender a varias dimensiones.

En el caso de coeficientes variable (especificamente q), existe un trabajo, [17] en el cual se muestra superconvergencia en el caso de DCO, se encuentra que es de $O\left(h^{2k}\right)$, para los esquemas continuos (polinomial y analítico), en el cual se usa una cuadratura numérica para el manejo de q, y para CCO (sin postprocesamiento) de $O\left(h^{k+1}\right)$; luego parece ser posible encontrar un CCOP más alto, ya que en un primer análisis presenta comportamientos similares al caso de coeficientes constantes.

CAPITULO 6

APENDICES

APENDICE A

TEOREMAS DE CONVERGENCIA

A.1 Para los Métodos de Momentos Continuos

Teorema 3. *La solución $u_h \in \mathcal{P}_k$, o $\mathcal{E}_k$ de las ecuaciones de momento continuas* (3.4) *tienen las siguientes propiedades de convergencia* :

1. *en el punto discreto x_i del intervalo discretizado $[a,b]$,*

$$|u(x_i) - u_h(x_i)| = O\left(h^{2k}\right), \tag{A.1}$$

donde h es el tamaño máximo de los intervalos usados sobre $[a,b]$.

2. *además*

$$\left|m_c^i(u) - m_c^i(u_h)\right| = O\left(h^{2k-i}\right), \qquad \text{para } i \leq k, \tag{A.2}$$

3. *finalmente, en cualquier punto $x \in [a,b]$,*

$$|u(x) - u_h(x)| = O\left(h^{k+1}\right). \tag{A.3}$$

Las referencias (A.1 *y* A.2) *son resultados de superconvergencia, observables sólo para $k > 1$.*

Demostración. Claramente u satisface las ecuaciones (3.2) para todo i, mientras que u_h, por construcción las satisface para $i = 0, \ldots, k-1$. Como consecuencia tenemos que

$$\int_{x_l}^{x_r} L(u - u_h) p_i(x)\, dx = 0, \qquad i = 0, \ldots, k-1. \tag{A.4}$$

donde

$$L(u) = \mu \frac{du}{dx} + qu.$$

Si llamamos $\epsilon \equiv u - u_h$, el *error*, entonces $L\epsilon = r$ donde r es el residual. El siguiente paso consiste en introducir la función de Green $G(x \mid \zeta)$ tal que

$$\begin{aligned} G(x \mid \zeta) &= 0, & x &\in [x_l, \zeta), \\ G(x \mid \zeta) &= \exp(\lambda(\zeta - x)), & x &\in (\zeta, x_r], \end{aligned}$$

con una discontinuidad de uno en $x = \zeta$ $(G(\zeta + 0 \mid \zeta) - G(\zeta - 0 \mid \zeta) = 1)$.

Suponiendo que $u(x_l) = u_h(x_l)$ o equivalentemente que $\epsilon(x_l) = 0$, llegamos a

$$\epsilon(x) = \int_{x_l}^{x_r} G(x \mid \zeta) r(\zeta) d\zeta. \tag{A.5}$$

Como $L\epsilon = r$ es ortogonal a p_i, $i = 0, \ldots, k-1$ (de (A.4), su expansión en series de polinomios de Legendre sobre $[x_l, x_r]$ sería

$$r(x) = \sum_{j=k}^{\infty} a_j p_j(x). \tag{A.6}$$

$G(x \mid \zeta)$, con su discontinuidad en $x = \zeta$ no es suave sobre $[x_l, x_r]$ pero $G_r(\zeta) \equiv G(x_r \mid \zeta)$ es $C^\infty[x_l, x_r]$. Con (A.6) y una expansión en polinomios de Legendre similar a la de r para $G_r(\zeta)$, (A.5) nos lleva a

$$\begin{aligned} \epsilon(x_r) &= \int_{x_l}^{x_r} G_r(\zeta) r(\zeta) d\zeta \\ &= \sum_{j=k}^{\infty} b_j \int_{x_l}^{x_r} p_j^2(\zeta) d\zeta. \end{aligned} \tag{A.7}$$

En la ecuación anterior, cada término de la expansión puede evaluarse por cuadratura numérica de la siguiente forma

$$\int_{x_l}^{x_r} p_j^2(\zeta)\, d\zeta = \sum_{p=1}^{P} \omega_p p_j^2(\zeta_p) + residuo,$$

donde ζ_p, ω_p, $p = 1, \ldots, P$ son los puntos y pesos de la cuadratura respectivamente. Si en el primer término del lado derecho de (A.7), escogemos como puntos de cuadratura los k puntos de Gauss-Legendre, es decir, los k ceros de $p_k(\zeta)$, la ecuación anterior nos queda

$$\int_{x_l}^{x_r} p_k^2(\zeta)\, d\zeta = O\left(h^{2k+1}\right), \tag{A.8}$$

ésto último se sigue de los resultados básicos de la integración de Gauss [15]. El término general en (A.7) es del $O(h^{2j+1})$, $j = k, \ldots, \infty$, así que $|\epsilon(x_r)| = O\left(h^{2k+1}\right)$. Con $\epsilon(x_l) = 0$, este es el error local, lo que implica que el correspondiente esquema es de orden global $2k$, lo cual prueba (A.1).

Para probar (A.2), tomaremos el i–ésimo momento de $\epsilon(x)$ sobre $[x_l, x_r]$:

$$\begin{aligned} m_c^i(u) - m_c^i(u_h) &= m_c^i(\epsilon) \\ &= \tfrac{1}{N_i} \int_{x_l}^{x_r} G^i(\zeta)\, r(\zeta)\, d\zeta, \end{aligned} \tag{A.9}$$

donde

$$G^i(\zeta) = \int_{x_l}^{x_r} p_i(x)\, G(x \mid \zeta)\, dx. \tag{A.10}$$

Para cualquier i, $G^i(\zeta) \in C^\infty[x_l, x_r]$ y usando la misma técnica que antes, llegamos a

$$m_c^i(u) - m_c^i(u_h) \propto O\left(h^{2k-i}\right) \quad \forall\, i \leq k, \tag{A.11}$$

donde $h = x_r - x_l$. La presencia del factor normalizador N_i en (A.9) hace que el exponente baje de $(2k+1)$ en (A.8) a $2k$. Sin embargo $p_i(x)$ introduce en (A.10) un término dominante en h^{-i} cuando $h \to 0$, por lo tanto el exponente final es $(2k - i)$ en (A.11).

Para obtener las cotas de error continuo (A.3), escribiremos loa solución u como su interpolación en $\mathcal{P}_k$ o $\mathcal{E}_k$ más un residuo R, de la forma

$$u(x) = u(x_l)\,u_l(x) + u(x_r)\,u_r(x) + \sum_{i=0}^{k-2} m_c^i(u)\,u_c^i(x) + R(x),$$

donde $R(x)$ es igual a cero en cada uno de los puntos de la malla. Consecuentemente,

$$\begin{array}{rl} \epsilon(x) & = u(x) - u_h(x) \\ & [u(x_l) - U_l]\,u_l(x) + [u(x_r) - U_r]\,u_r(x) + \\ & \sum_{i=0}^{k-2} [m_c^i(u) - m_c^i(u_h)]\,u_c^i(x) + R(x) \end{array} \qquad (A.12)$$

Asumiendo que $h\lambda \sim O(1)$ si $\mathcal{E}_k$ se usa en lugar de usar una exponencial degenerada ($\exp(\alpha)$ para α muy negativas), entonces existen constantes c_l, c_r y c_c^i tal que

$$|u_l(x)| \le c_l, \quad |u_r(x)| \le c_r \quad y \quad |u_c^i(x)| \le c_c^i, \quad \forall\, i,$$

y deducimos de (A.12) que

$$\begin{array}{rl} |\epsilon(x)| & = c_l\,|u(x_l) - U_l| + c_r\,|u(x_r) - U_r| + \\ & \sum_{i=0}^{k-2} c_c^i\,|m_c^i(u) - m_c^i(u_h)| + |R(x)|. \end{array}$$

Si x no es un punto malla, usando (A.1) y (A.2) así como el resultado de interpolación $|R(x)| = O\left(h^{k+1}\right)$, obtenemos (A.3). Este resultado de interpolación el cual es bien conocido para $\mathcal{P}_k$ es valido también para $\mathcal{E}_k$ ya que este espacio es un sistema de Tchebycheff Completo Extendido como $\mathcal{P}_k$ [16].

Y el otro teorema, que se refiere a momentos globlales es:

Teorema 4. *Sobre un intervalo de longitud fija, los momentos globales de la solución $u_h \in \mathcal{P}_k$, o $\mathcal{E}_k$, de las ecuaciones de momento continuas* (3.4) *tiene la siguiente propiedad de convergencia:*

$$\left|m_c^i(u) - m_c^i(u_h)\right| = O\left(h^{2k}\right).$$

A.2 Para los Métodos de Momentos Discontinuos

Teorema 5. *La solución $u_h \in \mathcal{P}_k$, o $\mathcal{E}_k$ de las ecuaciones de momento discontinuas* (3.20) *tienen las siguientes propiedades de convergencia* :

1. *en el punto discreto x_i del intervalo discretizado $[a,b]$,*

$$|u(x_i) - u_h(x_i)| = O\left(h^{2k+1}\right), \tag{A.13}$$

donde h es el tamaño máximo de los intervalos usados sobre $[a,b]$.

2. *además*

$$\left|m_c^i(u) - m_c^i(u_h)\right| = O\left(h^{2k+1-i}\right), \qquad \text{para } i \leq k, \tag{A.14}$$

finalmente, en cualquier punto $x \in [a,b]$, distinto de los x_i's

$$|u(x) - u_h(x)| = O\left(h^{k+1}\right). \tag{A.15}$$

donde (A.13) *y* (A.14) *son resultados de superconvergencia, que se observan sólo para $k > 0$.*

Demostración. Primeramente, para remover el término de frontera en $x = x_l$, en la ecuación (3.19), podemos escribirla alternativamente como

$$\int_{x_l}^{x_r} (x - x_l)(Lu_h - f)\, q_i(x)\, dx = 0, \quad i = 0, \ldots k,$$

donde $q_i(x)$ es el polinomio de Jacobi de orden i con respecto al peso $(x - x_l)$. Esta misma expresión es también cierta para la solución exacta u para cualquier i, tal que

$$\int_{x_l}^{x_r} (x - x_l)\, L(u - u_h)\, q_i(x)\, dx = 0, \quad i = 0, \ldots k. \tag{A.16}$$

Los primeros polinomios de Jacobi sobre $[-1,1]$ son

$$Q_0 = 1 = \frac{P_0 + P_1}{x+1},$$
$$Q_1 = \frac{3x-1}{2} = \frac{P_1 + P_2}{x+1},$$
$$Q_3 = \frac{5x^2 - 2x - 1}{2} = \frac{P_2 + P_3}{x+1} \ldots.$$

y sus ceros son los puntos izquierdos de Radau los cuales son reales, distintos y localizados en el interior de $[-1,1]$. Si llamamos $e \equiv u - u_h$ el error, entonces $Le = r$, donde r es el residual el cual por (A.16) es ortogonal a los primeros $k-1$ polinomios de Jacobi con respecto al producto interior

$$(f,g) = (g,f) = \int_{x_l}^{x_r} (x - x_l) f(x) g(x) dx.$$

Consecuentemente, su expansión en series de polinomios de Jacobi sobre el intervalo $[x_l, x_r]$ sería

$$r(x) = \sum_{j=k+1}^{\infty} a_j q_j. \tag{A.17}$$

El siguiente paso consiste en introducir la función de Green $G(x \mid \zeta)$ tal que

$$\begin{aligned} G(x \mid \zeta) &= 0, && x \in [x_l, \zeta) \\ G(x \mid \zeta) &= \tfrac{1}{\mu} \exp\left(\tfrac{q}{\mu}(\zeta - x)\right), && x \in (\zeta, x_r]. \end{aligned}$$

con una discontinuidad de $\frac{1}{\mu}$ en $x = \zeta$ $(G(\zeta + 0 \mid \zeta) - G(\zeta - 0 \mid \zeta)) = \frac{1}{\mu}$.

Asumiendo que $u(x_l) = u_h(x_l)$ o equivalentemente que $e(x_l) = 0$, tenemos que

$$\begin{aligned} e(x) &= e(x_l + 0) \exp\left\{-\tfrac{\mu}{q}(x - x_l)\right\} \\ &\quad + \int_{x_l}^{x_r} G(x \mid \zeta) r(\zeta) d\zeta. \end{aligned}$$

la cual difiere de (A.5) porque el error brinca a un valor finito en $x_l + 0$. $G(x \mid \zeta)$, con su discontinuidad en $x = \zeta$ no es suave sobre $[x_l, x_r]$ pero

$G_r(\zeta) \equiv G(x_r \mid \zeta)$ esta en $C^\infty[x_l, x_r]$. Con una expansión en series de $G_r(\zeta)$ alrededor de $\zeta = x_l$, nos lleva a

$$\begin{aligned} e(x) &= e(x_l + 0)\exp(-\epsilon) \\ &\quad + G_r(x_l)\int_{x_l}^{x_r} r(\zeta)\, d\zeta \\ &\quad + \int_{x_l}^{x_r} (\zeta - x_l)\, G_r^*(\zeta)\, r(\zeta)\, d\zeta. \end{aligned} \tag{A.18}$$

donde $G_r^*(\zeta)$es de la forma

$$G_r^*(\zeta) = DG_r(x_l) + (\zeta - x_l)\frac{D^2G_r(x_l)}{2!} + (\zeta - x_l)^2\frac{D^3G_r(x_l)}{3!} + \ldots$$

que puede desarrollarse en series de polinomios de Jacobi como

$$G_r^*(\zeta) = \sum_{i=0}^{\infty} b_i q_i(\zeta). \tag{A.19}$$

usando (A.17) y (A.19), el tercer término del lado derecho de (A.18) queda

$$\sum_{i=k+1}^{\infty} a_i b_i \int_{x_l}^{x_r} (x - x_l)\, q_i^2(x)\, dx.$$

Regresando a la ecuación (3.19) con $i = 0$, una ecuación la cual es común para todos los esquemas, y usando el hecho de que $Lu = S$ y que $u_h(x_l - 0) = U_l$, podemos escribir esto como

$$\begin{aligned} &\int_{x_l}^{x_r} r(x)\, dx + \mu\left[u_h(x_l + 0) - U_l\right] \\ &= \int_{x_l}^{x_r} r(x)\, dx + \mu e(x_l + 0) = 0, \end{aligned}$$

tal que $e(x_l + 0) = \frac{1}{\mu}\exp(-\epsilon)$. Consecuentemente los dos primeros términos del lado derecho de (A.18) se cancelan y solo nos quedamos con

$$e(x_r) = \sum_{i=k+1}^{\infty} a_i b_i \int_{x_l}^{x_r} (x - x_l)\, q_i^2(x)\, dx. \tag{A.20}$$

En la cual, cada término de la expansión puede evaluarse por cuadratura numérica como sigue

$$\int_{x_l}^{x_r} (x - x_l)\, q_i^2(x)\, dx = \sum_{p=1}^{P} \omega_p (\zeta_p - x_l)\, q_i^2(\zeta_p) + residuo,$$

donde ζ_p, ω_p, $p = 1, \ldots P$ son puntos y pesos de la cuadratura respectivamente. Si en el primer término del lado derecho de (A.20), escogemos como puntos de cuadratura los $(k-1)$ puntos izquierdos de Gauss-Radau los cuales reemplazan a los $(k+1)$ ceros de $q_k(\zeta)$ más x_l, la ecuación anterior nos queda

$$\int_{x_l}^{x_r} (x - x_l)\, q_i^2(x)\, dx = O(h^{2k+2}),$$

siguiendo los resultados acerca de las técnicas de integración de Gauss [15]. El término general en (A.20) es entonces del $O(h^{2i+2})$, $i = k, \ldots, \infty$. tal que $e(x_r) \sim O\left(h^{2k+2}\right)$. Con $e(x_l) = 0$, entonces es un error local. implicando que el correspondiente esquema es de orden global $2k+1$, lo cual prueba (A.13).

La demostración de las tasas de convergencia (A.14) para los momentos así como de las cotas de error continuas (A.15) son muy similar a las pruebas de los resultados análogos en el caso de momentos continuos (Teorema 3).

De la misma forma para los momentos globales

Teorema 6. *Sobre un intervalo de longitud fija, los momentos globales de la solución $u_h \in \mathcal{P}_k$, o $\mathcal{E}_k$, de las ecuaciones de momento continuas* (3.20) *tiene la siguiente propiedad de convergencia*:

$$\left| m_c^i(u) - m_c^i(u_h) \right| = O\left(h^{2k+1}\right).$$

APENDICE B

MOMENTOS GLOBALES

En el Capítulo 4 se mencionó la utilización de momentos globales en el sentido de que son calculados respecto a más de una celda, en particular se dijo que los momentos globales utilizados serían respecto a dos celdas adyacentes con el fin de obtener un postprocesamiento global, estos momentos son de la forma

$$m_c^i(u_h) = \frac{\int_{x_{2k-1}}^{x_{2k+1}} p_i\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) u_h(x)\,dx}{\int_{x_{2k-1}}^{x_{2k+1}} p_i^2\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) dx}$$

y dependiendo del momento que deseamos calcular, es el valor de i. En particular, necesitamos los $m_c^i(u_h)$ para $i = 1, 2$, y 3.

Para $i = 1$, primeramente calculamos la integral del denominador, pero para ello hacemos una transformación al intervalo $(-1, 1)$, la cual es $z = (2x - x_{2k-1} - x_{2k+1})/2h$, así que

$$\int_{x_{2k-1}}^{x_{2k+1}} p_1^2\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) dx = h\int_{-1}^{1} p_1^2(z)\,dz,$$

como $p_1(z) = z$, entonces

$$\int_{x_{2k-1}}^{x_{2k+1}} p_1^2\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) dx = \frac{2}{3}h.$$

Así que

$$m_c^1(u_h) = \frac{3}{2}h \int_{x_{2k-1}}^{x_{2k+1}} p_i\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) u_h(x)\, dx,$$

separando la integral en dos tenemos

$$\begin{aligned} m_c^1(u_h) = \frac{3}{2}h \Bigg[& \int_{x_{2k-1}}^{x_{2k}} p_i\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) u_h(x)\, dx \\ & + \int_{x_{2k}}^{x_{2k+1}} p_i\left(\frac{2x - x_{2k-1} - x_{2k+1}}{2h}\right) u_h(x)\, dx \Bigg], \end{aligned}$$

Haciendo de nuevo una transformación para $(x_{2k-1}, x_{2k}) \to (-1, 1)$ y para $(x_{2k}, x_{2k+1}) \to (-1, 1)$ llegamos a

$$\begin{aligned} m_c^1(u_h) = \frac{1}{4} \Big[& \left(m_{c,k}^1 - 3m_{c,k}^0\right) \\ & + \left(m_{c,k+1}^1 + 3m_{c,k+1}^0\right) \Big]. \end{aligned}$$

donde $m_{c,j}^i$ es el *momento de celda local* i respecto a la celda j, es decir cuando $j = k$ se refiere al intervalo (x_{2k-1}, x_{2k}) y cuando $j = k + 1$ se refiere a (x_{2k}, x_{2k+1}).

De manera análoga lo hacemos para los casos $i = 2$ y 3, para los cuales encontramos que

$$\begin{aligned} m_c^2(u_h) = \frac{1}{8} \Big[& \left(m_{c,k}^2 - 5m_{c,k}^1\right) \\ & + \left(m_{c,k+1}^2 + 5m_{c,k+1}^1\right) \Big], \\ m_c^3(u_h) = \frac{1}{16} \Big[& \left(m_{c,k}^3 - 7m_{c,k}^2 + 7m_{c,k}^1 + 7m_{c,k}^0\right) \\ & + \left(m_{c,k+1}^3 + 7m_{c,k+1}^2 + 7m_{c,k+1}^1 - 7m_{c,k+1}^0\right) \Big], \end{aligned}$$

donde $m_{c,j}^i$ está definido de la misma forma en que se definió arriba.

Nota. En el caso de los métodos de momentos discontinuos tenemos que $m_{c,k}^3 \equiv m_{c,k+1}^3 \equiv 0$.

APENDICE C

COMPARACION DE INTERPOLACIONES LOCALES

- Para CMP3.

1. Sin postprocesar, el polinomio es de grado 3, las funciones base son las siguientes

$$\begin{aligned} u_l &= \tfrac{1}{2}(P_2 - P_3) \\ u_r &= \tfrac{1}{2}(P_2 + P_3) \\ u_c^0 &= P_0 - P_1 \\ u_c^1 &= P_1 - P_3 \end{aligned}$$

y el polinomio de interpolación, de acuerdo a (2.15) nos queda de la forma

$$u_h = \tfrac{1}{2}(P_2 - P_3)\,U_l + \tfrac{1}{2}(P_2 + P_3)\,U_r + (P_0 - P_1)\,U_c^0 + (P_1 - P_3)\,U_c^1$$

1. Con postprocesamiento, el polinomio es de grado 5 y las funciones base quedan de la forma

$$\begin{aligned} u_l &= \tfrac{1}{2}(P_4 - P_5) \\ u_r &= \tfrac{1}{2}(P_4 + P_5) \\ u_c^0 &= P_0 - P_4 \\ u_c^1 &= P_1 - P_5 \\ u_c^2 &= P_2 - P_4 \\ u_c^3 &= P_3 - P_5 \end{aligned}$$

con los momentos 2 y 3 aproximados de la forma

$$U_c^2 = \frac{1}{2}(U_r + U_l) - U_c^0$$
$$U_c^3 = \frac{1}{2}(U_r - U_l) - U_c^1$$

así que el polinomio de interpolación, después de varias factorizaciones nos queda:

$$\tilde{u}_h = \frac{1}{2}(P_2 - P_3)U_l + \frac{1}{2}(P_2 + P_3)U_r + (P_0 - P_1)U_c^0 + (P_1 - P_3)U_c^1$$

Con ésto, de nuevo, se verifica que para el caso CMP3 una interpolación local sin postprocesar y con postprocesamiento es exactamente igual.

- Para los casos CMA2 y CMA3, se realiza de manera similar, y se puede verificar que efectivamente, hacer un postprocesamiento local es igual que si no se postprocesa.

- Para el caso DMP2

1. Sin postprocesamiento, el polinomio es de grado 2, las funciones base son las siguientes

$$u_r = P_2$$
$$u_c^0 = P_0 - P_2$$
$$u_c^1 = P_1 - P_2$$

y el polinomio de interpolación, de acuerdo a (2.59) nos queda de la forma

$$u_h = P_2 U_r + (P_0 - P_2)U_c^0 + (P_1 - P_2)U_c^1$$

1. Postprocesando, el polinomio es de grado 4 y las funciones base quedan de la forma

$$u_r = P_4$$
$$u_c^0 = P_0 - P_4$$
$$u_c^1 = P_1 - P_4$$
$$u_c^2 = P_2 - P_4$$
$$u_c^3 = P_3 - P_4$$

y los momentos

$$u_c^2 = U_r - U_c^0 - U_c^1$$
$$U_c^3 = 0$$

así que el polinomio de interpolación, después de varias factorizaciones nos queda:

$$\tilde{u}_h = P_2 U_r + (P_0 - P_2) U_c^0 + (P_1 - P_2) U_c^1$$

De nuevo, se verifica que para el caso DMP2 una interpolación local sin y con postprocesamiento son exactamente iguales.

- Por último, se puede verificar de manera análoga que para los casos DMA1 y DMA2, sucede exactamente lo mismo.

REFERENCIAS

[1] Fairweather, G. (1978), *Finite Element Galerkin Methods for Differential Equations*, Marcel Dekker, New York.

[2] Hennart, J. P. (1985), A general finite element framework for nodal methods. In *The Mathematics of Finite Elements and Applications V*, págs. 309-316, J. R. Whiteman, Ed., Academic Press, Londres.

[3] Hennart, J. P. (1986), A general family of nodal schemes. *SIAM J. Sci. Stat. Comp.* **7**, 264-287

[4] Lawrence, R. D. (1986), Progress in nodal methods for the solution of the neutron diffusion and transport equations, *Progress in Nuclear Energy* **17**, 271-301.

[5] Hennart, J. P. (1988), On the numerical analysis of analitical nodal methods, *Numerical Methods for Partial Differential Equations* **4**, 233-254.

[6] Hennart, J. P., Jaffré, J., y Roberts, J. E. (1988), A constructive method for deriving finite elements of nodal type, *Numerische Mathematik* **53**, 701-738.

[7] Johnson, C. (1991), *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge Univesity Press.

[8] Hennart, J. P. (1992), A mixed-hybrid finite element formulation of fast nodal elliptic solvers. In *Iterative Methods in Linear Algebra*, págs. 485-492, R. Beauwens y P. de Groen, Eds., Elsevier Science Publishers, Países Bajos.

[9] Hennart, J. P. (1992), Afinite element approach to pint- and mesh-centered finite difference schemes and over rectangular grids, *Annals of Nuclear Energy* **19**, 663-678.

[10] Hennart, J. P. (1993), On the relationship between nodal schemes and mixed-hybid finite elements, *Numerical Methods for Partial Differential Equations* **9**, 411-430.

[11] Hennart, J. P. y Del Valle, E. (1995), A Generalized Nodal Finite Element Formalism for Discrete Ordinates Equations in Slab Geometry, Part I: Theory in the Continuous Moment Case, *Transport Theory and Statistical Physics* **24**, 449-478.

[12] Hennart, J. P. y Del Valle, E. (1995), A Generalized Nodal Finite Element Formalism for Discrete Ordinates Equations in Slab Geometry, Part II: Theory in the Discontinuous Moment Case, *Transport Theory and Statistical Physics* **24**, 479-504.

[13] Del Valle, E. y Hennart, J. P. (1995), A Generalized Nodal Finite Element Formalism for Discrete Ordinates Equations in Slab Geometry, Part III: Numerical Results, *Transport Theory and Statistical Physics* **24**, 505-533.

[14] Larsen, E. W. y Miller, W. F. Jr. (1980), Convergence rates of spatial difference equations for the discrete-ordinates neutron transport equations in slab geometry, *Nucl. Sci. Eng.* **73**, 76.

[15] Davis, P. J. y Rabinowitz, P. (1984), *Methods of Numerical Integration*, Academic Press, Orlando.

[16] Schumaker, L. (1981), *Spline Functions: Basic Theory*, John Wiley and Sons, New York.

[17] Hennart, J. P. y Del Valle, E. (1995), Numerial Cuadrature for Slab Geometry Transport Equations, Preimpreso No. 35 del IIMAS y por aparecer en *Transactions of The American Nuclear Society*, American Nuclear Society 1995 Winter Meeting, San Francisco, California.

[18] Ciarlet P. G. (1978), *The Finite Element Methods for Eliptic Problems*, North-Holland.

[19] Richtmyer R. D. y Morton K. W. (1967), *Difference Methods for Initial-value Problemas*, Interscience.