Métodos Numéricos en Fenómenos de Transporte.

Norberto Nigro <nnigro@intec.unl.edu.ar>
Mario Storti <mstorti@intec.unl.edu.ar>

www: http://www.cimec.org.ar/cfd
Centro Internacional de Métodos Computacionales en Ingeniería
http://www.cimec.org.ar

(Document version: texstuff-1.0.36-58-gb08d323 'clean)
(Date: Wed Sep 21 09:07:51 2011 -0300)

Índice general

Introducción. Contenidos del curso

Este curso básico sobre *CFD* siguiendo los lineamientos del libro de C. Hirsch [Hirsch] se divide en 2 partes:

1. Fundamentos y técnicas generales aplicables a los fenómenos de transporte en general y al flujo de calor y de fluidos en particular

   a) MODELOS FISICOS Y MATEMATICOS EN CFD
   b) APROXIMACIONES DINAMICAS
   c) NATURALEZA MATEMATICA DE LAS ECUACIONES
   d) TECNICAS DE DISCRETIZACION GLOBAL
   e) METODOS ESPECTRALES
   f) TECNICAS DE DISCRETIZACION LOCAL
   g) METODOS DE ELEMENTOS FINITOS
   h) TECNICAS DE DISCRETIZACION LOCAL
   i) METODOS DE VOLUMENES FINITOS
   j) ANALISIS NUMERICO DE ESQUEMAS DISCRETOS
   k) RESOLUCION DE ECUACIONES DISCRETIZADAS
   l) APLICACIONES

2. Técnicas específicas aplicables a problemas de mecánica de fluidos y transferencia de calor.

   a) FLUJO INVISCIDO COMPRESIBLE
   b) FLUJO VISCOSO COMPRESIBLE
   c) FLUJO VISCOSO INCOMPRESIBLE
   d) TOPICOS ESPECIALES

La *primera parte* del curso consiste en presentar los principios generales sobre los que se apoyan los modelos físicos que interpretan muchas de las situaciones experimentales en mecánica de fluidos y transferencia de calor. Mediante una visión del material propia de la mecánica del continuo se obtiene posteriormente un modelo matemático que en general consiste de un conjunto de ecuaciones a derivadas parciales con o sin restricciones y con sus respectivos valores de contorno e iniciales que completan su definición. Dada la

complejidad matemática de estos modelos, salvo en situaciones muy particulares en las cuales se pueden obtener soluciones analíticas, requieren de su resolución numérica con lo cual se hace necesario presentar las diferentes técnicas de discretización habitualmente empleadas en problemas de transporte de calor y momento. Debido al diferente carácter de las ecuaciones diferenciales, tanto en su visión continua como en su contraparte discreta y a la presencia de ecuaciones adicionales en los contornos, tambien discretizadas, se requiere un minucioso análisis de los esquemas numéricos empleados previo a su resolución, con el fin de poder interpretar las técnicas numéricas desde el punto de vista de la precisión, la convergencia, la consistencia y la estabilidad. A continuación se aborda el tema de la resolución numérica delsistema algebraico/diferencial de ecuaciones que surge de la discretización empleada. Este tópico tiene alta incidencia en la factibilidad de resolver problemas numéricos ya que de acuerdo al problema en mano y a los recursos computacionales disponibles muestra las diferentes alternativas para su resolución. Esta primera parte finaliza con una serie de aplicaciones de los conceptos adquiridos a la resolución de las ecuaciones de convección difusión tanto en su version estacionaria como transiente, desde el simple caso unidimensional al multidimensional, considerando el caso lineal como el no lineal representado por la ecuación de Bürgers. Este modelo sencillo tiene especial interés dada la similitud que presenta con la estructura de las ecuaciones que conforman la mayoria de los modelos matemáticos mas frecuentemente usados en mecánica de fluidos y transferencia de calor. En esta primera parte del curso se introducirán en forma de trabajos prácticos y cuando la explicación teórica lo requiera algunos ejemplos a resolver tanto analítica como numéricamente. dado que esta parte es introductoria se verán modelos simplificados de aquellos comúnmente empleados en *CFD* pero que contienen muchas de las características matemático/numéricas propias de aquellos y que lo hacen atractivos en pos de ir incorporando conceptos, necesarios para abordar la segunda parte, en forma gradual. Paralelamente con el curso teórico se desarrollarán talleres sobre los aspectos prácticos a cubrir en esta primera parte. Debido a que el enfoque del curso está orientado hacia los fundamentos y el aprendizaje de las técnicas que están implícitas en todo código computacional se hace necesario programar por uno mismo algunas aplicaciones vistas en la sección teórica. Ya que esto difícilmente se encuentra en un paquete comercial y dado que el grado de avance que actualmente existe en el area de software educativo está bastante lejos de poder contar con herramientas aptas para la enseñanza se hace necesario elegir algún entorno que sea ameno para el usuario y potente para el ambicioso plan de aprender métodos numéricos desde cero. En este sentido consideramos que el uso de MatLab puede ser muy beneficioso por varias razones, a saber:

1. cuenta con muchas rutinas de alto nivel y otras de bajo nivel que permite ubicarse muchas veces en diferentes niveles o jerarquias con lo cual cada uno puede optar por el rol que mas le gusta,
2. es un lenguaje de programación, por lo tanto crear rutinas muy específicas,
3. gran y eficiente interacción entre cálculo y gráficos,
4. posibilidad de debugear aplicaciones en forma interactiva.

No obstante, por razones de eficiencia y para cuando la necesidad lo requiera es necesario contar con conocimientos de lenguajes de programación más orientados a simulaciones de gran escala, como por ejemplo el Fortran y el C o C++. Sin entrar en detalles acerca de la programación el curso incluye el manejo de un programa de elementos finitos para la resolución de algunos de los problemas incluidos en la primera parte del curso. Este software será utilizado en la segunda parte del curso para resolver problemas de flujos compresibles e incompresibles que requieren mucha mayor potencia de cálculo.

La *segunda parte* del curso trata acerca de las técnicas específicas empleadas en la resolución de problemas de mecánica de fluidos. Básicamente se tomará en primera instancia el caso de flujo invíscido compresible representado por el modelo de las ecuaciones de Euler y posteriormente se tratará el caso viscoso tanto compresible como incompresible modelado por las ecuaciones de Navier-Stokes. En cada uno de estos capítulos se volcarán los conceptos aprendidos en la primera parte del curso para diseñar y analizar esquemas numéricos que permitan resolver estos casos particulares. Dada la complejidad del problema surgen naturalmente restricciones muy severas en cuanto a la resolución numérica de las ecuaciones lo cual hace necesario explorar técnicas iterativas específicasa tal fin. Como las soluciones numéricas en los problemas de flujos de fluidos son altamente dependiente de la malla se hace necesario introducir nociones básicas sobre generación de mallas en *CFD* . Este tema forma parte del grupo de tópicos especiales. Otro de los temas especiales a tratar es el modelado de la turbulencia. Es bien sabido que la mayoría de los problemas de interés son gobernados por condiciones de flujo turbulento. Se verá a modo de introducción algunos modelos algebraicos típicos en los casos de flujos internos y externos asi como algunos modelos basados en ecuaciones a derivadas parciales como el caso del bien popular método $\kappa - \epsilon$. Finalmente cierra esta sección de tópicos especiales el tratamiento de problemas con dominios variables en el tiempo.

Capítulo 1

Modelos fisícos y matemáticos

En este capítulo se presentan los principios o leyes físicas que gobiernan el flujo de fluidos, las reglas que definen el comportamiento de los materiales involcucrados, los relaciones termodinámicas entre las variables que caracterizan el fenómeno y finalmente los modelos matemáticos conformados por sistemas de ecuaciones diferenciales que serán el punto de partida hacia la búsqueda de soluciones a diversos problemas de mecánica de fluidos y transferencia de calor.

1.1. Conceptos introductorios

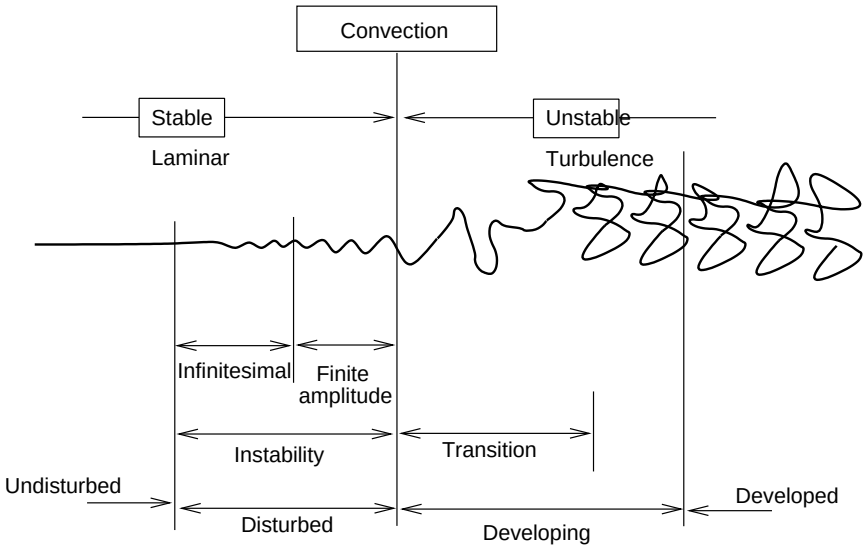
En esta sección mencionaremos algunos conceptos básicos necesarios para conformar el marco teórico para el tratamiento de los problemas a resolver.

1.1.1. Postulado del continuo

Partiendo de una descripción molecular de la materia podemos poner atención en el movimiento de ellas en forma individual o formar un *cluster* que agrupe a muchas de ellas y estudiar el movimiento del mismo. La idea del cluster equivale a una especie de promediación estadística que tiene sentido si las escalas de interés a ser resueltas son mucho mayores que el camino libre medio de las moléculas. Esta visión *fenomenológica* hace que el medio sea interpretado como un continuo a diferencia de la visión *microscópica* que mira al material desde una aproximación a las partículas. Desde la óptica del continuo las variables a resolver se asumen variar en forma continua respecto a las coordenadas espaciales y al tiempo. En la aproximación del continuo la operación de promediación antecede a la aplicación del principios termomecánicos. En la aproximación de partícula se suelen plantear las leyes físicas a la escala microscópica y estudiar los fenómenos que ocurren a esa escala. Si realizáramos la promediación después de aplicar los principios termomecánicos deberímos encontrar el mismo resultado que aquel que surge de la mecánica del continuo. De todos modos esto último no es muy práctico ya que a menudo la información microscópica que se dispone es muy escasa como para poder plantear un modelo a tan pequeña escala para después saltar mediante la promediación a la macroescala, de real interés a los fines ingenieriles.

1.1.2. Tipos de flujo

Compresible e incompresible Un fluido se considera incompresible si su densidad experimenta cambios despreciables frente a cambios apreciables en la temperatura y la presión. Despreciable es un término ambiguo y debe ser interpretado de acuerdo a la experiencia. Por ejemplo si un fluido varía su densidad un $5\,\%$ para un salto térmico de $\Delta T \approx 100^{\circ}\,C$ o en un $1\,\%$ para un salto de presión de $\Delta p \approx 100atm$ uno se inclinaría pensar que el fluido es incompresible. En realidad lo que interesa es el flujo que se establece con total independencia del fluido que lo experimenta. No importa si es agua o aire lo importante es en que medida es la compresibilidad del medio un factor importante por considerar. Un ejemplo es la convección natural en un medio como el agua. Este fenómeno es manejado por diferencia de densidades muchas veces producidas por gradientes térmicos. Si bien la densidad del agua tiene un comportamiento tal que podría pensarse como incompresible, es su compresibilidad la que permite poner en movimiento al sistema formando corrientes convectivas que describen por si mismas el fenómeno. No obstante este problema puede ser tratado como uno de flujo incompresible introduciendo efectos forzantes proporcionales al gradiente térmico mediante una aproximación debida a *Boussinesq*. En general el caso de flujo compresible es reservado para gases a alta velocidad, próximas o superior a las del sonido en donde los fenómenos ondulatorios son muy apreciables. No obstante el agua, un fluido qu a priori podría ser tildado como incompresible, al circular por una cañería y al experimentar el cierre abrupto de una válvula desarrolla ondas de presión que pueden ser analizadas por técnicas de flujos compresible. Resumiendo, la división entre compresible e incompresible debe ser analizada en término del flujo que produce y no del fluido que lo experimenta.



Flujo laminar y turbulento

Laminar y turbulento Mientras que el flujo laminar es caracterizado por un movimiento suave y determinístico de una lámina de fluido sobre otra, el turbulento es un movimiento aleatorio superpuesto sobre un movimiento medio del fluido. El humo de un cigarrillo es un experimento interesante que permite visualizar como el aire alrededor del mismo al calentarse se pone en movimiento de forma laminar. En dirección ascendente pronto se ve como esa corriente ordenada empieza a inestabilizarse formando remolinos de gran

escala que se returcen hasta que el desórden se amplifica y los remolinos se afinan en tamaño y crecen en cantidad retorciéndose más y mas alcanzando un régimen plenamente turbulento. Este fenómeno es visible a nuestros ojos por un efecto similar al utilizado en la técnica Schlieren en el que se hace atravezar rayos luminosos en un medio con densidad variable. De los muchos experimentos que se podrían mencionar acerca de flujos turbulentos no podemos dejar de mencionar aquel que hizo historia y que se debió al propio Reynolds. El pudo observar como el flujo que atravezaba un tubo de sección circular a medida que iba cambiando la velocidad de entrada experimentaba cambios despreciables en su patrón fluidodinámico hasta que al atravezar cierto valor límite se producía un cambio significativo en el régimen fluidodinámico. Ese valor crítico puede ser establecido mediante técnicas de análisis dimensional, que dan cuenta que cuando el número de Reynolds supera un valor próximo a los 2100 el flujo se transiciona y luego se transforma en turbulento. La transición se manifiesta porque se crea un movimiento vorticoso periódico que al crecer el Reynolds se desordena aún mas alcanzando un régimen turbulento. La figura 1.1.2 muestra a modo de esquema los diferentes regímenes que se presentan en un flujo desde uno laminar estable, pasando por inestabilidades hasta uno turbulento. El tema de la inestabilidad de flujos es toda un area de investigación aparte que merece mucha atención y que por razones de complejidad y espacio no será tratada en este curso. Aquellos interesados en el tema pueden recurrir a libros como Batchelor [Ba], Dreizin & Reid [DR] y a los trabajos de Taylor entre otros.

Estacionario y transiente En el caso laminar la diferencia entre un flujo estacionario y otro transiente es obvia, en el primero las variables de interes son independientes del tiempo mientras que en el último $p = p(t)$ y $\mathbf{v} = \mathbf{v}(t)$. Si el flujo es turbulento, por ser este siempre transiente, la diferencia debe establecerse sobre los valores medios, o sea $\overline{p} \neq \overline{p}(t)$ implica que el flujo es estacionario, siendo $\overline{p} = \frac{1}{T}\int_0^T p(t)dt$ el promedio temporal de la presión en un período de duración T.

Unidimensional y multidimensional En el punto anterior tratamos la dependencia o no de las variables dependientes sobre una variable independiente en particular, el tiempo, estableciendo las diferencias entre un movimiento estacionario y otro transiente. Ahora tomaremos otra variable independiente, las coordenadas espaciales y supongamos que las variables dependientes, la presión y la velocidad por ejemplo, sólo dependen de una de las coordenadas espaciales. En ese caso el movimiento es unidimensional siendo esta situación muy ventajosa a la hora de un tratamiento analítico. Lamentablemente estas situaciones difícilmente se encuentren en la realidad, siendo al menos 2D la clase de problemas que merecen atención. En estos casos como en el 3D los problemas deben ser generalmente abordados en forma numérica.

1.1.3. La solución a los problemas de mecánica de fluidos

El formalismo necesario para resolver problemas de mecánica de fluidos requiere de establecer los postulados fundamentales que gobiernan el movimiento de los mismos. Sin entrar en demasiado detalle en este tema podemos decir que la mayoría de los estudiantes aprenden en los cursos universitarios las leyes de Newton del movimiento y la aplican para resolver problemas de estática y dinámica en forma casi natural. Parecería natural que ellas deban incluirse como leyes o postulados fundamentales para tratar problemas de movimiento de fluidos. Sin embargo y desde el punto de vista de la mecánica del continuo son más adecuadas las dos leyes de Euler que dicen:

- 1.- la variación temporal de la cantidad de movimiento de un cuerpo es igual a la fuerza resultante actuando sobre el mismo.

- 2.- la variación temporal del momento de la cantidad de movimiento de un cuerpo es igual al torque resultante actuando sobre el mismo, considerando que tanto el momento angular como el torque son medidos respecto del mismo punto.

La primera ley de Euler es una generalización de la segunda ley de Newton mientras que la segunda ley de Euler es independiente de la primera ya que no solo incluye las fuerzas de volúmen sino también las fuerzas de superficie. Estas dos leyes son conocidas como el *principio del momento lineal (1)* y el *principio del momento angular (2)*. Ambas, junto con los *principios de conservación de la masa y la energía* forman las leyes fundamentales necesarias para definir el modelo físico utilizado en la mayoría de los fenómenos de transporte. Es más, los principios del momento lineal y angular pueden ser considerados como principios de conservación considerando que la variación temporal de la cantidad de movimiento lineal o angular son igualadas por la velocidad a la cual dicha cantidad de movimiento lineal o angular se suminstra al *cuerpo* mediante una fuerza o un torque respectivamente.

En lo anterior la palabra *cuerpo* se utiliza como una cantidad fija de material, un cuerpo siempre contiene la misma masa y algunas veces es referido como sistema. Considerando los 4 principios de conservación anteriores, cantidad de movimiento lineal y angular, masa y energía podemos ver que los dos primeros son principios sobre propiedades vectoriales mientras que los dos últimos son establecidos sobre cantidades escalares. Establecer los principios fundamentales es solo el comienzo de un largo camino en pos de obtener soluciones a problemas de mecánica de los fluidos. A continuación se requiere un detallado análisis matemático para transformar lo establecido por los principios físicos en ecuaciones matemáticas útiles. A posteriori se necesita introducir reglas sobre el comportamiento del material tanto desde un punto de vista mecánico como termodinámico ambas basadas en las observaciones o quizás en algunos casos deducibles de principios o leyes físicas aplicables a escalas mucho mas pequeñas. Finalmente es la intuición la que restringe aún más los modelos en pos de hacerlos tratables.

1.1.4. Unidades

En pos de normalizar el tratamiento tomaremos como unidades aquellas que surgen del sistema internacional de medidas y que se expresan en función de las siguientes magnitudes básicas o primarias:

$$\begin{aligned} M &= \text{masa (Kg)} \\ L &= \text{distancia (m)} \\ t &= \text{tiempo (seg)} \end{aligned} \tag{1.1}$$

con lo cual las cantidades cinemáticas como posición, velocidad y aceleración surgen de combinar distancias y tiempos en distintas potencias. La fuerza, como magnitud dinámica surge de aplicar el principio de conservación de la cantidad de movimiento lineal , entonces

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(M\mathbf{v}) &= \mathbf{F} \\ [=] \frac{1}{\mathrm{seg}}\,\mathrm{Kg}\,\frac{\mathrm{m}}{\mathrm{seg}} &= \mathrm{Newton} \end{aligned} \tag{1.2}$$

1.1.5. Propiedades de los fluidos

Para el caso de flujo incompresible a una fase solo se requiere conocer la densidad y la viscosidad si el fluido es newtoniano. En el caso que el fluido sea no newtoniano o si los efectos compresibles son importantes existen otras magnitudes a tener en cuenta.

Compresibilidad Con dos coeficientes tendremos en cuenta el efecto de la presión y la temperatura sobre la densidad. El primero se define como:

$$\kappa = \frac{1}{\rho}(\frac{\partial \rho}{\partial p})_T$$

mientras que el coeficiente de expansión β se define como:

$$\beta = -\frac{1}{\rho}(\frac{\partial \rho}{\partial T})_p$$

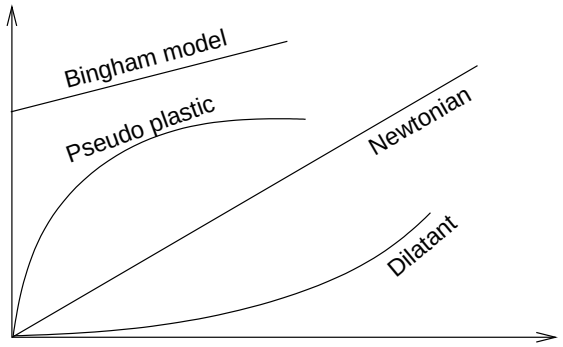
En el caso de los líquidos estos dos coeficientes son en general pequeños, especialmente κ. El coeficiente de expansión puede adquirir cierta importancia en el caso de fenómenos como la convección natural. En el caso de gases, un ejemplo es el caso de los gases ideales cuya ley de comportamiento viene comúnmente expresada por

$$\begin{aligned} \rho &= \frac{p}{RT} \\ \kappa &= \frac{1}{p} \\ \beta &= \frac{1}{T} \end{aligned} \tag{1.3}$$

En el caso de los gases reales el comportamiento es diferente especialmente cerca de los puntos críticos y se debe recurrir a la experiencia para poder determinar las leyes de comportamiento.

Viscosidad A diferencia de los materiales sólidos que ante un esfuerzo sufren una deformación en general independiente del tiempo, los fluidos no soportan determinados esfuerzos y se deforman continuamente. Esto muchas veces está asociado a la fluidez del medio. Un contraejemplo de los dicho en el caso de sólidos es el fenómeno de *creep* o fluencia lenta. En el caso de los sólidos la deformación es la variable cinemática de interés, definida como la variación relativa de la distancia entre dos puntos del cuerpo material. En los fluidos su análogo es la tasa o velocidad de deformación, o sea la variación relativa de la velocidad de dos puntos dentro del volúmen material. Lo que suele ser de interés es establecer cierta relación entre causa y efecto, o sea entre tensión y deformación en mecánica de sólidos o entre tensión y velocidad de deformación en fluidos. Esta funcionalidad puede asumir un rango lineal y luego uno no lineal, dependiendo del material el punto de corte entre uno y otro comportamiento. En sólidos es el módulo de Young el que vincula tensión con deformación mientras que en mecánica de fluidos es la viscosidad la que juega semejante rol. La *reología* es la ciencia que se encarga de establecer relaciones de este tipo válidas para ciertos materiales o fluidos. A su vez una vez establecido el ensayo de laboratorio aparecen los teóricos tratando de traducir los valores experimentales en alguna teoría que los explique. Entre estas teorías una de las más citadas es la de considerar

al fluido como Newtoniano, con eso queremos decir que la viscosidad es independiente del estado de deformación del fluido. La viscosidad puede variar con la posición y el tiempo pero por otras causas, por ejemplo calentamiento, pero no varía por su estado de deformación. Con el viscosímetro se pueden determinar valores para la viscosidad. En el caso más general la viscosidad puede depender del estado de deformación y en este caso el fluido exhibe un comportamiento no newtoniano. La figura 1.1.5 muestra 4 curvas tensión vs velocidad de deformación que muestran el caso newtoniano (recta por el orígen), el flujo de Bingham (recta desfasada del orígen), y dos casos de fluido no newtoniano, el caso dilatante con viscosidad proporcional a la deformación y el caso pseudo-plástico cuando la viscosidad disminuye cuanto más se deforma el fluido.



Comportamiento reológico de los fluidos

Existe un modelo muy usado llamado *modelo de la ley de potencia* o modelo de Ostwald-de Wael el cual trata de unificar el tratamiento definiendo una viscosidad aparente del tipo

$$\mu_{ap} \propto \mu_0 \parallel \frac{\mathrm{d}v_x}{\mathrm{d}y} \parallel^{n-1}$$

con μ_0 una viscosidad de referencia y n una potencia. En este caso el escurrimiento se piensa del tipo flujo paralelo.

Tensión superficial Esta aparece en general cuando existen interfaces entre dos o más fluidos o un fluido y un sólido y a veces suelen ser tan importantes que su omisión en las ecuaciones pone en peligro el realismo de la solución. Esta en general es función de la curvatura de la interface y de algún coeficiente de capilaridad. Su complejidad escapa los alcances de estas notas.

Presión de vapor

En algunas aplicaciones la presión local suele descender demasiado alcanzando la presión de saturación del vapor con lo cual aparece el fenómeno de *cavitación*. Este fenómeno es muy importante en el funcionamiento de máquinas hidráulicas como bombas y turbinas pero su tratamientoe escapa los alcances de estas notas.

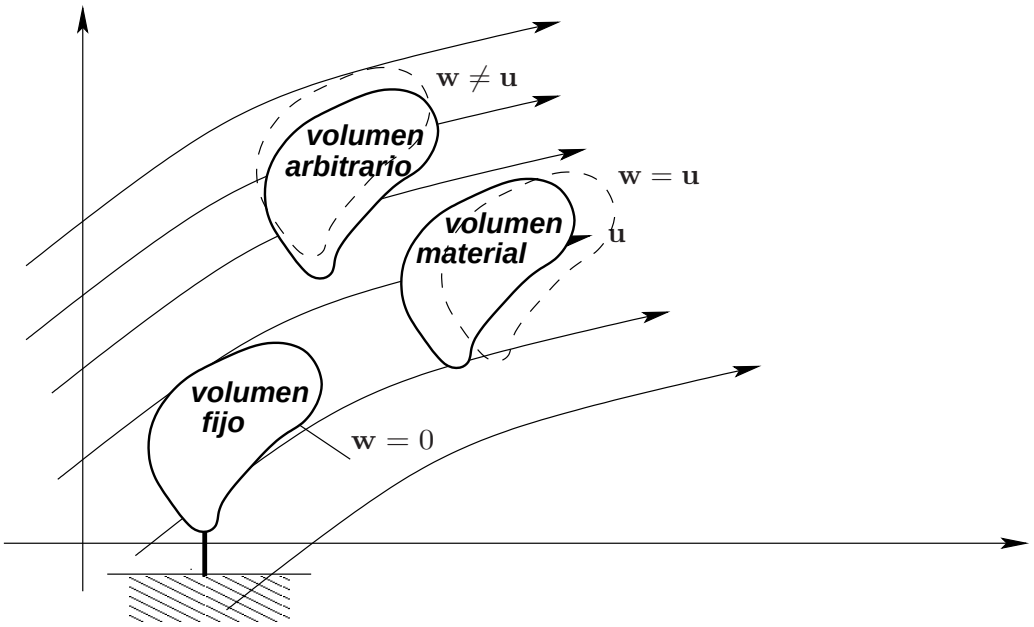
1.2. Cinemática de fluidos

En esta sección se presentan algunas definiciones necesarias para describir el movimiento de los fluidos y se muestran algunas reglas muy útiles que permiten manipular matemáticamente las ecuaciones de conser-

vación, pudiendo expresarlas de varias formas segun el tratamiento que se desee emplear. Empezaremos con la definición de los diferentes volúmenes de control comúnmente empleados en la descripción de las ecuaciones de movimiento y expresaremos en forma matemática el principio de conservación del momento lineal. Posteriormente trabajaremos sobre la cinemática del movimiento, el lado izquierdo de la ecuación, usaremos el teorema de la divergencia para poder transformar integrales de volúmen en integrales de area y viceversa, una ayuda para poder formular los problemas desde el punto de vista integral o diferencial, microscópico o macroscópico y finalmente se presenta el teorema del transporte para poder manipular volúmenes de control variables con el tiempo. De esta forma se arriba a las ecuaciones de conservación.

1.2.1. El volúmen material

Por lo previamente establecido el principio de momento lineal involucra la definición de *cuerpo*, diciendo que la variación temporal de la cantidad de movimiento de un cuerpo es igual a la fuerza resultante actuando sobre el mismo. Como cuerpo queremos decir un sistema con una cantidad fija de material. Debido a que los principios de conservación se aplican a los cuerpos y si consideramos el principio de conservación de la masa entonces se deduce que la masa de un volúmen material es constante. Al basar nuestro análisis en la mecánica del continuo y al asumir que nuestra escala de interés es mucho mayor que la del propio movimiento molecular, entonces, tiene sentido considerar que el volúmen material cambia de forma y posición con el tiempo de una manera continua sin intercambiar masa con el medio ambiente. Designaremos al volúmen material y a su respectiva area material como $\mathcal{V}_m$ y $\mathcal{A}_m$. En breve surgirá la necesidad de trabajar con volúmenes de control fijos en el espacio y a estos como a su respectiva area los designaremos sencillamente por $\mathcal{V}$ y $\mathcal{A}$. Finalmente presentamos una tercera posibilidad, aquella en la que el volúmen de control se mueve pero ya no siguiendo al sistema o cuerpo sino de una manera arbitraria y a estos y su respectiva area la simbolizamos como $\mathcal{V}_a$ y $\mathcal{A}_a$.



Distintas definiciones de volúmenes de control

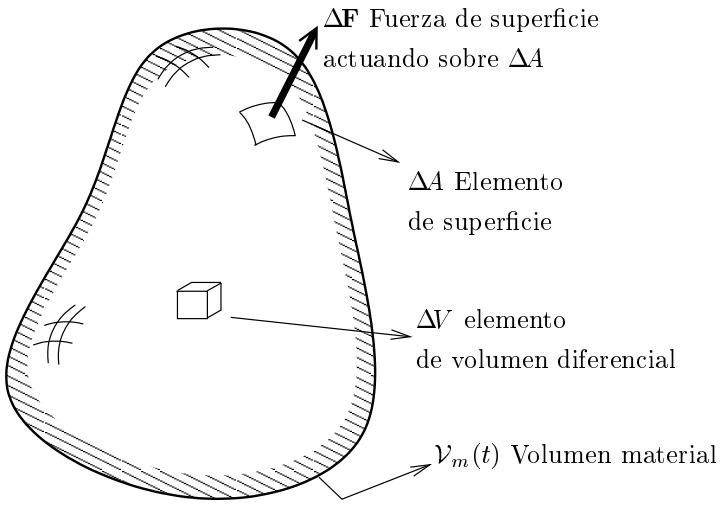
1.2.2. El principio de conservación de la cantidad de movimiento lineal

Consideremos un volúmen diferencial de fluido dV como el mostrado en la figura 1.2.2. La masa contenida en el mismo es $dM = \rho dV$ y la cantidad de movimiento del mismo es $\mathbf{v}dM = \rho \mathbf{v} dV$. Por definición la cantidad de movimiento del volúmen material se define como:

$$\int_{\mathcal{V}_m(t)} \rho \mathbf{v} dV \tag{1.4}$$

con lo cual el principio de la conservación de la cantidad de movimiento lineal puede escribirse como:

$$\frac{D}{Dt} \int_{\mathcal{V}_m(t)} \rho \mathbf{v} dV = \left\{ \begin{array}{c} \text{Fuerza actuando sobre} \\ \text{el volúmen material} \end{array} \right\} \tag{1.5}$$



Volúmen material

La derivada $\frac{D}{Dt}$ es llamada la *derivada material* y en breve será sometida a análisis junto con las otras dos derivadas temporales a considerar, la *derivada total* y la *derivada parcial*. Aquí simplemente lo que queremos indicar es que estamos tomando la variación temporal de una propiedad de un volúmen material. Del lado izquierdo de la anterior ecuación participa la cinemática del movimiento mientras que del derecho se hallan las causas del movimiento o fuerzas externas aplicadas al cuerpo. Estas pueden ser:

- a.- fuerzas de cuerpo o volúmen
- b.- fuerzas de superficie.

Mientras que las primeras actúan sobre la masa del sistema (fuerzas gravitatorias, electrostáticas, etc) las otras lo hacen sobre el contorno del sistema. Introduciendo estas dos fuerzas en la expresión anterior arribamos al principio de conservación de la cantidad de movimiento lineal, expresado como:

$$\frac{D}{Dt} \int_{\mathcal{V}_m(t)} \rho \mathbf{v} dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g} dV + \int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})} dA \tag{1.6}$$

donde $\mathbf{g}$ es la fuerza de cuerpo por unidad de masa y $\mathbf{t}_{(\mathbf{n})}$ es el vector tensión en el contorno.

Casos simples

En la mayoría de los cursos de mecánica de los fluidos de la carrera de de Ingeniería se hace especial énfasis entre otros temas a la resolución de problemas asociados con la estática de fluidos y el flujo en tubos y canales.

La estática de fluidos

El primer caso corresponde al caso particular de un flujo en reposo (estacionario) donde el término izquierdo de la expresión (1.6) se anula quedando una igualdad del tipo:

$$0 = \int_{\mathcal{V}_m(t)} \rho \mathbf{g} dV + \int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})} dA \tag{1.7}$$

Si además se acepta la definición que dice: *un fluido se deformará continuamente bajo la aplicación de un esfuerzo de corte* entonces, las únicas fuerzas de superficie posibles deben actuar en forma normal a la misma. Además se puede probar que el tensor de tensiones asociado a un fluido en reposo es isotrópico y por lo tanto permanecerá invariante con la dirección. Con todo esto las ecuaciones se simplifican demasiado y si asumimos cierta continuidad en los integrandos podemos cambiar a la forma diferencial del problema y arribar a la bien conocida expresión:

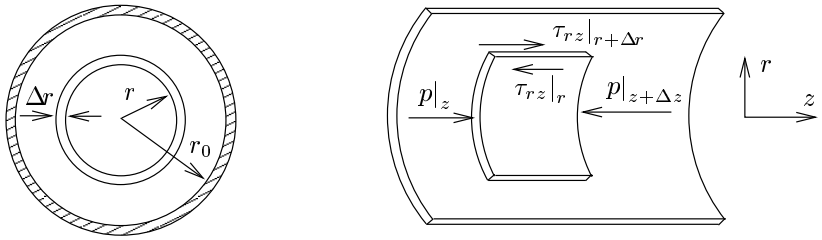
$$\begin{aligned} 0 &= \rho \mathbf{g} - \nabla p \\ 0 &= \rho g_i - \frac{\partial p}{\partial x_i} \\ \nabla &= \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z} \end{aligned} \tag{1.8}$$

donde hemos usado notación de Gibbs en la primera línea, notación indicial en la segunda y la definición del operador gradiente en la tercera. Esta expresiíon es muy familiar para los estudiantes de ingeniería ya que muchos problemas clásicos se resuelven con ella, a saber:

1. a.-barómetros
2. b.-manómetros
3. c.-fuerzas sobre cuerpos planos sumergidos
4. d.-fuerzas sobre cuerpos curvos sumergidos
5. e.-fuerzas de flotación
6. f.-hidrómetros, etc

En la parte I de la práctica 1 se presentan algunos problemas opcionales a resolver para aquellos que quieran ejercitarse sobre temas que no son centrales al curso pero que son necesarios conocer para poder analizar problemas de mecánica de fluidos. Opcionales significa que aquellos que se sientan confiados en que conocen la forma de resolverlos no están obligados a hacerlo.

Flujo laminar unidimensional



Flujo laminar unidimensional, definición del volúmen material

En este problema el fluido no esta en reposo pero si consideramos el caso de flujo laminar y líneas de corriente rectas entonces nuevamente el miembro izquierdo de (1.6) se anula. Esto tiene su explicación si tomamos un volúmen material como el de la figura 1.2.2,
allí su cantidad de movimiento es constante ya que la densidad es constante por tratarse de un flujo incompresible y la velocidad en ese volúmen de control es constante ya que la misma depende solamente del radio. Un ejemplo de flujo unidimensional que no tiene líneas de corriente rectas y por ende no anula el miembro izquierdo es el de flujo Couette entre dos cilíndricos concéntrico de longitud infinita. La razón es que allí existe una aceleración centrípeta necesaria para mantener el flujo en un movimiento circular. Nuevamente tomamos la expresión (1.7) pero en este caso por estar el fluido en movimiento no podemos despreciar los esfuerzos cortantes. Por lo tanto las fuerzas de superficie actuarán en la dirección normal (la presión) y en la dirección tangencial (la tensión de corte). Por la forma que tiene el volúmen material y debido a que el movimiento tiene una sola componente de velocidad según la dirección z entonces las fuerzas normales a las superficies solo pueden actuar sobre aquellas superficies cuya normal está alineada con el eje z. Como no existe flujo en la sección transversal del tubo los esfuerzos de corte en los mismos es nulo y solo puede haber esfuerzos de corte en los planos cuya normal coincide con la dirección radial como se muestra en la figura. Por lo tanto, si por el momento no consideramos las fuerzas de cuerpo la expresión (1.7) queda

$$0 = [(p2\pi r\Delta r)_z - (p2\pi r\Delta r)_{z+\Delta z}] + [-(\tau_{rz}2\pi r\Delta z)_r - (\tau_{rz}2\pi r\Delta z)_{r+\Delta r}] \tag{1.9}$$

Haciendo un poco de álgebra y tomando límites cuando $\Delta r \to 0$ y $\Delta z \to 0$ conduce a la siguiente expresión diferencial:

$$0 = -\frac{\partial p}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(r\tau_{rz}) \tag{1.10}$$

Esta ecuación se la conoce con el nombre de ecuación de tensión del movimiento porque viene expresada en términos de las componentes del tensor de tensiones. Si queremos expresar esta ecuación en término de las variables cinemáticas debemos usar alguna relación constitutiva. En este caso recurrimos a la ley de Newton de la viscosidad en la cual

$$\tau_{rz} = \mu\frac{\partial v_z}{\partial r} \tag{1.11}$$

con lo cual si consideramos que la viscosidad es constante la (1.10) se transforma en

$$\frac{\partial p}{\partial z} = \mu\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial v_z}{\partial r}) \tag{1.12}$$

Este flujo es denominado *flujo plano Couette* y representa uno de los casos simples con solución analítica y que ha tenido gran interés desde el punto de vista de las aplicaciones. Nosotros aquí solo lo hemos planteado, dejamos la resolución del mismo como parte de los trabajos prácticos.

Cinemática de fluidos en los casos generales

En la sección anterior nos volcamos hacia la resolución de dos casos muy particulares que permiten independizarse completamente del miembro izquierdo de la expresión (1.6) del principio de conservación de la cantidad de movimiento lineal . Ahora nos detendremos a analizar las variables que componen este miembro izquierdo de forma de adquirir las nociones elementales necesarias para su tratamiento. Este término expresa la cinemática del movimiento y como tal incluye la variación temporal de propiedades que se hallan distribuidas en el espacio de alguna manera continua. Por lo tanto aparece la necesidad de tratar a las dos variables independientes más importantes que incluyen los principios de conservación, las coordenadas espaciales y el tiempo. El objetivo está en conducir por una lados hacia la formulación diferencial de los principios de conservación con el propósito de analizarlos desde un punto de vista macroscópico local y por otro manipular la formulación integral para poder llevar a cabo balance

s macroscoópico globales, muy útiles por varias razones. Estos balances macroscópico globales permiten chequear soluciones obtenidas numéricamente mediante balances locales y por otro han sido de amplia difusión en la profesión del ingeniero para el cálculo y diseño de equipos e instalaciones.

Coordenadas espaciales y materiales

En esta sección pretendemos desarrollar las nociones básicas sobre lo que se entiende por *coordenadas materiales* y su relación con las *coordenadas espaciales* en pos de describir adecuadamente el movimiento de un fluido. El término coordenadas espaciales se refiere a un sistema de coordenadas fijo donde todos los puntos del espacio pueden ser localizados. Hay dos formas posibles de localizar o identificar una partícula de fluido que pertenece a un volúmen material diferencial $d\mathcal{V}_m(t)$. Una forma sería designar su posición mediante sus coordenadas espaciales x, y, z. Asumiendo que a un dado tiempo de referencia $t = 0$ las coordenadas espaciales se hallan localizadas en

$$x = X\ , \qquad y = Y\ , \qquad z = Z\ , \qquad t = 0 \tag{1.13}$$

Para un tiempo $t > 0$ la posición se puede expresar mediante

$$\begin{aligned} x &= X + \int_0^t \Big(\frac{\mathrm{d}x}{\mathrm{d}t}\Big) dt \\ y &= Y + \int_0^t \Big(\frac{\mathrm{d}y}{\mathrm{d}t}\Big) dt \\ z &= Z + \int_0^t \Big(\frac{\mathrm{d}z}{\mathrm{d}t}\Big) dt \end{aligned} \tag{1.14}$$

Compactando las tres expresiones anteriores en una sola mediante

$$\mathbf{r} = \mathbf{R} + \int_0^t \Big(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t}\Big) dt \tag{1.15}$$

entonces, $\mathbf{r}$ representa el vector posición espacial mientras que $\mathbf{R}$ es llamado el vector posición material. Este último identifica una partícula del sistema o cuerpo y en algun sentido la impone una marca al tiempo de referencia. Este conjunto específico de coordenadas no representa ningún sistema de coordenadas que

se mueve y se deforma con el cuerpo. De alguna manera las coordenadas espaciales se pueden expresar en función de las coordenadas materiales y el tiempo,

$$\mathbf{r} = \mathbf{r}(\mathbf{R}, t) \tag{1.16}$$

Una descripción *Lagrangiana* del movimiento es aquella expresada en término de las coordenadas materiales mientras que una descripción *Euleriana* se expresa según las coordenadas espaciales.

La derivada temporal del vector posición espacial para una partícula de fluido en particular es la velocidad de la misma. Ya que la derivada se evalúa con las coordenadas materiales fijas esta derivada es llamada *derivada material*,

$$(\frac{\mathrm{d}\mathbf{r}}{\mathrm{d}t})_{\mathbf{R}} = \frac{D\mathbf{r}}{Dt} = \mathbf{v} \tag{1.17}$$

Derivadas temporales

Sea $S = S(x, y, z, t)$ una función escalar, entonces su derivada temporal se define como:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \lim_{\Delta t \to 0} \Big[\frac{S(t + \Delta t) - S(t)}{\Delta t} \Big] \tag{1.18}$$

Si S fuera solo función del tiempo esta definición es directa mientras que si S depende de las coordenadas espaciales existe una ambiguedad respecto al punto que se toma en el instante t y en $t + \Delta t$. Para comenzar consideremos el movimiento de una partícula p de un sistema o cuerpo, acorde a (1.18) tenemos que la componente x de la velocidad de la misma será:

$$\frac{\mathrm{d}x_p}{\mathrm{d}t} = \lim_{\Delta t \to 0} \Big[\frac{x_p(t + \Delta t) - x_p(t)}{\Delta t} \Big] = v_x \tag{1.19}$$

Imaginemos que la medición la llevamos a cabo con un observador montado sobre la partícula, entonces para el observador las coordenadas materiales no cambian y por ende (1.18) se vuelve

$$\frac{Dx_p}{Dt} = (\frac{\mathrm{d}x_p}{\mathrm{d}t})_{\mathbf{R}} = \lim_{\Delta t \to 0} \Big[\frac{x_p(t + \Delta t) - x_p(t)}{\Delta t} \Big]_{\mathbf{R}} \tag{1.20}$$

Supongamos que queremos medir la temperatura de un flujo y como es habitual la medición la llevamos a cabo en una ubicación fija del laboratorio, entonces lo que medimos como derivada temporal de la temperatura es:

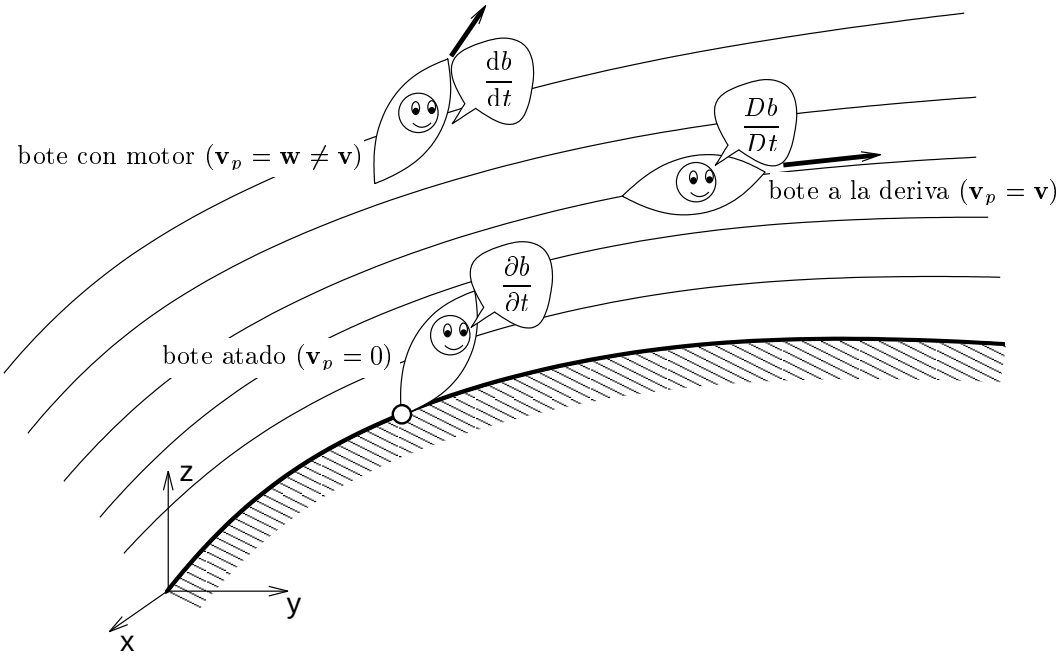
$$\frac{\partial T}{\partial t} = (\frac{\mathrm{d}T}{\mathrm{d}t})_{\mathbf{r}} = \lim_{\Delta t \to 0} \Big[\frac{T(t + \Delta t) - T(t)}{\Delta t} \Big]_{\mathbf{r}} \tag{1.21}$$

y a esta derivada se la suele llamar *derivada parcial* efectuada fijando las coordenadas espaciales del punto de medición en lugar de las coordenadas materiales como en (1.20).

Un tercer caso sería el de efectuar la medición montado sobre un dispositivo que se mueva de una forma arbitraria no manteniendo ni las coordenadas espaciales ni las materiales fijas. Esta derivada se la llama *derivada total* asumiendo que la velocidad del sistema de medición es $\mathbf{w} \neq \mathbf{v}$.

La figura 1.2.2 muestra una descripción de lo que acabamos de presentar como las tres derivadas temporales a utilizar en el resto del curso.

Para interpretar lo anterior pero desde un punto de vista matemático asumamos que tenemos cierta función escalar como la temperatura definida como:



Definición de las derivadas temporales

$$T = T(\mathbf{r}, t) \tag{1.22}$$

Si las coordenadas materiales se mantiene fijas, entonces podemos expresar las coordenadas espaciales en términos de las materiales y el tiempo como

$$\begin{aligned} T = T(\mathbf{r}, t) = T[\mathbf{r}(\mathbf{R}, t), t)] = \\ = T[x(\mathbf{R}, t), y(\mathbf{R}, t), z(\mathbf{R}, t), t] \end{aligned} \tag{1.23}$$

manteniendo $\mathbf{R}$ constante y diferenciando

$$(\frac{\mathrm{d}T}{\mathrm{d}t})_{\mathbf{R}} = \frac{DT}{Dt} = (\frac{\partial T}{\partial x})(\frac{\mathrm{d}x}{\mathrm{d}t})_{\mathbf{R}} + (\frac{\partial T}{\partial y})(\frac{\mathrm{d}y}{\mathrm{d}t})_{\mathbf{R}} + (\frac{\partial T}{\partial z})(\frac{\mathrm{d}z}{\mathrm{d}t})_{\mathbf{R}} + (\frac{\mathrm{d}T}{\mathrm{d}t})_{\mathbf{r}} \tag{1.24}$$

Por definición las derivadas de las coordenadas espaciales mantieniendo las coordenadas materiales $\mathbf{R}$ fijas son las componentes de la velocidad del fluido $\mathbf{v}$, mientras que la derivada manteniendo las coordenadas espaciales $\mathbf{r}$ fijas es la derivada parcial, entonces (1.24) se transforma en:

$$\frac{DT}{Dt} = (\frac{\partial T}{\partial x})v_x + (\frac{\partial T}{\partial y})v_y + (\frac{\partial T}{\partial z})v_z + (\frac{\partial T}{\partial t}) \tag{1.25}$$

mientras que en notación de Gibbs es

$$\frac{DT}{Dt} = (\frac{\partial T}{\partial t}) + \mathbf{v} \cdot \nabla T, \tag{1.26}$$

y en notación indicial:

$$\frac{DT}{Dt} = (\frac{\partial T}{\partial t}) + v_j(\frac{\partial T}{\partial x_j}). \tag{1.27}$$

En el caso en que la medición la efectuemos sobre un dispositivo que tiene su propio movimiento entonces ya las coordenadas materiales no son fijas y las derivadas totales de las coordenadas espaciales respecto al tiempo representan las componentes de la velocidad del dispositivo $\mathbf{w}$,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = (\frac{\partial T}{\partial x})w_x + (\frac{\partial T}{\partial y})w_y + (\frac{\partial T}{\partial z})w_z + (\frac{\partial T}{\partial t}) \tag{1.28}$$

en notación de Gibbs:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = (\frac{\partial T}{\partial t}) + \mathbf{w} \cdot \nabla T \tag{1.29}$$

y en notación indicial:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = (\frac{\partial T}{\partial t}) + w_j(\frac{\partial T}{\partial x_j}) \tag{1.30}$$

Como vemos comparando (1.21) e (1.27) con (1.30) vemos que esta última es la más general ya que (1.21) corresponde al caso $\mathbf{w} = \mathbf{0}$ mientras que como (1.27) sería cuando $\mathbf{w} = \mathbf{v}$.

Hasta aquí hemos analizado el caso de una función escalar y en particular hemos mencionado por razones de índole práctica el caso de la temperatura. A continuación veremos que sucede cuando la función a derivar es una cantidad vectorial. Tomemos como ejemplo el caso del vector velocidad cuya derivada temporal representa el vector aceleración. Por definición,

$$\mathbf{a} = (\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t})_{\mathbf{R}} = \frac{D\mathbf{v}}{Dt} \tag{1.31}$$

Haciendo el mismo análisis que con el caso escalar arribamos a que para un observador con coordenadas espaciales fijas este mide como aceleración

$$(\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t})_{\mathbf{r}} = \frac{\partial \mathbf{v}}{\partial t} \tag{1.32}$$

siendo la relación entre ambas

$$\mathbf{a} = \frac{D\mathbf{v}}{Dt} = (\frac{\partial \mathbf{v}}{\partial t}) + \mathbf{v} \cdot \nabla \mathbf{v} \tag{1.33}$$

en notación indicial

$$\frac{Dv_i}{Dt} = (\frac{\partial v_i}{\partial t}) + v_j(\frac{\partial v_i}{\partial x_j}) \tag{1.34}$$

Aquí vemos que la aceleración consiste de dos términos, uno es llamado la *aceleración local* y representa la variación temporal de la velocidad en un punto fijo en el espacio. La segunda es llamada la *aceleración convectiva* y depende tanto de la magnitud de la velocidad como de su gradiente. De la expresión anterior surge que aunque el flujo sea estacionario la aceleración puede no ser nula dependiendo de su componente convectiva. Un ejemplo de esto lo vemos en el caso del flujo entrando a un tubo desde un depósito. Hasta que se desarrolla el flujo experimenta una aceleración debido al término convectivo aun cuando para un observador ubicado justo enfrente de dicha entrada las condiciones parecen no variar. Sin embargo otro observador viajando con el fluido experimentará la aceleración convectiva.

Teorema de la divergencia

Esta herramienta matemática es muy útil en la discusión que sigue a continuación en este capítulo. Con ella podemos formular balances macroscópicos o desarrollar ecuaciones diferenciales a partir de los principios de conservación expresados en forma integral como por ejemplo el (1.6) .

Este teorema, conocido por aquellos alumnos que han tomado un curso de Cálculo de varias variables se puede expresar de la siguiente forma:

$$\int_{\mathcal{V}} \nabla \cdot \mathbf{G} dV = \int_{\mathcal{A}} \mathbf{G} \cdot \mathbf{n} dA \tag{1.35}$$

Nuestro objetivo no es mostrar una demostración del mismo, solamente presentarlo y tratar de aplicarlo en las secciones que siguen a esta. Para poder aplicarlo al caso de un campo escalar expresemos el campo vectorial anterior como $\mathbf{G} = S\,\mathbf{b}$ donde S es un escalar y $\mathbf{b}$ es un vector constante. Entonces se puede demostrar que

$$\int_{\mathcal{V}} \nabla S dV = \int_{\mathcal{A}} S \mathbf{n} dA \tag{1.36}$$

Teorema del transporte

Hasta el momento hemos presentado diferentes formas de derivar temporalmente una función tanto escalar como vectorial y a su vez hemos mencionado las dos formas de localizar un punto del cuerpo, mediante sus coordenadas espaciales o sus coordenadas materiales.

El objetivo de esta sección es desarrollar una expresión general para la derivada temporal de una integral de volúmen bajo condiciones tales que los puntos que pertenecen a la superficie del volúmen se mueven con una velocidad arbitraria $\mathbf{w}$. De acuerdo a lo ya presentado este volúmen arbitrario lo hemos denominado como $\mathcal{V}_a(t)$. Si la velocidad del mismo la fijamos igual a la velocidad de fluido $\mathbf{v}$ entonces el volúmen se transforma en un volúmen material $\mathcal{V}_m(t)$ y bajo estas condiciones nos referiremos al *teorema del transporte de Reynolds*.

Considerando el volúmen arbitrario $\mathcal{V}_a(t)$ como aquel ilustrado en la parte izquierda de la figura 1.1. Deseamos determinar la integral de volúmen de una cantidad escalar S, a saber:

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{V}_a(t)} S dV = \lim_{\Delta t \to 0} \left[\frac{\int_{\mathcal{V}_a(t+\Delta t)} S(t+\Delta t) dV - \int_{\mathcal{V}_a(t)} S(t) dV}{\Delta t} \right] \tag{1.37}$$

Para visualizar el teorema debemos pensar que el volúmen bajo consideración se mueve a través del espacio de forma tal que los puntos de su superficie se mueven con velocidad $\mathbf{w}$. Esta velocidad puede variar con el tiempo (aceleración) y con las coordenadas espaciales (deformación). En cada instante de tiempo la integral es evaluada y deseamos medir cual es la variación de esta medición. Observando la figura vemos que en un instante Δt el nuevo volúmen barrido por la superficie móvil es designado $d\mathcal{V}_{aII}$ mientras que el viejo volúmen dejado atrás por su movimiento se designa por $d\mathcal{V}_{aI}$.

El area de este volúmen arbitario $\mathcal{A}_a(t)$ se puede dividir en dos partes: $\mathcal{A}_{aI}(t)$ y $\mathcal{A}_{aII}(t)$ mediante una curva cerrada sobre la superficie tal que $\mathbf{n} \cdot \mathbf{w} = 0$. Sin entrar en demasiados detalles para su demostración, suponiendo que tal curva existe, entonces

$$\mathcal{V}_a(t+\Delta t) = \mathcal{V}_a(t) + \mathcal{V}_{aII}(\Delta t) - \mathcal{V}_{aI}(\Delta t) \tag{1.38}$$

lo cual nos permite escribir la integral en (1.37) como

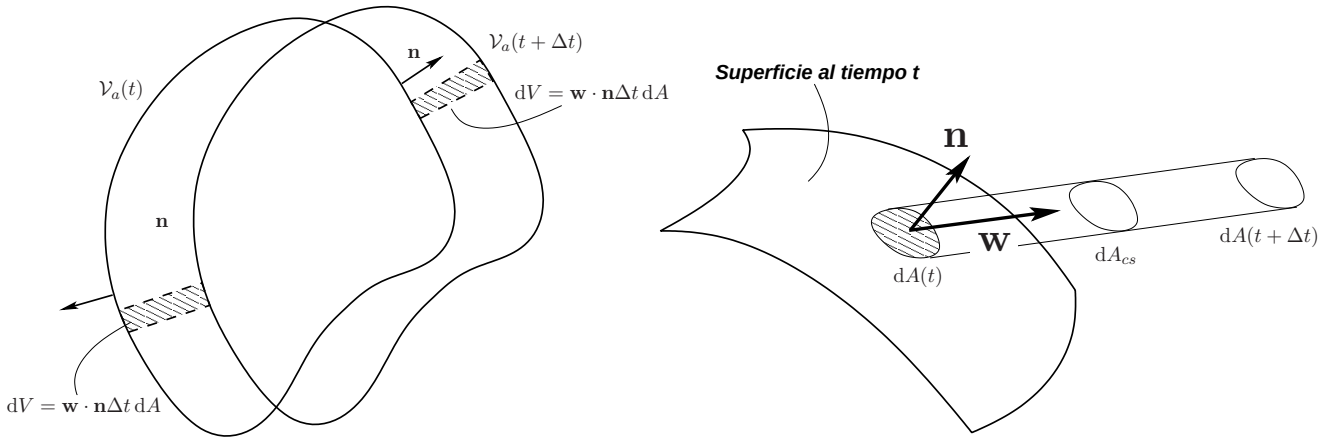


Figura 1.1: Teorema del transporte

$$\int_{\mathcal{V}_a(t+\Delta t)} S(t+\Delta t)dV = \int_{\mathcal{V}_a(t)} S(t+\Delta t)dV + \int_{\mathcal{V}_{aII}(\Delta t)} S(t+\Delta t)dV_{II} - \int_{\mathcal{V}_{aI}(\Delta t)} S(t+\Delta t)dV_I \quad (1.39)$$

que reemplazada en (1.37) produce

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} SdV = \lim_{\Delta t\to 0}\int_{\mathcal{V}_a(t)}[\frac{(S(t+\Delta t)-S(t))}{\Delta t}]dV + \\ \lim_{\Delta t\to 0}\Big[\frac{\int_{\mathcal{V}_{aII}(\Delta t)} S(t+\Delta t)dV_{II} - \int_{\mathcal{V}_{aI}(\Delta t)} S(t+\Delta t)dV_I}{\Delta t}\Big]\end{aligned} \quad (1.40)$$

Cambiando el orden de integración con el proceso límite en el primer término obtenemos

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} SdV = \int_{\mathcal{V}_a(t)}(\frac{\partial S}{\partial t})dV + \\ \lim_{\Delta t\to 0}\Big[\frac{\int_{\mathcal{V}_{aII}(\Delta t)} S(t+\Delta t)dV_{II} - \int_{\mathcal{V}_{aI}(\Delta t)} S(t+\Delta t)dV_I}{\Delta t}\Big]\end{aligned} \quad (1.41)$$

A continuación analizamos de que forma cambiar las integrales de volúmen por integrales de superficie.

Para ello consideremos nuevamente la figura 1.1 que en su parte derecha analiza lo que ocurre localizadamente sobre un diferencial de superficie del volúmen que se está moviendo. Este diferencial de volúmen al moverse una distancia L barre un cilindro oblicuo con respecto a la normal a la superficie siendo esta distancia

$$L = w\Delta t \quad (1.42)$$

con w la magnitud del vector $\mathbf{w}$ que define la velocidad con que se mueve el volúmen arbitrario, siendo su versor dirección representado en la figura mediante $\boldsymbol{\lambda}$, entonces

$$\mathbf{w} = w\boldsymbol{\lambda} \quad (1.43)$$

El volúmen barrido es

$$\mathrm{d}V = L\,\mathrm{d}A_{cs} \tag{1.44}$$

La relación entre la orientación del diferencial de area sobre la superficie con aquel generado por el barrido es:

$$\begin{aligned} a &= b \\ a &= \boldsymbol{\infty} \end{aligned} \tag{1.45}$$

$$\begin{aligned} \mathrm{d}A_{cs} &= \pm\cos(\theta)\,\mathrm{d}A \\ \cos(\theta) &= \boldsymbol{\lambda}\cdot\mathbf{n} \end{aligned} \tag{1.46}$$

Entonces

$$\mathrm{d}V = \pm\mathbf{w}\cdot\mathbf{n}\Delta t dA \tag{1.47}$$

y aplicando este resultado a los dos volúmenes que aparecen en la expresión (1.41) obtenemos lo siguiente:

$$\int_{\mathcal{V}_{aI}(\Delta t)} S(t+\Delta t)dV_I = -\Delta t\int_{\mathcal{A}_{aI}(t)} S(t+\Delta t)\mathbf{w}\cdot\mathbf{n}dA_I \quad \int_{\mathcal{V}_{aII}(\Delta t)} S(t+\Delta t)dV_{II} \quad = +\Delta t\int_{\mathcal{A}_{aII}(t)} S(t+\Delta t)\mathbf{w}\cdot\mathbf{n} \tag{1.48}$$

cuya sustitución en (1.41) produce

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} SdV = \int_{\mathcal{V}_a(t)}(\frac{\partial S}{\partial t})dV + \\ &\lim_{\Delta t\to 0}\Big[\int_{\mathcal{A}_{aII}(t)} S(t+\Delta t)\mathbf{w}\cdot\mathbf{n}dA_{II} + \int_{\mathcal{A}_{aI}(t)} S(t+\Delta t)\mathbf{w}\cdot\mathbf{n}dA_I\Big] \end{aligned} \tag{1.49}$$

Tomando el límite vemos que $\mathcal{A}_{aI}(t)+\mathcal{A}_{aII}(t)=\mathcal{A}_a(t)$ obteniendo finalmente el **Teorema general del transporte**:

$$\boxed{\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} SdV = \int_{\mathcal{V}_a(t)}(\frac{\partial S}{\partial t})dV + \int_{\mathcal{A}_a(t)} S\mathbf{w}\cdot\mathbf{n}dA} \tag{1.50}$$

En el caso de un volúmen fijo en el espacio tenemos que $\mathcal{V}_a(t)=\mathcal{V}$ y $\mathbf{w}=0$ con lo que lo anterior se simplifa a

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}} SdV = \int_{\mathcal{V}}(\frac{\partial S}{\partial t})dV \tag{1.51}$$

El uso más frecuente del teorema general del transporte es cuando se lo aplica a un volúmen material en cuyo caso este resultado se lo conoce como el *Teorema del transporte de Reynolds*, cuya expresión viene dada como:

$$\boxed{\frac{D}{Dt}\int_{\mathcal{V}_m(t)} SdV = \int_{\mathcal{V}_m(t)}(\frac{\partial S}{\partial t})dV + \int_{\mathcal{A}_m(t)} S\mathbf{v}\cdot\mathbf{n}dA} \tag{1.52}$$

La extensión de los resultados anteriores al caso de una función vectorial es directa, simplemente podemos aplicarlo a cada componente, luego multiplicarlo por su respectivo versor dirección y luego sumar, con lo que (1.50) aplicado por ejemplo al campo de velocidades produce el siguiente resultado:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} \mathbf{v}dV = \int_{\mathcal{V}_a(t)} (\frac{\partial \mathbf{v}}{\partial t})dV + \int_{\mathcal{A}_a(t)} \mathbf{v}\mathbf{w}\cdot\mathbf{n}dA \tag{1.53}$$

Conservación de la masa

Hasta este punto hemos tratado de obtener las herramientas necesarias para manipular el miembro izquierdo del principio de conservación de la cantidad de movimiento lineal con lo cual en pos de generalizar su uso deberíamos a esta altura completarlo con el manejo del miembro derecho, aquel representado por las fuerzas de volúmen y las de superficie. Antes de ello y en pos de ir conduciendo al estudiante hacia la utilización más importante de todos los conceptos hasta aquí presentados vamos a considerar el caso particular de la conservación de la masa, ya que este no presenta miembro derecho. Definamos la propiedad a medir, la masa de un volúmen material como:

$$M = \int_{\mathcal{V}_m} (t)\rho dV \tag{1.54}$$

Ya que el principio de conservación requiere que la misma se mantenga constante, entonces:

$$(\frac{\mathrm{d}M}{\mathrm{d}t})_{\mathbf{R}} = \frac{DM}{Dt} = \frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho dV = 0 \tag{1.55}$$

Aplicando el teorema del transporte a la derivada material de la integral obtenemos:

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho dV = \int_{\mathcal{V}_m(t)} \frac{\partial \rho}{\partial t}dV + \int_{\mathcal{A}_m(t)} \rho\mathbf{v}\cdot\mathbf{n}dA = 0 \tag{1.56}$$

que surge de reemplazar $S = \rho$ en (1.52) .

De esta forma fue posible introducir el operador derivada dentro de la integral. Ahora a continuación el teorema de la divergencia nos ayudará para transformar la integral de area en una de volúmen con el fin de poder juntar todos los términos bajo un mismo signo integral y de esta forma pasar de la formulación integral a una diferencial. Esto produce finalmente:

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho dV = \int_{\mathcal{V}_m(t)} \Big[\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v})\Big]dV = 0 \tag{1.57}$$

Asumiendo continuidad en los integrandos podemos deshacernos de la integral y obtener la tan ansiada forma diferencial del principio de conservación de la masa

$$\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) = 0 \tag{1.58}$$

Existen algunas otras formas de expresar el mismo principio de conservación, como por ejemplo si aplicamos la definición de derivada material y alguna manipulación algebraica básica llegamos a:

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0 \tag{1.59}$$

Otra forma particular de la ecuación de continuidad se obtiene si consideramos el caso de un flujo incompresible. Como la densidad es constante de (1.59) surge que

$$\nabla \cdot \mathbf{v} = 0 \tag{1.60}$$

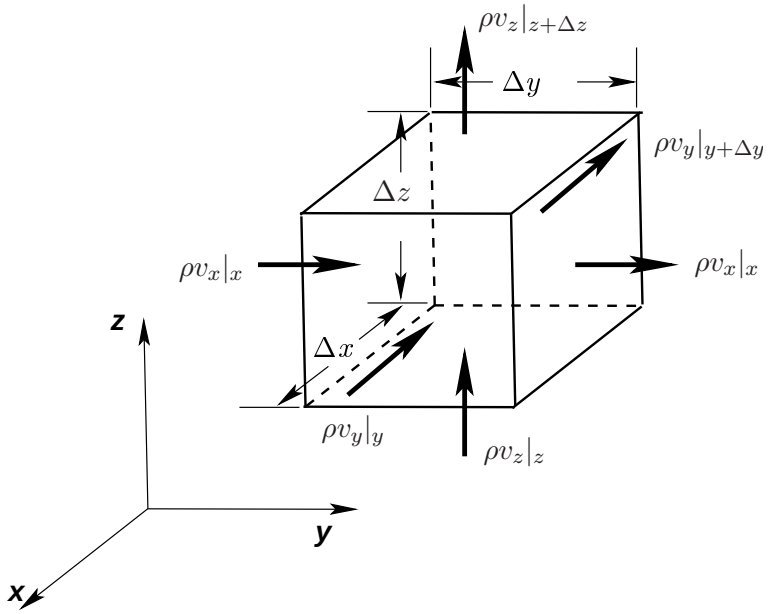
Una forma particular del teorema del transporte de Reynolds se puede lograr si expresamos la propiedad escalar S como producto de la densidad por su propiedad específica,

$$S = \rho s \tag{1.61}$$

En ese caso aplicando todo lo visto hasta aquí se puede arribar a la siguiente igualdad, llamada **forma especial del teorema del transporte de Reynolds**

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho s dV = \int_{\mathcal{V}_m(t)} \rho \frac{Ds}{Dt} dV \tag{1.62}$$

Una forma alternativa de obtener la ecuación de continuidad mucho más práctica y con menos manipuleo matemático es posible mediante el concepto de balance de flujos.



Balance de masa

En general para un diferencial de volúmen como el mostrado en la figura 1.2.2 podemos expresar el siguiente principio de conservación de la masa de la siguiente forma:

$$\left\{\begin{array}{c}\text{variación temporal de}\\ \text{la masa del volúmen control}\end{array}\right\} = \left\{\begin{array}{c}\text{Flujo másico entrante}\\ \text{al volúmen material}\end{array}\right\} - \left\{\begin{array}{c}\text{Flujo másico saliente}\\ \text{del volúmen material}\end{array}\right\} \tag{1.63}$$

El término *flujo* es muy frecuentemente usado en referencia al transporte de masa, momento y energía y significa una especie de caudal de alguna propiedad en la unidad de tiempo. Aplicando el caudal másico,

por tratarse en este caso del balance de masa, a cada cara del cubo elemental y asumiendo una variación continua de las propiedades encontramos que:

$$\frac{\partial}{\partial t}(\rho\Delta x\Delta y\Delta z) = \\ = \underbrace{[-\rho v_x|_{x+\Delta x} + \rho v_x|_x]\Delta y\Delta z}_{\text{caudal másico a través de la superficie orientada según } x} + \underbrace{[-\rho v_y|_{y+\Delta y} + \rho v_y|_y]\Delta x\Delta z}_{\text{caudal másico a través de la superficie orientada según } y} + \underbrace{[-\rho v_z|_{z+\Delta z} + \rho v_z|_z]\Delta x\Delta y}_{\text{caudal másico a través de la superficie orientada según } z} \quad (1.64)$$

Dividiendo por el volúmen del cubo y tomando límites cuando las dimensiones tienden a cero arribamos a una expresión idéntica a (1.58) . Como vemos en lo anterior solo hemos usado un concepto de balance de flujos a través de la superficie con términos de incremento de la propiedad en el volúmen considerando a éste fijo en el espacio por lo que aparece la derivada temporal parcial.

Lineas de corriente, lineas de camino, trazas y función de corriente

Este importante tema perteneciente a la cinemática de fluidos por razones de espacio y por no estar dentro de los objetivos del curso será dejado al lector para su estudio. En general la mayoría de los libros introductorios de mecánica de fluidos lo tratan en forma extensiva. Sugerimos a aquellos interesados en el tema los libros de Whitaker [Wi], Batchelor [Ba] y White [Wh] entre otros. Por razones de completitud de las presentes notas en un futuro está previsto incluirlo al tema en esta sección.

Fuerzas de superficie en un fluido

Volvamos por un momento a la expresión del principio de conservación de la cantidad de movimiento lineal , ecuación (1.6),

$$\underbrace{\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho\mathbf{v}dV}_{\left\{\begin{smallmatrix}\text{variación temporal de}\\ \text{la cantidad de movimiento}\end{smallmatrix}\right\}} = \underbrace{\int_{\mathcal{V}_m(t)} \rho\mathbf{g}dV}_{\text{fuerzas de masa}} + \underbrace{\int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})}dA}_{\text{fuerzas de superficie}} \quad ((1.6))$$

En la sección anterior hemos tratado la cinemática de los fluidos en el caso general, o sea el miembro izquierdo de la ecuación (1.6) y hemos hecho una mención aparte dentro de esta al caso especial de la estática de los fluidos. Nosotros ahora volcamos nuestra atención al miembro derecho, o sea a las causas del movimiento y en especial al término de las fuerzas de superficie ya que las fuerzas de cuerpo o de masa en general no presentan mayor dificultad.

Vector tensión y tensor de tensiones

A continuación trataremos al vector tensión a pesar de que ya lo hemos presentado cuando tratamos el caso particular de la estática de los fluidos o el caso simple de flujo desarrollado en un tubo (movimiento unidimensional). Como antes consideramos la hipótesis del continuo, o sea asumimos que el vector tensión varía en forma continua con las variables independientes del problema, al igual que las otras variables incluidas en el problema. No obstante en el caso del vector tensión también debemos agregar que varía en forma suave o continua con la orientación en la cual está calculado, $\mathbf{n}$. Sin entrar en detalles acerca de la prueba de esto [Wi] establecemos las siguientes hipótesis:

1. 1.- el vector tensión actuando sobre lados opuestos de la misma superficie en un dado punto es igual en magnitud y opuesto en dirección, $\mathbf{t}_{(\mathbf{n})} = -\mathbf{t}_{(\mathbf{n})}$.

2. 2.- el vector tensión puede escribirse en términos del tensor de tensiones del cual proviene como $\mathbf{t}_{(\mathbf{n})} = \mathbf{n} \cdot \mathbf{T}$.

3. 3.- el tensor de tensiones es simétrico, o sea $T_{ij} = T_{ji}$

El primer punto se demuestra planteando el principio de conservación de la cantidad de movimiento lineal , el teorema del valor medio y un procedimiento de ir al límite cuando la longitud característica tiende a cero. El segundo punto requiere plantear el equilibrio de un volúmen tetraédrico sobre el cual el vector tensión actuando sobre un plano puede descomponerse o formarse según aquellos pertenecientes a los otros tres planos orientados según los ejes cartesianos. Los tres vectores tensión actuando sobre los tres planos cartesianos son:

$$\begin{aligned}
\mathbf{t}_{(\mathbf{i})} &= \mathbf{i}T_{xx} + \mathbf{j}T_{xy} + \mathbf{k}T_{xz} \quad \left\{ \begin{array}{c} \text{Fuerza por unidad de area actuando en una} \\ \text{superficie con normal orientada según } x \end{array} \right\} \\
\mathbf{t}_{(\mathbf{j})} &= \mathbf{i}T_{yx} + \mathbf{j}T_{yy} + \mathbf{k}T_{yz} \quad \left\{ \begin{array}{c} \text{Fuerza por unidad de area actuando en una} \\ \text{superficie con normal orientada según } y \end{array} \right\} \\
\mathbf{t}_{(\mathbf{k})} &= \mathbf{i}T_{zx} + \mathbf{j}T_{zy} + \mathbf{k}T_{zz} \quad \left\{ \begin{array}{c} \text{Fuerza por unidad de area actuando en una} \\ \text{superficie con normal orientada según } z \end{array} \right\}
\end{aligned} \tag{1.65}$$

siendo el vector tensión expresado según estos tres vectores tensión:

$$\begin{aligned}
\mathbf{t}_{(\mathbf{n})} &= \Big[(\mathbf{n} \cdot \mathbf{i})\mathbf{t}_{(\mathbf{i})} + (\mathbf{n} \cdot \mathbf{j})\mathbf{t}_{(\mathbf{j})} + (\mathbf{n} \cdot \mathbf{k})\mathbf{t}_{(\mathbf{k})} \Big] \\
\mathbf{t}_{(\mathbf{n})} &= \mathbf{n} \cdot \Big[\mathbf{i}\mathbf{t}_{(\mathbf{i})} + \mathbf{j}\mathbf{t}_{(\mathbf{j})} + \mathbf{k}\mathbf{t}_{(\mathbf{k})} \Big]
\end{aligned} \tag{1.66}$$

con lo cual se alcanza el resultado esperado

$$\mathbf{t}_{(\mathbf{n})} = \mathbf{n} \cdot \mathbf{T} \tag{1.67}$$

donde el tensor de tensiones está compuesto de términos $\mathbf{i}\mathbf{t}_{(\mathbf{i})}$ que no representan ni un producto escalar ni uno vectorial, este producto es a menudo llamado *producto diádico*, base del algebra tensorial. Entonces surge que el vector tensión (tensor de primer orden) es la contracción del tensor de tensiones (tensor de segundo orden) en la dirección normal. Este resultado no es tan obvio para aquellos no familiarizados con el algebra tensorial y se recomienda volcar la atención a este tema para aquellos que pretender lograr un mejor entendimientos de las ecuaciones de movimiento que más adelante se definen. En resúmen, el tensor de tensiones en tres dimensiones expresado por simplicidad según las coordenadas cartesianas es:

Las nueve componentes escalares en (1.65) representan o caracterizan al tensor de tensiones, que en notación matricial puede escribirse como:

$$\mathbf{T} = \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \tag{1.68}$$

El primer índice representa la orientación del plano sobre el cual la tensión actúa mientras que el segundo índice está relacionada con la dirección en la cual la tensión está actuando.

En notación indicial (1.67) puede escribirse como:

$$t_{(n)\,i} = n_j T_{ji} \tag{1.69}$$

La forma de arribar al tercer punto, la simetría del tensor, es mediante el principio de conservación del momento de la cantidad de movimiento lineal aplicado a un volúmen diferencial. Esta simetría a su vez conduce al siguiente resultado:

$$\mathbf{n} \cdot \mathbf{T} = \mathbf{T} \cdot \mathbf{n} \tag{1.70}$$

Ecuaciones de movimiento expresada según el tensor de tensiones

El objetivo es plantear las ecuaciones de movimiento en término del tensor de tensiones siendo esta forma una herramienta muy útil para plantear un analísis macroscópico global de la mecánica de los fluidos o incluso como paso previo a la formulación diferencial del problema permitiendo cerrar el análisis mediante la introducción de las ecuaciones constitutivas del material, siendo esta forma bastante general. Comenzamos planteando el principio de conservación de la cantidad de movimiento lineal ,

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho \mathbf{v} dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g} dV + \int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})} dA \tag{1.71}$$

La idea es encontrar la formulación diferencial al problema tal como hicimos previamente al tratar la conservación de la masa. Allí usamos el teorema de la divergencia sobre un integrando formado por el producto escalar del vector velocidad con la normal. Aquí la dificultad está en que el integrando tiene al vector tensión y aún no presentamos como aplicar el teorema de la divergencia a un integrando vectorial que no puede ser considerado constante. Para salvar esta dificultad Whitaker multiplica la anterior ecuación escalarmente por un vector constante $\mathbf{b}$ y por ser constante puede ser introducida sin dificultad dentro del símbolo diferencial:

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho \mathbf{v} \cdot \mathbf{b} dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g} \cdot \mathbf{b} dV + \int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})} \cdot \mathbf{b} dA \tag{1.72}$$

y mediante el teorema del transporte de Reynolds y algunas manipulaciones algrebraica simples llegamos a:

$$\int_{\mathcal{V}_m(t)} \rho \frac{D}{Dt}(\mathbf{v} \cdot \mathbf{b}) dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g} \cdot \mathbf{b} dV + \int_{\mathcal{A}_m(t)} \mathbf{n} \cdot (\mathbf{T} \cdot \mathbf{b}) dA \tag{1.73}$$

como $(\mathbf{T} \cdot \mathbf{b})$ es un vector se puede aplicar el teorema de la divergencia tal como lo vimos anteriormente y obtener

$$\int_{\mathcal{V}_m(t)} \Big\{ \rho \frac{D}{Dt}(\mathbf{v} \cdot \mathbf{b}) - \rho \mathbf{g} \cdot \mathbf{b} - \nabla \cdot (\mathbf{T} \cdot \mathbf{b}) \Big\} dV = 0 \tag{1.74}$$

Otra vez, al ser los límites de integración arbitrarios podemos removerlos y el integrando se satisface en todo punto del dominio. La eliminación del vector constante $\mathbf{b}$ es trivial en los dos primeros términos. En

cuanto al tercero una forma de resolverlo es recurrir o a la notación indicial o sino a la definición matricial del tensor (1.68)

$$\begin{aligned}\nabla \cdot \Big(\mathbf{T} \cdot \mathbf{b}\Big) &= \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{pmatrix} \left\{ \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} \right\} = \\ &= \begin{pmatrix} \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \end{pmatrix} \left\{ \begin{pmatrix} T_{xx} & T_{xy} & T_{xz} \\ T_{yx} & T_{yy} & T_{yz} \\ T_{zx} & T_{zy} & T_{zz} \end{pmatrix} \right\} \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} = \\ &= (\nabla \cdot \mathbf{T}) \cdot \mathbf{b} \end{aligned} \tag{1.75}$$

de esta forma surge que la integral de superficie del vector tensión se transforma en la integral de volúmen de la divergencia del tensor de tensiones, una generalización del teorema de la divergencia. Por lo tanto la ecuación de movimiento (1.6) se transforma en:

$$\rho \frac{D\mathbf{v}}{Dt} = \rho \mathbf{g} + \nabla \cdot \mathbf{T} \tag{1.76}$$

Entonces partiendo de la formulación integral o macroscópica global arribamos a la formulación diferencial o macroscópica local del principio de conservación de la cantidad de movimiento lineal . Esta expresión tiene la ventaja que el miembro izquierdo es expresado en las variables dependientes del problema mientras que el miembro derecho contienen los flujos de estas variables o su cantidades duales. Incorporando las ecuaciones constitutivas del material es como transformamos esta ecuación en otra equivalente pero expresada solo en las variables de estado.

Ecuaciones diferenciales de movimiento de un fluido

En la sección anterior derivamos la ecuación vectorial de movimiento expresada según el tensor de tensiones y junto al principio de conservación de la masa forman un sistema de 4 ecuaciones en 3D con 9 incógnitas, las tres componentes del vector velocidad y los 6 coeficientes del tensor de tensiones simétrico. Para salvar esta indeterminación introducimos información adicional via las ecuaciones constitutivas del material que relacionan las componentes del tensor de tensiones con la presión y los gradientes del vector velocidad llegando finalmente a expresar las anteriores 4 ecuaciones en función de las 4 incógnitas, la presión y el vector velocidad.

$$\rho \frac{D\mathbf{v}}{Dt} = \rho \mathbf{g} + \nabla \cdot \mathbf{T} \qquad ((1.76,1.58))$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.77}$$

Tensor de tensiones viscoso

En la sección 2 hemos visto que en el caso de la estática de fluidos no podían existir esfuerzos cortantes y que los únicos esfuerzos que el fluido es capaz de resistir son aquellos normales. Habímos mencionado que estos eran isotrópicos y que en definitiva estaban caracterizados por la presión:

$$\mathbf{t}_{(\mathbf{n})} = \mathbf{n} \cdot \mathbf{T} = -\mathbf{n}p \tag{1.78}$$

Analizando la anterior vemos que para el caso de un fluido en reposo el tensor de tensiones se reduce a un escalar, en realidad asumiendo que la isotropía se manifiesta mediante un tensor múltiplo de la identidad podemos extender lo anterior diciendo que $\mathbf{T} = -p\mathbf{I}$. En general podemos dividir al tensor de tensiones en

dos partes, una isotrópica como la anterior y una no isotrópica, denominada muchas veces *deviatórica* y que corresponde al tensor viscoso,

$$\mathbf{T} = -p\mathbf{I} + \boldsymbol{\tau} \tag{1.79}$$

siendo $\boldsymbol{\tau} = 0$ para el caso particular de la estática de fluidos. Ya que $\mathbf{T}$ e $\mathbf{I}$ son simétricos, entonces $\boldsymbol{\tau}$ es simétrico y dado que el término isotrópico es un tensor diagonal, entonces se satisface que:

$$T_{ij} = \boldsymbol{\tau}_{ij} \qquad \forall i \neq j \tag{1.80}$$

Reemplazando en las 4 ecuaciones arriba planteadas (1.76,1.58) se obtiene:

$$\rho\Big(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\Big) = -\nabla p + \rho\mathbf{g} + \nabla\cdot\boldsymbol{\tau} \tag{1.81}$$

donde solo se requiere establecer ecuaciones para las componentes del tensor viscoso en término de las velocidades o mas generalmente en términos del tensor velocidad de deformación $\mathbf{d}$. Si suponemos que el material se comporta como un fluido newtoniano entonces es muy habitual usar la siguiente ley:

$$\begin{aligned} \boldsymbol{\tau} &= 2\mu\mathbf{d} + [(\kappa - \frac{2}{3}\mu)\nabla\cdot\mathbf{v}]\mathbf{I} \\ \tau_{ij} &= 2\mu d_{ij} + [(\kappa - \frac{2}{3}\mu)(\frac{\partial v_k}{\partial x_k})]\delta_{ij} \end{aligned} \tag{1.82}$$

expresada en notación de Gibbs e indicial y en término de dos coeficientes de viscosidad, μ y κ.

En pos de poder entender mejor la forma en la que surgen las ecuaciones constitutivas es necesario introducir varios conceptos relacionados con la deformación y la velocidad de deformación de un fluido. No obstante esto es dejado para futuras versiones de estas notas debido a lo extenso del análisis, aquellos interesados pueden consultar la bibliografia básica del tema [Wi],[Ba]. Por el momento nuestro análisis lo enfocamos hacia derivar las ecuaciones diferenciales de movimiento que serán la que posteriormente resolveremos via métodos numéricos. Para ello nos conformamos con la definición de alguna ley constitutiva. Ya que por lo general siempre se comienza por los casos más simples o más observados experimentalmente recurrimos a la hipótesis newtoniana, con la definición (1.82), que reemplazada en (1.81) nos da:

$$\begin{aligned} \rho\Big(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\Big) &= -\nabla p + \rho\mathbf{g} + \nabla\cdot\Big[2\mu\mathbf{d} + (\kappa - \frac{2}{3}\mu)\nabla\cdot\mathbf{v}\mathbf{I}\Big] \\ \mathbf{d} &= \tfrac{1}{2}(\nabla\mathbf{v} + \nabla^T\mathbf{v}) \end{aligned} \tag{1.83}$$

donde $\nabla^T\mathbf{v}$ es el gradiente del vector velocidad traspuesto, un tensor de segundo orden.

Casos particulares como el de flujo incompresible donde la ecuación de continuidad según (1.60) equivale a $\nabla\cdot\mathbf{v} = 0$ simplifica la anterior a:

$$\rho\Big(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\Big) = -\nabla p + \rho\mathbf{g} + \nabla\cdot\Big[\mu(\nabla\mathbf{v} + \nabla^T\mathbf{v})\Big] \tag{1.84}$$

Si la viscosidad fuera constante sale fuera del símbolo divergencia y entonces la anterior se simplifica a:

$$\rho\Big(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\Big) = -\nabla p + \rho\mathbf{g} + \mu\nabla^2\mathbf{v} \tag{1.85}$$

Formulación vorticidad-función de corriente

Definimos la vorticidad como:

$$\boldsymbol{\omega} = \nabla \wedge \mathbf{v} \tag{1.86}$$

Si tomamos la ecuación de conservación del momento lineal para el caso incompresible a viscosidad constante y si le aplicamos el rotor a la misma obtenemos

$$\begin{aligned} \nabla \wedge \left\{ (\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}) = \mathbf{g} + -\frac{1}{\rho} \nabla p + \nu \nabla^2 \mathbf{v} \right\} \\ \frac{D\boldsymbol{\omega}}{Dt} = \boldsymbol{\omega} \cdot \nabla \mathbf{v} + \nabla \wedge \mathbf{g} + \nu \nabla^2 \boldsymbol{\omega} \end{aligned} \tag{1.87}$$

donde $\nabla \wedge \mathbf{g}$ representa el rotor del campo de fuerzas de volúmen en general, ya sea fuerzas gravitatorias como de flotación térmica o electromagnéticas, etc. El término $\boldsymbol{\omega} \cdot \nabla \mathbf{v}$ representa matemáticamente un término fuente proporcional a la vorticidad y físicamente está asociado a la deformación por contracción o elongación del volúmen material. Como se alcanza a ver en el caso bidimensional este término no existe porque la vorticidad es un vector orientado en la dirección normal al plano del movimiento y el producto escalar con ella es nulo. Solo existe en el escurrimiento 3D. En el caso 2D la vorticidad puede tratarse como un escalar ya que su orientación permanecerá fija y apuntando en la normal al plano del movimiento. Si consideramos ausencia de fuerzas de volúmen la anterior se simplifica a:

$$\frac{D\omega}{Dt} = \nu \nabla^2 \omega \tag{1.88}$$

Introduciendo la definición de función de corriente,

$$\begin{aligned} u = & \frac{\partial \psi}{\partial y} \\ v = & -\frac{\partial \psi}{\partial x} \end{aligned} \tag{1.89}$$

y reemplazándola en la definición de la vorticidad llegamos a

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = -\omega \tag{1.90}$$

Resumiendo la formulación vorticidad-función de corriente en 2D se puede escribir como:

$$\begin{aligned} \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = -\omega \\ \frac{\partial \omega}{\partial t} + \mathbf{v} \cdot \nabla \omega = \nu \nabla^2 \omega \end{aligned} \tag{1.91}$$

donde

$$\mathbf{v} = \begin{pmatrix} \frac{\partial \psi}{\partial y} \\ -\frac{\partial \psi}{\partial x} \end{pmatrix} \tag{1.92}$$

El planteamiento de las condiciones de contorno en esta formulación merece un tratamiento cuidadoso y algunos detalles de los mismos se incluyen en el capítulo de las aplicaciones.

Flujo compresible

Hasta el momento hemos hecho bastante hincapié en el caso de flujo incompresible en el cual asumimos que la densidad es constante. En esta sección trataremos de extender el tratamiento para incluir algunos conceptos introductorios acerca del flujo compresible. En el caso compresible la densidad aparece como una incógnita a resolver y en general a partir de argumentos termodinámicos está ligada a otras variables, por ejemplo a la presión y temperatura a través de la ecuación de estado.

$$\begin{aligned} \rho &= \text{ constante} && \text{INCOMPRESIBLE} \\ \rho &= \rho(p,T) && \text{COMPRESIBLE} \end{aligned} \tag{1.93}$$

Con todo esto es posible cerrar el sistema. Al respecto presentaremos la ecuación de conservación de energía que debe ser acoplada al resto de las ecuaciones de conservación para luego hacer algunos comentarios acerca de la entropía.

Conservación de la energía

El postulado fundamental de la conservación de la energía establece:

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho(e + \tfrac{1}{2} \parallel \mathbf{v} \parallel^2) dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g}\cdot\mathbf{v} dV + \int_{\mathcal{A}_m(t)} \mathbf{t}_{(\mathbf{n})}\cdot\mathbf{v} dA - \int_{\mathcal{A}_m(t)} \mathbf{q}\cdot\mathbf{n} dA \tag{1.94}$$

donde el término a la izquierda representa la variación temporal de la energía interna y la cinética dentro del volúmen material o cuerpo, el primer término de la derecha y el segundo representan trabajos por unidad de tiempo de las fuerzas de gravedad y las de superficie, mientras que el tercer término contiene el calor intercambiado. De alguna forma esta versión integral equivale al primer principio de la termodinámica aplicado a sistemas cerrados que en general se expresa en forma diferencial o en diferencias:

$$\Delta(U + KE + PE) = Q - W' \tag{1.95}$$

Usando la igualdad derivada anteriormente (forma especial del teorema del transporte de Reynolds) (1.62)

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)} \rho s dV = \int_{\mathcal{V}_m(t)} \rho \frac{Ds}{Dt} dV \tag{1.96}$$

y expresando el vector tensión como la contracción del tensor de tensiones en la dirección normal llegamos a:

$$\int_{\mathcal{V}_m(t)} \rho \frac{D}{Dt}(e + \tfrac{1}{2} \parallel \mathbf{v} \parallel^2) dV = \int_{\mathcal{V}_m(t)} \rho \mathbf{g}\cdot\mathbf{v} dV + \int_{\mathcal{A}_m(t)} (\mathbf{n}\cdot\mathbf{T})\cdot\mathbf{v} dA - \int_{\mathcal{A}_m(t)} \mathbf{q}\cdot\mathbf{n} dA \tag{1.97}$$

y aplicando el teorema de la divergencia sobre las dos integrales de area para llevarlas a formas equivalentes sobre integrales de volúmen lo anterior se puede escribir como:

$$\rho \frac{D}{Dt}(e + \tfrac{1}{2} \parallel \mathbf{v} \parallel^2) = \rho \mathbf{g}\cdot\mathbf{v} + \nabla\cdot(\mathbf{T}\cdot\mathbf{v}) - \nabla\cdot\mathbf{q} \tag{1.98}$$

y si expresamos las fuerzas gravitatorias como conservativas y provenientes del gradiente de un potencial $(\nabla\phi)$, podemos simplificar lo anterior a:

$$\rho \frac{D}{Dt}(e + \tfrac{1}{2} \parallel \mathbf{v} \parallel^2 + \phi) = \nabla\cdot(\mathbf{T}\cdot\mathbf{v}) - \nabla\cdot\mathbf{q} \tag{1.99}$$

donde hemos agrupado la energia interna, la cinética y la potencial en el miembro izquierdo y si aplicamos el balance de masa se llega a:

$$\frac{D}{Dt}\Big(\rho(e+\tfrac{1}{2}\parallel \mathbf{v} \parallel^2 +\phi)\Big) = \nabla\cdot(\mathbf{T}\cdot\mathbf{v}) - \nabla\cdot\mathbf{q} \tag{1.100}$$

En general es común a partir de esta expresión plantear el balance en un volúmen de control arbitrario que se mueve a una velocidad distinta a la del fluido y de esta forma obtener el balance macroscópico global. También es habitual extender lo anterior al caso de sistemas abiertos donde es usual definir la entalpía y plantear el balance en términos de esta función.

Ecuación de la energía térmica

Una forma de poner la anterior expresión en términos de la energía térmica es remover de (1.97) la ecuación correspondiente al balance de energía mecánica. Este se puede hallar tomando las ecuaciones de movimiento y multiplicarla escalarmente por la velocidad,

$$\begin{aligned}\rho\frac{D\tfrac{1}{2}\parallel \mathbf{v} \parallel^2}{Dt} &= \rho\mathbf{g}\cdot\mathbf{v} + \mathbf{v}\cdot(\nabla\cdot\mathbf{T}) = \\ &= \rho\mathbf{g}\cdot\mathbf{v} + \nabla\cdot(\mathbf{T}\cdot\mathbf{v}) - \nabla\mathbf{v}:\mathbf{T}\end{aligned} \tag{1.101}$$

Restando de (1.95) hallamos

$$\int_{\mathcal{V}_m(t)}\rho\frac{De}{Dt}dV = \int_{\mathcal{A}_m(t)}\nabla\mathbf{v}:\mathbf{T}dA - \int_{\mathcal{A}_m(t)}\mathbf{q}\cdot\mathbf{n}dA \tag{1.102}$$

Separando $\mathbf{T} = -p\mathbf{I} + \boldsymbol{\tau}$ en su componente normal y en la deviatórica y pasando a la versión diferencia mediante el teorema de la divergencia encontramos :

$$\rho\frac{De}{Dt} = -p\nabla\cdot\mathbf{v} + \Phi - \nabla\cdot\mathbf{q} \tag{1.103}$$

Φ representa la disipación viscosa que aumenta la energía interna y por ende la temperatura.

Ecuación de la entropia

Partiendo de las funciones características de la termodinámica encontramos la relación entre energía interna, entropía y densidad,

$$e = e(s,\rho) \tag{1.104}$$

Tomando la derivada material y multiplicando por la densidad

$$\begin{aligned}\rho\frac{De}{Dt} &= \rho\Big[\Big(\frac{\partial e}{\partial s}\Big)_\rho\frac{Ds}{Dt} + \Big(\frac{\partial e}{\partial \rho}\Big)_s\frac{D\rho}{Dt}\Big] = \\ &= \rho T\frac{Ds}{Dt} + \frac{p}{\rho}\frac{D\rho}{Dt}\end{aligned} \tag{1.105}$$

y escribiendo la ecuación de continuidad en la forma

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{v} = 0 \tag{1.106}$$

y usando la igualdad (1.97) llegamos a

$$\begin{aligned} \rho\frac{De}{Dt} &= \rho T\frac{Ds}{Dt} - p\nabla\cdot\mathbf{v} \\ \rho\frac{Ds}{Dt} &= \frac{1}{T}(-\nabla\cdot\mathbf{q}+\Phi) \end{aligned} \tag{1.107}$$

Haciendo el proceso inverso, desde la formulación diferencial llegar a la integral implica en este caso obtener:

$$\frac{D}{Dt}\int_{\mathcal{V}_m(t)}\rho s dV = -\int_{\mathcal{A}_m(t)}\frac{\mathbf{q}\cdot\mathbf{n}}{T}dA + \int_{\mathcal{V}_m(t)}(\frac{-\mathbf{q}\cdot\nabla T}{T^2}+\frac{\Phi}{T})dV \tag{1.108}$$

De este balance integral de la entropía surge que si un proceso es isentrópico entonces la entropía del volúmen material o cuerpo debe permanecer constante y esta condición se cumple si:

1. 1.- $\mathbf{q}\cdot\mathbf{n}=0$ sobre la superficie, o sea es adiabático,
2. 2.- $\nabla T = 0$ sobre el sistema, proceso reversible,
3. 2.- $\Phi = 0$ sobre el sistema, efectos de fricción despreciables.

1.3. TP.I.- Trabajo Práctico #1

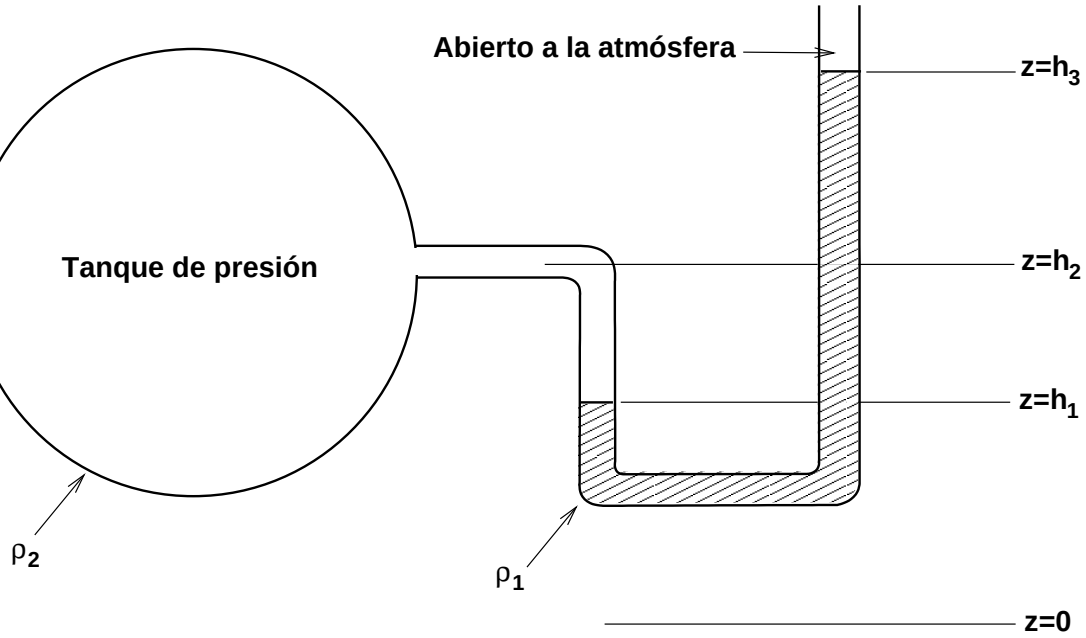
1. Demostrar que para un fluido en reposo (estática de fluidos) el tensor de tensiones es isotrópico.*Ayuda:* Considere un tetraedro diferencial sumergido en un fluido en reposo.
2. Dado el manómetro de la figura encuentre la expresión que relaciona la presión relativa del interior del tanque respecto a la atmosférica en función de las alturas de las columnas usando el principio de conservación de la cantidad de movimiento lineal .
3. Calcule la fuerza y el torque aplicado sobre la placa plana de la siguiente figura:
4. Calcule la fuerza a la que se halla sometida una esfera sumergida sobre la que actúa un desnivel como el de la siguiente figura:
5. Deduzca la expresión de un flujo Couette plano laminar y calcule: (a) la velocidad máxima y la velocidad promedio en la sección transversal al flujo

   (b) el caudal en función de la caida de presión
6. Deduzca la expresión de un flujo Couette plano laminar pero ahora utilizando una ley de viscosidad no newtoniana. Para ello tome aquella del modelo de fluido de ley de potencia:

$$\tau_{yx} = \mu_0|\frac{\mathrm{d}v_x}{\mathrm{d}y}|^{n-1}\frac{\mathrm{d}v_x}{\mathrm{d}y}$$

   y considere: (a) el caso de un fluido pseudoplástico $(n<1)$

   (b) el caso de un fluido dilatante $(n>1)$



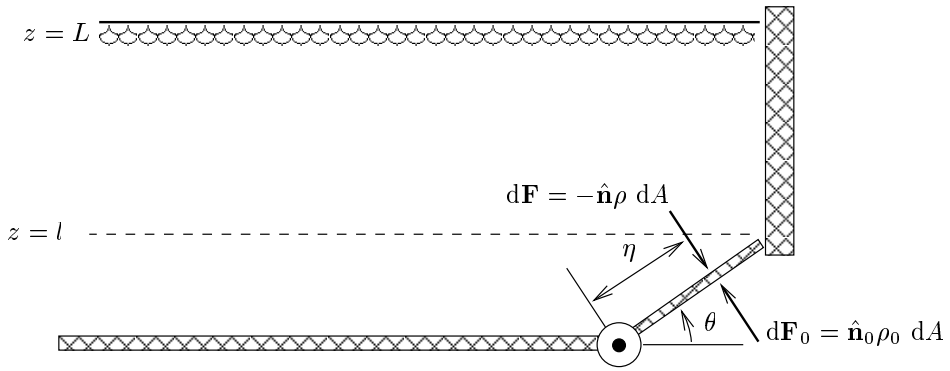
Manómetro

7. Deduzca la expresión de un flujo Couette laminar para un escurrimiento longitudinal a lo largo de un tubo anular donde el radio interior es r_1 y el exterior es r_2.

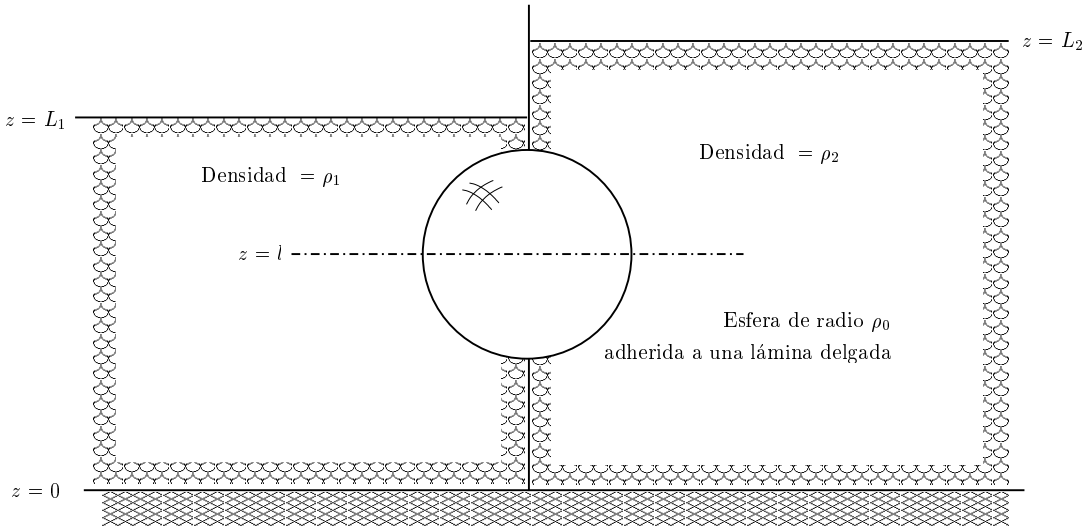
8. Repita el Ej 7 pero utilizando un fluido no newtoniano con una ley como la presentada en el ej 6 y calculo la relación entre el caudal y la caida de presión.

9. Partiendo de la definición vectorial del teorema de la divergencia demostrar la versión escalar del mismo.*Ayuda:* Un campo escalar puede ser definido por uno vectorial con una orientación fija del campo.

10. Aplique el teorema de la divergencia a la expresión del principio de conseración de cantidad de movimiento lineal (1.6) para el caso de estática de los fluidos para obtener la expresión diferencial (1.8).



Fuerza sobre un cuerpo plano sumergido



Fuerza sobre un cuerpo curvo sumergido

11. A partir del teorema del transporte de Reynolds (1.52) aplicado a una función escalar del tipo $S = \rho s$ hallar la forma especial del teorema expresado por (1.62). *Ayuda: utilice el principio de conservación de la masa para alcanzar el resultado*

12. Probar que si $\mathbf{w}$ es independiente de las coordenadas espaciales, entonces

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathcal{V}_a(t)} S dV = \int_{\mathcal{V}_a(t)} \frac{\mathrm{d}S}{\mathrm{d}t} dV$$

13. Si la velocidad de un fluido viene dada por:

$$\mathbf{v} = u_0 e^{-at}[\mathbf{i}bx + \mathbf{j}cy^2]$$

obtener una expresión para la derivada material de la velocidad $\frac{D\mathbf{v}}{Dt}$ en término de los coeficientes a, b, c y u_0.

Capítulo 2

Niveles dinámicos de aproximación

2.0.1. Introducción

- las ecuaciones de Navier-Stokes
- la dependencia de la viscosidad y la conductividad térmica con otras variables
- leyes constitutivas

describen completamente el fenómeno fluidodinámico en régimen laminar.

No obstante en casi todas las situaciones reales en la naturaleza y en la tecnología existe una particular forma de inestabilidad conocida con el nombre de *turbulencia*. Esta ocurre cuando en ciertas situaciones un número adimensional llamado número de *Reynolds*, definido por $Re = \rho U_{char} L/\mu$, alcanza o supera determinado valor que depende del problema en particular. L representa una longitud característica y U_{char} una velocidad característica. Este fenómeno se caracteriza por la presencia de fluctuaciones estadísticas en todas las variables del flujo que se agregan a los valores medios, llegando en algunos casos a alcanzar valores de hasta el $10\,\%$ del promedio.

Dado que la escala que imponen los fenómenos turbulentos están fuera de los alcances de los recursos computacionales actuales, es la tarea de estos tiempos tratar de resolver las ecuaciones de Navier-Stokes en su versión promediada y suplementada por alguna descripción externa de las tensiones de Reynolds. Esta información es suministrada por los modelos de turbulencia, cubriendo ellos un rango muy amplio, desde los más simples basados en definir una longitud de mezcla y una viscosidad para los *eddies* hasta aquellos que plantean un balance para el transporte de cantidades como la energía cinética turbulenta y la disipación, modelo denominado $k - \epsilon$, o formas más complicadas de calcular las componentes del tensor de Reynolds. A continuación presentamos en forma resumida cuales serían los distintos niveles de aproximación que se pueden plantear sobre la base de los efectos dinámicos presentes.

- **Navier-Stokes tridimensional laminar**
- **Navier-Stokes tridimensional turbulento**
- En el caso de no existir extensas regiones viscosas podemos plantearnos *Thin shear layer* (TSL). Esta aproximación desprecia los fenómenos de difusión molecular y turbulenta en la dirección de transporte reduciendo el número de términos de tensión viscosa a calcular.

- **Navier-Stokes parabolizado:** En esta aproximación el carácter elíptico de las ecuaciones es responsabilidad de la presión mientras que el resto de las variables se parabolizan, reduciendo los fenómenos de difusión a la dirección transversal.
- **Capa límite** Para altos Re podemos introducir una suerte de separación de los efectos viscosos e invíscidos, desacoplando la presión de los efectos viscosos y confinando éstos a una región cercana a los cuerpos mientras que lejos el flujo se comporta como invíscido.
- **Aproximación invíscida (ecuaciones de Euler)**. Aquí despreciamos todos los efectos viscosos y nos acercamos a los cuerpos tanto como el espesor de la capa límite, no resolviendo lo que sucede dentro de ellas.
- **Modelo de pérdidas distribuidas** Es un modelo usado en problemas de flujo interno, en especial en turbomáquinas. Se ubica en un nivel intermedio entre los modelos total o parcialmente viscosos y aquellos puramente invíscidos. Debido a la existencia de filas de álabes las capas límites y las estelas que deja una fila de ellas se mezclan y son vista por la siguiente como una fuerza de fricción distribuida.
- **Modelo de flujo potencial** Este está asociado a flujo irrotacional y por su carácter isoentrópico presenta problemas de unicidad cuando aparecen discontinuidades debiendo imponerse las condiciones de Rankine-Hugoniot para evitarlas.

2.1. Las ecuaciones de Navier-Stokes

$$\frac{\partial}{\partial t}\begin{pmatrix}\rho\\ \rho\vec{v}\\ \rho E\end{pmatrix}+\vec{\nabla}\cdot\begin{pmatrix}\rho\vec{v}\\ \rho\vec{v}\otimes\vec{v}+p\overline{\overline{I}}-\overline{\overline{\tau}}\\ \rho\vec{v}H-\overline{\overline{\tau}}\cdot\vec{v}-k\vec{\nabla}T\end{pmatrix}=\begin{pmatrix}0\\ \rho\vec{f_e}\\ W_f+q_H\end{pmatrix} \tag{2.1}$$

donde $W_f=\rho\vec{f_e}\cdot\vec{v}$. Estas forman un sistema de 5 ecuaciones en el caso 3D expresado aquí en variables conservativas U definidas como:

$$U=\begin{pmatrix}\rho\\ \rho\vec{v}\\ \rho E\end{pmatrix}=\begin{pmatrix}\rho\\ \rho u\\ \rho v\\ \rho w\\ \rho E\end{pmatrix} \tag{2.2}$$

No debe confundirse lo que se llama una forma conservativa de lo que son las variables conservativas. La matriz de los vectores flujos $\vec{F}$ tiene dimensión (5×3)

$$\vec{F}=\begin{pmatrix}\rho\vec{v}\\ \rho\vec{v}\otimes\vec{v}+p\overline{\overline{I}}-\overline{\overline{\tau}}\\ \rho\vec{v}H-\overline{\overline{\tau}}\cdot\vec{v}-k\vec{\nabla}T\end{pmatrix} \tag{2.3}$$

y puede dividirse en tres vectores columnas (f,g,h) de 5 componentes cada uno.

El lado derecho contiene los términos fuentes, Q, un vector de (5×1)

$$Q = \begin{pmatrix} 0 \\ \rho \vec{f_e} \\ W_f + q_H \end{pmatrix} \tag{2.4}$$

De esta forma podemos arribar a una forma condensada de las ecuaciones que suele ser muy útil en ciertas situaciones

$$\frac{\partial U}{\partial t} + \vec{\nabla} \cdot \vec{F} = Q \tag{2.5}$$

que expresada en sus tres coordenadas cartesianas da lugar a

$$\frac{\partial U}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z} = Q \tag{2.6}$$

Las ecuaciones de Navier-Stokes deben ser suplementadas por leyes constitutivas y por la definición del tensor de tensiones viscosas en función de otras variables del flujo. Consideraremos las hipótesis Newtonianas ya vistas. Las leyes termodinámicas definen la relación entre la energía o la entalpía y otras variables termodinámicas, tales como la temperatura, la densidad o la presión.

$$e = e(p, T) \qquad \text{o} \qquad h = h(p, T) \tag{2.7}$$

Además deben especificarse leyes de dependencia de la viscosidad y la conductividad térmica con otras variables del flujo como la temperatura o eventualmente la presión. En particular es muy usual en gases expresar la dependencia de la viscosidad con la temperatura a partir de la ley de Sutherland

$$\mu = \frac{1.45\ T^{3/2}}{T + 110} 10^{-6}, \ \ [T]\text{K}elvin \tag{2.8}$$

Para la conductividad térmica puede plantearse algo similar conociendo la relación que vincula $C_p = C_p(T)$. En el caso de líquidos k es asumida como constante.

Muchas veces se usa la hipótesis de gases perfectos que plantea una expresión particular de (2.7), como $e = C_v T$ y una relación para las tres variables termodinámicas a través de la ecuación de estado de los gases

$$p = \rho R_{\text{g}as} T$$

donde $R_{\text{g}as}$ es la constante universal de los gases.

2.1.1. Modelo de fluido incompresible

Las ecuaciones de Navier-Stokes se simplifican considerablemente para el caso de fluidos incompresibles para el cual se asume que la densidad permanece invariable tanto en la coordenada espacial como en el tiempo. Si además el flujo permanece isotérmico esto separa la ecuación de energía del resto y el sistema se reduce en una ecuación. En el caso de flujos que involucren variaciones de temperatura el acoplamiento entre la temperatura y el movimiento ocurre a través de :

- variación de la viscosidad con T

- variación de la conductividad con T
- fuerzas de flotación
- fuentes de calor de orígen mecánico, químico o eléctrico

entre otras.

En el caso incompresible la ecuación de masa se transforma en:

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{2.9}$$

que aparece como una especie de restricción a la ecuación del movimiento que en forma no conservativa se puede escribir como:

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} = -\frac{1}{\rho}\vec{\nabla}p + \nu\Delta\vec{v} + \vec{f_e} \tag{2.10}$$

La ecuación de vorticidad de Helmholtz, presentada previamente, se transforma en

$$\frac{\partial \vec{\zeta}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{\zeta} = (\vec{\zeta} \cdot \vec{\nabla})\vec{v} + \vec{\nabla}p \times \vec{\nabla}\frac{1}{\rho} + \nu\Delta\vec{\zeta} + \vec{\nabla} \times \vec{f_e} \tag{2.11}$$

Para flujos donde la densidad no se estratifica el término en presión desaparece de las ecuaciones y el primer término del lado derecho se anula si el flujo es plano.

El caso incompresible presenta una situación muy particular, una de las 5 incógnitas, la presión, no tiene una forma dependiente del tiempo debido al carácter no evolutivo de la ecuación de continuidad. Esto redunda en una dificultad numérica que ha dado y continua dando lugar a muchas especulaciones.

Una ecuación para la presión puede obtenerse tomando la divergencia de las ecuaciones de momento y asumiendo un campo de velocidades *solenoidal* , conduce a:

$$\frac{1}{\rho}\Delta p = -\vec{\nabla} \cdot (\vec{v} \cdot \vec{\nabla})\vec{v} + \vec{\nabla} \cdot \vec{f_e} \tag{2.12}$$

que puede ser considerada como una ecuación de Poisson para la presión cuando el campo de velocidades es especificado. El término derecho contiene solo derivadas de primer órden para la velocidad.

2.1.2. Las ecuaciones de Navier-Stokes promediadas

El proceso de promediación en el tiempo para un flujo turbulento se introduce para alcanzar las leyes de movimiento para las cantidades medias. La promediación temporal se define de forma de remover la influencia de las fluctuaciones turbulentas sin destruir la dependencia temporal con otros fenómenos no estacionarios con escala de tiempo diferente a aquellas asociadas con la turbulencia.

La promediación se define como:

$$\begin{aligned} A &= \overline{A} + A' \\ \overline{A}(\vec{x}, t) &= \frac{1}{T}\int_{-T/2}^{T/2} A(\vec{x}, t+\tau)d\tau \end{aligned} \tag{2.13}$$

donde T se define como un tiempo suficientemente largo comparado con la escala de sucesos de la turbulencia pero pequeño comparado a la escala temporal del problema no estacionario. En flujos compresibles

este proceso de promediación conduce a productos de fluctuaciones entre cantidades como la densidad y otras variables como la velocidad y la energía interna. Para evitarlas se recurre a un promedio pesado con la densidad.

$$\begin{aligned} \tilde{A} &= \frac{\overline{\rho A}}{\overline{\rho}} \\ A &= \tilde{A} + A'' \\ \overline{\rho A''} &= 0 \end{aligned} \tag{2.14}$$

De esta forma se remueven todos los productos necesarios de las fluctuaciones de la densidad con las fluctuaciones de las otras variables. Por ejemplo, la ecuación de continuidad se transforma en

$$\frac{\partial}{\partial t}\tilde{\rho} + \vec{\nabla}\cdot(\tilde{\rho}\tilde{\vec{v}}) = 0 \tag{2.15}$$

Aplicada a las ecuaciones de momento llegamos a:

$$\frac{\partial}{\partial t}(\tilde{\rho}\tilde{\vec{v}}) + \vec{\nabla}\cdot(\rho\tilde{\vec{v}}\otimes\tilde{\vec{v}} + \tilde{p}\overline{\overline{I}} - \overline{\overline{\tau}}^{v} - \overline{\overline{\tau}}^{R}) = 0 \tag{2.16}$$

donde las tensiones de Reynolds se definen como:

$$\overline{\overline{\tau}}^{R} = \overline{-\rho\vec{v}''\otimes\vec{v}''} \tag{2.17}$$

y se agregan a las tensiones viscosas promediadas $\overline{\overline{\tau}}^{v}$.
En coordenadas cartesianas tenemos que

$$\tau_{ij}^{R} = -\overline{\rho v_i'' v_j''} \tag{2.18}$$

La relación entre las tensiones de Reynolds y las variables debe ser especificada en forma externa y se realiza a través de modelos que contienen una dosis de especulación teórica combinada con evidencias experimentales.

2.1.3. Aproximación "Thin shear layer" (TSL)

A elevados números de Reynolds las capas de corte próximas a las paredes y las que se producen en las estelas serán de un tamaño limitado y si la extensión de las zonas viscosas permanece limitada durante la evolución del flujo entonces la influencia dominantes de estas capas estará dada por los gradientes en la dirección transversal a la corriente fluida. Entonces, la aproximación "Thin shear layers" consiste en despreciar las derivadas en las direcciones de la superficie que delimitan estas capas de corte y considerar solo las derivadas en la dirección normal a ellas. Esta aproximación se sustenta aún más si hechamos un vistazo a las mallas discretas que se confeccionan para flujos con $Re > 10^4$ que son muy densas en la dirección normal a los cuerpos y son muy gruesas en el plano tangente al cuerpo, alcanzándose muy baja precisión en las direcciones contenidas en el plano tangente.

La forma condensada de las ecuaciones de Navier-Stokes (2.5) permanecen inalteradas pero el vector flujos $\vec{F}$ se simplifica por el hecho que

$$(\vec{\nabla}\cdot\overline{\overline{\tau}})^{i} \simeq \frac{\partial\tau^{in}}{\partial n} = \frac{\partial}{\partial n}(\vec{\tau}^{n})^{i} \tag{2.19}$$

Entonces si tomamos los flujos en cada una de las tres direcciones locales (tangente, normal y binormal) vemos que

$$
\begin{aligned}
f &= \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ \rho uH \end{pmatrix} \\
g &= \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ \rho vH \end{pmatrix} \\
h &= \begin{pmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ \rho wH \end{pmatrix} + \begin{pmatrix} 0 \\ -\mu \frac{\partial u}{\partial z} \\ -\mu \frac{\partial v}{\partial z} \\ -\frac{4}{3}\mu \frac{\partial w}{\partial z} \\ -\mu(u\frac{\partial u}{\partial z} + v\frac{\partial v}{\partial z} + \frac{4}{3}w\frac{\partial w}{\partial z}) - k\frac{\partial T}{\partial z} \end{pmatrix}
\end{aligned}
\tag{2.20}
$$

Como vemos esta aproximación asume flujo invíscido para las direcciones del plano tangente y flujo viscoso sólo en la dirección normal. A diferencia de la aproximación de capa límite, TSL no asume que la presión sea constante dentro de la capa límite, con lo cual representa una aproximación tipo capa límite de mayor órden.

2.1.4. Aproximación Navier-Stokes parabolizada

Esta aproximación se basa sobre consideraciones similares a TSL pero se aplica solamente al caso estacionario. Esta aproximación va dirigida hacia aquellas situaciones donde existe una dirección predominante como sería el caso de flujo en canales. Además se supone que las regiones viscosas cerca de bordes sólidos son dominadas por gradientes en el plano normal y por lo tanto la difusión de momento y energía en la dirección principal se desprecian. Esta aproximación es solo válida mientras no existan zonas con variaciones importantes tal como recirculaciones. Analizando la ecuación de momento en la dirección principal (asumimos la x) sería :

$$
\frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) + \frac{\partial}{\partial z}(\rho uw) = \frac{\partial}{\partial y}(\mu \frac{\partial u}{\partial y}) + \frac{\partial}{\partial z}(\mu \frac{\partial u}{\partial z}) \tag{2.21}
$$

Esta ecuación por haber perdido el término de derivada segunda según la dirección x cambió de tipo elíptico a parabólico ya que según "x" la derivada primera es la de mayor órden. Entonces la variable "x" puede ser vista como un pseudo tiempo y se resuelven problemas en 2D avanzando de a capas en "x". Similarmente la ecuación de energía parabolizada se vuelve

$$
\frac{\partial \rho uH}{\partial x} + \frac{\partial \rho vH}{\partial y} + \frac{\partial \rho wH}{\partial z} = \frac{\partial}{\partial y}(\overline{\overline{\tau}} \cdot \vec{v})_y + \frac{\partial}{\partial z}(\overline{\overline{\tau}} \cdot \vec{v})_z + \frac{\partial}{\partial y}(k\frac{\partial}{\partial y}) + \frac{\partial}{\partial z}(k\frac{\partial}{\partial z}) \tag{2.22}
$$

2.1.5. Aproximación de capa límite

Fue un gran descubrimiento alcanzado por Prandtl quien reconoció que a altos Re las regiones viscosas permanecen acotadas a una extensión δ (del órden de $\delta/L \simeq \sqrt{\nu/UL}$ para un cuerpo de longitud L) a lo largo de cuerpos sólidos o superficies inmersas o limitando el flujo. En estos casos el cálculo de la presión puede separarse de aquel del campo de velocidades. Comparando esta aproximación con la TSL se puede decir que aquí además de las suposiciones hechas con la TSL se agrega aquella que la velocidad en la dirección normal a la superficie se desprecia por lo cual no se produce caida de presión en esa dirección. De esta forma la presión en la capa límite se asume igual a la externa $p_e(x, y)$ computada a partir de una aproximación invíscida.

Si bien uno puede separar el cómputo de la zona o región invíscida del de la capa límite existe muchas situaciones donde existe interacción entre ambas y deben resolverse en forma iterativa. Este tipo de aproximación cae dentro de lo que suele llamarse *interacción viscosa-invíscisa*

2.2. Modelo de flujo invíscido

La configuración de flujo más general para el caso no viscoso donde se desprecian los efectos de conducción del calor se describen mediante las ecuaciones de Euler, obtenidas a partir de las ecuaciones de Navier-Stokes despreciando todos los términos de tensiones viscosas y los de conducción del calor. Esta aproximación vale para flujos a elevados número de Reynolds y fuera de las regiones donde se concentran los efectos viscosos. Este modelo respecto al de Navier-Stokes introduce un cambio importante en el conjunto de ecuaciones, el sistema de segundo órden se transforma en uno de primer órden, modificando no solo la aproximación física y numérica sino tambien la especificación de las condiciones de contorno. Las ecuaciones de Euler no estacionarias tienen una forma condensada similar a aquella presentada en (2.5)

$$\frac{\partial U}{\partial t} + \vec{\nabla} \cdot \vec{F} = Q \tag{2.23}$$

para el caso de Navier-Stokes , con la modificación puesta en la forma de definir el vector flujo $\vec{F}$ En este caso los flujos se definen como

$$f = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ \rho uH \end{pmatrix} \qquad g \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ \rho vH \end{pmatrix} \qquad \begin{pmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ \rho wH \end{pmatrix} \tag{2.24}$$

Una consideración importante le cabe a la entropía. Ya que el modelo de Euler asume la no existencia de conducción del calor ni de esfuerzos viscosos, si tomamos la ecuación (2.6.f)

$$\rho T \frac{\mathrm{d}s}{\mathrm{d}t} = \epsilon_v + \vec{\nabla} \cdot (k\vec{\nabla}T) + q_H \tag{2.25}$$

vemos que el segundo miembro es nulo, por lo cual ésta se reduce a:

$$T(\frac{\partial s}{\partial t} + \vec{v} \cdot \vec{\nabla} s) = 0 \tag{2.26}$$

expresando que la entropía es constante a lo largo de las líneas de corriente. Por lo tanto las ecs. de Euler describen flujos isentrópicos en ausencia de discontinuidades. No obstante la entropía puede variar de una línea de corriente a otra y esto es visto por una forma especial atribuida a Crocco de describir las ecuaciones de momento , cuando se la aplica al caso invíscido y estacionario en ausencia de fuerzas externas. Esta se expresa como:

$$-\vec{v}\times\vec{\zeta}=T\vec{\nabla}s-\vec{\nabla}H \tag{2.27}$$

Si por un momento asumimos un campo uniforme de entalpías vemos que la derivada normal de la entropía (en el plano tangente no varía) equivale a la componente normal en la terna de Frenet del producto vectorial entre la velocidad y la vorticidad y que es equivalente al producto de la componente binormal de la vorticidad con la velocidad normal, o sea

$$w\zeta_b=T\frac{\partial s}{\partial n} \tag{2.28}$$

Por lo tanto gradientes de entropia y vorticidad están directamente vinculados. Las ecuaciones de Euler también permiten discontinuidades en capas vorticosas, en ondas de choque o del tipo discontinuidades de contacto, pero ellas deben obtenerse usando la forma integral del las ecuaciones ya que la forma diferencial asume continuidad de la solución.

2.2.1. Propiedades de las soluciones discontinuas

Si dentro de un volúmen V existe una superficie Σ que se mueve con una velocidad $\vec{C}$ donde la solución presenta una discontinuidad debemos recurrir a la forma integral de las ecuaciones de Euler, que en ausencia de términos fuentes se puede escribir como:

$$\begin{aligned}\frac{\partial}{\partial t}\int_\Omega U d\Omega+\oint_S \vec{F}\cdot d\vec{S}=0\\ \int_\Omega\frac{\partial U}{\partial t}d\Omega+\int_\Omega U\frac{\partial}{\partial t}d\Omega+\oint_S\vec{F}\cdot d\vec{S}=0\end{aligned} \tag{2.29}$$

Ya que debemos seguir el movimiento de la superficie para poder estudiar su evolución, debemos aplicar el teorema del transporte de Reynolds sobre el término temporal, que dice:

Para cualquier función $f(\vec{x},t)$ tenemos que

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{V(t)}f d\Omega=\int_{V(t)}\Big(\frac{\partial f}{\partial t}+\vec{\nabla}\cdot(f\vec{C})\Big)d\Omega \tag{2.30}$$

Entonces usando (2.30) con $f=1$ y el teorema de la divergencia sobre (2.29) produce lo siguiente:

$$\int_\Omega\frac{\partial U}{\partial t}d\Omega-\oint_S U\vec{C}\cdot d\vec{S}+\oint_S\vec{F}\cdot d\vec{S}=0 \tag{2.31}$$

Asumiendo que el volúmen tiende a cero el término del flujo puede ponerse como:

$$\lim_{V\to 0}\oint_S\vec{F}\cdot d\vec{S}=\int_\Sigma(\vec{F}_2-\vec{F}_1)\cdot d\vec{\Sigma}=\int_\Sigma\Big[\vec{F}\cdot\vec{1}_n\Big]d\Sigma \tag{2.32}$$

donde $d\vec{\Sigma}$ es la normal a la superficie de la discontinuidad Σ y donde definimos el salto de una variable que cruza una discontinuidad $\Big[A\Big]$ como

$$\Big[A\Big] = A_2 - A_1$$

Combinando (2.29) con (2.32) llegamos a

$$\int_\Sigma (\Big[\vec{F}\Big] - \vec{C}\Big[U\Big]) \cdot d\vec{\Sigma} = 0 \tag{2.33}$$

que conduce a la forma local de las leyes de conservación sobre una discontinuidad, llamadas las relaciones de *Rankine-Hugoniot*:

$$\Big[\vec{F}\Big] \cdot \vec{1}_n - \vec{C}\Big[U\Big] \cdot \vec{1}_n = 0 \tag{2.34}$$

Si $\Sigma(\vec{x}, t) = 0$ define la superficie de discontinuidad, entonces

$$\frac{\mathrm{d}\Sigma}{\mathrm{d}t} = \frac{\partial \Sigma}{\partial t} + \vec{C} \cdot \vec{\nabla}\Sigma = 0$$

y si definimos la normal a la superficie como

$$\vec{1}_n = \frac{\vec{\nabla}\Sigma}{|\vec{\nabla}\Sigma|}$$

entonces, reemplazando en (2.34) obtenemos

$$\Big[\vec{F}\Big] \cdot \vec{\nabla}\Sigma + \frac{\partial \Sigma}{\partial t}\Big[U\Big] = 0 \tag{2.35}$$

Distintas formas de discontinuidad

- *shocks:* todas las variables del flujo son discontinuas
- *de contacto y capas vorticosas ("slip lines"):*
  - no ha transferencia de masa a través de ellas
  - densidad y velocidad tangencial discontinuas
  - presión y velocidad normal continuas

Si vemos las propiedades del sistema desde una terna que se mueve junto a la discontinuidad, las relaciones de Rankine Hugoniot para las ecuaciones de Euler se transforman en:

$$\begin{aligned} \Big[\rho\vec{v} \cdot \vec{1}_n\Big] &= 0 \\ \rho\vec{v}\Big[\vec{v}\Big] \cdot \vec{1}_n + \Big[p\Big] \cdot \vec{1}_n &= 0 \\ \rho\vec{v} \cdot \vec{1}_n\Big[H\Big] &= 0 \end{aligned} \tag{2.36}$$

Discontinuidad de contacto Al no haber transporte de masa a través de ellas

$$v_{n1} = v_{n2} = 0$$

y de (2.36) surge que

$$\left[p\right] = 0$$

y en general

$$\left[\rho\right] \neq 0 \qquad \text{y} \qquad \left[v_t\right] = 0$$

Capas vorticosas ("Slip lines") Como antes

$$\begin{aligned} v_{n1} &= v_{n2} = 0 \\ \left[p\right] &= 0 \end{aligned} \tag{2.37}$$

permitiendo saltos en la velocidad tangencial y en la densidad

$$\left[\rho\right] \neq 0 \qquad \text{y} \qquad \left[v_t\right] \neq 0$$

Ondas de choque Ellas permiten el transporte de masa a través y por lo tanto la velocidad normal y la presión pueden ser discontinuas mientras que la velocidad tangencial permanece continua. Entonces

$$\begin{aligned} \left[\rho\right] &\neq 0 \\ \left[p\right] &\neq 0 \\ \left[v_n\right] &\neq 0 \\ \left[v_t\right] &= 0 \end{aligned} \tag{2.38}$$

La llamada *condición de entropía* da la información necesaria para poder filtrar de todas las posibles soluciones aquellas que no tengan sentido físico. Esto está fuera de los alcances de este curso. Del mismo modo no será tratado aquía el tema de la formulación de flujo invíscido rotacional mediante la representación de *Clebsch* que permite una descripción más económica en término de cómputo pero más compleja en cuanto a su interpretación.

2.3. Flujo potencial

Si a la hipótesis del flujo invíscido le agregamos la *irrotacional* lo cual matemáticamente podemos expresar como:

$$\vec{\zeta} = \vec{\nabla} \times \vec{v} = 0$$

el campo tridimensional de velocidades puede ser descripto por una única función escalar ϕ, llamada función potencial y definida por:

$$\vec{v} = \vec{\nabla}\phi$$

lo cual reduce notoriamente el cálculo. Si las condiciones iniciales son compatibles con un campo de entropía uniforme, luego para flujos continuos la entropía será constante en todo el dominio y las ecuaciones de momento se transforman en

$$\frac{\partial}{\partial t}(\vec{\nabla}\phi) + \vec{\nabla}H = 0 \tag{2.39}$$

o

$$\frac{\partial \phi}{\partial t} + H = H_0 \qquad \text{constante}$$

donde la constante H_0 es la entalpía de todas las líneas de corriente. De esta forma la ecuación de energía ya no será independiente del resto. La ecuación para flujo potencial surge de la ecuación de continuidad tomando en cuenta la hipótesis de flujo isoentrópico para poder expresar la densidad como una función de la velocidad, o sea del gradiente del potencial.

En forma de conservación tenemos

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{\nabla}\phi) = 0 \tag{2.40}$$

que junto con la relación entre densidad y función potencial

$$\frac{\rho}{\rho_A} = (\frac{h}{h_A})^{1/(\gamma-1)} = \Big[\Big(H_0 - \vec{v}^2/2 - \frac{\partial \phi}{\partial t}\Big)/h_A\Big]^{1/(\gamma-1)} \tag{2.41}$$

completan el modelo. ρ_A y h_A son valores de referencia para la densidad y la entalpía que a menudo corresponden con las condiciones de estancamiento.

Caso estacionario

$$\vec{\nabla} \cdot (\rho \vec{\nabla}\phi) = 0 \tag{2.42}$$

y

$$\frac{\rho}{\rho_A} = \Big(1 - \frac{(\vec{\nabla}\phi)^2}{2H_0}\Big)^{1/(\gamma-1)} \tag{2.43}$$

surgen como el modelo a resolver.

Algunos temas a agregar más adelante son:
Flujo irrotacional con circulación - condición de Kutta-Joukowski
Limitaciones del modelo de flujo potencial para flujos transónicos

2.3.1. Aproximación de pequeñas pertubaciones

En flujos estacionarios o no estacionarios alrededor de perfiles alares con un espesor mucho menor que su cuerda se puede asumir que las velocidades en el plano normal son despreciables simplificando aún más el modelo. Asi surge el modelo de *pequeñas perturbaciones* que matemáticamente se escribe como:

$$(1 - M_\infty^2)\phi_{xx} + \phi_{yy} + \phi_{zz} = \frac{1}{a^2}(\phi tt + 2\phi_x\phi_{xt}) \tag{2.44}$$

2.3.2. Flujo potencial linealizado

Si el flujo es incompresible hemos visto que la ecuación de continuidad se expresa como

$$\vec{\nabla} \cdot \vec{v} = 0 \tag{2.45}$$

y por la hipótesis del modelo potencial $\vec{v} = \vec{\nabla}\phi$ entonces (2.45) se transforma en

$$\Delta\phi = 0 \tag{2.46}$$

la ecuación de Laplace.

Otra forma de linealizar el problema es usando superposición de flujos simples, como fuentes, sumideros y vórtices calibrando los coeficientes con que cada uno participa de forma de ajustar la condición de contorno de velocidad normal nula sobre cuerpos sólidos.

Los métodos basados en singularidad provienen de plantear el problema en forma integral, usar núcleos (“ kernels “) que representen la influencia espacial y resolver numéricamente. Estos métodos son llamados singulares porque casualmente producen integrales singulares cuyo tratamiento no es trivial y está fuera del alcance del curso. Métodos como el de los *paneles* o el método de los *elementos de borde* son los máximos representantes de ésta técnica.

Capítulo 3

Naturaleza matemática de las ecuaciones de la mecánica de fluidos

3.1. Introducción

Para comenzar resumimos las varias aproximaciones vistas en el capítulo anterior de forma tal de mostrar las ecuaciones que surgen de los modelos vistos.

En general las leyes de conservación plantean una ecuación del tipo

$$\frac{\partial U}{\partial t} + \vec{\nabla} \cdot \vec{F} = Q \qquad (II.1.e)$$

que expresada en sus tres coordenadas cartesianas dan lugar a

$$\frac{\partial U}{\partial t} + \frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + \frac{\partial h}{\partial z} = Q \qquad (II.1.f)$$

Las ecuaciones de Navier-Stokes se pueden expresar en su forma general como

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho\vec{v} \\ \rho E \end{pmatrix} + \vec{\nabla} \cdot \begin{pmatrix} \rho\vec{v} \\ \rho\vec{v} \otimes \vec{v} + p\overline{\overline{I}} - \overline{\overline{\tau}} \\ \rho\vec{v}H - \overline{\overline{\tau}} \cdot \vec{v} - k\vec{\nabla}T \end{pmatrix} = \begin{pmatrix} 0 \\ \rho\vec{f_e} \\ W_f + q_H \end{pmatrix} \qquad (3.1)$$

En el caso incompresible se transforman en

$$\vec{\nabla} \cdot \vec{v} = 0 \qquad (II.2.a)$$

$$\frac{\partial \vec{v}}{\partial t} + (\vec{v} \cdot \vec{\nabla})\vec{v} = -\frac{1}{\rho}\vec{\nabla}p + \nu\Delta\vec{v} + \vec{f_e} \qquad (II.2.b)$$

donde es muy común desacoplar el problema en una solución para la velocidad y otra para la presión. Para esta última surge

$$\frac{1}{\rho}\Delta p = -\vec{\nabla} \cdot (\vec{v} \cdot \vec{\nabla})\vec{v} + \vec{\nabla} \cdot \vec{f_e} \qquad (II.2.d)$$

En la aproximación *"Thin shear layer (TSL)"* hemos visto

$$f = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ \rho uH \end{pmatrix}$$

$$g = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ \rho vH \end{pmatrix} \tag{3.2}$$

$$h = \begin{pmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ \rho wH \end{pmatrix} + \begin{pmatrix} 0 \\ -\mu \frac{\partial u}{\partial z} \\ -\mu \frac{\partial v}{\partial z} \\ -\frac{4}{3}\mu \frac{\partial w}{\partial z} \\ -\mu(u\frac{\partial u}{\partial z} + v\frac{\partial v}{\partial z} + \frac{4}{3}w\frac{\partial w}{\partial z}) - k\frac{\partial T}{\partial z} \end{pmatrix}$$

mientras que en el caso parabolizado cambia solamente la ecuación de momento según la dirección preferencial del flujo por otra del tipo

$$\frac{\partial}{\partial x}(\rho u^2 + p) + \frac{\partial}{\partial y}(\rho uv) + \frac{\partial}{\partial z}(\rho uw) = \frac{\partial}{\partial y}(\mu \frac{\partial u}{\partial y}) + \frac{\partial}{\partial z}(\mu \frac{\partial u}{\partial z}) \tag{II.5.1}$$

y la de energía por la siguiente

$$\frac{\partial \rho uH}{\partial x} + \frac{\partial \rho vH}{\partial y} + \frac{\partial \rho wH}{\partial z} = \frac{\partial}{\partial y}(\overline{\overline{\tau}} \cdot \vec{v})_y + \frac{\partial}{\partial z}(\overline{\overline{\tau}} \cdot \vec{v})_z + \frac{\partial}{\partial y}(k\frac{\partial}{\partial y}) + \frac{\partial}{\partial z}(k\frac{\partial}{\partial z}) \tag{II.5.2}$$

En el modelo invíscido

$$f = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ \rho uw \\ \rho uH \end{pmatrix} \qquad g = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ \rho vw \\ \rho vH \end{pmatrix} \qquad h = \begin{pmatrix} \rho w \\ \rho uw \\ \rho vw \\ \rho w^2 + p \\ \rho wH \end{pmatrix} \tag{II.6.1}$$

y en flujo potencial tenemos el caso general

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{\nabla} \phi) = 0 \tag{II.8.2}$$

junto con la relación entre densidad y función potencial

$$\frac{\rho}{\rho_A} = (\frac{h}{h_A})^{1/(\gamma-1)} = \Big[\Big(H_0 - \vec{v}^2/2 - \frac{\partial \phi}{\partial t}\Big)/h_A\Big]^{1/(\gamma-1)} \tag{II.8.3}$$

y los casos de pequeñas perturbaciones

$$(1 - M_\infty^2)\phi_{xx} + \phi_{yy} + \phi_{zz} = \frac{1}{a^2}(\phi tt + 2\phi_x \phi_{xt}) \tag{II.8.6}$$

o el caso linealizado con

$$\Delta\phi = 0 \qquad (II.8.7)$$

Examinando las anteriores ecuaciones provenientes de los modelos físicos asumiendo distintos niveles de aproximación dinámica podemos decir que todas se representan por un conjunto de ecuaciones a derivadas parciales cuasi-lineales de primer o a lo sumo segundo órden.

Por lo visto al comienzo el modelo matemático refleja un balance entre flujos convectivos, difusivos y diversas fuentes internas y externas que surgen a partir del modelo planteado por la física.

Los *flujos difusivos* aparecen con operadores de segundo órden como una consecuencia de la ley generalizada de Fick, con una tendencia a suavizar gradientes.

Los *flujos convectivos* aparecen con operadores de primer órden y expresan el transporte de la propiedad.

Por lo tanto la competencia entre ellos influenciará sobre la naturaleza matemática de las ecuaciones, desde las ecuaciones elípticas, parabólicas hasta las hiperbólicas. Del mismo modo es la persona que modela la que al introducir una aproximación está forzando sobre el tipo de ecuación con que se enfrentará.

Tomemos a modo de ejemplo de esto la componente x de la ecuación de momento de las ecuaciones de Navier-Stokes asumiendo flujo laminar e incompresible.

$$\rho\frac{\partial u}{\partial t} + \rho(\vec{v}\cdot\vec{\nabla})u = -\frac{\partial p}{\partial x} + \mu\Delta u \qquad (3.3)$$

Adimensionalizar estas ecuaciones es un modo de ver mejor aquello de la competencia. Para ello tomemos una longitud de referencia L, un tiempo característico T, una escala de velocidades V y una de presiones ρV^2. Reemplazando como es habitual en (4.1.1) surge

$$\frac{VT}{L}\rho\frac{\partial u}{\partial t} + (\vec{v}\cdot\vec{\nabla})u = -\frac{\partial p}{\partial x} + \frac{1}{Re}\Delta u \qquad (3.4)$$

donde $Re = \frac{\rho VL}{\mu} = \frac{VL}{\nu}$ es el número de Reynolds y todas las variables, las independientes como las dependientes, se expresan en la versión adimensionalizada.

Para $Re \to 0$, flujos fuertemente viscosos, denominados reptantes, el término convectivo y no lineal es despreciable y surgen las ecuaciones de *Stokes*

$$-\frac{V^2T}{\nu}\rho\frac{\partial u}{\partial t} + \Delta u = Re\frac{\partial p}{\partial x} \qquad (3.5)$$

Asumimos que el gradiente de presión está fijado para poder reducir este caso ejemplificatorio al caso de una ecuación y no entar en las complejidades que presupone un sistema de ecuaciones con varias incógnitas. Asi como están escritas son de naturaleza parabólica debido a la existencia de un operador de primer órden para una de las variables independientes, el tiempo, y uno de segundo órden para la otra, la coordenada x. Si agregamos la hipótesis de flujo estacionario entonces desaparece el término temporal y la ecuación se transforma en elíptica *(ec. de Poisson)*

Si $Re \to \infty$ y si nos ubicamos fuera de la capa límite, los términos viscosos tendrán muy poca influencia haciendo que el flujo sea dominado por los términos convectivos (ecs. de Euler).

$$\rho\frac{\partial u}{\partial t} + \rho(\vec{v}\cdot\vec{\nabla})u = -\frac{\partial p}{\partial x} \qquad (3.6)$$

Si reducimos el problema al caso unidimensional la anterior se vuelve

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\frac{1}{\rho}\frac{\partial p}{\partial x} \tag{3.7}$$

que es una ecuación hiperbólica que describe un fenómeno de propagación.

Es de gran importancia la distinción entre fenómenos de difusión (elípticos) y de propagación (hiperbólicos), ya que en los primeros la información se propaga en todas direcciones mientras que en los últimos existe una dirección preferencial y la información se propaga solo a determinadas regiones. Entre ambos existen las ecuaciones parabólicas que amortiguan en el tiempo los efectos difusivos. Las ecuaciones de Navier-Stokes son de carácter parabólicas en el espacio y en el tiempo y cambian de tipo al caso elíptico en el estado estacionario. De todas maneras es importante resaltar que la ecuación de continuidad al no contar con difusión es de carácter hiperbólico lo cual hace que, estrictamente hablando, las ecuaciones de Navier-Stokes sean incompletamente parabólicas.

3.2. Superficies características. Soluciones del tipo ondas

Las ecuaciones a derivadas parciales que describen los niveles de aproximación vistos en el capítulo anterior son cuasi-lineales y a lo sumo de segundo órden. No obstante se sabe que en muchos casos existe una forma no degenerada de llevar un sistema de segundo órden a otro de primer órden. Asumimos que contamos con la transformación necesaria y que tenemos entre manos un sistema cuasi-lineal de primer órden. La clasificación de las ecuaciones está íntimamente relacionada con el concepto de *características*, definida como hipersuperficies a lo largo de la cual ciertas propiedades del flujo permanecen invariables o ciertas derivadas pueden volverse discontinuas.

Un sistema de ecuaciones a derivadas parciales de primer órden se dice hiperbólico si su parte homogénea admite solución del tipo ondulatoria. Ya veremos que esto está asociado con el hecho que los autovalores del sistema son reales. Por otro lado si el sistema admite solución del tipo ondas amortiguadas (evanescentes) el sistema será parabólico y si no admite solución del tipo ondulatorio será elíptico y dominado por difusión.

3.3. Ecuaciones diferenciales parciales de segundo órden

Sea el siguiente operador cuasi-lineal de segundo órden

$$a\frac{\partial^2 \phi}{\partial x^2} + 2b\frac{\partial^2 \phi}{\partial x \partial y} + c\frac{\partial^2 \phi}{\partial y^2} = 0 \tag{3.8}$$

con $a, b, c = f(x, y, \phi, \vec{\nabla}\phi)$. Podemos escribirla como un sistema introduciendo las siguientes variables :

$$\begin{gathered} u = \frac{\partial \phi}{\partial x} \qquad ; \qquad v = \frac{\partial \phi}{\partial y} \\ a\frac{\partial u}{\partial x} + 2b\frac{\partial u}{\partial y} + c\frac{\partial v}{\partial y} = 0 \\ \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0 \end{gathered} \tag{3.9}$$

que en forma matricial se escribe como

$$\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 2b & c \\ -1 & 0 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} u \\ v \end{pmatrix} = 0$$
$$A^1 \frac{\partial U}{\partial x} + A^2 \frac{\partial U}{\partial y} = 0 \tag{3.10}$$

Si la solución es expresable como una onda plana que se propaga en la dirección $\vec{n}$, ésta tendrá la forma

$$U = \hat{U} e^{I(\vec{n}\cdot\vec{x})} = \hat{U} e^{I(n_x x + n_y y)} \tag{3.11}$$

con $I = \sqrt{-1}$.

Reemplazando la solución(4.2.3) en (4.2.2) y tomando la parte homogénea nos queda

$$(A^1 n_x + A^2 n_y)\hat{U} = 0 \tag{3.12}$$

que admitirá solución no trivial solo si el determinante del sistema es nulo,

$$\mathsf{d}et|A^1 n_x + A^2 n_y| = 0 \tag{3.13}$$

O sea que las raices de

$$a(\frac{n_x}{n_y})^2 + 2b(\frac{n_x}{n_y}) + c = 0 \tag{3.14}$$

nos darán la respuesta a si la solución es expresable como una onda, o sea si es hiperbólica o si se trata de una onda pero amortiguada (parabólica) o si no existe solución real en cuyo caso es elíptica.

Resolviendo,

- $b^2 - ac > 0$ dos soluciones reales, dos ondas (hiperbólico)
- $b^2 - ac = 0$ una única solución (parabólico)
- $b^2 - ac < 0$ dos soluciones complejas conjugadas, (elíptico)

En lo anterior hemos asumido una solución del tipo onda plana. Esto puede generalizarse a hipersuperficies generales con una representación no lineal de la onda. Esta consiste en definir una superficie que funciona como frente de onda $S(x, y)$ y la solución se representa por una onda general del tipo

$$U = \hat{U} e^{IS(x,y)} \tag{3.15}$$

A partir de aquí la operatoria es la misma que en el caso de onda plana. Un sistema es hiperbólico si (4.2.7) es una solución para valores reales de $S(x, y)$

Esto se logra calculando

$$\mathsf{d}et|A^1 S_x + A^2 S_y| = 0 \tag{3.16}$$

donde S_x, S_y son las derivadas de la superficie respecto a las direcciones que caen sobre ella.

Las superficies $S(x, y)$ que hacen el anterior determinante nulo son superficies características con una normal que apunta segun $\vec{\nabla} S$.

Ejemplo 1: Flujo potencial estacionario Llamemos c a la velocidad del sonido. Si tomamos la ecuación del modelo de flujo potencial y trabajamos algebraicamente sobre la relación entre densidad y función potencial asumiendo transformaciones isoentrópicas llegamos a

$$(\delta_{ij} - M_i M_j)\frac{\partial^2 \phi}{\partial x_i dx_j} = \frac{1}{c^2}\Big[\frac{\partial^2 \phi}{\partial t^2} + \frac{\partial}{\partial t}(\vec{\nabla}\phi)^2\Big] \tag{3.17}$$

que en el caso 2D estacionario se simplifica a

$$(1 - \frac{u^2}{c^2})\frac{\partial^2 \phi}{\partial x^2} - \frac{2uv}{c^2}\frac{\partial^2 \phi}{\partial x dy} + (1 - \frac{v^2}{c^2})\frac{\partial^2 \phi}{\partial y^2} = 0 \tag{3.18}$$

que bajo la forma de una ecuación general de segundo órden (4.2.2) posee los siguientes coeficientes:

$$\begin{aligned} a &= 1 - \frac{u^2}{c^2} \\ b &= -\frac{uv}{c^2} \\ c &= 1 - \frac{v^2}{c^2} \end{aligned} \tag{3.19}$$

Entonces el discriminante $b^2 - 4ac$ se vuelve

$$b^2 - 4ac = \frac{u^2 + v^2}{c^2} - 1 = M^2 - 1$$

lo cual es negativo (elíptico) en el caso subsónico y positivo (hiperbólico) en el caso supersónico. En el caso transónico el problema se transforma en parabólico. Como vemos el problema potencial tiene una naturaleza bastante variada dependiendo del régimen de velocidades que estemos resolviendo. Una complicación adicional a esto aparece por el hecho que en el caso transónico y supersónico pueden aparecen discontinuidades como ondas de choque con un tratamiento numérico bien particular.

Ejemplo 2: La ecuación potencial en pequeñas perturbaciones Como hemos visto si la componente transversal al flujo del vector velocidad es despreciable, la ecuación potencial estacionaria se reduce a

$$(1 - M_\infty^2)\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0 \tag{3.20}$$

entonces las raices de la ecuación (4.2.6) son

$$\frac{n_y}{n_x} = \pm\sqrt{M_\infty^2 - 1}$$

que definen las normales a las dos características para flujos supersónicos. Estas se obtienen como

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \pm 1/\sqrt{(M_\infty^2 - 1)} = \pm \tan\mu$$

Estas características son idénticas a las líneas de Mach que se hallan a un ángulo μ respecto a la dirección del vector velocidad, con

$$\sin\mu = 1/M_\infty$$

3.4. Definición general de superficie característica

Sea un sistema de n ecuaciones diferenciales parciales de primer órden para las n incógnitas $u^i, i = 1, \ldots, n$ en el espacio m dimensional $x^k, k = 1, \ldots, m$ incluyendo eventualmente la variable tiempo. Escrita en forma de conservación y usando la convención de suma sobre índices repetidos tenemos

$$\frac{\partial}{\partial x^k} F_i^k = Q_i \qquad\qquad k = 1, \ldots, m; \qquad i = 1, \ldots, n \tag{3.21}$$

El análisis de las propiedades del sistema recae sobre la forma cuasi-lineal obtenida después de introducir las matrices jacobianas A^k definidas por

$$A_{ij}^k = \frac{\partial F_i^k}{\partial u^j} \tag{3.22}$$

Por lo tanto el sistema (4.5.1) toma la forma cuasi-lineal

$$A_{ij}^k \frac{\partial u^j}{\partial x^k} = Q_i, \quad k = 1, \ldots, m; \quad i = 1, \ldots, n \tag{3.23}$$

que en forma condensada se escribe como

$$A^k \frac{\partial U}{\partial x^k} = Q, \quad k = 1, \ldots, m$$

donde el vector columna U es de $(n \times 1)$ y contiene las incógnitas u^j, mientras que A^k son matrices de $n \times n$ y Q un vector de términos fuentes. Las matrices A^k y el vector Q pueden depender de x^k y de U pero no del gradiente de U.

Una solución expresada como una onda plana de la forma (4.2.3) o en general (4.2.7) existirá si el sistema homogéneo

$$A^k \frac{\partial S}{\partial x^k} \hat{U} = A^k n_k \hat{U} = 0 \qquad\qquad k = 1, \ldots, m$$

admite soluciones no triviales, lo cual como antes redunda en que

$$\det |A^k n_k| = 0$$

Esta ecuación tiene a los sumo n soluciones, las n superficies características. El sistema se dice *hiperbólico* si todas las características son reales y si las soluciones son linealmente independientes. Si todas son complejas es *elíptico* y si hay de ambas es híbrido. Además si el rango de $A^k n_k$ no es completo el sistema se dice *parabólico*. Esto ocurrirá cuando al menos una de las variables no posea derivadas respecto a alguna de las coordenadas espaciales.

Ejemplo 3 : Sistema de 2 ecs de primer órden en 2D Sea el siguiente sistema de ecuaciones en 2D

$$\begin{aligned} a\frac{\partial u}{\partial x} + c\frac{\partial u}{\partial y} &= f_1 \\ b\frac{\partial u}{\partial x} + d\frac{\partial u}{\partial y} &= f_2 \end{aligned} \tag{3.24}$$

que en forma matricial se puede escribir como

$$\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 & c \\ d & 0 \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

donde

$$A^1 = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad A^2 = \begin{pmatrix} 0 & c \\ d & 0 \end{pmatrix}$$

Planteando el determinante e igualando a cero conduce a las raices:

$$|\frac{n_y}{n_x}|^2 = \frac{cd}{ab}$$

Si $cd/ab > 0$ el sistema es hiperbólico, por ejemplo tomemos $a = b = c = d = 1$, en este caso el sistema de ecuaciones es la conocida *ecuación de las ondas* que escrita como una ecuación de segundo órden luce como

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0$$

Si $cd/ab < 0$ el sistema es elíptico, por ejemplo tomemos $a = b = 1 \; ; \; c = -d = -1$, en este caso el sistema de ecuaciones es la conocida *ecuación de difusión o de Laplace*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Finalmente si $b = 0$ existe una sola característica y el sistema es parabólico. Con $a = 1 \; ; b = 0 \; ; c = -d = -1 \; ; f_1 = 0 \; ; f_2 = v$ obtenemos la conocida *ecuación del calor*

$$\frac{\partial u}{\partial x} = \frac{\partial^2 u}{\partial y^2}$$

Ejemplo 4: Ecuaciones en aguas poco profundas estacionaria Este sistema, conocido como *"shallow water equations"*, describe la distribución espacial de las alturas de la superficie libre de una corriente de agua que posee un campo de velocidades (u, v). Si llamamos g a la aceleración de la gravedad, entonces:

$$\begin{aligned} u\frac{\partial h}{\partial x} + v\frac{\partial h}{\partial y} + h\frac{\partial u}{\partial x} + h\frac{\partial v}{\partial y} &= 0 \\ u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + g\frac{\partial h}{\partial x} &= 0 \\ u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + g\frac{\partial h}{\partial y} &= 0 \end{aligned} \tag{3.25}$$

Introduciendo el vector de incógnitas

$$U = \begin{pmatrix} h \\ u \\ v \end{pmatrix}$$

el sistema se puede escribir en forma matricial como

$$\begin{pmatrix} u & h & 0 \\ g & u & 0 \\ 0 & 0 & u \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} h \\ u \\ v \end{pmatrix} + \begin{pmatrix} v & 0 & h \\ 0 & v & 0 \\ g & 0 & v \end{pmatrix} \frac{\partial}{\partial y} \begin{pmatrix} h \\ u \\ v \end{pmatrix} = 0$$

o en forma compacta

$$A^1 \frac{\partial U}{\partial x} + A^2 \frac{\partial U}{\partial y} = 0$$

Las tres superficies características se obtienen como la solución de plantear el siguiente determinante nulo, con $\lambda = n_x/n_y$

$$\det \begin{vmatrix} u\lambda + v & h\lambda & h \\ g\lambda & u\lambda + v & 0 \\ g & 0 & u\lambda + v \end{vmatrix} = 0$$

Trabajando sobre el determinante conduce a las soluciones :

$$\begin{aligned} \lambda^{(1)} &= -\frac{v}{u} \\ \lambda^{(2),(3)} &= \frac{-uv \pm \sqrt{u^2 + v^2 - gh}}{u^2 - gh} \end{aligned} \tag{3.26}$$

Como vemos $\sqrt{gh}$ juega el mismo rol que la velocidad del sonido en el caso de las ecuaciones de flujo compresible, y como allí , existe un valor crítico donde el sistema cambia de tipo. Este valor se llama velocidad supercrítica y en el caso que $\vec{v}^2 = u^2 + v^2 > gh$ el sistema se vuelve *hiperbólico*.

En el caso de velocidades subcrítico el sistema es *híbrido* ya que habrá 2 soluciones complejas conjugadas y $\lambda^{(1)}$ que siempre es real. La superficie característica asociada con $\lambda^{(1)}$ es la línea de corriente, ya que $n_x^{(1)} = -v$; $n_y^{(1)} = u$.

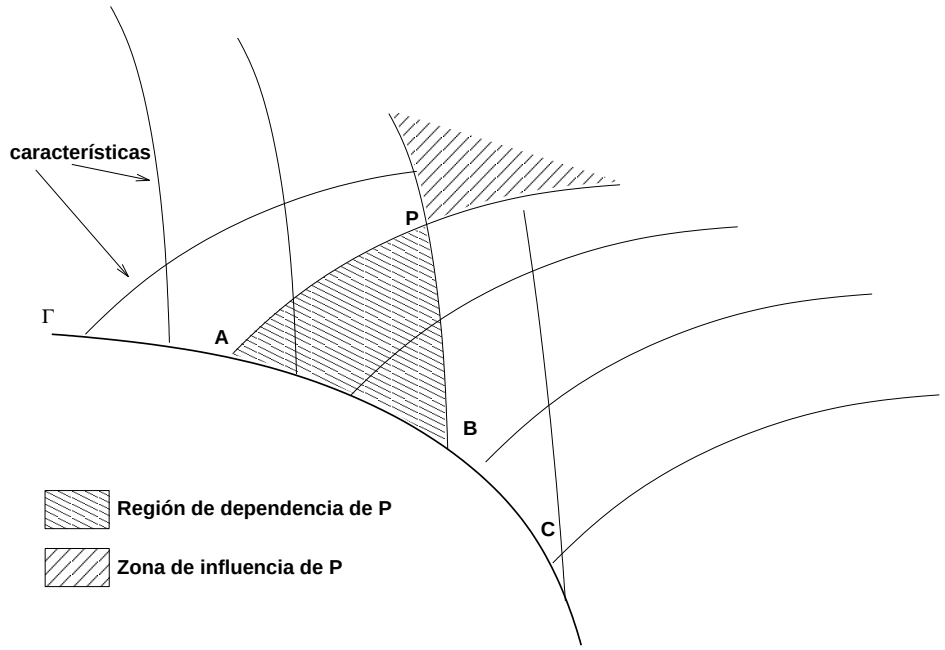
3.5. Dominio de dependencia - zona de influencia

Las propiedades de propagación de los problemas hiperbólicos tiene importantes consecuencias respecto a la forma en que la información se propaga o transmite por todo el dominio.

Si consideramos el caso escalar bidimensional presentado en (4.2.1) este generará dos características en el caso hiperbólico donde cada una se representa por una familia de curvas. Si tomamos 1 miembro de cada familia y tomamos un punto P donde se intersectan, vemos que ambas características dejan una región aguas arriba que afectará la solución en el punto P y una zona aguas abajo que dependerá del valor de la función en el punto P. Mirándolo desde el punto P, la primera se llama *zona de dependencia* del punto P y la segunda *zona de influencia* del punto P. Este hecho es muy importante y tiene muchas consecuencias matemáticas, físicas y numéricas. (figura 3.5)

En el caso de problemas parabólicos ambas características degeneran en una (el rango del sistema no es completo) y en este caso la zona de dependencia cae sobre la característica mientras que la de influencia es la región completa aguas abajo de la característica.

En el caso elíptico no existe superficie característica que separe el dominio en zonas de dependencia e influencia. Esto produce que la zona de influencia, la de dependencia y el dominio coincidan y la información se propaga en todas direcciones.



Dominio de dependencia e influencia

3.6. Condiciones de contorno e iniciales

Para que los anteriores sistemas de ecuaciones que surgen del modelo físico estén finalmente bien planteados en el sentido matemático es necesario que se impongan ciertas condiciones sobre los datos. Si alguna de las variables independientes del problema es el tiempo es necesario establecer una condición a tiempo $t = 0 \; \forall \vec{x}$, problema llamado *de Cauchy* o de *valores iniciales*. Cuando la variable espacial x está acotado en el espacio por un borde o contorno, el problema suele llamarse de *valores de frontera e iniciales*.

Sin entrar en detalles acerca del tema podemos decir que la correcta especificación de las condiciones de contorno e iniciales requiere de un detallado estudio que incluye a las autofunciones del sistema.

De todos modos este es un tema abierto ya que en la mayoría de los casos que trata la mecánica de fluidos no existen resultados contundentes acerca de como tratar estas condiciones.

Sin ir tan lejos terminamos esta sección diciendo que en los problemas independientes del tiempo de carácter elíptico es usual imponer algunos valores sobre algunos nodos *(Dirichlet)* o sobre su derivada *(Neumann)*. En el primer caso puede tratarse de imponer el vector velocidad sobre una pared sólida o la temperatura, mientras que en el caso Neumann estamos hablando de imponer que el flujo térmico asuma alguna ley de transferencia en el contorno (convección, radiación,etc).

En casos donde el sistema se transforma en primer órden (Euler), la velocidad no puede ser fijada en un contorno sólido, recurriendo a fijar solo la componente normal, dejando la tangencial como parte del cálculo (condición slip). En el caso de superficies libres es usual establecer la continuidad de las tensiones normales y tangenciales.

INTRODUCCION AL USO DE MATLAB

Objetivos: Aprender el uso del lenguaje *MatLab* muy poderoso en cuanto a sus posibilidades de uso como herramienta de cálculo numérico y visualización gráfica asi como de programar en este entorno extendiendo su aplicación al análisis numérico y simulación en Ingeniería.

3.6.1. Introducción

MatLab es un lenguaje de programación interpretado de bajo nivel que a diferencia de los lenguajes compilados permite gran flexibilidad para programar, debuggear y correr pequeñas aplicaciones respecto a los tradicionalmente rápidos pero muy rígidos lenguajes compilados. Además es un software de aplicación y como su nombre lo indica es un laboratorio de matrices (Mat=Matrices / Lab =Laboratorio). Otra de las importantes características es que incluye poderosas facilidades gráficas que se pueden correr run-time, mientras que en el caso de los lenguajes compilados uno debe o bien generar un software gráfico completo o sino adaptar algunas rutinas para los casos particulares de aplicación, lo cual lo transforma en una gran complicación. Además el hecho que los cálculos y las gráficas se vayan generando en tiempo de ejecución lo hace muy flexible. Otras de la características interesantes de este lenguaje interpretado es que equivale a un entorno en sí mismo, o sea permite ejecutar comandos en forma muy interactiva lo cual lo hace apto para ir avanzando en lo que uno está haciendo, viendo resultados en pantalla de lo que uno está obteniendo, corrigiendo rumbos, etc. Además cuenta con una enorme biblioteca de rutinas de cálculo que lo transforma en un lenguaje de bajo nivel sin necesidad de tener que programar demasiado, salvo aplicaciones especiales. Hemos dicho al comienzo que una de las principales diferencias entre un lenguaje compilado y uno interpretado es la velocidad de procesamiento de datos, especialmente para grandes problemas. El programa compilado arma un ejecutable habiendo interpretado cada sentencia previamente en la etapa de compilación y genera un archivo binario optimizado. En cambio el lenguaje interpretado realiza la interpretación de cada sentencia a medida que corre lo cual lo hace lento. Por ejemplo si uno realiza un loop de 1000 iteraciones el intérprete debe trabajar 1000 veces haciendo lo mismo. MatLab soporta vectorización lo cual hace que en lugar de manejar escalares e interpretar operaciones entre escalares pueda manejar vectores e interpretar vectores. Por ejemplo si tenemos un vector de 1000 elementos y queremos hacer cálculos con cada uno de sus coeficientes, en lugar de hacer un loop de 1000 iteraciones podemos invocar directamente la operación sobre todo el vector. Por ejemplo, si tenemos una matriz A de (1000×2), donde cada columna representa un vector de 1000 elementos y queremos calcular la suma de ellos podemos hacer:

```
for k=1:1000,
   c(k) = A(k,1)+A(k,2);
end
```

o sencillamente

```
c = A(:,1)+A(:,2);
```

Todo esto y muchas cosas más que surgen cuando uno lo utiliza hacen a MatLab una muy interesante herramienta para la investigación en torno a los métodos numéricos.

MatLab cuenta con muchas rutinas propias llamadas *built in functions*, solo basta con recorrer el help para ver la enorme cantidad de ellas. Además cuenta con lo que se llaman *Toolkits* que son paquetes de rutinas para aplicaciones particulares. Aquí no entraremos en detalle sobre estas últimas sino que sólo mencionamos

su existencia para aquellos curiosos que puedan interesarse. A grandes rasgos podemos decir que MatLab es un lenguaje de programación integrado con un software de aplicación. Entonces podemos hacer una división inicial en:

$$\text{Matlab} \begin{cases} \text{software de aplicación} \begin{cases} \text{cálculo numérico} \\ \text{visualización gráfica} \end{cases} \\ \text{lenguaje de programación} \end{cases}$$

A continuación planeamos hacer un pequeño *tour* por las galerías de MatLab tratando de tocar sólo algunos de sus puntos con el objetivo de motivar a los turistas a volver a ellos en el futuro.

3.6.2. MatLab como software de aplicación

En esta parte de la visita a MatLab investigaremos sus capacidades como software de aplicación, sin entrar por ahora en detalles de programación. Como fue dicho en la introducción este software es un *laboratorio de matrices* y como tal debemos saber como definir matrices, vectores como un caso particular de una matriz, recordar algunos aspectos del álgebra de espacios vectoriales para poderlas utilizar en forma coherente y finalmente ver las potencialidades que tiene esto. Yendo más allá en cuanto a aplicaciones hemos mencionado las capacidades gráficas del software. A este respecto debemos primero saber como convertir matrices en mapas redefiniendo el álgebra y viendo como de esta forma es posible visualizar funciones en varias variables y hacer distintos tipos de gráficos muy útiles en general.

(1) Manejo standard de matrices, vectores y escalares

Una matriz es el átomo de este software, es algo asi como la menor porción divisible del cálculo, siendo los escalares y los vectores sólo casos particulares. Por ejemplo para ingresar la matriz

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \tag{3.27}$$

debemos ingresar sus coeficientes como una lista de números separados por un blanco al menos que corresponden a los elementos de cada fila, separando las filas entre si con un ; o simplemente con `<Enter>`, o sea

```
A = [  1   2   3 ;
       4   5   6 ;
       7   8   9 ];
```

El `;` despues del `]` finaliza la sentencia sin mostrar por pantalla lo ingresado. Si uno quiere ver lo que ingresa hay que remover el `;`. Su uso es muy importante cuando las matrices son muy grandes ya que la salida por pantalla puede tomar bastante tiempo.

Un vector es un caso particular de lo anterior, este puede ser un vector fila o un vector columna. Sean

$$\begin{aligned} b &= \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} \\ c &= \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} \end{aligned} \tag{3.28}$$

En estos casos el ingreso es

```
b = [  1  2  3 ];

c = [ 4 ;
      5 ;
      6 ];
```

o simplemente podemos definir

```
b = [  4  5  6 ];
c = b';
```

donde el símbolo ' significa la transpuesta, de un vector como en este caso pero su uso es en general para cualquier matriz.

Todas los cálculos que se van haciendo si se asignan a variables permanecen en la memoria de trabajo, sino se asignan son eliminadas permaneciendo solo en memoria el último resultado almacenado en una variable denominada `ans`. Por ejemplo la instrucción siguiente genera un vector fila pero al no estar asignado solo queda en memoria bajo la variable `ans`.

```
[  4  5   6 ];
```

Si a continuación emito otra instrucción pierdo el resultado anterior, salvo que previamente lo haya almacenado en una variable, como por ejemplo

```
b = ans
```

Si hacemos `who` podemos ver las variables en memoria hasta el momento almacenadas y si hacemos `whos` podemos ver más detalles de las mismas.

Como es sabido del algebra lineal una matriz puede generarse como el producto tensorial de vectores. Por ejemplo si en el caso anterior hacemos

```
A1 = c*b
```

el resultado es una matriz de 3×3 porque es el producto de un vector de 3×1 por otro de 1×3. En este caso el resultado será:

```
A1 = [ 4     8    12;
       5    10    15;
       6    12    18;
```

Del mismo modo podemos definir el producto interno entre vectores de una forma trivial. La definición de producto interno es:

$$\Big(b, c\Big) = \sum_{i=1}^{N} b_i c_i$$

En este caso si hacemos

```
b*c
```

obtenemos el producto interno anterior y tal como hemos definido la instrucción no está alojado en ninguna variable por lo cual corre peligro de ser destruido después de la próxima instrucción.

Otro aspecto importante es que existen formas de generar matrices muy especiales. Por ejemplo

- `ones(10)` genera una matriz de (10×10) con elementos unitarios.
- `ones(10,3)` genera una matriz de (10×3) con elementos unitarios.
- `zeros(5,3)` genera una matriz de (5×3) con elementos nulos.
- `(-3:2:15)` genera un vector fila cuyo primer elemento es -3, generando el resto avanzando de a 2 y no supera el valor 15.

Otra forma sería

```
kron((1:3)',(2:5))
```

que produce el producto tensorial del vector columna `(1:3)'` con el vector fila `(2:5)`, lo cual arroja la siguiente matriz como resultado

```
 2     3     4     5
 4     6     8    10
 6     9    12    15
```

Existen muchas otras formas, por ejemplo:

- `rand(4)` produce una matriz de (4×4) con números aleatorios
- `eye(5)` produce la matriz identidad de (5×5)
- `diag(v,idiag)` produce una matriz poniendo el vector `v` en la codiagonal `idiag`, determinando el tamaño de la matriz la longitud del vector v y la codiagonal en la cual se ubica, ya que debe entrar el vector entero en ella. Por ejemplo si el vector es de (5×1) y lo queremos ubicar en la segunda codiagonal entonces la matriz será de (7×7) y tendrá la siguiente forma:

```
>> diag(ones(5,1),2)

 ans =
     0     0     1     0     0     0     0
     0     0     0     1     0     0     0
     0     0     0     0     1     0     0
     0     0     0     0     0     1     0
     0     0     0     0     0     0     1
     0     0     0     0     0     0     0
     0     0     0     0     0     0     0
```

Si no especificamos la codiagonal asume un valor cero lo cual significa la propia diagonal y si especificamos un valor negativo asume las codiagonales inferiores en lugar de las superiores. Es importante resaltar que si el argumento del comando `diag` es una matriz entonces devuelve la diagonal o la codiagonal de la misma como un vector. Vea el help (`help diag`).

A esta altura debemos asegurarnos de salvar lo hecho por si ocurriera algún imprevisto o por si tuviéramos que detener temporariamente nuestro trabajo. El comando `save` salva las variables en memoria en un archivo. Si omitimos el nombre (solo invocamos el comando `save`) guarda el contenido de la memoria en un archivo llamado `matlab.mat`, caso contrario, si especificamos el nombre , por ejemplo `save tempo`, guarda la memoria entera en el archivo `tempo.mat` Tambien podemos guardar parte de la memoria, para ello debemos invocar `save <filename> variables`.

Cuando uno desea continuar con el trabajo interrumpido y quiere recuperar la memoria salvada con el comando `save` debe ejecutar `load <filename>` o simplemente `load` si salvó las variables sin especificar el fichero.

Con `clear` se limpia todo el espacio de trabajo (memoria) y con `clear variables` aquellas variables que uno desea.

Con `size(<arreglo>)` o `length(<arreglo>)` se puede obtener la dimensión del arreglo. El primer caso es general y devuelve dos valores mientras que el segundo es exclusivamente para vectores y devuelve solo un valor.

El software contiene una serie de operadores de matrices tal como suma, resta, producto, potencia, división por derecha e izquierda, etc. Todos estos operadores se definen tal como lo establece el álgebra de los espacios vectoriales. Una aclaración va con la división.

Que significa dividir matrices para MatLab?

Sean A, B dos matrices de $(m \times m)$, entonces

$$\begin{aligned} \texttt{A / B} &= A\,B^{-1} \\ \texttt{A \ B} &= A^{-1}\,B \end{aligned} \tag{3.29}$$

Como vemos la división involucra matrices inversas, pero especial cuidado hay que tener cuando tratamos grandes matrices ya que el tiempo de cálculo se achica mucho en estos casos con los comandos de división comparado al comando de inversión `inv(A)` o A^{-1}

Otro detalle interesante en el manipuleo de matrices es que uno puede extraer submatrices de otras matrices o armar matrices con bloques de otras submatrices. Por ejemplo,

```
>> A  = eye(4) ;
>> B  = ones(4,2) ;
>> C  = zeros(2,4) ;
>> D  = eye(2) ;
>> E = [A B ; C D]

ans =
     1   0   0   0   1   1
     0   1   0   0   1   1
     0   0   1   0   1   1
     0   0   0   1   1   1
     0   0   0   0   1   0
     0   0   0   0   0   1
```

Del mismo modo podemos tomar un bloque de la matriz E, por ejemplo el de las filas 1,3,5 con las columnas 2,3 haciendo:

```
E([1;3;5],[2;3])
```

En cuanto a las capacidades de cálculo además de existir toda una gama de funciones standard para escalares existen otras más para vectores y matrices. Por ejemplo es muy común hablar del seno trigonométrico de un número real. Como se calcula el seno de una matriz?

Bueno existen varias formas de calcularlo pero hay que tener cuidado al hacerlo con MatLab.

Si nosotros definimos por ejemplo una matriz A de (3×3) y ejecutamos la instrucción `sin(A)` el resultado será diferente a si nosotros usamos la definición de una función aplicada a una matriz

$$f(A) = V f(\Lambda) V^{-1} \tag{3.30}$$

donde V es la matriz de autovectores y Λ es la matriz diagonal con los autovalores en la diagonal. En este caso

$$\sin(A) = V \sin(\Lambda) V^{-1} \tag{3.31}$$

Donde está la diferencia?

La diferencia está en que MatLab soporta otro tipo de operadores además de los standard para matrices, llamados operadores de arreglos. Sin entrar por ahora en detalles porque lo abordaremos más adelante, un operador de arreglo realiza una operación sobre cada elemento del arreglo como si fuera una lista de elementos. Entonces si a MatLab le decimos `B = sin(A)` él entiende:

$$B_{ij} = sin(A_{ij}) \tag{3.32}$$

Conclusión, cuando queremos realizar una función especial sobre una matriz debemos o realizar la descomposición en autovalores y autovectores utilizando la función de MatLab `eig(A)` o sino usar otra función llamada `funm` que en este caso se usa de la siguiente forma : `funm(A,'sin')`.

MatLab soporta muchas funciones para matrices, por ejemplo:

- cálculo de determinantes →`det`
- cálculo de autovalores →`eig`
- cálculo de la traza →`trace`
- cálculo del número de condición →`cond`
- cálculo de la norma → `norm`
- cálculo del rango →`rank`
- factorización Choleski →`chol`
- factorización LU →`lu`
- ortogonalización →`orth`

entre otras.

Consulte el help de matrices.

Matrices como listas y tablas

Las matrices también pueden ser interpretadas como listas o tablas de datos. Recién habíamos mencionado que hay que tener cuidado en el uso de MatLab cuando se aplican funciones a matrices porque su resultado puede ser diferente al esperado. Esto se debe a que como dijimos una matriz puede ser una matriz en el sentido estricto o una lista o una tabla. En el caso de listas son como elementos independientes sobre los cuales se pueden aplicar operaciones individualmente con el fin de obtener algún resultado específico. Por ejemplo, una matriz puede representar la distribución de temperatura en un dominio rectangular o mapeable a un rectángulo, donde cada elemento de la matriz puede equivaler a un punto de esa grilla rectangular. Entonces planteando operaciones sobre ellos podemos lograr efectos interesantes.

Por ejemplo sea una matriz A de $(m \times n)$ elementos. Haciendo:

- `diff A` $\Rightarrow B_{i,j} = A_{i,j+1} - A_{i,j}$
- `gradient A` $\Rightarrow \left(\frac{\partial A}{\partial x}, \frac{\partial A}{\partial y}\right)$
- `interp2 A` interpolación de datos en 2D

Del mismo modo dado un conjunto de datos en forma de lista en 1,2 o 3 dimensiones, podemos aplicarle ciertas operaciones e incluso visualizar usando instrucciones gráficas.

En el caso de tablas se puede pensar a una matriz como una planilla de cálculo y realizar toda una serie de análisis estadísticos sobre los mismos y calcular promedios, dispersiones, histogramas, sumas y productos acumulados, medianas, etc.

Es importante a esta altura resaltar que existen operaciones sobre arreglos del mismo modo que existen operaciones sobre matrices. Por ejemplo para multiplicar matrices simplemente invocamos el símbolo del producto. Si queremos multiplicar arreglos esto se entiende como un producto elemento a elemento en forma individual. Por ejemplo, sean A y B dos arreglos, entonces el producto de arreglo se define como:

$$C_{ij} = A_{ij} * B_{ij}$$

Para poder diferenciar esto de la habitual multiplicación de matrices MatLab soporta las operaciones precedidas por un punto.

Por ejemplo

- `.*` equivale al producto de arreglos
- `./` equivale al cociente de arreglos
- `.^n` equivale a elevar todos los coeficientes del arreglo a una potencia `n`

Visualización gráfica

El hecho de permitir un manejo de matrices como si fuera una lista de datos ordenados según la estructura de la matriz o como una tabla de valores permite altas facilidades de graficación.

Visite el help de los comandos `plot`, `mesh`, `surf`, etc para ver todas las capacidades gráficas que ofrece MatLab.

Programación

Finalmente cerramos este breve paseo con otra facilidad muy importante a la hora de pretender ir un poco más alla de lo que ofrece MatLab. Programando en MatLab es posible generar programas de propósitos particulares o generales, como por ejemplo construir un generador de mallas, mejores formas de visualizar soluciones, visualizar mallas de elementos finitos, generar un programa de algún método numérico, etc. Aquí no entraremos en detalle de los aspectos de programación. Para aquellos intersados visitar el help o la bibliografía especializada.

- **[Ejercicio 1.]** Dadas las grandes capacidades que tiene Matlab para el tratamiento de matrices uno de los principales objetivos es aprender como manipularlas. Existen muchas formas de ingresar matrices. En este ejemplo se pretende que Ud explore diferentes formas de hacerlo. Por ejemplo, dada la siguiente matriz de (10×10) con los siguientes elementos:

$$\mathbf{A} = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{3.33}$$

  explique una forma sencilla de cargarla sin necesidad de entrar cada uno de los 100 coeficientes que la conforman.

  A continuación calcule:

  1.- la traspuesta

  2.- el producto consigo misma

  3.- su raiz cuadrada

  4.- sume la matriz identidad del mismo orden

- **[Ejercicio 2.]** A continuación se desea armar una matriz como la siguiente:

$$\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{0}_{10\times 10} \\ \mathbf{0}_{10\times 10} & \mathbf{A} \end{pmatrix} \tag{3.34}$$

  donde $\mathbf{A}$ es la matriz del ejemplo anterior. Muestre una forma sencilla de hacerlo.

- **[Ejercicio 3.]** Tome la matriz del ejemplo 2 y arme una nueva matriz de 3×2 que contenga los elementos de las filas 1,3,5 y columnas 2,4 de la matriz $\mathbf{B}$.

- **[Ejercicio 4.]** Se sabe de la necesidad de contar con un buen manipulador de matrices a la hora de resolver problemas de álgebra lineal. Un ejemplo de ello es la resolución de sistemas de ecuaciones.

Sea la matriz del ejemplo 1 y un vector miembro derecho $\mathbf{b}$ que representa la función $sin(\pi/2 * x)$ donde x es un vector de 10 componentes cuyos valores varían entre 0 y 1. Resolver el sistema $\mathbf{Au} = \mathbf{b}$. Grafique la solución $\mathbf{u}$ y el vector $\mathbf{b}$ en la misma figura.

- **[Ejercicio 5.]** Calcule el determinante de la matriz anterior $\mathbf{A}$ del ejemplo 1 y sus autovalores. A continuación grafique los autovalores en el plano complejo. Posteriormente ingrese la matriz

$$\mathbf{C} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \tag{3.35}$$

  calcule el determinante y sus autovalores y muestre su distribución en el plano complejo. Use el comando `spy(A)` y `spy(C)` y explique que es lo que hace.

- **[Ejercicio 6.]** Para las matrices $\mathbf{A}$ y $\mathbf{C}$ de los ejemplos anteriores calcule la solución del sistema lineal $(\mathbf{A}+\mathbf{C})\mathbf{u} = \mathbf{b}$ con el miembro derecho similar al usado en el Ej. 4. Grafique la solución. A continuación resuelva el sistema$(\mathbf{A} + 10\mathbf{C})\mathbf{u} = \mathbf{b}$ y grafique la solución.

- **[Ejercicio 7.]** Dada las matrices

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \\ \mathbf{B} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \end{aligned} \tag{3.36}$$

  generar con Matlab la matriz

$$\mathbf{C} = \begin{pmatrix} A & B & B & A \\ -A & B & -A & A \\ B & A & -B & A \end{pmatrix} \tag{3.37}$$

  y extraiga de la matriz C aquella correspondiente a las filas 2 y 3 y columnas 5,8,11,12 y muestre su estructura.

- **[Ejercicio 8.]** Dado el vector $x = j^2, j = 1, \ldots, 10$ y el vector $y = j^3, j = -3, \ldots, 3$, calcule el producto tensorial de ambos vectores y muestre la matriz obtenida con el producto tensorial en forma gráfica. *Ayuda:* La matriz es tal que sus elementos se calculan de la siguiente forma:

$$z_{ij} = x_i y_j$$

  y una forma de graficarla es mediante el comando `mesh`

A continuación genere una grilla en 2D con $x \in (-2, 2)$ y $y \in (0, 3)$ y calcule la función $z = x^2 + \sqrt{y}$ y grafique la solución. *Ayuda:* Explore el comando `meshgrid`

- **[Ejercicio 9.]** Confeccione un programa en Matlab que realice el producto vectorial de vectores en tres dimensiones. Pruébelo con un ejemplo no trivial. Extiéndalo al caso de varios vectores. Para esto utilice primero un archivo tipo script y luego uno tipo función. Si no recuerda las diferencias entre ambos revise las notas entregadas (Primer de Matlab).

- **[Ejercicio 10.]** Confeccione un programa en Matlab que resuelva la siguiente ecuación diferencial ordinaria usando la función `ode23` y `ode45`,

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}t} &= ye^{-t} \\ y(t=0) &= 1 \end{aligned} \tag{3.38}$$

- **[Ejercicio 11.]** Resuelve el siguiente sistema no lineal de ecuaciones

$$\begin{aligned} sin(x) + y^2 + log(z) - 7 &= 0 \\ 3 * x + 2^y - z^3 + 1 &= 0 \\ x + y + z - 5 &= 0 \end{aligned} \tag{3.39}$$

usando como estimación inicial

$$\begin{aligned} x &= 1 \\ y &= 1 \\ z &= 1 \end{aligned} \tag{3.40}$$

utilizando la función de Matlab `fsolve` o `fsolve2`

- **[Ejercicio 12.]** Construya un polinomio de cuarto orden

$$p = \sum_{i=0}^{4} a_i x^i$$

y encuentre las raices del mismo usando un conjunto de coeficientes a_i a elección. Grafique el polinomio en el rango donde se hallan todas sus raices.

- **[Ejercicio 13.]** Utilice las rutinas del proyecto sobre sistemas dinámicos masa resorte amortiguador para el caso de 1 grado de libertad lineal y verifique el valor del amortiguamiento crítico del sistema.

- **[Ejercicio 14.]** Genere una aplicación para resolver el ejemplo del péndulo doble y trate de animar el movimiento del sistema

Capítulo 4

Método de diferencias finitas

4.1. Diferencias finitas en 1D

Queremos resolver el campo de temperaturas a través de una pared de material ($-\infty \leq y, z \leq +\infty$, $0 \leq x \leq L_x$). La temperatura en $x = 0, L_x$ es mantenida a $\bar{\phi}_0$, $\bar{\phi}_{L_x}$ y hay una fuente de calor repartida $Q(x)$:

$$\begin{aligned} k\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} &= -Q(x) \\ \phi(0) &= \bar{\phi}_0 \\ \phi(L_x) &= \bar{\phi}_{L_x} \end{aligned} \tag{4.1}$$

Dividimos el intervalo en L segmentos de longitud $\Delta x = L_x/L$ y llamaremos *"nodos"* o *"puntos de la grilla"* a los extremos de los segmentos:

$$x_l = l\Delta x, \quad l = 0, 1, 2, \ldots, L, \quad x_0 = 0, \;\; x_L = L_x \tag{4.2}$$

4.1.1. Desarrollo en Serie de Taylor

Si bien la ecuación que debemos resolver es de segundo orden, empezaremos, por simplicidad, por desarrollar aproximaciones en diferencias para la derivada de primer orden,

$$\begin{aligned} \phi(x_{l+1}) &= \phi(x_l + \Delta x) \\ &= \phi(x_l) + \Delta x \frac{\mathrm{d}\phi}{\mathrm{d}x}\bigg|_{x=x_l} + \frac{\Delta x^2}{2}\left.\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right|_{x=x_l+\theta_1\Delta x} \quad 0 \leq \theta_1 \leq 1 \end{aligned} \tag{4.3}$$

Indicando $f_l = f(x_l)$ para cualquier función $f(x)$, tenemos:

$$\phi_{l+1} = \phi_l + \Delta x \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l + \frac{\Delta x^2}{2}\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_{l+\theta_1} \tag{4.4}$$

Donde el subíndice $l + \theta_1$ es una extensión de la notación que indica evaluación en $(l + \theta_1)\Delta x$. Despejando la derivada de primer orden:

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l = \frac{\phi_{l+1} - \phi_l}{\Delta x} - \frac{\Delta x}{2}\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_{l+\theta_1} \tag{4.5}$$

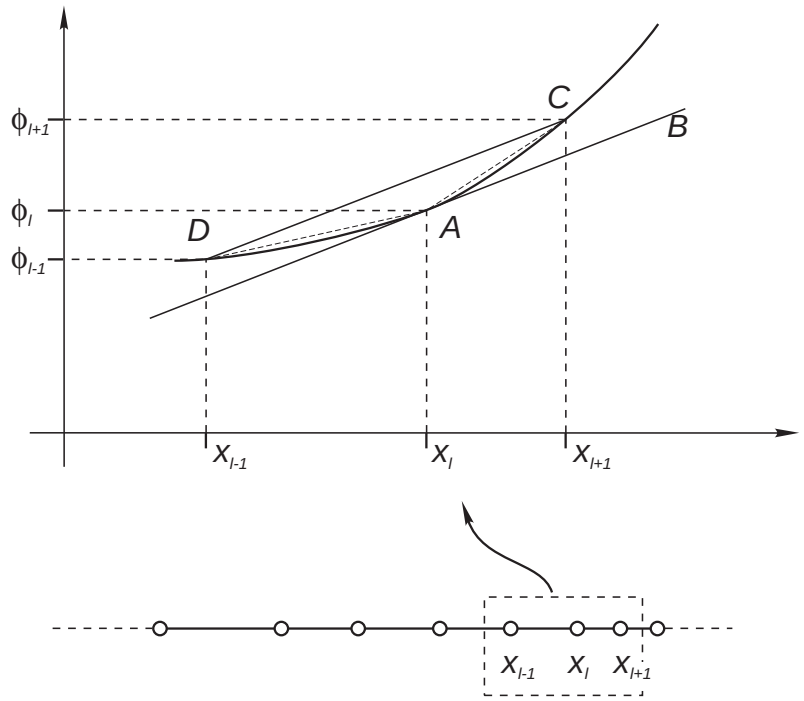


Figura 4.1: Interpretación geométrica de las diferentes aproximaciones por diferencias finitas a la derivada.

A esta aproximación para la derivada de primer orden la llamamos *"por diferencia hacia adelante"* (*forward difference*):

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l \approx \frac{\phi_{l+1}-\phi_l}{\Delta x} \tag{4.6}$$

Véase la figura 4.1 para una interpretación gráfica de esta aproximación. Hemos aproximado la derivada en el punto x_l por la pendiente de la secante a la curva que pasa por los puntos x_l y x_{l+1} (segmento AC). El error de esta aproximación es:

$$\begin{aligned} E &= -\frac{\Delta x}{2}\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_{l+\theta_1} \\ |E| &\leq \frac{\Delta x}{2}\max_{[x_l,x_{l+1}]}\left|\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right| \leq C\Delta x, \quad \Delta x \to 0 \end{aligned} \tag{4.7}$$

Similarmente, la aproximación hacia atrás (*"backward difference"*)da:

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l \approx \frac{\phi_l-\phi_{l-1}}{\Delta x} \tag{4.8}$$

$$|E| \leq \frac{\Delta x}{2}\max_{[x_{l-1},x_l]}\left|\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right| \leq C'\Delta x, \quad \Delta x \to 0 \tag{4.9}$$

que corresponde a la pendiente del segmento DA en la figura.

4.1.2. Aproximaciones de mayor orden

Haciendo desarrollos de mayor orden:

$$\phi_{l\pm1} = \phi_l \pm \Delta x \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l + \frac{\Delta x^2}{2}\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_l \pm \frac{\Delta x^3}{6}\left(\frac{\mathrm{d}^3\phi}{\mathrm{d}x^3}\right)_{l\pm\theta_{3,4}}, \quad 0 \leq \theta_{3,4} \leq 1 \tag{4.10}$$

de donde:

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l \approx \frac{\phi_{l+1} - \phi_{l-1}}{2\Delta x} \tag{4.11}$$

que corresponde al segmento DC de la figura 4.1. La correspondiente estimación del error es:

$$|E| \leq \frac{\Delta x^2}{6} \max_{[x_{l-1}, x_{l+1}]} \left|\frac{\mathrm{d}^3\phi}{\mathrm{d}x^3}\right| \leq C''\Delta x^2 \tag{4.12}$$

A esta la llamamos una *"aproximación centrada"*, ya que involucra a los dos nodos vecinos del nodo l. Notar que, al contrario de las otras, esta es simétrica con respecto al punto en cuestión. Notar también que el error resulta ser un orden mayor. Con lo cual en principio se pueden obtener mejores aproximaciones con menos puntos usando una aproximación de mayor orden como esta. Pero, como veremos más adelante, en la sección §4.1.7, el hecho de que la aproximación sea de un mayor orden no es una condición *"suficiente"* para obtener un orden de convergencia mayor.

4.1.3. Aproximación de derivadas de orden superior

Para obtener una estimación de la derivada segunda comenzamos haciendo una expansión hasta cuarto orden de $\phi_{l\pm1}$:

$$\begin{aligned}\phi_{l\pm1} \quad &= \phi_l \pm \Delta x \left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_l + \frac{\Delta x^2}{2}\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_l \pm \frac{\Delta x^3}{6}\left(\frac{\mathrm{d}^3\phi}{\mathrm{d}x^3}\right)_l \\ &\quad + \frac{\Delta x^4}{24}\left(\frac{\mathrm{d}^4\phi}{\mathrm{d}x^4}\right)_{l\pm\theta_{5,6}}, 0 \leq \theta_{5,6} \leq 1\end{aligned} \tag{4.13}$$

de donde:

$$\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_l = \frac{\phi_{l+1} - 2\phi_l + \phi_{l-1}}{\Delta x^2} - \frac{\Delta x^2}{24}\left[\left(\frac{\mathrm{d}^4\phi}{\mathrm{d}x^4}\right)_{l+\theta_5} + \left(\frac{\mathrm{d}^4\phi}{\mathrm{d}x^4}\right)_{l+\theta_6}\right] \tag{4.14}$$

o sea que:

$$\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_l \approx \frac{\phi_{l+1} - 2\phi_l + \phi_{l-1}}{\Delta x^2} \tag{4.15}$$

$$|E| \leq \frac{\Delta x^2}{12} \max_{[x_{l-1}, x_{l+1}]} \left|\frac{\mathrm{d}^4\phi}{\mathrm{d}x^4}\right| \tag{4.16}$$

Esta es una *"aproximación centrada"*.

4.1.4. Número de puntos requeridos

Nos cuestionamos cuantos puntos son necesarios para obtener una aproximación de un dado orden (digamos $O(\Delta x)$) para una derivada de orden k. Por ejemplo para obtener una aproximación a la derivada primera es obvio que necesitamos al menos dos puntos ya que por dos puntos pasa una recta y la recta es el polinomio de bajo orden que posee una derivada de primer orden no nula. El mismo razonamiento nos dice que se necesitan tres puntos para aproximar una derivada segunda. Por otra parte, parece también obvio que si queremos una aproximación de mayor orden entonces necesitaremos más puntos. La expresión que relaciona

- N el número de puntos,
- p la precisión del método y
- k el orden de la derivada a aproximar,

es

$$N \geq k + p \tag{4.17}$$

Es decir, con N puntos o más podemos desarrollar una aproximación de orden p (es decir $|E| \leq C\Delta x^p$) para $\left(\mathrm{d}^k\phi/\mathrm{d}x^k\right)$. Verificamos esto en los desarrollos anteriores,

Tipo de aproximación	(precisión)	(orden de la derivada)	Nro. de puntos utilizados	(nro. de puntos requeridos de acuerdo a (4.17))
hacia atrás/adelante	1	1	2	2
centrada	2	1	3	3
centrada	2	2	3	4

Cuadro 4.1: Tabla Número de puntos utilizados en las diferentes aproximaciones por diferencias utilizadas.

4.1.5. Solución de la ecuación diferencial por el método de diferencias finitas

Consideremos primero, por simplicidad, el caso de condiciones Dirichlet $\phi(0) = \bar{\phi}_0$, $\phi(L_x) = \bar{\phi}_{L_x}$. Evaluando la ecuación diferencial en x_l:

$$k\left(\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2}\right)_l = -Q_l \tag{4.18}$$

y aproximando la derivada segunda por diferencias finitas de segundo orden centradas:

$$k\frac{\phi_{l+1} - 2\phi_l + \phi_{l-1}}{\Delta x^2} = -Q_l \tag{4.19}$$

hay una ecuación para cada $l = 1, 2, \ldots, L-1$, que son los *puntos interiores*.
Concretamente, las ecuaciones resultan ser:

$$\begin{array}{cccc} & +2\phi_1 & -\phi_2 & = \frac{\Delta x^2 Q_1}{k} \quad +\bar{\phi}_0 \\ -\phi_1 & +2\phi_2 & -\phi_1 & = \frac{\Delta x^2 Q_2}{k} \\ -\phi_2 & +2\phi_3 & -\phi_4 & = \frac{\Delta x^2 Q_3}{k} \\ \vdots & \vdots & \vdots & \vdots \\ -\phi_{L-3} & +2\phi_{L-2} & -\phi_{L-1} & = \frac{\Delta x^2 Q_{L-2}}{k} \\ -\phi_{L-2} & +2\phi_{L-1} & & = \frac{\Delta x^2 Q_{L-1}}{k} \quad +\bar{\phi}_{L_x} \end{array} \tag{4.20}$$

Definiendo un vector de incógnitas $\boldsymbol{\phi}$, que contiene sólo los nodos interiores:

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{L-1} \end{bmatrix}, \quad \boldsymbol{\phi} \in \mathbb{R}^{L-1} \tag{4.21}$$

Tenemos el siguiente sistema de ecuaciones:

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{f} \tag{4.22}$$

con:

$$\mathbf{K} = \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & \ldots & \ldots & \ldots \\ -1 & 2 & -1 & 0 & \ldots & \ldots & \ldots & \ldots \\ 0 & -1 & 2 & -1 & \ldots & \ldots & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & \\ \vdots & \vdots & \vdots & \vdots & 0 & -1 & 2 & -1 \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 & -1 & 2 \end{bmatrix} \tag{4.23}$$

$$\mathbf{f} = \begin{bmatrix} \Delta x^2 Q_1/k + \bar{\phi}_0 \\ \Delta x^2 Q_2/k \\ \vdots \\ \Delta x^2 Q_{L-1}/k + \bar{\phi}_{L_x} \end{bmatrix} \tag{4.24}$$

Nótese la analogía:

$$\begin{array}{ccccc} \left(\mathrm{d}^2/\mathrm{d}x^2\right) & \phi & = & -Q/k \\ \downarrow & \downarrow & & \downarrow \\ \mathbf{K} & \boldsymbol{\phi} & = & \mathbf{f} \end{array} \tag{4.25}$$

El operador del continuo $\left(\mathrm{d}^2/\mathrm{d}x^2\right)$ es reemplazado por la matriz $\mathbf{K}$, el campo del continuo ϕ es reemplazado por el conjunto de valores nodales $\boldsymbol{\phi}$ y, finalmente, la fuente interna $Q(x)$ es reemplazada por el vector

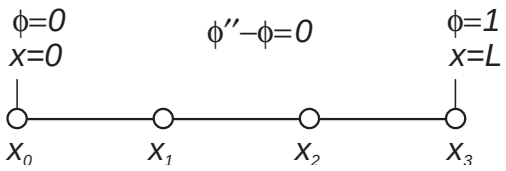


Figura 4.2: Problema unidimensional con condiciones de contorno Dirichlet en ambos extremos

miembro derecho $\mathbf{f}$. A su vez, la condiciones de contorno tipo Dirichlet también generan un término en el miembro derecho.

La matriz $\mathbf{K}$ es simétrica ($K_{ij} = K_{ji}$ para todos i, j) y definida postitiva ($\mathbf{v}^T\mathbf{K}\mathbf{v} > 0$ para todo $\mathbf{v} \in \mathbb{R}^{L-1}$). Esto tiene mucha importancia: se puede demostrar la existencia y unicidad de la solución discreta y además tiene importancia práctica ya que permite el desarrollo de rutinas de resolución especialmente diseñadas.

4.1.6. Ejemplo

Resolver:

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} - \phi = 0, \quad \phi(0) = 0,\ \phi(1) = 1 \tag{4.26}$$

Usaremos $\Delta x = 1/3$ (ver figura 4.2). Tenemos 4 valores nodales ϕ_0, ϕ_1, ϕ_2 y ϕ_3, de los cuales ϕ_0 y ϕ_3 son conocidos de las condiciones de contorno y sólo restan dos incógnitas ϕ_1 y ϕ_2. La ecuación en los nodos interiores es:

$$\phi_{l+1} - 2\phi_l + \phi_{l-1} - \Delta x^2\phi_l = 0 \tag{4.27}$$

para $l = 1$:

$$\phi_2 - 2\phi_1 + \phi_0 - \Delta x^2\phi_1 = 0 \tag{4.28}$$

pero $\phi_0 = 0$ por la condición de contorno, de manera que la ecuación resultante es:

$$\phi_2 - 19/9\,\phi_1 = 0 \tag{4.29}$$

Para $l = 2$ la ecuación resultante es:

$$-19/9\,\phi_2 + \phi_1 = -1 \tag{4.30}$$

Resolviendo el sistema (4.29-4.30) obtenemos:

$$\begin{aligned} \phi_1 &= \frac{1}{(19/9)^2 - 1} = 0.2893\ldots \\ \phi_2 &= 19/9\,\phi_1 = 0.6107\ldots \end{aligned} \tag{4.31}$$

La solución exacta a este problema puede ser encontrada fácilmente. Proponiendo soluciones de la forma $e^{-\alpha x}$, se obtiene la ecuación característica $\alpha^2 = 1$ de donde $\alpha = \pm 1$. Proponemos entonces soluciones de la forma $\phi = ae^x + be^{-x}$. a y b se obtienen de imponer las condiciones de contorno, y resulta ser:

$$\phi = \frac{\sinh(x)}{\sinh(1)} \tag{4.32}$$

	Exacta	$\Delta x = 1/3$		$\Delta x = 1/6$	
x	ϕ	ϕ	Error	ϕ	Error
1/3.	0.28892	0.28929	3.6×10^{-4} (0.12 %)	0.28901	9.2×10^{-5} (0.032 %)
2/3	0.61024	0.61071	4.7×10^{-4} (0.077 %)	0.61036	1.2×10^{-4} (0.019 %)

Cuadro 4.2: Tabla : Errores para $\phi'' - \phi = 0,\ \phi(0) = 0,\ \phi(1) = 1$, para $\Delta x = 1/3, 1/6$

Los resultados numéricos y exactos son comparados en la tabla 4.2, se incluyen también los resultados obtenidos para $\Delta x = 1/6$:

Nótese que tanto en $x = 1/3$ como en $x = 2/3$ el error ha bajado en un factor $1/4$ al reducir el paso de la malla en un factor $1/2$. Primero vemos que, cualitativamente, el esquema de discretización es *convergente*, es decir el error, en alguna norma predeterminada, se reduce, al reducir el paso de la malla. Cuantitativamente, podemos decir que el error tiene un comportamiento $\propto \Delta x^2$. Tambien se dice que el *"orden de convergencia"* es $\propto h^2$, donde h representa el paso de la malla (Δx en nuestro caso). Esto es consistente con el error de truncamiento que encontramos en el desarrollo de Taylor usado.

4.1.7. Análisis de error. Teorema de Lax

Si denotamos por ϕ_{ex} los valores nodales de la solución exacta, es decir

$$\phi_{\text{ex},i} = \phi(x_i) \tag{4.33}$$

entonces ϕ_{ex} satisface (4.19) pero con un error de truncamiento, es decir

$$k\frac{\phi_{\text{ex},l+1} - 2\phi_{\text{ex},l} + \phi_{\text{ex},l-1}}{\Delta x^2} = -Q_l - E_{\text{trunc},l} \tag{4.34}$$

donde $E_{\text{trunc},l}$ es el error correspondiente a la ecuación l-ésima. Sabemos que el error de truncamiento es

$$|E_{\text{trunc},l}| \leq k\frac{\Delta x^2}{12}\max_{[x_{l-1},x_{l+1}]}\left|\frac{\mathrm{d}^4\phi}{\mathrm{d}x^4}\right| \leq C\Delta x^2 \tag{4.35}$$

y tenemos, en forma matricial

$$\mathbf{K}\,\boldsymbol{\phi}_{\text{ex}} = \mathbf{f} + \mathbf{E}_{\text{trunc}} \tag{4.36}$$

Restando (4.22) de (4.36) obtnemos una ecuación para el error $\mathbf{E}_\phi = \boldsymbol{\phi} - \boldsymbol{\phi}_{\text{ex}}$ en los valores nodales

$$\mathbf{K}\,\mathbf{E}_\phi = -\mathbf{E}_{\text{trunc}} \tag{4.37}$$

y entonces,

$$\mathbf{E}_\phi = -\mathbf{K}^{-1}\mathbf{E}_{\text{trunc}} \tag{4.38}$$

Al hecho de que el error de truncamiento $E_{\text{trunc},i} \to 0$ para $\Delta x \to 0$ lo llamamos *"consistencia"* del método. Queremos saber bajo que condiciones, el esquema es *"convergente"*, es decir $\|\mathbf{E}_\phi\| \to 0$ para $\Delta x \to 0$. Por lo que vemos, esto está relacionado con alguna propiedad de $\mathbf{K}$, es decir que de alguna forma la inversa de $\mathbf{K}$ "se mantenga acotada", para $\Delta x \to 0$.

Tomando normas

$$\|\mathbf{E}_\phi\| \leq \|\mathbf{K}^{-1}\| \, \|\mathbf{E}_{\text{trunc}}\| \tag{4.39}$$

Recordemos que la norma (Euclídea) de un vector está definida como

$$\|\mathbf{v}\|^2 = \sum_{i=1}^{L-1} v_i^2 \tag{4.40}$$

y la norma de una matriz está definida como el máximo autovalor de la misma. Puede verse entonces que $\|\mathbf{K}^{-1}\|$ es la inversa del mínimo autovalor de $\mathbf{K}$, y usando (4.12)

$$\sqrt{\sum (\phi_{\text{ex},i} - \phi_i)^2} \leq \frac{1}{\lambda_{\min}} \sqrt{\sum E_{\text{trunc},i}^2} \tag{4.41}$$

$$\leq \frac{1}{\lambda_{\min}} \sqrt{L-1}\, C\Delta x^2 \tag{4.42}$$

de manera que

$$\sqrt{\frac{1}{L-1}\sum_{i=1}^{L-1} (\phi_{\text{ex},i} - \phi_i)^2} \leq \frac{1}{\lambda_{\min}} \sqrt{\sum E_{\text{trunc},i}^2} \tag{4.43}$$

$$\leq \frac{1}{\lambda_{\min}} C\Delta x^2 \tag{4.44}$$

El miembro izquierdo de esta ecuación representa el *"error cuadrático medio"* de la solución numérica. Esta expresión nos dice que este error es del mismo orden que el error de truncamiento, con la condición de que $\lambda_{\min}$ se mantenga acotado (es decir, que no tienda a cero) cuando Δx tiende a cero. Si se cumple esta condición decimos que la discretización es *"estable"*. Entonces, la condición para la convergencia es

$$\text{Consistencia} \; + \; \text{Estabilidad} \implies \text{Convergencia} \tag{4.45}$$

Este resultado es conocido como *Teorema de Lax* y es la base del análisis de error para el método de diferencias finitas.

4.1.8. Condiciones de contorno tipo Neumann ("flujo impuesto")

El problema a resolver es:

$$k\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} = -Q(x), \tag{4.46}$$

$$\phi(0) = 0 \tag{4.47}$$

$$-k\frac{\mathrm{d}\phi}{\mathrm{d}x}\bigg|_{L_x} = \bar{q} \tag{4.48}$$

Ahora ϕ_L pasa a ser una incógnita más:

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{L-1} \\ \phi_L \end{bmatrix}, \quad \boldsymbol{\phi} \in \mathbb{R}^L \tag{4.49}$$

	Exacta	$\Delta x = 1/3$		$\Delta x = 1/6$	
x	ϕ	ϕ	Error	ϕ	Error
$1/3$	0.2200	0.2477	0.0277 (12 %)	0.2340	0.0140 (6 %)
$2/3$	0.4648	0.5229	0.0582 (12 %)	0.4942	0.0294 (6 %)
1	0.7616	0.8563	0.0947 (12 %)	0.8097	0.0481 (6 %)

Cuadro 4.3: Tabla : Errores para $\phi'' - \phi = 0,\ \phi(0) = 0,\ \phi'(1) = 1$, para $\Delta x = 1/3, 1/6$

Para determinar esta incógnitas tenemos $L - 1$ ecuaciones que provienen de aproximar el operador diferencial por diferencias finitas de segundo orden en puntos interiores como en la ecuación (4.19). Para dar cuenta de la condición de contorno tipo Neumann en L_x agregamos la siguiente aproximación:

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_L \approx \frac{\phi_L - \phi_{L-1}}{\Delta x} = -\frac{\bar{q}}{k} \tag{4.50}$$

Ahora sí, tenemos L ecuaciones en L incógnitas. Volviendo al ejemplo anterior del problema (4.26), pero ahora con condición de tipo Neumann en $x = 1$: $\phi'(1) = 1$ el sistema resultante es:

$$\begin{array}{cccc} & -\phi_2 & +19/9\phi_1 & = 0 \\ -\phi_3 & +19/9\phi_2 & -\phi_1 & = 0 \\ \phi_3 & -\phi_2 & & = 3 \end{array} \tag{4.51}$$

Resolviendo el sistema, se encuentran los valores discretos que pueden observarse en la tabla 4.3. (se incluyen también los resultados obtenidos para $\Delta x = 1/6$):

La solución exacta puede obtenerse de la misma forma que en el caso Dirichlet y resulta ser:

$$\phi(x) = \frac{\sinh(x)}{\cosh(1)} \tag{4.52}$$

Notamos que los errores resultan ser notablemente mayores que en el caso Dirichlet puro, concretamente son dos órdenes de magnitud mayores. Un indicio de la causa del problema la da el hecho de que, al reducir el paso de la malla a la mitad, el error no ha bajado en un factor $1/4$ como antes, sino que apenas ha bajado un factor $1/2$, exhibiendo una convergencia $\propto \Delta x$. La causa es que hemos usado una expansión orden $O(\Delta x)$ para la condicion de Neumann, ecuación (), si bien la expansión en los nodos interiores es $O(\Delta x^2)$. Podemos deducir una regla muy importante que es que *el orden de convergencia está dictado por el más bajo orden de las expansiones utilizadas, tanto para el interior del dominio como para las condiciones de contorno*.

En consecuencia, si queremos recuperar el orden de convergencia $\propto \Delta x^2$, necesariamente debemos desarrollar una aproximación $O(\Delta x^2)$ para la condición tipo Neumann. Una forma de hacer esto es introduciendo un *"nodo ficticio"* (ver figura 4.3) x_{L+1} y aproximando la condición de contorno como:

$$\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_L \approx \frac{\phi_{L+1} - \phi_{L-1}}{2\Delta x} = -\frac{\bar{q}}{k} \tag{4.53}$$

Pero hemos introducido una incógnita más: ϕ_{L+1}, de manera que agregamos la ecuación para nodos interiores en el nodo de contorno L:

$$\frac{\mathrm{d}^2\phi}{\mathrm{d}x^2} \approx \frac{-\phi_{L+1} + 2\phi_L - \phi_{L-1}}{\Delta x^2} = -\frac{Q_L}{k} \tag{4.54}$$

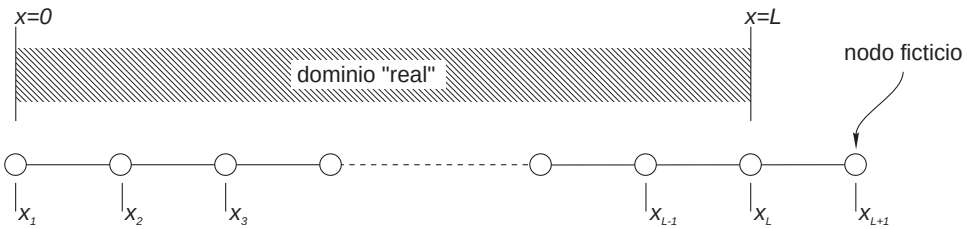


Figura 4.3: Inclusión de un nodo ficticio para obtener una discretización más precisa de la condición de contorno tipo Neumann.

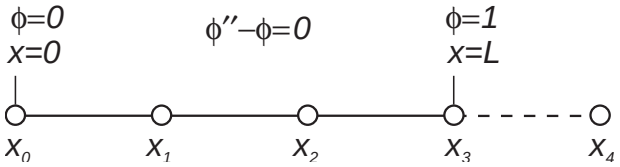


Figura 4.4: Malla 1D con $\Delta x = 1/3$ y un nodo ficticio.

El sistema es:

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{f} \tag{4.55}$$

$$\mathbf{K} = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & \dots & \dots & \dots \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots & \dots \\ 0 & -1 & 2 & -1 & \dots & \dots & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & \\ \vdots & \vdots & \vdots & \vdots & 0 & -1 & 2 & -1 \\ \vdots & \vdots & \vdots & \vdots & 0 & -1 & 0 & 1 \end{bmatrix} \tag{4.56}$$

$$\mathbf{f} = \begin{bmatrix} \Delta x^2 Q_1/k + \bar{\phi}_0 \\ \Delta x^2 Q_2/k \\ \vdots \\ \Delta x^2 Q_L/k \\ -2\bar{q}\Delta x/k \end{bmatrix} \tag{4.57}$$

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_L \\ \phi_{L+1} \end{bmatrix}, \quad \boldsymbol{\phi} \in \mathbb{R}^{L+1} \tag{4.58}$$

Volviendo al ejemplo de la ecuación $\phi'' - \phi = 0$ con condiciones $\phi(0) = 0$, $\phi'(1) = 1$ y discretizado con $\Delta x = 1/3$ (ver figura 4.4), los resultados con este nuevo método pueden observarse en la tabla 4.4.

	Exacta	$\Delta x = 1/3$		$\Delta x = 1/6$	
x	ϕ	ϕ	Error	ϕ	Error
1/3	0.2200	0.2168	0.0033 (1.5 %)	0.2192	0.0008 (0.4 %)
2/3	0.4648	0.4576	0.0071 (1.5 %)	0.4629	0.0018 (0.4 %)
1	0.7616	0.7493	0.0123 (1.6 %)	0.7585	0.0031 (0.4 %)

Cuadro 4.4: Tabla : *Errores para* $\phi'' - \phi = 0,\ \phi(0) = 0,\ \phi'(1) = 1$*, para* $\Delta x = 1/3, 1/6$*, con esquema* $O(\Delta x^2)$

El error es ahora un orden de magnitud menor y se recupera la convergencia cuadrática. Sin embargo el sistema ha dejado de ser simétrico. Para recuperar la simetría de la matriz del sistema, podemos obtener una ecuación para ϕ_L y ϕ_{L-1}, eliminando ϕ_{L+1} de las dos últimas ecuaciones:

$$\phi_L - \phi_{L-1} = \frac{\Delta x^2 Q_L}{2k} - \frac{\bar{q}\Delta x}{k} \tag{4.59}$$

Nótese que esta ecuación es la misma que la (4.50) pero con el agregado del término $\Delta x^2 Q_L/2k$ en el miembro derecho. De hecho, esta misma ecuación puede obtenerse planteando un balance de energía en el intervalo $[L - 1/2, L]\Delta x$. El sistema total es:

$$\mathbf{K} = \begin{bmatrix} 2 & -1 & 0 & 0 & \dots & \dots & \dots & \dots \\ -1 & 2 & -1 & 0 & \dots & \dots & \dots & \dots \\ 0 & -1 & 2 & -1 & \dots & \dots & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & & \\ \vdots & \vdots & \vdots & \vdots & 0 & -1 & 2 & -1 \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 & -1 & 1 \end{bmatrix} \tag{4.60}$$

$$\mathbf{f} = \begin{bmatrix} \Delta x^2 Q_1/k + \bar{\phi}_0 \\ \Delta x^2 Q_2/k \\ \vdots \\ \Delta x^2 Q_L/2k - \bar{q}\Delta x/k \end{bmatrix}, \tag{4.61}$$

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_L \end{bmatrix}, \quad \boldsymbol{\phi} \in \mathbb{R}^L \tag{4.62}$$

4.2. Problemas no-lineales

Consideremos el siguiente problema uni-dimensional, no-lineal debido a la dependencia de k con la temperatura:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[k(\phi)\frac{\mathrm{d}\phi}{\mathrm{d}x}\right] = -Q(x) \tag{4.63}$$

Mencionaremos a continuación algunos otros problemas de la mecánica del continuo que presentan no-linealidad:

- *Material hiperélastico:* $E = E(\epsilon)$, $G = G(\epsilon)$, coeficientes elásticos dependientes de las deformaciones

- *Flujo potencial subsónico (compresible)*: $\nabla \cdot [\rho(|\nabla\phi|)\nabla\phi] = 0$, donde ϕ es el potencial $\mathbf{v} = \nabla\phi$, $\mathbf{v}$=velocidad, ρ=densidad. Aquí ρ juega el papel de la conductividad en el problema térmico. En contraste con áquel, ρ depende de las derivadas de ϕ, y no del valor de ϕ.

La regla en estos casos es diferenciar en forma *conservativa*. Llamando ψ al flujo:

$$\psi = k(\phi)\frac{\mathrm{d}\phi}{\mathrm{d}x} \tag{4.64}$$

La ecuación de balance para el nodo l, puede ser escrita a segundo orden como:

$$\frac{\psi_{l+1/2} - \psi_{l-1/2}}{\Delta x} = -Q_l \tag{4.65}$$

donde $\psi_{l+1/2}$ indica el valor de ψ en nodos ubicados en el punto medio entre los nodos reales: $x_{l+1/2} = (x_l + x_{l+1})/2$. Una aproximación de segundo orden para los flujos es:

$$\begin{aligned}\psi_{l+1/2} &= k(\phi_{l+1/2})\left(\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)_{l+1/2} \\ &= k(\phi_{l+1/2})\,\frac{\phi_{l+1} - \phi_l}{\Delta x} + O(\Delta x^2) \\ &= k(1/2[\phi_l + \phi_{l+1}])\,\frac{\phi_{l+1} - \phi_l}{\Delta x} + O(\Delta x^2)\end{aligned} \tag{4.66}$$

La ecuación resultante para nodos interiores es:

$$-k(\phi_{l+1/2})\phi_{l+1} + \left[k(\phi_{l+1/2}) + k(\phi_{l-1/2})\right]\phi_l - k(\phi_{l-1/2})\phi_{l-1} = \Delta x^2 Q_l, l = 1, 2, \ldots, L-1 \tag{4.67}$$

Sumando sobre las ecuaciones sobre l obtenemos un *principio de conservación discreta*:

$$\psi_{L-1/2} - \psi_{1/2} = -\Delta x(Q_1 + Q_2 + \ldots + Q_{L-1}) \tag{4.68}$$

de allí viene el nombre de *"esquema conservativo"*, ya que reproduce el balance de energía que satisface la ecuación del continuo (ver figura 4.5):

$$\int_{x=\Delta x/2}^{L_x - \Delta x/2} Q(x)\,\mathrm{d}x = -\int_{x=0}^{L_x} \frac{\mathrm{d}}{\mathrm{d}x}\left(k(x)\frac{\mathrm{d}\phi}{\mathrm{d}x}\right)\,\mathrm{d}x = \bar{q}_0 + \bar{q}_{L_x} \tag{4.69}$$

Esto es muy importante en problemas donde puede llegarse a esperar variaciones abruptas de las variables en intervalos muy pequeños, como es el caso de las *ondas de choque* (*"shock waves"*) en fluidos compresibles.

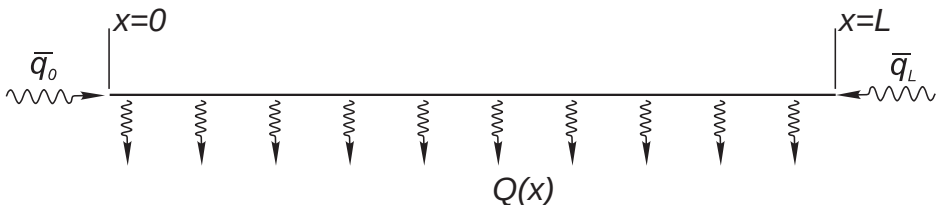


Figura 4.5: Balance global de calor para el problema unidimensional.

El sistema de ecuaciones es ahora no-lineal:

$$\mathbf{K}(\boldsymbol{\phi})\boldsymbol{\phi} = \mathbf{f} \tag{4.70}$$

y no puede resolverse, en general, en forma cerrada, pero puede resolverse en forma aproximada generando una sucesión de valores $\boldsymbol{\phi}^0$, $\boldsymbol{\phi}^1$, ..., $\boldsymbol{\phi}^n$, ... de tal forma que converja a la solución exacta del sistema (4.70):

$$\boldsymbol{\phi}^n \to \phi, \quad \text{para } n \to \infty \tag{4.71}$$

Una forma apropiada de generar estas sucesiones es llevando el sistema anterior a una forma de *"punto fijo"*:

$$\boldsymbol{\phi} = \mathbf{O}(\boldsymbol{\phi}) \tag{4.72}$$

donde $\mathbf{O}$ es un mapeo de $\mathbb{R}^{L-1}$ en sí mismo, fácilmente evaluable. Además, debe elegirse un vector de inicialización $\boldsymbol{\phi}^0$ y la secuencia se genera recursivamente como:

$$\boldsymbol{\phi}^{n+1} = \mathbf{O}(\boldsymbol{\phi}^n) \tag{4.73}$$

En la práctica, esta forma de recurrencia es ejecutada un número finito de iteraciones N, de tal manera que $\boldsymbol{\phi}^N$ esté suficientemente cerca de $\boldsymbol{\phi}$. Obviamente, como no conocemos $\boldsymbol{\phi}$, el criterio para detener el proceso iterativo, no puede basarse en una evaluación directa de $\|\boldsymbol{\phi}^N - \boldsymbol{\phi}\|$. Aquí, $\|\mathbf{v}\| = \sqrt{\sum_{i=1}^{n} v_i^2}$ es la norma L_2 del vector. Una posibilidad es analizar la diferencia entre dos iteraciones consecutivas de la sucesión:

$$\|\boldsymbol{\phi}^{N+1} - \boldsymbol{\phi}^N\| < \epsilon \tag{4.74}$$

donde ϵ es la *"tolerancia"* deseada. Este criterio puede hacerse adimensional poniendo:

$$\frac{\|\boldsymbol{\phi}^{N+1} - \boldsymbol{\phi}^N\|}{\|\boldsymbol{\phi}^0\|} < \epsilon \tag{4.75}$$

Este criterio puede ser engañoso, por ejemplo, si la sucesión está convergiendo muy lentamente. Un criterio más robusto se basa en el *residuo* del sistema de ecuaciones:

$$\frac{\|\mathbf{R}(\boldsymbol{\phi}^N)\|}{\|\mathbf{f}\|} < \epsilon \tag{4.76}$$

donde el residuo $\mathbf{R}$ se define como:

$$\mathbf{R}(\boldsymbol{\phi}^j) = \mathbf{K}(\boldsymbol{\phi}^j)\boldsymbol{\phi}^j - \mathbf{f} \tag{4.77}$$

y el factor $\|\mathbf{f}\|$ en el denominador ha sido agregado para hacer el criterio adimensional.
El valor de la tolerancia ϵ, en ambos casos, depende de múltiples factores:

- Obviamente, cuanto menor sea la tolerancia mayor es el costo computacional. La mayoría de los métodos exhiben *convergencia lineal*, de manera que veremos que el costo computacional es proporcional a $\log \epsilon$.

- Debido a errores de redondeo, es de esperarse que, ambos criterios no puedan bajar más allá de un cierto valor ϵ_{mach}. Con esto queremos decir que $\|\mathbf{R}(\boldsymbol{\phi}^n)\| > \epsilon_{\text{mach}}$ para $n \to \infty$, en vez de tender a cero, como sería en una máquina de precisión infinita. Para problemas bien condicionados, y si el criterio ha sido convenientemente adimensionalizado la precisión de la máquina está en $\epsilon_{\text{mach}} = 10^{-13}, 10^{-16}$, si todos los cáculos se hacen en doble precisión. Si los cáculos se hacen en simple precisión entonces la precisión cae a: $\epsilon_{\text{mach}} = 10^{-6}, 10^{-8}$.

- Debe recordarse siempre que $\boldsymbol{\phi}$ (la solución exacta al problema discreto) posee un error debido a la discretización. Llamando $\boldsymbol{\phi}^*$ a los valores nodales de la solución del continuo, tenemos que:

$$(\boldsymbol{\phi}^N - \boldsymbol{\phi}^*) = (\boldsymbol{\phi}^N - \boldsymbol{\phi}) + (\boldsymbol{\phi} - \boldsymbol{\phi}^*) \tag{4.78}$$

  El error que interesa es el miembro izquierdo de esta desigualdad, es decir, el error con respecto a al solución del continuo. El primer término del miembro derecho es el *error en la resolución del sistema no-lineal*, mientras que el segundo es el *error de discretización*. A medida que se itera, el error de resolución se va reduciendo, hasta anularse, en el límite:

$$\lim_{N\to\infty} \|\boldsymbol{\phi}^N - \boldsymbol{\phi}^*\| = \|\boldsymbol{\phi} - \boldsymbol{\phi}^*\| \tag{4.79}$$

  Este expresión quiere decir que, por más que se itere, el error con respecto a la solución exacta no va a bajar del error propio de discretización. De poder estimarse, el error de discretización, puede fijarse la tolerancia en una fracción (digamos $1/10$) del error de discretización. Poner una tolerancia más baja, aumenta el costo sin mejorar notablemente la solución.

4.2.1. Ejemplo

El esquema de iteración puede ponerse simplemente como:

$$\boldsymbol{\phi}^{j+1} = \mathbf{K}(\boldsymbol{\phi}^j)^{-1}\mathbf{f} \tag{4.80}$$

La implementación es muy simple. En la figura 4.6 vemos el esquema aplicado a la resolución de una ecuación unidimensional $k(\phi)\phi = f$, con $k(\phi) = \phi^m$, $m = -0.7$ y $f = 1$. Nótese que el esquema puede ser también puesto de la forma:

$$\phi^{j+1} = \frac{\phi^j}{k(\phi^j)\phi^j} f = \frac{f}{s^j} \tag{4.81}$$

Dado el valor de ϕ^j, se puede obtener una pendiente aproximada entre el punto $P_j = (\phi^j, k(\phi^j)\phi^j)$ y el origen O. La intersección de esta recta con $y = f$ marca el nuevo punto ϕ^{j+1}. El esquema es convergente si $k(\phi)\phi$ es monótona creciente. En caso contrario es divergente (ver figura 4.7, para el caso $m = -1.5$. Esta condición se cumple en muchos casos físicos, ya que está asociado a una *condición de estabilidad* del sistema. Pensemos por ejemplo en un resorte no-lineal, para el cual k es la constante del resorte, ϕ la

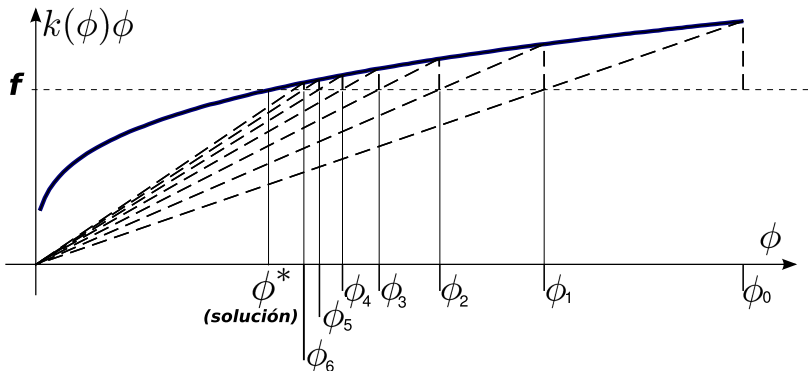


Figura 4.6: El esquema de punto fijo (4.80) aplicado a un problema undimensional con $k(\phi) = \phi^{-0.7}$.

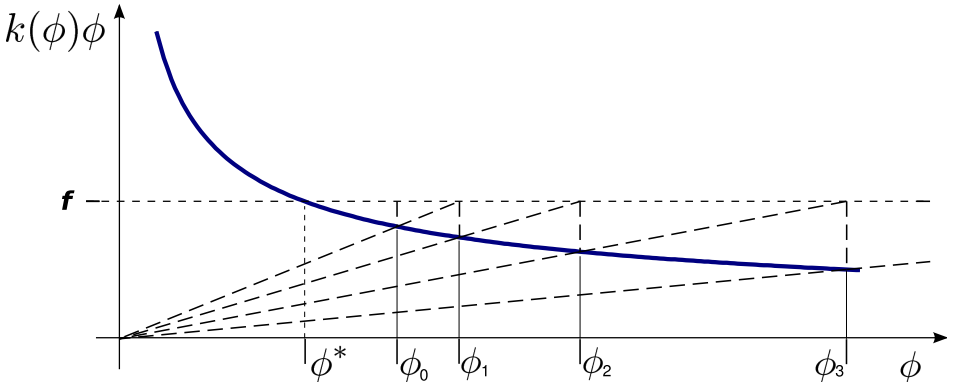


Figura 4.7: Idem, para $k(\phi) = \phi^{-1.5}$.

elongación y f la fuerza aplicada. Entonces $k(\phi)\phi$ creciente quiere decir: *a mayor elongación, mayor fuerza*, lo cual coincide con el criterio de estabilidad. Sin embargo, la convergencia puede ser muy lenta, como por ejemplo en la figura 4.8, que corresponde a $m = -0.9$, $f = 1$, $\phi_0 = 3$.

4.2.2. Método secante

Consideremos por simplicidad un problema de un sólo grado de libertad:

$$\phi^{j+1} = O(\phi^j) \tag{4.82}$$

Supongamos que ϕ^j está suficientemente cerca de la solución ϕ, de maner que podemos hacer un desarrollo de Taylor alrededor de ϕ:

$$\phi^{j+1} = O(\phi) + O'(\phi^j - \phi) + \tfrac{1}{2}O''(\phi^j - \phi)^2 \tag{4.83}$$

pero ϕ es un punto fijo de O, de manera que $O(\phi) = \phi$. Reemplazando en (4.82):

$$\phi^{j+1} - \phi = O'(\phi^j - \phi) + \tfrac{1}{2}O''(\phi^j - \phi)^2 \tag{4.84}$$

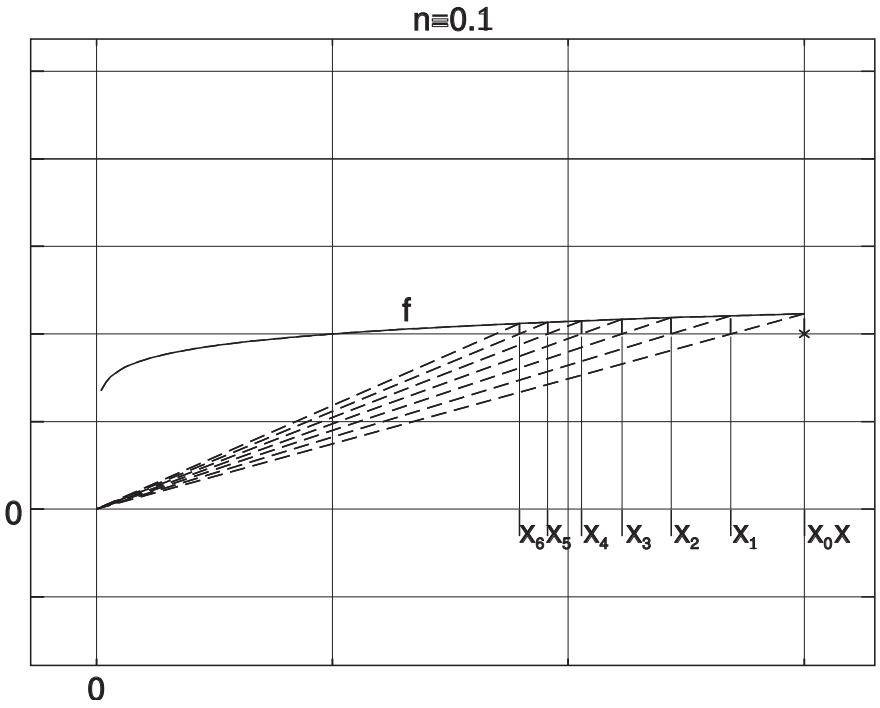


Figura 4.8: Idem, para $k(\phi) = \phi^{-0.9}$.

Si $|O'| < 1$ entonces podemos ϕ^{j+1} estará todavía más cerca de la solución y podremos despreciar el término cuadrático para todo los j.

$$|\phi^{j+1} - \phi| \leq C|\phi^j - \phi| \tag{4.85}$$

con $C = |O'|$. A este tipo de convergencia se le llama, "convergencia lineal". Usando en forma recursiva esta relación:

$$|\phi^j - \phi| \leq C|\phi^{j-1} - \phi| \leq C^2|\phi^{j-2} - \phi| \leq \ldots \leq C^j|\phi^0 - \phi| \tag{4.86}$$

Además, podemos poner:

$$R(\phi^j) \approx R(\phi) + R'(\phi^j - \phi) = R'(\phi^j - \phi) \tag{4.87}$$

De manera que (4.86) puede ponerse como:

$$|R(\phi^j)| \leq C^j|R(\phi^0)| \tag{4.88}$$

Es común graficar $\log|R|$ en función de j:

$$\log\left|\frac{R(\phi^j)}{R(\phi^0)}\right| \leq j\log C \tag{4.89}$$

De manera que, asintóticamente, es una recta de pendiente $\log C$. Cuanto más pequeño es C, más empinada es la pendiente y se llega a una dada precisión con menos iteraciones.

4.2.3. Método tangente

El sistema puede ser también puesto en forma de punto fijo como:

$$\phi = \phi - \frac{k(\phi)\phi - f}{J(\phi)} \tag{4.90}$$

donde $J(\phi)$ se definirá más adelante. La constante C es, en este caso:

$$\begin{aligned} C &= \left| \frac{\mathrm{d}}{\mathrm{d}\phi} \left\{ \phi - \frac{k(\phi)\phi - f}{J(\phi)} \right\} \right| \\ &= \left| 1 - \frac{R'J - RJ'}{J^2} \right| \end{aligned} \tag{4.91}$$

En ϕ vale que $R(\phi) = 0$, de manera que:

$$C = |1 - R'/J| \tag{4.92}$$

Si J se parece mucho a R' entonces C es muy pequeño y la tasa de convergencia es muy alta. El caso óptimo es poner $J = R'$, en cuyo caso la convergencia deja de ser lineal y pasa a ser *"cuadrática"*:

$$|\phi^{j+1} - \phi| \leq C|\phi^j - \phi|^2 \tag{4.93}$$

Esta estrategia es el *"método de Newton"* o *"tangente"*. Cuando $J \neq R'$ entonces decimos que es un método *"secante"*, si bien por supuesto se debe tratar que J se parezca lo más posible a R', ($J \approx R'$). La estimación correspondiente para residuos es:

$$R(\phi^{j+1}) \leq CR(\phi^j)^2 \tag{4.94}$$

Usando esta estimación en forma recursiva:

$$\begin{aligned} R(\phi^n) &\leq CR(\phi^{n-1})^2 && (4.95)\\ &\leq C^{1+2}R(\phi^{n-2})^4 && (4.96)\\ &\leq C^{1+2+4}R(\phi^{n-2})^{2^3} && (4.97)\\ &\leq C^{1+2+4+\ldots+2^{n-1}}R(\phi^0)^{2^n} && (4.98)\\ &= C^{2^n-1}R(\phi^0)^{2^n} && (4.99)\\ &= \frac{1}{C}\left[CR(\phi^0)\right]^{2^n} && (4.100) \end{aligned}$$

$$\log R(\phi^n) \leq -\log C - 2^n \log(CR(\phi^0)) \tag{4.101}$$

Esto es una exponencial cuando se grafica como $\log R$ en función de j (ver figura 4.9).

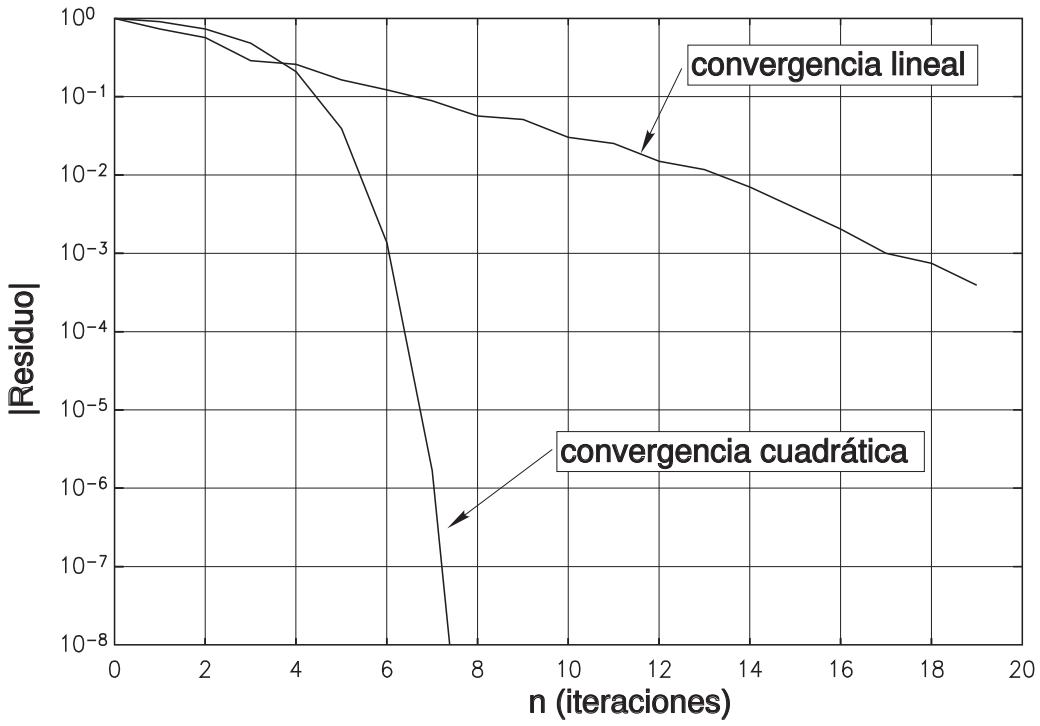


Figura 4.9: Tipos de convergencia lineal y cuadrático.

4.3. Precisión y número de puntos en el esquema de diferencias finitas

Ahora veremos como generar esquemas en diferencias de una forma general. Veremos que para una aproximación para un cierto orden de derivación se requiere un número mínimo de puntos y que cada vez que querramos aumentar el orden de aproximación, deberemos incrementar el número de puntos en el stencil.

Consideremos N puntos arbitrarios $\{x_l\}_{l=1}^{N}$, y expandamos los valores de $\phi_l = \phi(x_l)$ alrededor de un cierto punto x_0 (que no necesariamente debe coincidir con uno de los x_l, $l \geq 1$).

$$\begin{aligned} \phi_l \quad &= \phi_0 + \phi_0' (x_l - x_0) + \phi_0'' \,½(x_l - x_0)^2 + \\ &\quad \cdots + \phi_l^{(N-1)} \frac{(x_l - x_0)^{N-1}}{(N-1)!} + O(|x_l - x_0|^N) \end{aligned} \tag{4.102}$$

Ahora suponemos que el stencil se va reduciendo, hacia el punto x_0 pero manteniendo las distancias relativas invariantes, es decir:

$$\Delta x_l = x_l - x_0 = \epsilon \xi_l \tag{4.103}$$

con ξ_l=cte y $\epsilon \to 0$. Poniendo el sistema anterior en forma matricial:

$$\boldsymbol{\phi} = \mathbf{A}\mathbf{d} + \epsilon^N \mathbf{e} \tag{4.104}$$

donde:

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_L \end{bmatrix}, \tag{4.105}$$

$$\mathbf{d} = \begin{bmatrix} \phi_0 \\ \phi_0' \, \epsilon \\ \vdots \\ \phi_l^{(N-1)} \, \epsilon^{N-1} \end{bmatrix} \tag{4.106}$$

$$\mathbf{A} = \begin{bmatrix} 1 & \xi_1 & \xi_1^2/2 & \cdots & \xi_1^{N-1}/(N-1)! \\ 1 & \xi_2 & \xi_2^2/2 & \cdots & \xi_2^{N-1}/(N-1)! \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \xi_N & \xi_N^2/2 & \cdots & \xi_N^{N-1}/(N-1)! \end{bmatrix} \tag{4.107}$$

y $\mathbf{e}$ es un vector que depende en principio de ϵ pero cuyas componentes se mantienen acotadas al hacer tender $\epsilon \to 0$. Resolviendo el sistema lineal, podemos obtener una expresion para cualquiera de las derivadas, digamos, por ejemplo, la $k-$ésima:

$$\phi_0^{(k)} \, \epsilon^k = \sum_{l=1}^{N} c_{kl} (\phi_l - \epsilon^N e_l) \tag{4.108}$$

Ahora bien, $\mathbf{A}$ no depende de ϵ de manera que tampoco dependen los coeficientes de su inversa c_{kl}, de manera que:

$$\phi_0^{(k)} = \epsilon^{-k} \sum_{l=1}^{N} c_{kl} \phi_l + O(\epsilon^{N-k}) \tag{4.109}$$

De manera que llegamos a la siguiente regla simple que vincula la el orden de la derivada k, el número de puntos del stencil N y el orden de la aproximación p:

$$p = N - k \tag{4.110}$$

En particular, si se quiere aproximar una derivada k-ésima, entonces es necesario al menos $k+1$ puntos para que la aproximación sea "convergente" es decir $p \geq 1$. Por ejemplo, podemos obtener la derivada de primer orden $k = 1$ con precisión $O(\Delta x)^2$ ($p = 2$), con $N = 3$ puntos. Por el contrario, para la derivada segunda ($k = 2$) sólo podemos esperar una precisión $O(\Delta x)$ ($p = 1$) con $N = 3$ puntos. En el caso de una malla de paso constante, y con diferencias centradas se puede obtener una precisión un orden mayor por cuestiones de simetría.

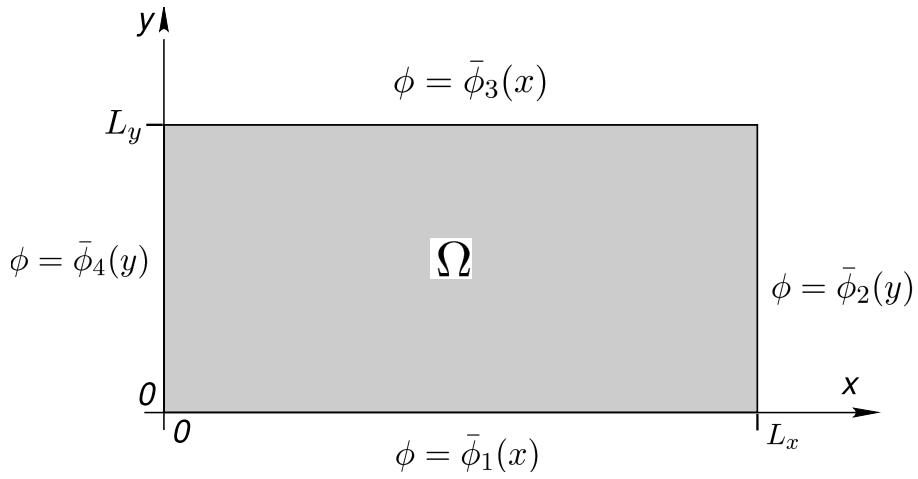


Figura 4.10: Problema bidimensional de conducción del calor rectángulo.

4.4. Método de diferencias finitas en más de una dimensión

Empecemos por un dominio rectangular con condiciones Dirichlet (ver figura 4.10):

$$k\left(\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}\right)=-Q(x,y), \quad \text{en } \Omega=\{x,y \,/\, 0\le x\le L_x,\ 0\le y\le L_y\} \tag{4.111}$$

$$\phi(0,y)=\bar{\phi}_1 \tag{4.112}$$

$$\phi(x,0)=\bar{\phi}_2 \tag{4.113}$$

$$\phi(L_x,y)=\bar{\phi}_3 \tag{4.114}$$

$$\phi(x,L_y)=\bar{\phi}_4 \tag{4.115}$$

Generamos una malla de paso constante Δx, Δy (ver figura 4.11):

$$\Delta x=\frac{L_x}{L}, \qquad \Delta y=\frac{L_y}{M} \tag{4.116}$$

llamamos *"nodo lm"* al punto de coordenadas:

$$(x_l,y_m)=(l\Delta x, m\Delta y), \quad 0\le l\le L, \quad 0\le m\le M \tag{4.117}$$

4.5. Aproximación en diferencias finitas para derivadas parciales

Consideramos una expansión en x de la forma (ver figura 4.12):

$$\phi_{l+1,m} = \phi(x_l+\Delta x, y_m) \tag{4.118}$$

$$= \phi(x_l,y_m)+\Delta x\left(\frac{\partial\phi}{\partial x}\right)_{lm}+\tfrac{1}{2}\Delta x^2\left(\frac{\partial^2\phi}{\partial x^2}\right)_{l+\theta_l,m}, \quad 0\le\theta_l\le 1 \tag{4.119}$$

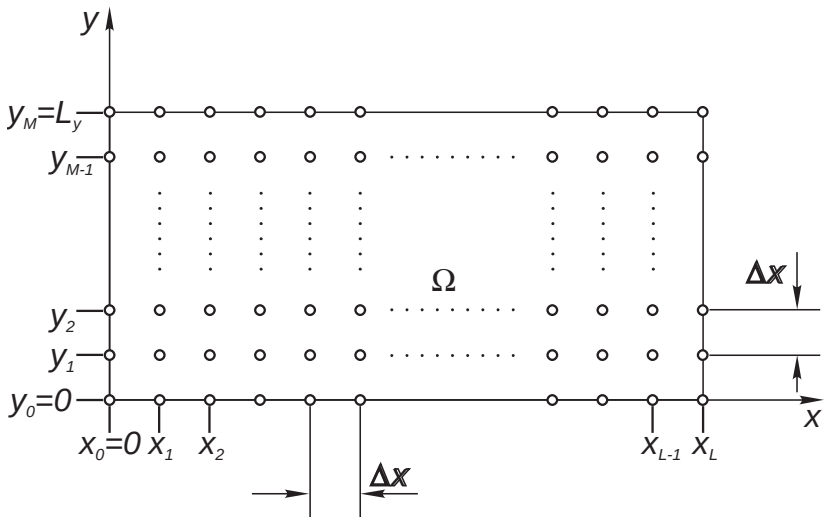


Figura 4.11: Malla homoǵenea de diferencias finitas.

Igual que en 1D, se obtienen expresiones aproximadas para las derivadas de primer y segundo orden:

$$\left(\frac{\partial\phi}{\partial x}\right)_{lm} = \frac{\phi_{l+1,m}-\phi_{lm}}{\Delta x} + O(\Delta x) \tag{4.120}$$

$$\left(\frac{\partial\phi}{\partial x}\right)_{lm} = \frac{\phi_{lm}-\phi_{l-1,m}}{\Delta x} + O(\Delta x) \tag{4.121}$$

$$\left(\frac{\partial\phi}{\partial x}\right)_{lm} = \frac{\phi_{l+1,m}-\phi_{l-1,m}}{2\Delta x} + O(\Delta x^2) \tag{4.122}$$

$$\left(\frac{\partial^2\phi}{\partial x^2}\right)_{lm} = \frac{\phi_{l+1,m}-2\phi_{lm}+\phi_{l-1,m}}{\Delta x^2} + O(\Delta x^2) \tag{4.123}$$

y expresiones similares en y.

La discretización se hace remplazando las derivadas segundas por diferencias de segundo orden para los nodos interiores:

$$k\left(\frac{\phi_{l+1,m}-2\phi_{lm}+\phi_{l-1,m}}{\Delta x^2}+\frac{\phi_{l,m+1}-2\phi_{lm}+\phi_{l,m-1}}{\Delta y^2}\right) = -Q_{lm} \text{ para } \begin{array}{l} l=1,\ldots,L-1 \\ m=1,\ldots,M-1 \end{array} \tag{4.124}$$

y las condiciones de contorno:

$$\begin{aligned} \phi_{0m} &= \bar{\phi}_4(y_m) \quad m=0,\ldots,M \\ \phi_{l0} &= \bar{\phi}_1(x_l) \quad l=0,\ldots,L \\ \phi_{Lm} &= \bar{\phi}_2(y_m) \quad m=0,\ldots,M \\ \phi_{lM} &= \bar{\phi}_3(x_l) \quad l=0,\ldots,L \end{aligned} \tag{4.125}$$

El sistema es, como siempre:

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{f} \tag{4.126}$$

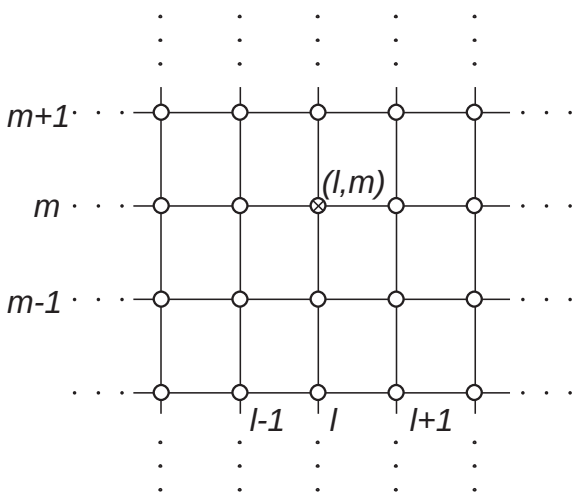


Figura 4.12: Nodo típico en una malla estructurada bidimensional.

donde $\mathbf{K}$ es ahora una matriz tri-diagonal por bloques:

$$\mathbf{K} = \frac{k}{\Delta x^2} \begin{bmatrix} \bar{\mathbf{K}} & -\mathbf{I} & \mathbf{0} & \mathbf{0} & \dots & \dots & \dots & \dots \\ -\mathbf{I} & \bar{\mathbf{K}} & -\mathbf{I} & \mathbf{0} & \dots & \dots & \dots & \dots \\ \mathbf{0} & -\mathbf{I} & \bar{\mathbf{K}} & -\mathbf{I} & \dots & \dots & \dots & \dots \\ & & \ddots & \ddots & \ddots & & & \\ & & & & & -\mathbf{I} & \bar{\mathbf{K}} & -\mathbf{I} \\ & & & & & \mathbf{0} & -\mathbf{I} & \bar{\mathbf{K}} \end{bmatrix} \tag{4.127}$$

siendo $\bar{\mathbf{K}}$:

$$\bar{\mathbf{K}} = \begin{bmatrix} 4 & -1 & 0 & 0 & \dots & \dots & \dots & \dots \\ -1 & 4 & -1 & 0 & \dots & \dots & \dots & \dots \\ 0 & -1 & 4 & -1 & \dots & \dots & \dots & \dots \\ & & \ddots & \ddots & \ddots & & & \\ & & & & & -1 & 4 & -1 \\ & & & & & 0 & -1 & 4 \end{bmatrix} \tag{4.128}$$

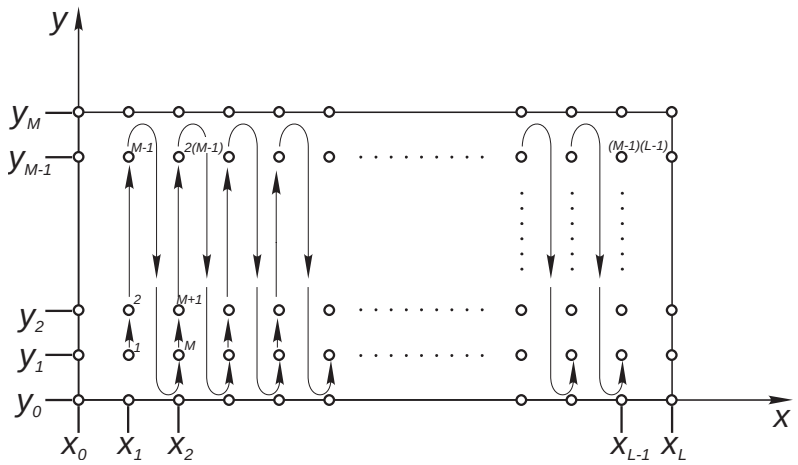


Figura 4.13: Orden de numeración de los grados de libertad en el sistema lineal.

(esto es para el caso especial $\Delta x = \Delta y$). El vector de incógnitas es:

$$\phi = \begin{bmatrix} \phi_{11} \\ \phi_{12} \\ \phi_{13} \\ \vdots \\ \phi_{1,M-1} \\ \phi_{21} \\ \phi_{22} \\ \phi_{23} \\ \vdots \\ \phi_{2,M-1} \\ \vdots \\ \phi_{L-1,1} \\ \phi_{L-1,2} \\ \phi_{L-1,3} \\ \vdots \\ \phi_{L-1,M-1} \end{bmatrix} \tag{4.129}$$

que corresponde a haber numerado las incógnitas primero según y y después según x (ver figura 4.13).

4.5.1. Stencil del operador discreto

Consideremos la fila de la matriz correspondiente a la ecuación para el nodo lm. De todos los coeficientes sólo cinco son no nulos, correspondiente al nodo en cuestión y los cuatro vecinos. Estos coeficientes son los mismos para todos los nodos de la malla, y entonces podemos caracterizar el operador discreto por una sola

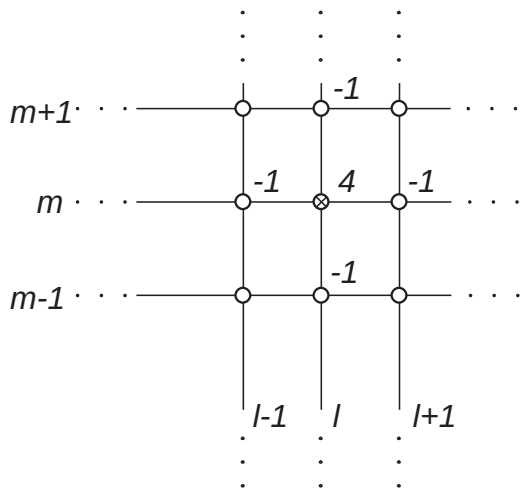


Figura 4.14: Stencil de 5 puntos para el operador de Laplace en 2D.

de las líneas. Para ser más grafico aún, podemos poner los coeficientes en los nodos asociados. A esto se le llama el *stencil* o *estrella* del operador discreto. Para el caso de la ecuación de Poisson que estamos tratando, con $\Delta x = \Delta y$, el stencil obtenido consta de 5 puntos involucrados y tiene como coeficientes 4 para el nodo central y -1 para todos los otros (ver figura 4.14)

4.6. Resolución del sistema de ecuaciones

4.6.1. Estructura banda

Consideremos el caso $L_x = 3$, $L = 3$, $L_y = 5$, $M = 5$, $\Delta x = \Delta y = 1$, por simplicidad (ver figura 4.15). La matriz es:

$$\mathbf{K} = \begin{bmatrix} 4 & -1 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 4 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & -1 & 4 \end{bmatrix} \tag{4.130}$$

vemos que los elementos no nulos se encuentran sólo sobre la diagonal y sus cuatro codiagonales superiores e inferiores, es decir (ver figura 4.16):

$$K_{ij} = 0, \text{para } |i-j| > a \tag{4.131}$$

con $a = 4$. Toda matriz que satisface (4.131) para algún a es llamada una *matriz banda* y a el *ancho de banda* de la matriz. Obviamente, el interés surge cuando a es mucho menor que la dimensión N de la matriz,

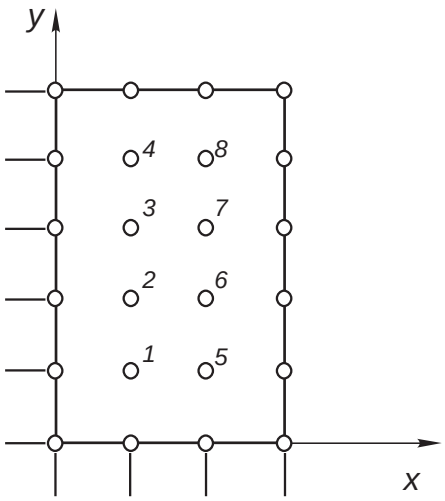


Figura 4.15: Malla de diferencias finitas para un problema con condiciones Dirichlet.

que es el caso para las matrices de diferencias finitas. Veremos que en ese caso se puede ganar mucho tanto en requerimientos de memoria para factorizar la matriz, como en tiempo de procesamiento. Podemos ver que en el caso que nos ocupa $a = M - 1$.

4.6.2. Requerimientos de memoria y tiempo de procesamiento para matrices banda

Consideremos una matriz simétrica de $N \times N$, con ancho de banda a. Puede verse que al factorizar la matriz por un método de eliminación tipo Gauss o Cholesky los elementos fuera de la banda no se "llenan", esto es, siguen siendo nulos después del proceso de factorización. Mediante algoritmos especialmente diseñados, se puede trabajar sólo sobre las diagonales activas de manera que sólo se requieren almacenar $N + (N-1) + (N-2) + \cdots + (N-a) \approx Na$ elementos, contra los $N(N+1)/2$ elementos que hace falta almacenar, si no se tiene en cuenta la estructura banda de la matriz. El factor de ganancia en memoria es:

$$\frac{(\text{almacenamiento matriz banda})}{(\text{almacenamiento matriz llena})} = \frac{2a}{N} \tag{4.132}$$

Consideremos ahora el costo computacional en términos de tiempo de CPU del método de eliminación de Gauss para una matriz llena. Para eliminar la primera columna, debemos hacer $N-1$ operaciones de fila, cada una de las cuales tiene N elementos, lo cual requiere $N(N-1)e$ operaciones, donde e es el número de operaciones necesarios para eliminar un elemento. Usualmente se necesita una suma y una multiplicación, de manera que $e = 2$, sin embargo, dependiendo de la máquina y de detalles de implementación, debe tenerse en cuenta las operaciones de traer los elementos desde la RAM el procesador y de incrementar los contadores. Para eliminar la segunda columna, debemos realizar $N-2$ operaciones de $N-1$ elementos,

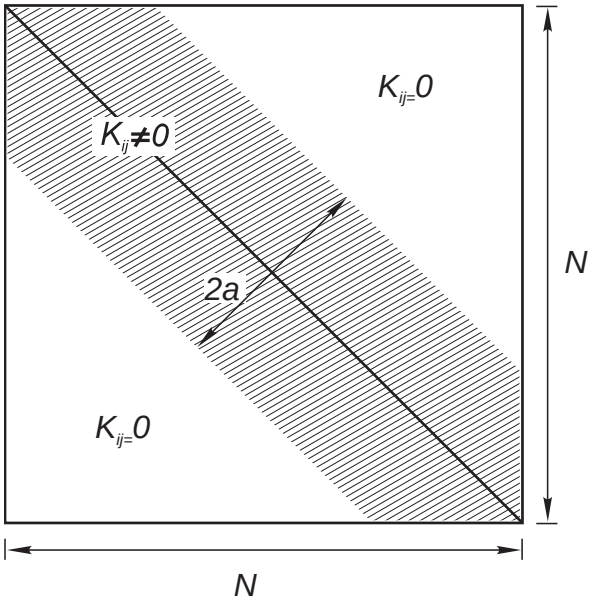


Figura 4.16: Definición del ancho de banda de una matriz.

es decir $(N-1)(N-2)e$ operaciones. El número total de operaciones es de:

$$\begin{aligned}\text{(Costo computacional matriz llena)} = \\ &= [N(N-1)+(N-1)(N-2)+...+3\times 2+2\times 1]\,e \\ &= e\sum_{j=1}^{N} j(j-1) = eN^3/3+O(N^2)\end{aligned} \tag{4.133}$$

Si la matriz es banda, entonces en la primera columna solo hay $a+1$ elementos no nulos. Por lo tanto, solo debemos hacer a operaciones de fila. Además, en cada una de las filas sólo los $2a$ primeros elementos están dentro de la banda, de manera que deben efectuarse $2ea^2$ operaciones para eliminar la primera columna. Para las otras columnas, ocurre algo parecido y el costo total es:

$$\text{(Costo computacional matriz banda)} = N\times 2ea^2 = 2eNa^2 \tag{4.134}$$

La relación de costos es:

$$\frac{\text{(Costo computacional matriz banda)}}{\text{(Costo computacional matriz llena)}} = \frac{2eNa^2}{eN^3} = 2\left(\frac{a}{N}\right)^2 \tag{4.135}$$

Para fijar ideas, consideremos el caso de una malla de $L=M=200$ nodos. El número total de grados de libertad es $N=(L-1)\times(M-1)\approx LM=40000$. El número total de elementos a almacenar como matriz llena es $\approx N^2/2=8.0\times 10^8$ elementos. En el caso de utilizar doble precisión esto equivale a 6.4Gbytes de memoria RAM. Utilizando almacenamiento banda tenemos $a=200$ y el número de elementos a almacenar es $Na=200\times 40000=8000000$, 64.0Mbyte de memoria en doble precisión.

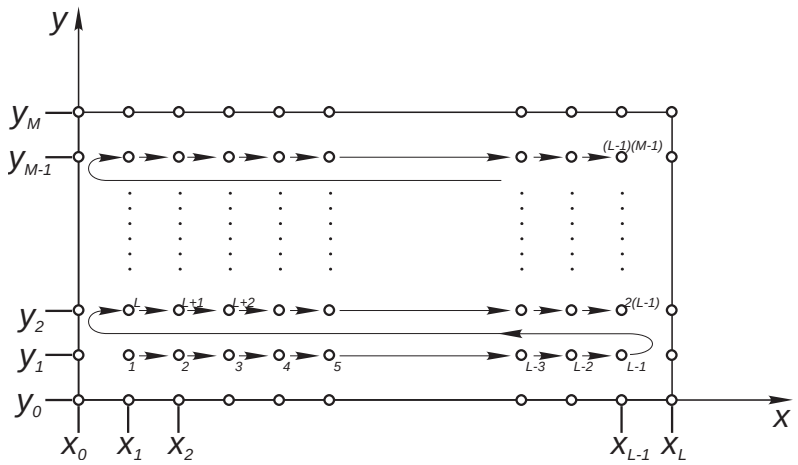


Figura 4.17: Numeración alternativa para reducir el ancho de banda.

Con respecto al costo computacional, para la matriz llena es $40000^3/3e \approx 2.1\times10^{13}e$ operaciones, mientras que para matriz banda es de sólo $2\times200^2\times40000e = 3.2\times10^9 e$ operaciones. Considerando $e=2$ y una velocidad de procesamiento de 1Gflops (1 Gflop=10^9 operaciones de punto flotante por segundo) los tiempos de procesamiento resultan ser de:

$$\text{(tiempo de CPU matriz llena)} = \frac{4.3\times10^{13}\text{flops}}{1\times10^9\text{flops/sec}} = 11.9 \text{ horas} \tag{4.136}$$

$$\text{(tiempo de CPU matriz banda)} = \frac{3.2\times10^{9}\text{flops}}{1\times10^6\text{flops/sec}} = 6.4 \text{ segundos} \tag{4.137}$$

4.6.3. Ancho de banda y numeración de nodos

El ancho de banda es altamente dependiente de la *numeración de los nodos*, esto es, del orden en que las incoǵnitas son puestas en el vector ϕ. Por ejemplo si la numeración se hace primero en x y después en y (ver figura 4.17):

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \phi_{31} \\ \vdots \\ \phi_{L-1,1} \\ \vdots \\ \phi_{1,M-1} \\ \phi_{2,M-1} \\ \phi_{3,M-1} \\ \vdots \\ \phi_{L-1,M-1} \end{bmatrix} \tag{4.138}$$

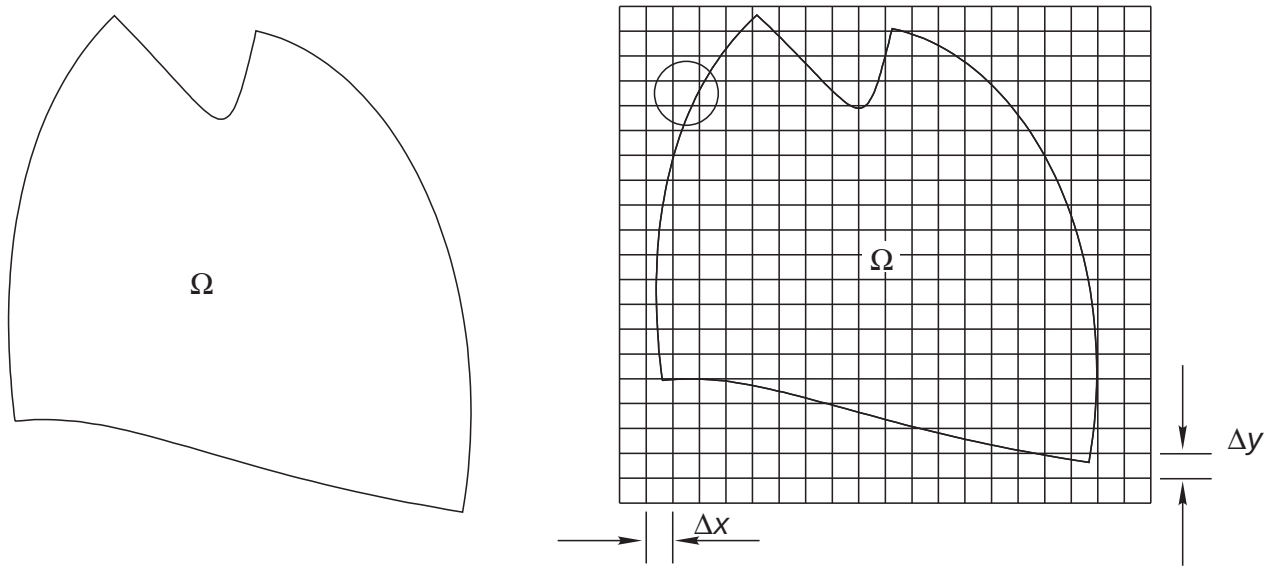


Figura 4.18: El dominio de resolución de resolución Ω es "embebido" en una malla homogénea.

entonces el ancho de banda pasa a ser de $a = L - 1$. La regla es, entonces, *numerar siempre primero en aquella dirección en la cual hay menos nodos*.

4.7. Dominios de forma irregular

El método de diferencias finitas sería de muy poca utilidad si sólo pudiera aplicarse a dominios de forma rectangular. Existen básicamente dos posibilidades para extender el método a un dominio de forma arbitraria como en la figura 4.18. (, indican ventajas y desventajas del método, respectivamente):

- Considerar al dominio *inmerso en una malla homogénea* (ver figura 4.18). En este caso el esquema para los nodos interiores es el mismo como el que consideramos hasta ahora.
  - La generación de la malla es muy sencilla ( )
  - Deben generarse ecuaciones especiales para los nodos en el contorno ( )
  - En principio no hay posibilidad de *"refinar"* la malla en ciertas partes ( )
- *Ajuste del contorno ("boundary fitting")* (ver figura 4.19): La idea es encontrar una transformación de coordenadas que lleve el dominio irregular en cuestión a un rectángulo. La ecuación original es transformada siguiendo las reglas clásicas y finalmente se resuelven las ecuaciones transformadas en el dominio transformado.
  - La generación de la malla es relativamente compleja, incluso para dominios relativamente simples ( )

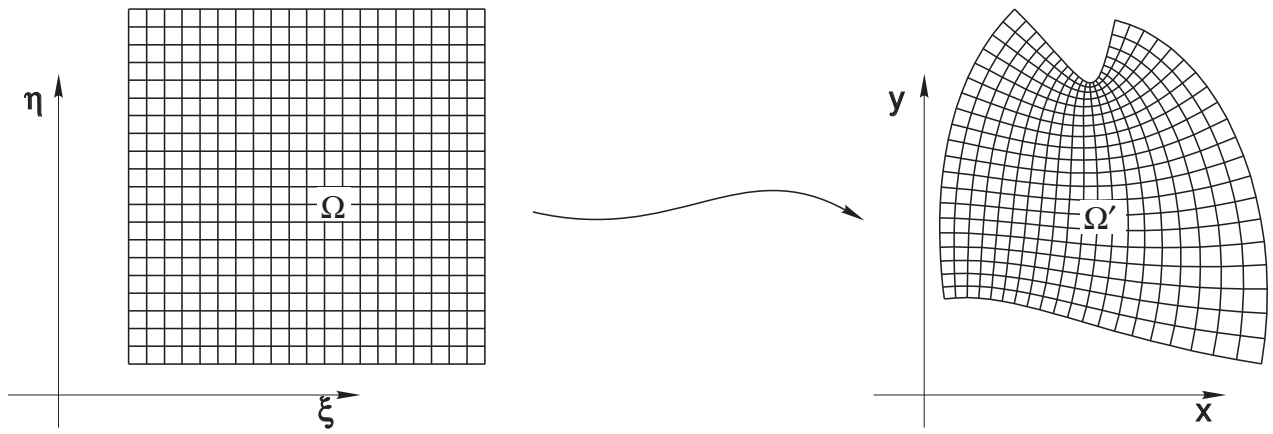


Figura 4.19: La malla es generada por mapeo de una malla homogénea sobre un rectángulo al dominio Ω.

- Los esquemas en diferencias, tanto para nodos interiores como nodos de contorno se obtienen fácilmente por los métodos estándar ( )
- La malla suele estar más densa en ciertas partes que en otras. Esto puede ser una ventaja o desventaja dependiendo de si el lugar donde la malla está mas densa es un punto donde se necesita mayor precisión o no (?)
- Admite cierto grado de refinamiento ( )

4.7.1. Inmersión del dominio irregular en una malla homogénea

Como se mencionara, aquí el punto delicado es el hallar fórmulas en diferencias para los nodos en contacto con el contorno. Consideremos por ejemplo la figura 4.18. Haciendo un zoom de una región cerca del contorno, tenemos una disposición como en la figura 4.20. Suponiendo que la condición de contorno es de tipo Dirichlet, debemos encontrar ecuaciones en diferencias para un nodo como el P. Considerando que la malla es homogénea ($PT = PS = \Delta x$, $PQ = PR = \Delta y$) y definiendo:

$$\lambda = \frac{PU}{PQ} \leq 1, \quad \mu = \frac{PV}{PS} \leq 1 \tag{4.139}$$

podemos hacer un desarrollo de Taylor para ϕ_T y ϕ_V:

$$\phi_T = \phi_P + \Delta x \left(\frac{\partial \phi}{\partial x}\right)_P + \tfrac{1}{2}\Delta x^2 \left(\frac{\partial^2 \phi}{\partial x^2}\right)_P + \tfrac{1}{6}\Delta x^3 \left(\frac{\partial^3 \phi}{\partial x^3}\right)_{P_1} \tag{4.140}$$

$$\phi_V = \phi_P - \mu\Delta x \left(\frac{\partial \phi}{\partial x}\right)_P + \tfrac{1}{2}\mu^2\Delta x^2 \left(\frac{\partial^2 \phi}{\partial x^2}\right)_P - \tfrac{1}{6}\mu^3\Delta x^3 \left(\frac{\partial^3 \phi}{\partial x^3}\right)_{P_2} \tag{4.141}$$

con

$$P_1 \in [P, T], \quad P_2 \in [P, V] \tag{4.142}$$

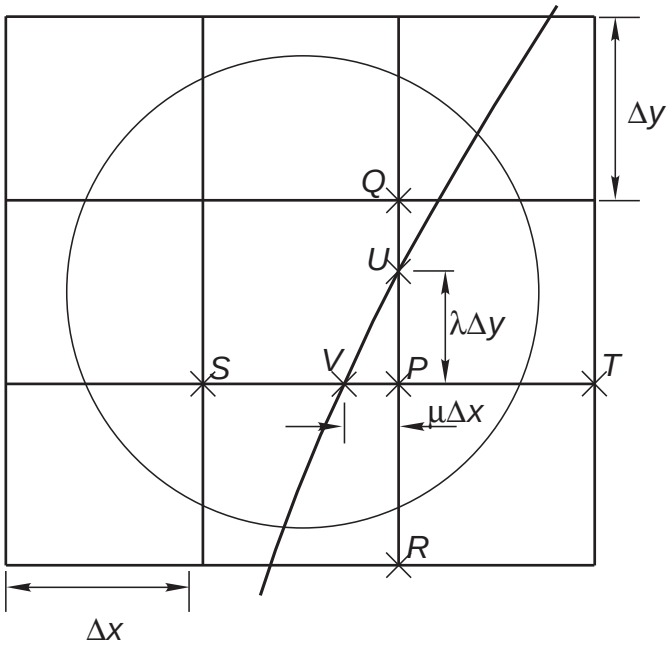


Figura 4.20: Diagrama general de los nodos sobre la malla regular en la zona cercana al contorno del dominio de resolución. Fírmulas especiales en diferencias deben ser desarrolladas para los nodos como el P

La derivada de primer orden se puede obtener, a orden Δx^2, haciendo una combinación lineal de ϕ_T y ϕ_V de forma de que se cancelen los términos que contienen la derivada segunda.

$$\mu^2\phi_T - \phi_V = (\mu^2-1)\phi_P + (\mu^2+\mu)\Delta x\left(\frac{\partial\phi}{\partial x}\right)_P + O(\Delta x^3) \tag{4.143}$$

de manera que:

$$\left(\frac{\partial\phi}{\partial x}\right)_P = \frac{\mu^2\phi_T - \phi_V - (\mu^2-1)\phi_P}{\Delta x\mu(\mu+1)} + O(\Delta x^2) \tag{4.144}$$

La derivada segunda se puede aproximar de:

$$\begin{aligned}
½\Delta x^2\left(\frac{\partial^2\phi}{\partial x^2}\right)_P &= \phi_T - \phi_P - \Delta x\left(\frac{\partial\phi}{\partial x}\right)_P + O(\Delta x^3) && (4.145)\\
&= \phi_T - \phi_P - \Delta x\frac{\mu^2\phi_T - \phi_V - (\mu^2-1)\phi_P}{\Delta x\mu(\mu+1)} + O(\Delta x^3) && (4.146)\\
&= \frac{\mu\phi_T + \phi_V - (\mu+1)\phi_P}{\mu(\mu+1)} + O(\Delta x^3) && (4.147)
\end{aligned}$$

de donde:

$$\left(\frac{\partial^2\phi}{\partial x^2}\right)_P = 2\frac{\mu\phi_T + \phi_V - (\mu+1)\phi_P}{\mu(\mu+1)\Delta x^2} + O(\Delta x) \tag{4.148}$$

Nótese que para $\mu = 1$ ($PV = PS$) se recupera la fórmula centrada de segundo orden. Por el contrario, la expresión anterior es de primer orden y, de hecho *es imposible generar una aproximación de segundo orden para la derivada segunda con 3 puntos en una malla irregular* (Confiérase a la sección anterior donde se explica la relación entre precisión, número de puntos y orden de la derivada).

Finalmente, el sistema de ecuaciones se halla planteando las ecuaciones en diferencias para los puntos interiores sobre una malla regular, mientras que para los puntos como el P sobre el contorno, las ecuaciones son de tipo:

$$\frac{\mu\phi_T + \phi_V - (\mu+1)\phi_P}{\mu(\mu+1)\Delta x^2} + \frac{\lambda\phi_R + \phi_U - (\lambda+1)\phi_P}{\lambda(\lambda+1)\Delta y^2} = -\frac{Q_P}{2k} \tag{4.149}$$

Debe recordarse que esta expresión es $O(\Delta x)$ de manera que sólo puede esperarse una tal convergencia.

Para condiciones de tipo Neumann, el problema es aún más complicado.

4.7.2. Mapeo del dominio de integración

Consideremos, por ejemplo, resolver el problema:

$$\nabla \cdot (k\nabla\phi) = -Q \tag{4.150}$$

En notación tensorial:

$$\frac{\partial}{\partial x_i}\left(k\frac{\partial\phi}{\partial x_i}\right) = -Q \tag{4.151}$$

Como fue adelantado, aquí se trata de encontrar una transformación de coordenadas:

$$\mathbf{x} = (x_1, x_2, x_3) \rightarrow \boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3) \tag{4.152}$$

de tal forma que en la variable $\boldsymbol{\eta}$, el dominio de integración sea un retángulo. La transformación de las ecuaciones se hace por la *"regla de la cadena"*:

$$\frac{\partial\phi}{\partial x_i} = \frac{\partial\phi}{\partial\eta_j}\frac{\partial\eta_j}{\partial x_i} = [\mathbf{J}\nabla_\eta\phi]_i \tag{4.153}$$

donde $\mathbf{J}$ denota la matriz del jacobiano de la transformación y ∇_η es el operador gradiente con respecto a las coordenadas $\boldsymbol{\eta}$. Usando cáculo tensorial, podemos plantear la ecuación diferencial en términos de:

- derivadas de las incógnitas y los datos (propiedades del material) con respecto a las coordenadas transformadas, e.g.: $(\partial\phi/\partial\eta_j)$, etc...
- propiedades de la transformación como: jacobianos, tensores métricos, factores de escala, (conocidos).

Continuando con la ecuación de Poisson, su transformación es:

$$g^{-1/2}\frac{\partial}{\partial\eta_i}\left(kg_{ij}g^{1/2}\frac{\partial\phi}{\partial\eta_j}\right) = -Q \tag{4.154}$$

donde:

$$g^{ij} = [\mathbf{g}]_{ij} = \frac{\partial x_k}{\partial \eta_i}\frac{\partial x_k}{\partial \eta_j} \quad \text{(tensor métrico covariante)} \tag{4.155}$$

$$g_{ij} = [\mathbf{g}^{-1}]_{ij} \quad \text{(tensor métrico contravariante)} \tag{4.156}$$

$$g^{1/2} = (\det \mathbf{g})^{1/2} \quad \text{(determinante de la transformación)} \tag{4.157}$$

La ecuación transformada puede ser reescrita como una ecuación quasi-ármonica con una conductividad anisotrópica:

$$\frac{\partial}{\partial \eta_i}\left(\tilde{k}_{ij}\frac{\partial \phi}{\partial \eta_j}\right) = -\tilde{Q} \tag{4.158}$$

donde:

$$\tilde{k}_{ij} = k g_{ij} g^{1/2} \quad \text{(conductividad anisotrópica equivalente)} \tag{4.159}$$

$$\tilde{Q} = g^{1/2} Q \quad \text{(término fuente equivalente)} \tag{4.160}$$

4.7.3. Coordenadas curvilíneas ortogonales

Un caso particular es cuando la transformación es tal que las superficies $\eta_i =$cte son ortogonales entre sí en el sistema x_i. Puede verse que la condición para que esto ocurra es que el tensor métrico de la transformación satisfaga:

$$g^{ij} = \begin{bmatrix} h_1^2 & 0 & 0 \\ 0 & h_2^2 & 0 \\ 0 & 0 & h_3^2 \end{bmatrix} \tag{4.161}$$

h_i es llamado el *factor de escala* de la coordenada transformada η_i. Se desprende que $g^{1/2} = h_1 h_2 h_3$.

La ecuación del calor en coordenadas cúrvilineas ortogonales es:

$$\left[\frac{\partial}{\partial \eta_1}\left(\frac{h_2 h_3}{h_1}\frac{\partial \phi}{\partial \eta_1}\right) + \frac{\partial}{\partial \eta_2}\left(\frac{h_1 h_3}{h_2}\frac{\partial \phi}{\partial \eta_2}\right) + \frac{\partial}{\partial \eta_3}\left(\frac{h_1 h_2}{h_3}\frac{\partial \phi}{\partial \eta_3}\right)\right] = h_1 h_2 h_3\, Q \tag{4.162}$$

4.7.4. Ejemplo

Sea resolver la ecuación del calor en una cáscara esférica:

$$\Omega = \{r, \theta, \varphi\} \text{tales que } \begin{cases} r_{\text{int}} \leq & r \leq r_{\text{ext}} \\ -\pi/2 \leq & \theta \leq \pi/2 \\ -\pi \leq & \varphi \leq \pi \end{cases} \tag{4.163}$$

con condiciones de frontera Dirichlet en la cáscara exterior e interior:

$$\phi\left(r_{\text{ext,int}}, \theta, \varphi\right) = \bar{\phi}_{r_{\text{ext}}, r_{\text{int}}}(\theta, \varphi), \quad -\pi/2 \leq \theta \leq \pi/2,\ -\pi \leq \varphi \leq \pi, \tag{4.164}$$

donde las coordenadas esféricas están dadas por la transformación usual:

$$x = r\cos\theta\,\cos\varphi \tag{4.165}$$
$$y = r\cos\theta\,\sin\varphi \tag{4.166}$$
$$z = r\sin\theta \tag{4.167}$$

Los factores de escala son:

$$\begin{aligned} h_r &= 1 \\ h_\theta &= r \\ h_\varphi &= r\cos\theta \end{aligned} \tag{4.168}$$

de manera que la ecuación transformada es:

$$\frac{\partial}{\partial r}\left(r^2\cos\theta\frac{\partial\phi}{\partial r}\right)+\frac{\partial}{\partial\theta}\left(\cos\theta\frac{\partial\phi}{\partial\theta}\right)+\frac{\partial}{\partial\varphi}\left(\frac{1}{\cos\theta}\frac{\partial\phi}{\partial\varphi}\right)=r^2\cos\theta\,Q \tag{4.169}$$

Se contruye una malla homogénea en coordenadas transformadas, dada por una serie de $(I+1)\times(J+1)\times(K+1)$ puntos P_{ijk} cuyas coordenadas (r_i,θ_j,φ_k) están dadas por:

$$\begin{aligned} r_i &= r_{\rm int}+(i/I)(r_{\rm ext}-r_{\rm int}) && i=0,\ldots,I \\ \theta_j &= \pi/2\,(1-\epsilon_\theta)\,[-1+2(j/J)] && j=0,\ldots,J \\ \varphi_k &= \pi[-1+2\,(k/K)] && k=0,\ldots,K \end{aligned} \tag{4.170}$$

donde $0\le\epsilon_\theta\ll\pi$ tiene el fin de evitar la singularidad en los polos $\theta=\pm\pi/2$. La ecuación para el nodo ijk es:

$$\frac{q_{r,i+1/2\,jk}-q_{r,i-1/2\,jk}}{\Delta r}+\frac{q_{\theta,i\,j+1/2\,k}-q_{\theta,i\,j-1/2\,k}}{\Delta\theta}+\frac{q_{\varphi,ij\,k+1/2}-q_{\varphi,ij\,k-1/2}}{\Delta\varphi}=r_i^2\cos\theta_j\,Q_{ijk} \tag{4.171}$$

donde:

$$\begin{aligned} q_{r,i+1/2\,jk} &= r_{i+1/2}^2\cos\theta_j\,\frac{\phi_{i+1\,jk}-\phi_{ijk}}{\Delta r} \\ q_{\theta,i\,j+1/2\,k} &= \cos\theta_{j+1/2}\,\frac{\phi_{i\,j+1\,k}-\phi_{ijk}}{\Delta\theta} \\ q_{\varphi,ij,k+1/2} &= \frac{1}{\cos\theta_j}\,\frac{\phi_{ij\,k+1}-\phi_{ijk}}{\Delta\varphi} \end{aligned} \tag{4.172}$$

y

$$\begin{aligned} r_{i+1/2} &= r_i+\frac{\Delta r}{2} \\ \theta_{j+1/2} &= \theta_i+\frac{\Delta\theta}{2} \end{aligned} \tag{4.173}$$

La ecuaciones (4.171-4.173) son conservativas y precisas de segundo orden.

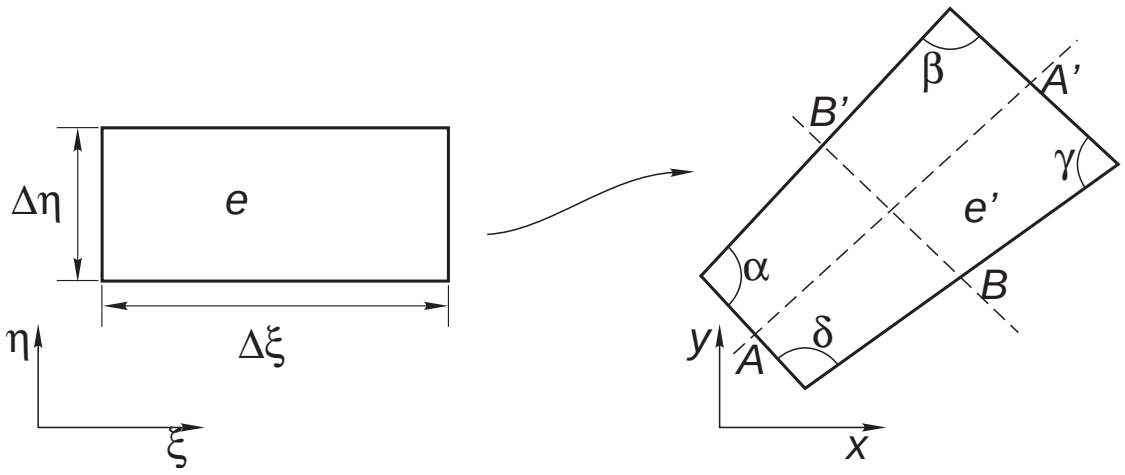


Figura 4.21: Transformación de un "elemento" de volumen $\Delta\xi$, $\Delta\eta$ por una transformación conforme. No sólo los ángulos interiores permanecen aproximadamente rectos, sino que la relación de aspecto permanece inalterado.

4.7.5. Mallas generadas por transformación conforme

Una clase especial de transformaciones pueden generarse mediante la teoría de variable compleja llamada *transformación conforme*. Si bien está es válida sólo para 2D, puede combinarse con otras transformaciones para generar mallas tridimensionales.

Una transformación conforme $(x, y) \rightarrow (\xi, \eta)$, se basa en definir dos variables complejas $z = x + iy$ y $w = \xi + i\eta$ y poner la transformación en forma de una función analítica: $w = f(z)$ con f analítica. Las transformaciones conformes son un subconjuto de las transfomaciones obtenidas por coordenadas curvílineas ortogonales. No sólo las lineas ξ =cte, η =cte son ortogonales entre sí, sino que los factores de escala son iguales $h_\xi = h_\eta$. Esto significa que un elemento rectangular e en coordenadas ξ, η (ver figura 4.21) es mapeado por la transformación en otro e' de forma aproximadamente rectangular con ángulos interiores $\alpha, \beta\gamma\delta \approx 90^\circ$) y con, aproximadamente, la misma relación de aspecto ($BB'/AA' \approx \Delta\eta/\Delta\xi$).

Por ejemplo, la transformación de coordenadas $z = \exp(w)$, equivale a la transformación:

$$\begin{aligned} x &= e^\xi \cos\eta \\ y &= e^\xi \sin\eta \end{aligned} \tag{4.174}$$

y es muy similar al cambio de coordenadas de cartesianas a polares. Por ejemplo, un dominio rectangular como el $ABCDEF$ (ver figura 4.22) es transformado en una corona circular. Nótese que el Δr entre sucesivas capas de nodos va aumentando con el radio. Lo interesante es que esta variación es tal que la relación de aspecto se mantiene constante, como fue mencionado en un marco más general.

Las transformaciones conformes pueden ser concatenadas de forma de poder obtener dominios bastantes más complicados. Por ejemplo, a la transformación (4.174) se puede aplicar una transformación lineal para correr el centro del círculo interior O de $z(O) = 0$ a $z_1(O) = 1 + e$, manteniendo el punto B en $z = 1$

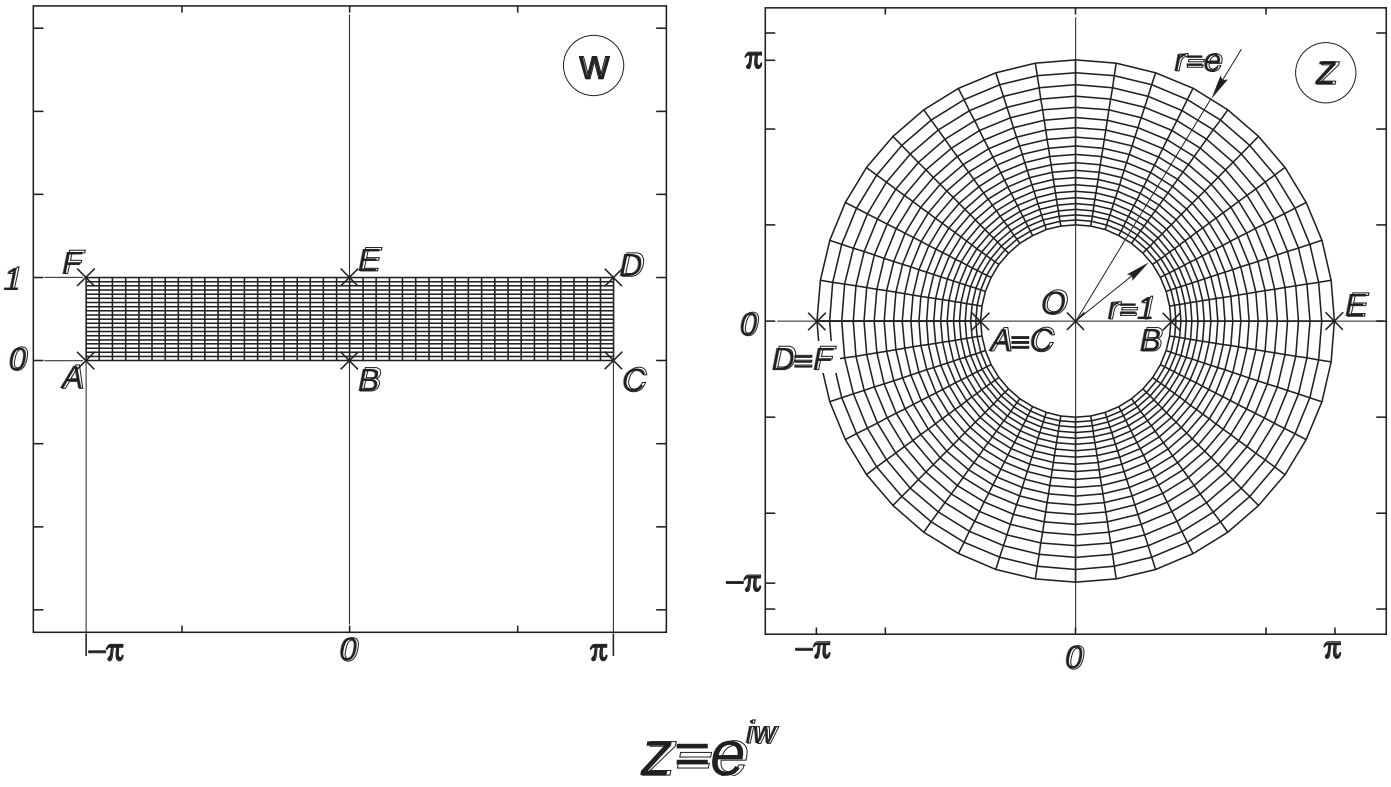


Figura 4.22: Un rectángulo es transformado en una corona circular usando la transformación exponencial $z = e^w$.

(ver figura 4.24). La transformación es:

$$z_1 = 1 + (1 - e)(z - 1) \tag{4.175}$$

A continuación una *"transformación de Joukowski"* lleva la corona circular al exterior de un perfil aerodinámico (ver figura 4.23):

$$z_2 = \tfrac{1}{2}\left(z_1 + \frac{1}{z_1}\right) \tag{4.176}$$

La circunferencia interior ABC es transformada sobre el contorno del perfil, mientras que el círculo exterior DEF es transformado en una curva cercana a una circunferencia. Usualmente, se supone que sobre esta circunferencia el flujo no está perturbado para poder imponer las condiciones de contorno apropiadas. El punto B en $z_1 = 1$ es transformado en el borde de fuga (TE = *"trailing edge"*, mientras que el punto $A \equiv C$ es transformado en el borde de ataque LE (*"Leading Edge"*).

Otro ejemplo de transformación podemos observarlo en las figuras 4.25, 4.26, donde el dominio rectangular $ABCDEF$ en el plano w es mapeado por la transformación $z = w^{3/2}$ en el dominio indicado en la figura 4.26.

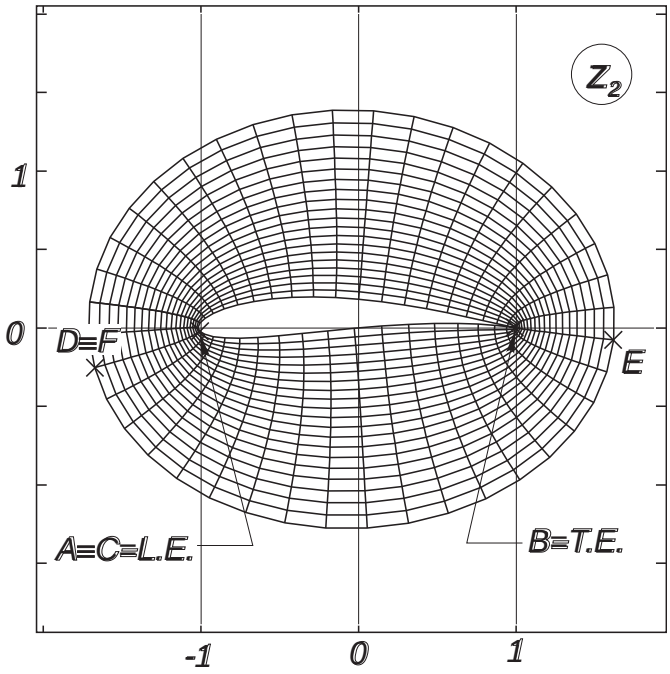


Figura 4.23: Transformación de Joukowski. El círculo interior es mapeado sobre el perfil mientras que el exterior es mapeado sobre un cuasi-círculo al infinito.

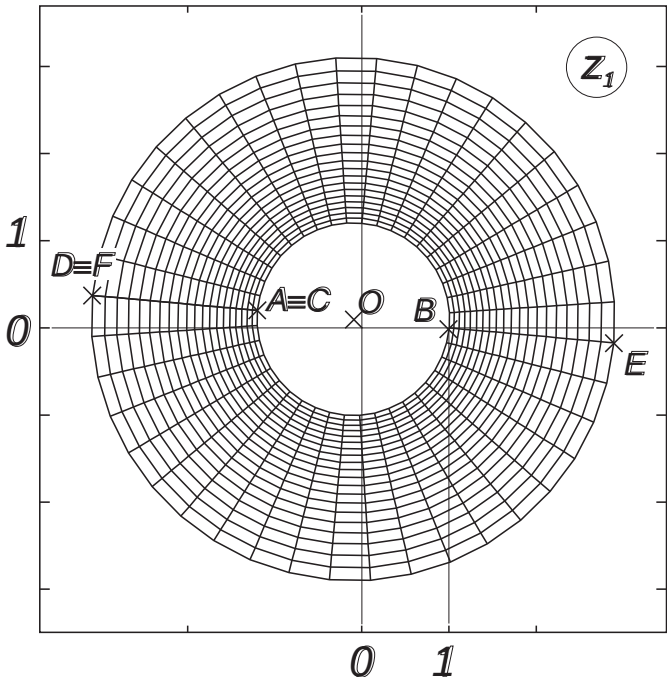


Figura 4.24: Transformación intermedia para correr el centro del círculo, manteniendo el punto $z = 1$ fijo.

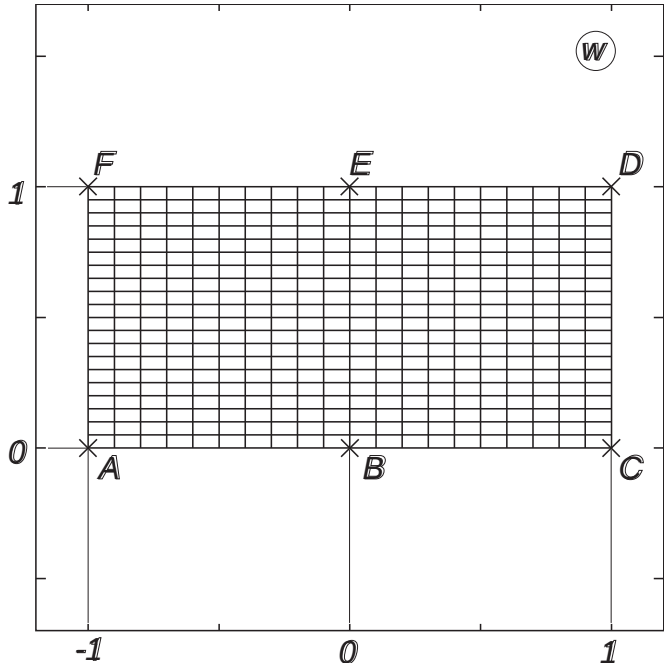


Figura 4.25: Dominio rectangular en el plano w.

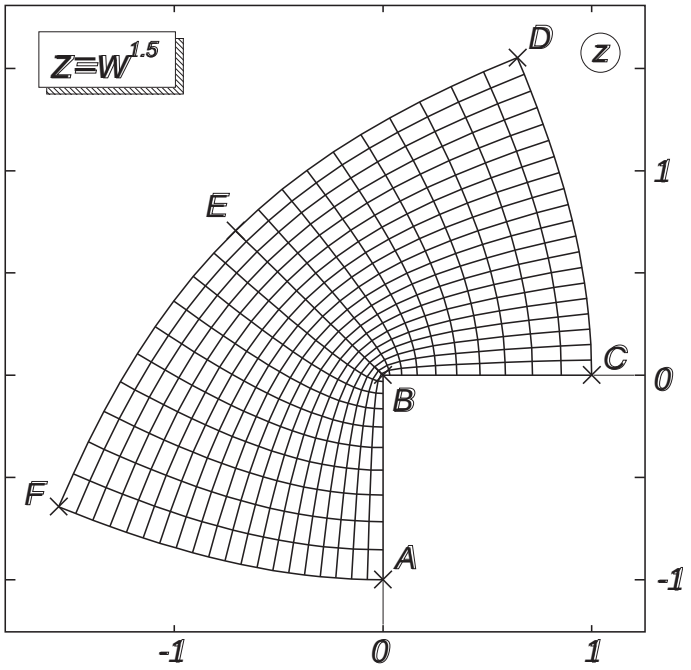


Figura 4.26: Dominio transformado en el plano $z = w^{3/2}$.

4.8. La ecuación de convección-reacción-difusión

Consideremos el transporte de una sustancia de concentración ϕ en un medio fluido con campo de velocidades $\mathbf{v}$ y difusividad $k > 0$. Además consideramos que ϕ s consume con una reacción química de cinética de primer orden, con constante $c > 0$ y que hay una producción de ϕ dada por una densidad de producción g

$$\begin{aligned} \frac{D\phi}{Dt} &= k\Delta\phi - c\phi + g \\ \frac{\partial\phi}{\partial t} + \mathbf{v}\cdot\nabla\phi &= k\Delta\phi - c\phi + g \end{aligned} \tag{4.177}$$

con condiciones de contorno

$$\begin{aligned} \phi &= \bar{\phi}, \quad \text{en } \Gamma_\phi \\ -k\frac{\partial\phi}{\partial n} &= q, \quad \text{en } \Gamma_q \\ -k\frac{\partial\phi}{\partial n} &= h(\phi - \phi_{\text{fl}}), \quad \text{en } \Gamma_h \end{aligned} \tag{4.178}$$

Los términos involucrados en (4.177) se denominan

$$\begin{aligned} \frac{\partial\phi}{\partial t} &= \text{Término temporal} \\ \mathbf{v}\cdot\nabla\phi &= \text{convectivo o de transporte} \\ k\Delta\phi &= \text{difusivo} \\ c\phi &= \text{reacción} \\ g &= \text{producción} \end{aligned} \tag{4.179}$$

Tanto $\mathbf{v}$ como c, g y k pueden ser funciones de la posición y del tiempo $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$, etc... Las dimensiones de las costantes es

$$\begin{aligned} \mathbf{v}[=]&\mathrm{m/sec} \\ k[=]&\mathrm{m}^2/\mathrm{sec} \\ c[=]&\mathrm{sec}^{-1} \\ g[=]&[\phi]/\mathrm{sec} \end{aligned} \tag{4.180}$$

4.8.1. Interpretación de los diferentes términos

Los términos de reacción y producción pueden agruparse como $-c(\phi - \phi_{\text{eq}})$, donde $\phi_{\text{eq}} = g/c$ es la concentración de ϕ que esta en *"equilibrio local"* con la producción. En estado estacionario y con condiciones homogéneas tal que ϕ no depende de x entonces $\phi \to \phi_{\text{eq}}$. (Nota: si $g = f(\phi)$ entonces los zeros de f son puntos de equilibrio.)

Para entender mejor el significado de los diferentes términos involucrados vamos a considerar algunos casos particulares.

No hay dependencia espacial. Si consideramos que ϕ no depende de $\mathbf{x}$ ($\phi \neq \phi(\mathbf{x})$) entonces los términos convectivo y difusivo son nulos y llegamos a una ODE para el valor de ϕ (constante en todo el dominio)

$$\frac{\partial \phi}{\partial t} + c(\phi - \phi_{\text{eq}}) = 0 \tag{4.181}$$

cuya solución es

$$\phi = \phi_{\text{eq}} + (\phi(t=0) - \phi_{\text{eq}})\mathrm{e}^{-ct} \tag{4.182}$$

y vemos que ϕ decae exponencialmente hacia ϕ_{eq}. Podemos deducir de esto que en zonas donde los gradientes son bajos ϕ tiende a aproximarse a ϕ_{eq}.

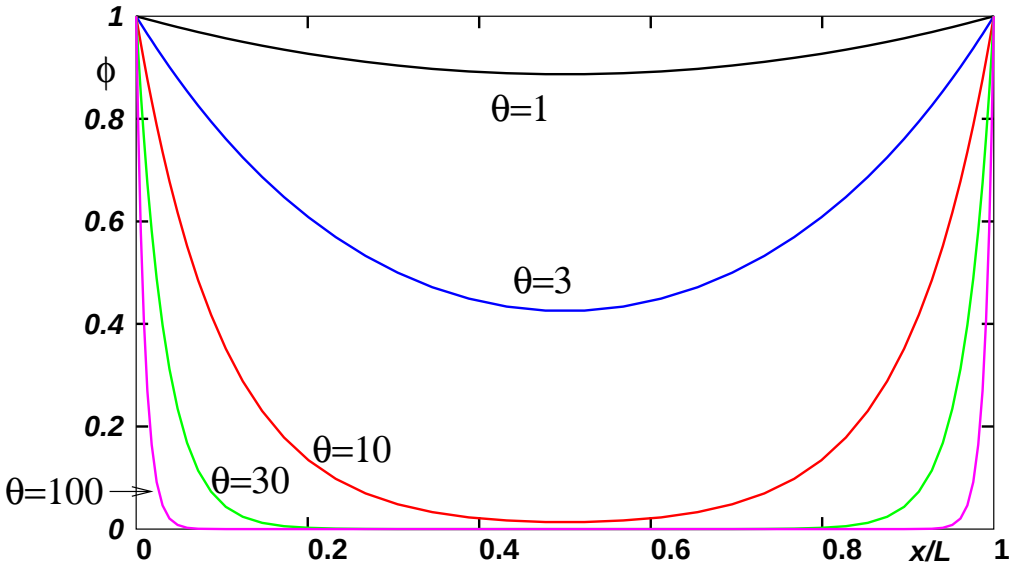


Figura 4.27: Concentración para diferentes valores del módulo de Thiele

Reacción difusión. Ahora, si incluimos la difusión, pero estacionario y con $\mathbf{v} = 0$, entonces la ecuación es

$$k\Delta\phi - c(\phi - \phi_{\text{eq}}) = 0 \tag{4.183}$$

Si $\phi_{\text{eq}} = \text{cte}$, pero la condición de contorno es $\phi_w \neq \phi_{\text{eq}}$ entonces ϕ va a tratar de estar cerca de ϕ_{eq} en el interior del dominio, yendo suavemente hacia ϕ_w en los contornos. La rapidez con la cual el valor interior empalma con la condición de contorno dependerá de la importancia relativa entre c y k. Consideremos, para fijar ideas, el siguiente problema

$$\begin{aligned} &k\Delta\phi - c\phi = 0, \quad \text{en } 0 \leq x \leq L \\ &\phi = 1, \quad \text{en } x = 0, L \end{aligned} \tag{4.184}$$

entonces la solución es

$$\phi = \frac{\cosh(\theta(x/L - 1/2))}{\cosh(\theta/2)}, \;\; \theta = \sqrt{c/k}L \tag{4.185}$$

donde θ es el *"módulo de Thiele"* o *"número de reacción"*. La solución se observa en la figura 4.27 y vemos que para números de Thiele altos la solución se pega al valor de equilibrio en el interior del dominio y empalma con la condición de contorno en una capa límite de esperor $\sqrt{k/c} = L/\theta$. Estas capas límites representan grandes gradientes para la solución y pueden aparejar problemas (falta de convergencia) para los métodos numéricos.

En el caso térmico ϕ es la temperatura, k la difusividad térmica y g una fuente de calor distribuida. c puede provenir de una reacción endotérmica proporcional a la temperatura. Tambien en el caso 2D el término $-c(\phi - \phi_{\text{eq}})$ puede pensarse como un término de enfriamiento Newtoniano.

Notar que si $\mathbf{v} = 0$, los restantes términos sólo contienen derivadas espaciales de orden par (0 ó 2) y por lo tanto son invariantes ante inversión de coordenadas ($x \to -x$). Esto dará lugar después a que las matrices de los métodos numéricos resulten simétricas.

Advección difusión. En el caso estacionario ($(\partial\phi/\partial t) = 0$) y si no hay reacción ni producción ($c, g = 0$) queda la *"ecuación de advección-difusión"*. Consideremos el caso 1D en un intervalo de longitud L con condiciones Dirichlet

$$\begin{aligned} v\frac{\partial\phi}{\partial x} - k\frac{\partial^2\phi}{\partial x^2} &= 0, \quad \text{en } 0 \le x \le L \\ \phi &= 0, \quad \text{en } x = 0 \\ \phi &= 1, \quad \text{en } x = L \end{aligned} \tag{4.186}$$

Esto representa el transporte de temperatura ϕ por un fluido con difusividad k y velocidad v. La solución puede encontrarse por métodos operacionales estándar, proponiendo soluciones de la forma $\mathrm{e}^{\lambda x}$, resolviendo el polinomio característico en λ y buscando la combinación lineal que satisface las condiciones de contorno. La solución resulta ser

$$\phi = \frac{\mathrm{e}^{2\mathrm{Pe}\,(x/L)} - 1}{\mathrm{e}^{2\mathrm{Pe}} - 1}, \quad \mathrm{Pe} = \frac{vL}{2k} \tag{4.187}$$

(ver figura 4.28). Para valores de v muy pequeños la soución se aparta poco de la solución de conducción pura

$$\phi = x/L \tag{4.188}$$

A medida que v aumenta las temperaturas bajan ya que el movimiento del fluido tiende a contrarrestar el efecto de la condición de contorno en $x = L$ y refrigera más que cuando el fluido está quieto. La importancia relativa de ambos términos (difusivo y convectivo) se puede cuantificar a través del *"número de Pèclet"* Pe dado por (4.187). A medida que el Pe aumenta, el gradiente de ϕ se concentra más y más cerca de la pared $x = 1$, formando una capa límite de espesor

$$\delta = O(k/v) = O(L/\mathrm{Pe}) \tag{4.189}$$

De nuevo, estos altos gradientes son una fuente de problemas para los métodos numéricos.

Si la velocidad se invierte entonces la discontinuidad se produce en $x = 0$ que es la nueva salida ($x = L$ es la entrada).

Una forma diferente de ver este fenómeno es considerar que, a medida que $k \to 0$, el problema se hace cada vez más advectivo (Pe $\to \infty$). Ahora bien, para el problema advectivo puro la ecuación es de primer orden y por lo tanto requiere de una sola condición de contorno. La teoría de sistemas hiperbólicos indica que la condición de contorno debe aplicarse donde las líneas características entran. El valor de la variable en un contorno donde las características salen resulta de la integración de la ecuación dentro del dominio. Si se pretende imponer un valor diferente como condición de contorno Dirichlet, la diferencia se absorbe en una capa límite.

Debido a que la ecuación de advección pura propaga los valores a lo largo de las características, también se pueden producir altos gradientes en el interior del dominio. Por ejemplo consideremos advección pura en u dominio rectangular $ADD'A'$ como se muestra en la figura 4.30. El flujo entra por el lado AA' y sale por DD'. La condición a la entrada es $\phi = \bar{\phi}$ donde $\bar{\phi}$ contiene un salto cerca del punto W. El transporte convectivo tiende a propagar esta discontinuidad a lo largo de la característica WW', pero debido a la difusión la discontinuidad se va suavizando y termina en un escalón suavizado a la salida DD'.

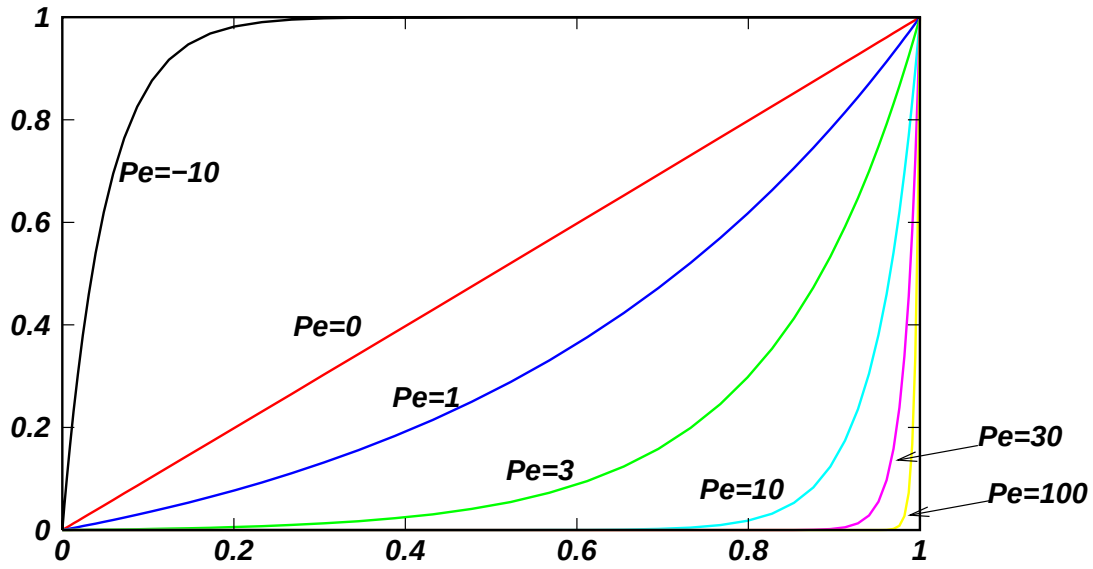


Figura 4.28: Solución al problema de advección pura 1D con coeficientes constantes

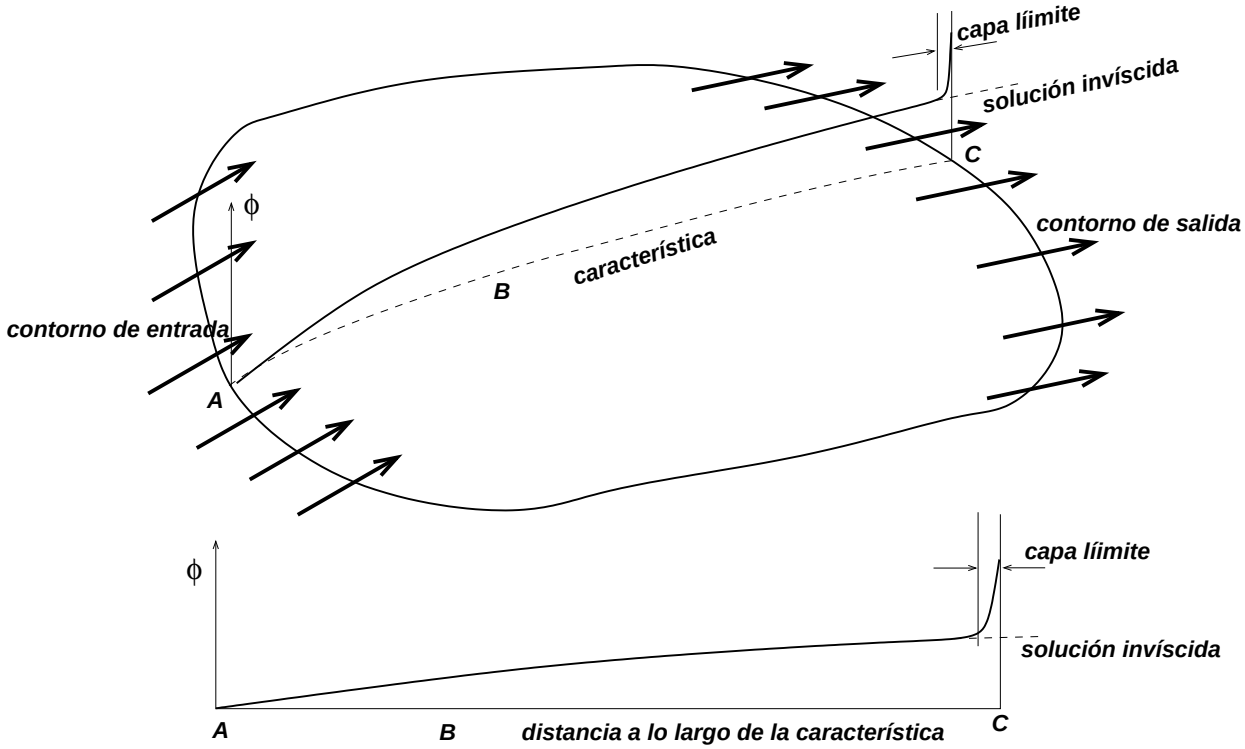


Figura 4.29: Sistemás fuertemente advectivos en 2D

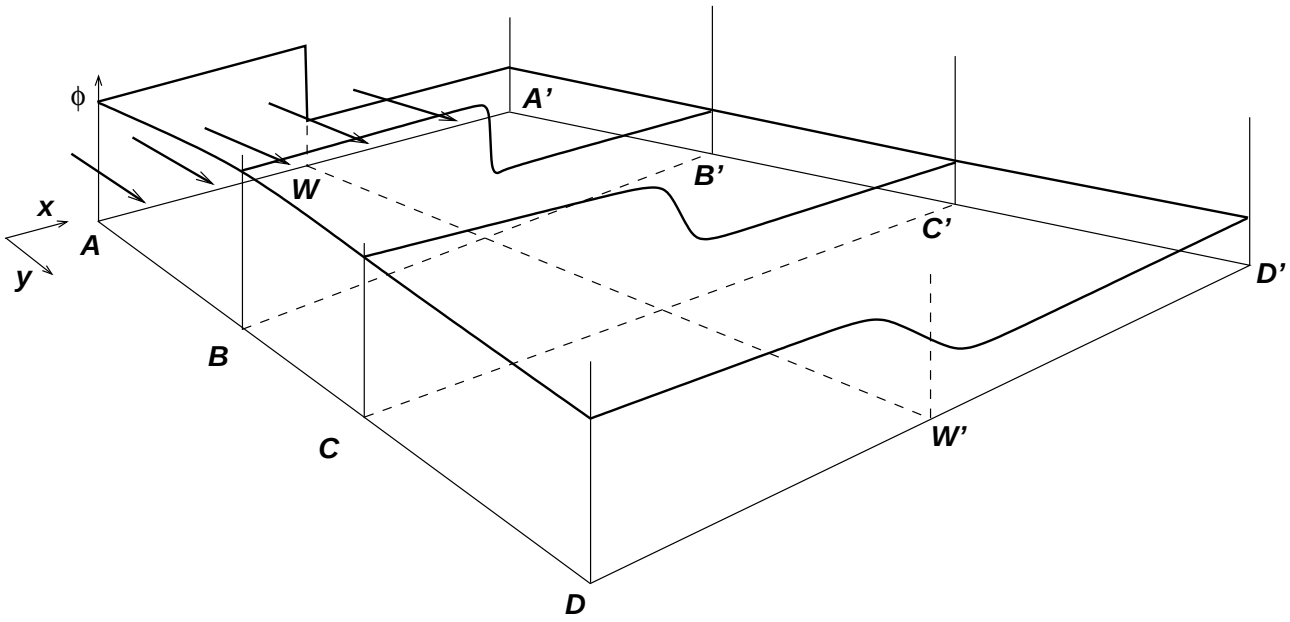


Figura 4.30: Discontinuidad interna propagada desde la condición de en un sistema fuertemente advectivo

4.8.2. Discretización de la ecuación de advección-difusión

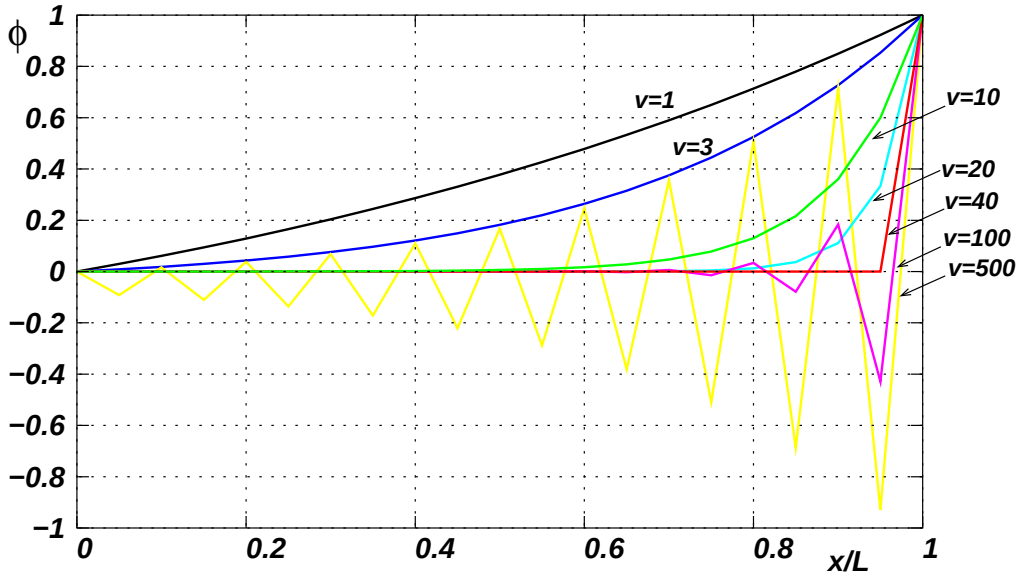


Figura 4.31: Solución numérica al problema de advección difusión con un esquema centrado

Consideremos el problema 1D de advección difusión a coeficientes constantes (4.186). Consideremos una malla uniforme de N segmentos de longitud $\Delta x = L/N$. Los nodos son $x_j = (j-1)\Delta x$, para $j = 1, \ldots, N+1$. Una discretización centrada de segundo orden es

$$v\,\frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} - k\,\frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{\Delta x^2} = 0 \tag{4.190}$$

Notar que la derivada de primer orden introduce un término antisimétrico. Este esquema funciona bien mientras la velocidad se mantenga por debajo de un cierto límite, que resulta ser $v_{\text{crit}} = 2k/\Delta x$ (En la figura $k = 1$ y el número de puntos es $N = 20$, de manera que $\Delta x = 0.05$ y $v_{\text{crit}} = 40$). Para velocidades mayores la solución numérica se vuelve oscilatoria y para velocidades mucho más grandes que la crítica las oscilaciones contaminan todo el dominio. Nótese que la velocidad crítica se produce cuando el Pèclet de la malla es

$$\text{Pe}_{\Delta x} = \frac{v\Delta x}{2k} = 1 \tag{4.191}$$

Estas oscilaciones pueden asociarse a un falta de estabilidad del esquema numérico, sin embargo el esquema es estable, estrictamente hablando, ya que si refinamos suficientemente entonces $\text{Pe}_{\Delta x}$ pasará a ser menor que uno y se recupera la convergencia $O(\Delta x^2)$. Sin embargo vemos que si tomamos la estimación de error estándar

$$\|\mathbf{E}_\phi\| \leq C\Delta x^2, \tag{4.192}$$

la constante C depende de Pe. O sea, no existe un C independiente de Pe tal que (4.192) valga para todo Δx y Pe, es decir *no se puede obtener una cota de error uniforme sobe Pe*. Bajo esta definición de estabilidad más estricta el esquema es inestable.

4.8.3. Desacoplamiento de las ecuaciones

Si consideramos las ecuaciones discretas del esquema centrado (4.190) para Pe $\to \infty$, es decir $k = 0$, obtenemos

$$v\,\frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} = 0, j = 2, \ldots, N \tag{4.193}$$

Esta ecuación dice que

$$\begin{aligned} &\phi_1 = \phi_3 = \phi_5 = \cdots = \phi_{2n+1} = \ldots \\ &\phi_{N+1} = \phi_{N-1} = \phi_{N-3} = \ldots \end{aligned} \tag{4.194}$$

Entonces, si N es impar existe una solución que es

$$\phi_j = \begin{cases} 0 & ; \text{ si } j = \text{ impar} \\ 1 & ; \text{ si } j = \text{ par} \end{cases} \tag{4.195}$$

y por otro lado no existe solución si N es par.

Se dice que hay un *"desacoplamiento"* de las ecuaciones para los nodos pares e impares, lo cual es asociado normalmente a una falta de estabilidad del esquema numérico.

4.8.4. Esquemas de diferencias contracorriente (upwinded)

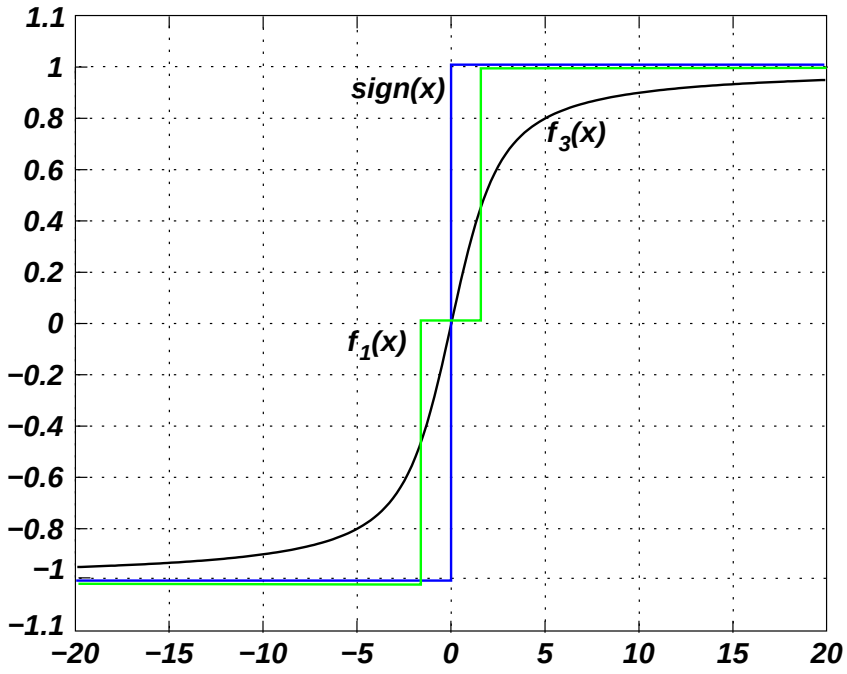


Figura 4.32: Diferentes propuestas para el parámetro de estabilidad

Notemos que, en el caso de advección pura (Pe $= \infty$) la solución numérica debería ser

$$\phi_j = \begin{cases} 0 & ; \text{ si } j <= N \\ 1 & ; \text{ si } j = N + 1 \end{cases} \tag{4.196}$$

y esto se lograría si reemplazamos la derivada centrada en (4.193) por una derivada lateral a izquierda (también llamada *"contracorriente"* o *"upwinded"*)

$$v\,\frac{\phi_j-\phi_{j-1}}{\Delta x}=0, j=2,\ldots,N \tag{4.197}$$

Pero esto se puede reescribir como

$$v\,\frac{\phi_{j+1}-\phi_{j-1}}{2\Delta x}-\left(\frac{v\Delta x}{2}\right)\frac{\phi_{j+1}-2\phi_j+\phi_{j-1}}{\Delta x^2}=0 \tag{4.198}$$

o, poniendo $k_{\text{num}}=v\Delta x/2$,

$$v\,\frac{\phi_{j+1}-\phi_{j-1}}{2\Delta x}-k_{\text{num}}\,\frac{\phi_{j+1}-2\phi_j+\phi_{j-1}}{\Delta x^2}=0 \tag{4.199}$$

donde k_{num} es una difusión *"numérica artificial"* que *"estabiliza"* el esquema. Notar que, si $v<0$, entonces el esquema debe estar decentrado a derecha

$$v\,\frac{\phi_{j+1}-\phi_j}{\Delta x} \tag{4.200}$$

y entonces debe ser $k_{\text{num}}=-v\Delta x/2$, de manera que, en general

$$k_{\text{num}}=\frac{|v|\Delta x}{2} \tag{4.201}$$

Ahora bien, para $\text{Pe}<\infty$ podemos tomar

$$v\,\frac{\phi_{j+1}-\phi_{j-1}}{2\Delta x}-(k+k_{\text{num}})\,\frac{\phi_{j+1}-2\phi_j+\phi_{j-1}}{\Delta x^2}=0 \tag{4.202}$$

Este esquema no presenta más inestabilidades y para $\text{Pe}\gg 1$ se aproxima al upwindado pero para Pe pequeños sigue agregando difusión, incluso para $\text{Pe}_{\Delta x}<1$ cuando sabemos que el esquema es estable, resultando en un esquema demasiado difusivo. Entonces surge la idea de tratar de reducir la difusión numérica para valores de $\text{Pe}_{\Delta x}$ pequeños. Notemos que (4.201) puede reescribirse como

$$k_{\text{num}}=\frac{|v|\Delta x}{2}=k|\text{Pe}_{\Delta x}|=k\text{Pe}_{\Delta x}\,f(\text{Pe}_{\Delta x})=\frac{v\Delta x}{2}\,f(\text{Pe}_{\Delta x}) \tag{4.203}$$

con

$$f(x)=\text{sign}(x) \tag{4.204}$$

de manera que podríamos reemplazar la *"función de upwinding"* $\text{sign}()$ por otra que anule la difusión numérica para $|\text{Pe}_{\Delta x}|<=1$, por ejemplo podemos usar

$$k_{\text{num}}=\frac{v\Delta x}{2}\,f_1(\text{Pe}_{\Delta x}) \tag{4.205}$$

con

$$f_1(x)=\begin{cases}\text{sign}(x) & ;\ \text{si } |x|>1\\ 0 & ;\ \text{si } |x|<=1\end{cases} \tag{4.206}$$

o también, para hacerlo más continuo

$$f_2(x) = \begin{cases} \text{sign}(x) & ;\ \text{si } |x| > 1 \\ x & ;\ \text{si } |x| \leq 1 \end{cases} \tag{4.207}$$

Puede demostrarse que, si elegimos $f(x)$, como la siguiente *"función mágica"*

$$f_3(x) = \alpha(x) = \frac{1}{\tanh(x)} - \frac{1}{x} \tag{4.208}$$

entonces, en este caso simple se obtiene la solución exacta (4.187) (de ahí el nombre de función mágica). Sin embargo, esto sólo vale mientras el problema sea 1D, con coeficientes constantes, paso de la malla constante y sin término fuente. De todas formas es una función de upwinding interesante, ya que en forma muy suave satisface todos los límites necesarios.

Como $(\mathrm{d}\alpha/\mathrm{d}x)\,|_{x=0} = 1/3$ otra función comunmente utilizada es

$$f_4(x) = \begin{cases} x/3 & ;|x| < 3 \\ \text{sign}(x) & ;|x| > 3 \end{cases} \tag{4.209}$$

4.8.5. El caso 2D

El primer intento por extender el esquema upwindado a 2D (o 3D) es agregar una viscosidad numérica según cada una de las direcciones principales de la malla. Sea un dominio rectangular $0 \leq x \leq L_x$, $0 \leq y \leq L_y$, dividido en N_x, N_y intervalos de longitud $\Delta x = L_x/N_x$, $\Delta y = L_y/N_y$. Los puntos de la malla están ubicados en $x_{jk} = ((j-1)\Delta x, (k-1)\Delta y)$ y sea $\phi(x_{jk}) \approx \phi_{j,k}$. Ademas, consideremos $\mathbf{v} = (v_x, v_y) = \text{cte}$,

$$\begin{aligned} v_x \frac{\phi_{j+1,k} - \phi_{j-1,k}}{2\Delta x} + v_y \frac{\phi_{j,k+1} - \phi_{j,k-1}}{2\Delta y} - \\ - (k + k^x_{\text{num}}) \frac{\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}}{\Delta x^2} - (k + k^y_{\text{num}}) \frac{\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}}{\Delta y^2} = 0 \end{aligned} \tag{4.210}$$

donde

$$\begin{aligned} k^x_{\text{num}} &= \frac{v_x \Delta x}{2} f(\text{Pe}_x) \\ k^y_{\text{num}} &= \frac{v_y \Delta y}{2} f(\text{Pe}_y) \end{aligned} \tag{4.211}$$

Este esquema resulta ser estable. Si el flujo esta alineado con la malla, es decir $v_x = 0$ o $v_y = 0$, entonces el esquema reproduce exactamente sus virtudes en el caso 1D. En general la viscosidad numérica es *"anisotrópica"*, por ejemplo, si $\mathbf{v} = (v, 0)$ entonces

$$\begin{aligned} k^x_{\text{num}} &= \frac{v \Delta x}{2} f(\text{Pe}_x) \\ k^y_{\text{num}} &= 0 \end{aligned} \tag{4.212}$$

o, en forma tensorial

$$\mathbf{k}_{\text{num}} = \begin{bmatrix} k^x_{\text{num}} & 0 \\ 0 & 0 \end{bmatrix} \tag{4.213}$$

y, en general, $\mathbf{k}_{\text{num}}$ es un tensor anisotrópico, con ejes principales a lo largo de los ejes principales de la malla.

Si consideramos una velocidad cruzada con la malla, entonces el esquema resulta ser demasiado disipativo. Por ejemplo, consideremos $\mathbf{v} = v\,(1,1)/\sqrt{2}$, y $\Delta x = \Delta y = h$. Entonces

$$k^x_{\text{num}} = k^y_{\text{num}} = k_{\text{num}} = \frac{v_x \Delta x}{2}\, f(\text{Pe}_x) \tag{4.214}$$

es decir que el tensor de difusión numérica es escalar

$$\mathbf{k}_{\text{num}} = \begin{bmatrix} k_{\text{num}} & 0 \\ 0 & k_{\text{num}} \end{bmatrix} = k_{\text{num}}\, \mathbf{I}_{2\times 2} \tag{4.215}$$

donde $\mathbf{I}_{2\times 2}$ es el tensor identidad de 2 por 2. Ahora bien, si consideramos un sistema de coordenadas alineado con la velocidad, es decir $x' \parallel \mathbf{v}$ e $y' \perp \mathbf{v}$, entonces el tensor difusividad numérica sigue siendo

$$\mathbf{k}'_{\text{num}} = \begin{bmatrix} k_{\text{num}} & 0 \\ 0 & k_{\text{num}} \end{bmatrix} = k_{\text{num}}\, \mathbf{I}_{2\times 2} \tag{4.216}$$

mientras que lo deseable sería tener una difusividad según x' pero no según y'

$$\mathbf{k}'_{\text{num}} = \begin{bmatrix} k'^x_{\text{num}} & 0 \\ 0 & 0 \end{bmatrix} \tag{4.217}$$

con

$$k'^x_{\text{num}} = \frac{v h_s}{2}\, f(\text{Pe}_s) \tag{4.218}$$

donde el subíndice s indica valores según la línea de corriente, así por ejemplo

$$\begin{aligned} h_s &= \sqrt{2}h \\ \text{Pe}_s &= \frac{v h_s}{2k} \end{aligned} \tag{4.219}$$

Antitransformando el tensor (4.217) a los ejes $x - y$ obtenemos

$$k_{ij} = \frac{v h_s}{2}\, s_i s_j\, f(\text{Pe}_s) \tag{4.220}$$

donde $\hat{\mathbf{s}} = \mathbf{v}/v$ es el versor según la línea de corriente.

Finalmente el esquema es

$$\begin{aligned} v_x \frac{\phi_{j+1,k} - \phi_{j-1,k}}{2h} + v_y \frac{\phi_{j,k+1} - \phi_{j,k-1}}{2h} - \\ - (k + k_{\text{num},xx}) \frac{\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}}{h^2} \\ - (k + k_{\text{num},yy}) \frac{\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}}{h^2} \\ - 2k_{\text{num},xy} \frac{\phi_{j+1,k+1} - \phi_{j+1,k-1} - \phi_{j-1,k+1} + \phi_{j-1,k-1}}{4h^2} = 0. \end{aligned} \tag{4.221}$$

Para el caso $\Delta x \neq \Delta y$ tenemos

$$v_x \frac{\phi_{j+1,k} - \phi_{j-1,k}}{2\Delta x} + v_y \frac{\phi_{j,k+1} - \phi_{j,k-1}}{2\Delta y} - \\ - (k + k_{\text{num},xx}) \frac{\phi_{j+1,k} - 2\phi_{j,k} + \phi_{j-1,k}}{\Delta x^2} \\ - (k + k_{\text{num},yy}) \frac{\phi_{j,k+1} - 2\phi_{j,k} + \phi_{j,k-1}}{\Delta y^2} \\ - 2k_{\text{num},xy} \frac{\phi_{j+1,k+1} - \phi_{j+1,k-1} - \phi_{j-1,k+1} + \phi_{j-1,k-1}}{4\Delta x \Delta y} = 0. \quad (4.222)$$

Notar la expresión en diferencias de la última fila que es una aproximación $O(\Delta x \Delta y)$ para $\left(\partial^2\phi/\partial x \partial y\right)$.

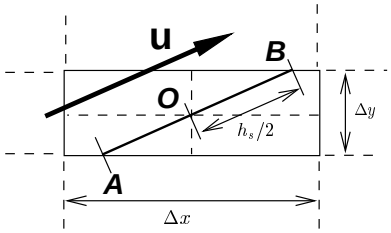


Figura 4.33: Definición del tamaño de la celda según la línea de corriente

Falta definir la expresión general para h_s en función del ángulo que forma $\mathbf{v}$ con la malla. Existen varias propuestas para esto, no siendo ninguna de ellas totalmente satisfactoria. La más simple podría ser tomar alguna media de los parámetros de la malla $h = (\Delta x + \Delta y)/2$ o $h = \sqrt{\Delta x \Delta y}$. Notar que en realidad estas definiciones no son *"según la línea de corriente"*. Consecuentemente, es de esperarse que puedan ser muy sobre- o sub-difusivas en ciertos casos. Una posibilidad mejor es tomar la mayor distancia dentro de la celda a lo largo de una dirección paralela a $\mathbf{v}$ como el segmento AB en la figura 4.33. La expresión resulta ser

$$h_s = \min_j \frac{\Delta x_j}{|s_j|} \quad (4.223)$$

4.8.6. Resolución de las ecuaciones temporales

Hasta ahora vimos como discretizar las ecuaciones espacialmente para obtener un estado estacionario. Consideremos ahora un problema no estacionario. La idea es primero discretizar en el espacio, tal cual como se ha hecho hasta ahora. Por ejemplo consideremos la ecuación de advección-reacción-difusión lineal 1D,

$$\frac{\partial \phi}{\partial t} + v \frac{\partial \phi}{\partial x} = k\Delta\phi - c\phi + g \quad (4.224)$$

entonces un esquema estabilizado posible es

$$\dot{\phi}_j + v \frac{\phi_{j+1} - \phi_{j-1}}{2\Delta x} - (k + k_{\text{num}}) \frac{\phi_{j+1} - 2\phi_j + \phi_{j-1} + c\phi_j}{\Delta x^2} = g_j \quad (4.225)$$

donde $\dot{\phi}$ significa la derivada temporal de ϕ. Estas ecuaciones representan un *"sistema de ecuaciones diferenciales ordinarias"* (ODE's) acoplado

$$\dot{\boldsymbol{\phi}} + \mathbf{K}\boldsymbol{\phi} = \mathbf{f}. \tag{4.226}$$

Al mismo puede aplicarse una serie de esquemas de integración temporal, por ejemplo

$$\begin{aligned} \frac{\boldsymbol{\phi}^{n+1} - \boldsymbol{\phi}^n}{\Delta t} + \mathbf{K}\,\boldsymbol{\phi}^n &= \mathbf{f}^n, \text{ forward Euler} \\ \frac{\boldsymbol{\phi}^{n+1} - \boldsymbol{\phi}^n}{\Delta t} + \mathbf{K}\,\boldsymbol{\phi}^{n+1} &= \mathbf{f}^{n+1}, \text{ backward Euler} \\ \frac{\boldsymbol{\phi}^{n+1} - \boldsymbol{\phi}^n}{\Delta t} + \mathbf{K}\,\frac{\boldsymbol{\phi}^{n+1} + \boldsymbol{\phi}^n}{2} &= \mathbf{f}^{n+1/2}, \text{ Crank Nicholson} \end{aligned} \tag{4.227}$$

Donde $\boldsymbol{\phi}(t^n) \approx \boldsymbol{\phi}^n$, $t^n = n\Delta t$. El método de forward Euler permite avanzar un paso de tiempo con un costo en tiempo de procesamiento y memoria de almacenamiento muy bajos, ya que no necesita resolver ningún sistema lineal. Por el contrario, en los otros dos casos sí se necesita resolver un sistema. Si el problema es lineal, entonces podemos notar que la matriz del sistema $\mathbf{K}$ es constante (no varía con el tiempo) de manera que podemos factorizarla una vez y posteriormen hacer solamente una retrosustitución. De todas formas el tiempo y memoria necesarios para avanzar un paso de tiempo en el backward Euler es mucho mayor que para el forward. Pero el forward tiene la limitación de que, para pasos de tiempo grandes la solución se hace inestable y diverge. El paso de tiempo crítico viene dado por

$$\Delta t_{\text{crit}} < \min\left(\frac{h}{v}, \frac{h^2}{4k}\right). \tag{4.228}$$

Notar que estos criterios pueden ponerse como

$$\text{Co} = \frac{v\Delta t}{h} < 1, \;\; \text{Fo} = \frac{k\Delta t}{4h^2} < 1 \tag{4.229}$$

donde Co, Fo son los números adimensionales de Courant y Fourier.

4.9. Solución exacta al problema de conducción del calor con generación constante en un cuadrado

Ecuaciones de gobierno

$$\begin{aligned} \Delta\phi = -1, \;\; 0 \le x, y \le 1 \\ \phi = 0, \;\; x, y = 0, 1 \end{aligned} \tag{4.230}$$

Proponemos un desarrollo

$$\phi(x, y) = \sum_{j,k=1}^{\infty} a_{jk}\, \sin(j\pi x)\, \sin(l\pi y). \tag{4.231}$$

Este desarrollo satisface las condiciones de contorno y representa una base completa. Por simetría podemos descartar los términos para j, k pares, ya que involucrarían funciones antisimétricas con respecto a $x, y = 12$. Aplicando el operador de Laplace obtenemos

$$\sum_{j,k=1}^{\infty} \pi^2(j^2 + k^2)\, a_{jk}\, \sin(j\pi x)\, \sin(l\pi y) = 1 \tag{4.232}$$

Multiplicamos miembro a miembro por $\sin(l\pi x)\,\sin(m\pi y)$ e integrando sobre el cuadrado. Usando que

$$\int_0^1 \sin(l\pi x)\,\sin(j\pi x)\,\mathrm{d}x = \tfrac{1}{2}\,\delta_{jl}$$
$$\int_0^1 \sin(l\pi x)\,\mathrm{d}x = \begin{cases} 2/(\pi l) & ;\ l \text{ impar} \\ 0 & ;\ l \text{ par} \end{cases} \tag{4.233}$$

De manera que

$$\tfrac{1}{4}\pi^2(j^2+k^2)\,a_{jk} = \frac{4}{\pi^2 lm}, \qquad (l, m \text{ impares}). \tag{4.234}$$

Es decir que

$$a_{jk} = \frac{16}{\pi^4 lm(l^2+m^2)}, \quad (l, m \text{ impares}). \tag{4.235}$$

Pero

$$\sin((2p+1)\pi x)|_{x=\frac{1}{2}} = \sin((p+\tfrac{1}{2})\pi) = s_p = \begin{cases} +1 & ; \quad p \text{ par} \\ -1 & ; \quad p \text{ impar} \end{cases} \tag{4.236}$$

De manera que el valor máximo de ϕ, que ocurre en el centro del cuadrado, es

$$\max_{0\le x,y\le 1} \phi(x,y) = \phi(\tfrac{1}{2},\tfrac{1}{2}) = \sum_{l,m=1,\ \text{impar}}^{\infty} \frac{16 s_l s_m}{\pi^4 lm(l^2+m^2)} \approx 0.073671 \pm 10^{-6} \tag{4.237}$$

Capítulo 5

Técnicas de discretización

Habiendo presentado en la primera parte los modelos matemáticos que rigen el movimiento de los fluidos y estableciendo el conjunto de ecuaciones que gobiernan el caso general y algunos otros particulares ahora estamos en condiciones de pasar a tratar algunos aspectos numéricos relacionados con el diseño de los esquemas más comúnmente empleados en fluidodinámica computacional. Una de las técnicas más empleadas en fluidodinámica computacionalha sido la de diferencias finitas. Esta fue una de las primeras en aparecer y conserva vigencia a pesar de algunas restricciones propias de la técnica. Su alta eficiencia para la resolución de problemas definidos en geometrías sencillas lo hace muy atractivo y es muchas veces la mejor opción cuando existe la posibilidad de mapear el dominio real en otro completamente regular y estructurado. Su definición se basa en aproximar los operadores diferenciales por otros denominados *operadores en diferencias* que se aplican a un vector de datos que representa la solución en un conjunto finito de puntos en el dominio. Esta forma de discretizar un operador diferencial es una alternativa y no la única para tal fin. Desventajas claras del método como su difícil implementación en problemas gobernados por geometrías arbitrarias hizo que en los últimos años gran cantidad de investigación en el area de fluidodinámica computacionalse volcase al uso de otras técnicas alternativas. La mayoría de las mismas se pueden presentar bajo un método general conocido como el *método de los residuos ponderados* (WRM).

5.1. Método de los residuos ponderados

5.1.1. Introducción

Este método es conceptualmente diferente a aquel empleado en diferencias finitas ya que asume que la solución a un problema planteado puede ser analíticamente representable mediante una expansión del tipo:

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M} a_m N_m \tag{5.1}$$

donde en general a_m forman un conjunto finito de coeficientes con M la dimensión del espacio finito dimensional empleado y N_m las funciones analíticas conocidas y elegidas para representar o expandir a la solución del problema. Estas funciones son comúnmente denominadas *soluciones de prueba* y su elección hace a la diferencia entre una vasta cantidad de métodos numéricos, como veremos en breve. La función ψ

es introducida con el propósito de satisfacer las condiciones de contorno del problema. Sea Γ el contorno del dominio Ω del problema, entonces la elección de la función ψ es tal que

$$\begin{aligned} \psi|_\Gamma &= \phi|_\Gamma \\ N_m|_\Gamma &= 0 \qquad \forall m \end{aligned} \tag{5.2}$$

De la elección de ψ y N_m dependerá la calidad de la solución y la convergencia del método numérico a medida que se refina la discretización, o sea que $M \to \infty$. Este requisito denominado *completitud* está sustentado fuertemente en bases matemáticas con lo que a diferencia del método de las diferencias finitas esta clase de técnica goza con el apoyo de sólidos conceptos matemáticos de teoría de operadores, análisis funcional y análisis numérico. Si bien nuestra aplicación será la de obtener soluciones a ecuaciones a derivadas parciales un buen ejercicio para introducir los conceptos de aproximación es el caso simple de aproximar una función conocida a priori con una expansión del tipo (5.1) .

Aproximación puntual de funciones

Este caso simple consiste en dada una función ϕ aproximarla por $\hat{\phi}$ bajo el requisito que ambas coincidan en M puntos distintos elegidos arbitrariamente sobre Ω. Este requisito conduce a un sistema de ecuaciones algebraicas lineales en el conjunto de parámetros incógnita $\{a_m; m = 1, 2, \ldots, M\}$.

Para graficar la explicación supongamos una función a aproximar como
$\phi = \sin(-1.8\pi x) + x$ graficada en la parte superior izquierda de la figura 5.1 en trazo lleno.
En la misma figura se muestra la función lineal ψ que satisface las condiciones de contorno, o sea

$$\begin{aligned} \psi(x = 0) &= \phi(x = 0) = 0 \\ \psi(x = 1) &= \phi(x = 1) = \sin(-1.8\pi) + 1 \end{aligned} \tag{5.3}$$

La rutina Ej_2_0.m permite definir una aproximación a ϕ usando una cantidad M de términos a elección del alumno. En la figura 5.1 se grafica en la parte superior izquierda en linea de puntos la solución numérica obtenida con 2 términos y a su derecha el valor absoluto del error donde como vemos este vale cero en un conjunto de puntos equidistribuidos cuya cantidad coincide con M. Las dos gráficas del medio corresponden a $M = 3$ mientras que las dos de abajo representan solamente el error que se comete cuando $M = 4$ y $M = 9$.

La forma de construir el sistema de ecuaciones algebraicas a resolver es bastante simple. Se debe reemplazar (5.1) para los M puntos interiores al dominio e igualarlos a los valores de la función ϕ en esos nodos. Esto, para el caso de $M = 2$ genera lo siguiente:

$$\begin{aligned} \psi(x_1) + N_1(x_1)a_1 + N_2(x_1)a_2 &= \phi(x_1) \\ \psi(x_2) + N_1(x_2)a_1 + N_2(x_2)a_2 &= \phi(x_2) \\ \begin{pmatrix} N_1(x_1) & N_2(x_1) \\ N_1(x_2) & N_2(x_2) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} &\begin{pmatrix} \phi(x_1) - \psi(x_1) \\ \phi(x_2) - \psi(x_2) \end{pmatrix} \end{aligned} \tag{5.4}$$

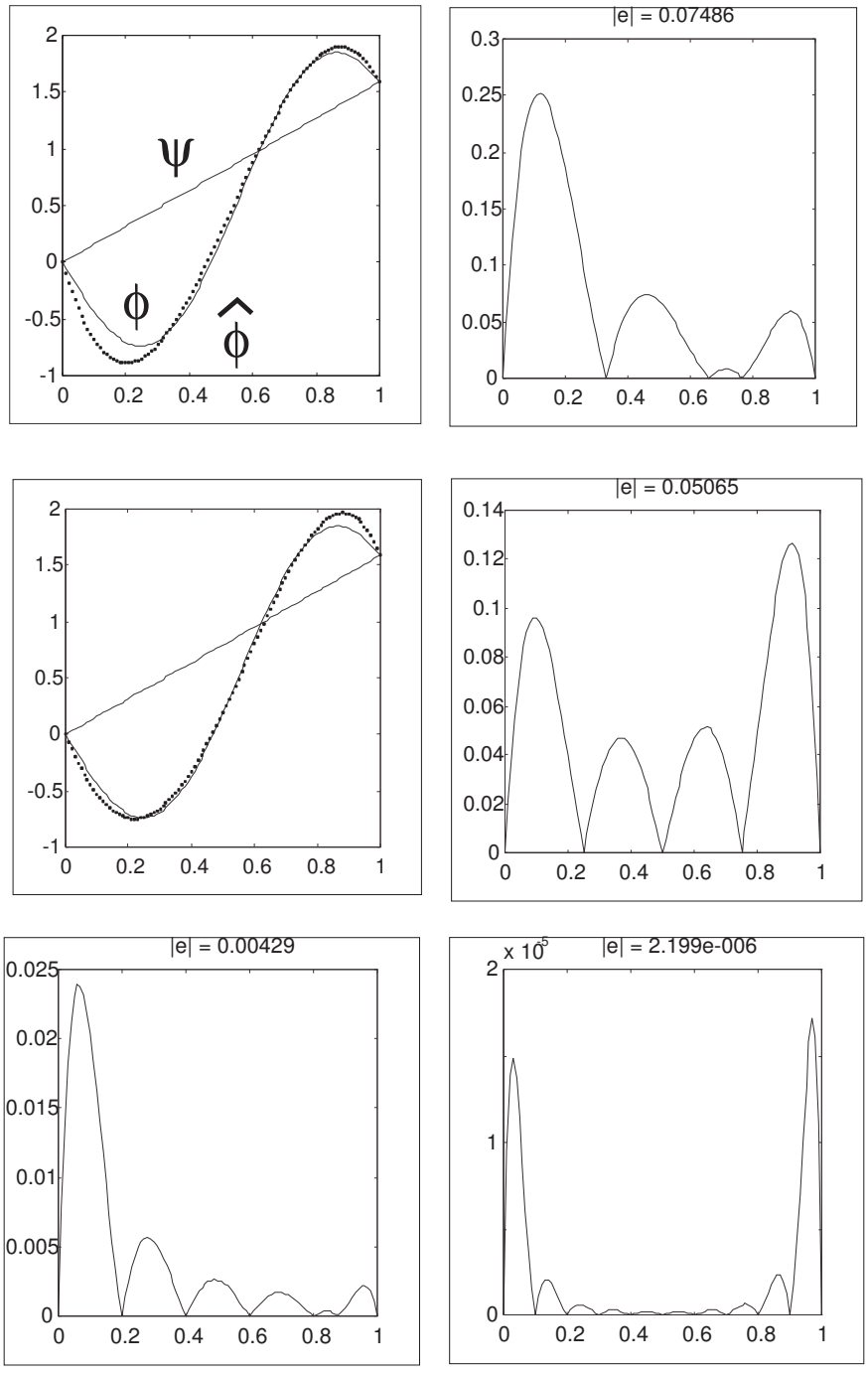


Figura 5.1: Aproximación de una función por ajuste puntual

La figura 5.2 muestra como varía el error en un gráfico semilogarítmico siendo su convergencia del tipo espectral. En general el error en la aproximación se puede escribir como:

$$\begin{aligned}(|e|) &= Ch^M \\ \log(|e|) &= log(C) + Mlog(h)\end{aligned} \tag{5.5}$$

por lo que la pendiente en el gráfico nos da una idea de la convergencia en función de la discretización.

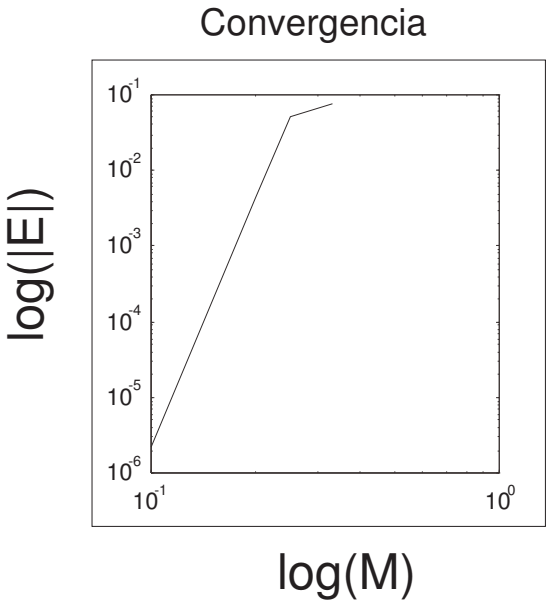


Figura 5.2: Convergencia de la aproximación

Aproximación por series de Fourier

Usando la teoría de series de Fourier es posible aproximar una función ϕ arbitraria siempre que esta cuente con un número finito de discontinuidades y de extremos locales, cosa que casi siempre ocurre en las aplicaciones. Entonces la aproximación se escribe como:

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M} a_m \sin \frac{m\pi x}{L_x} \qquad 0 \leq x \leq L_x \tag{5.6}$$

La inherente completitud de que gozan las series de Fourier le confiere la propiedad que al incrementar la dimensión del espacio de trabajo la precisión mejora.

5.1.2. Aproximación por residuos ponderados

A continuación se presenta un método general que permite hallar los coeficientes de (5.1) siendo las dos formas anteriores casos particulares del mismo. Definamos el error o residuo R_Ω en la aproximación como:

$$R_\Omega = \phi - \hat{\phi} \tag{5.7}$$

siendo ésta una función de la posición en el dominio Ω. En la figura chapV-1 hemos presentado funciones de este tipo. La idea es que en lugar de pedir que esta función R_Ω sea idénticamente nula en todo el dominio le pedimos que integrada mediante alguna función de peso esta sea nula, es decir:

$$\int_\Omega W_l(\phi - \hat{\phi})d\Omega = \int_\Omega W_l R_\Omega d\Omega = 0 \qquad\qquad l = 1, 2, \ldots, M \tag{5.8}$$

con W_l un conjunto de funciones de peso independientes. Entonces en lugar de pedir que $\hat{\phi} \to \phi$ como $M \to \infty$ le pedimos que chapV-7 se satisfaga para todo l como $M \to \infty$. De alguna manera se puede verificar que esto último equivale a asumir que $R_\Omega \to 0$ en todo el dominio. Reemplazando $\hat{\phi}$ de (5.1) en (5.7) obtenemos un conjunto o sistema de ecuaciones algebraicas lineales para los coeficientes incógnitas a_m que puede ser escrito en forma genérica como:

$$\begin{aligned} \mathbf{Ka} &= \mathbf{f} \\ \mathbf{a}^T &= \begin{pmatrix} a_1 & a_2 & \ldots & a_M \end{pmatrix} \\ K_{lm} &= \int_\Omega W_l N_m d\Omega \qquad\qquad 1 \leq l, m \leq M \\ f_l &= \int_\Omega W_l(\phi - \psi)d\Omega \qquad\qquad 1 \leq l \leq M \end{aligned} \tag{5.9}$$

Una vez que la función ϕ se conoce la aproximación se define mediante la elección de ψ y las funciones de peso W_l y de prueba N_m. Diferentes funciones de peso dan orígen a diferentes métodos, todos del tipo de residuos ponderados.

Método de colocación puntual

En este método las funciones de peso son de la forma:

$$W_l = \delta(x - x_l) \tag{5.10}$$

con $\delta(x - x_l)$ la función delta de Dirac. Esto equivale a anular el residuo en un conjunto finito de puntos x_l, o sea es similar a la aproximación por ajuste puntual. La matriz $\mathbf{K}$ y el vector derecho se calculan como:

$$K_{lm} = N_m(x_l) \qquad\qquad f_l = [\phi - \psi]_{x=x_l} \tag{5.11}$$

Método de colocación por subdominios

$$W_l = \begin{cases} 1 & x_l < x < x_{l+1} \\ 0 & x < x_l, x > x_{l+1} \end{cases} \tag{5.12}$$

donde en este caso se requiere que el error integrado sobre cada una de estas subregiones sea nulo.

$$K_{lm} = \int_{x_l}^{x_{l+1}} N_m dx \qquad\qquad f_l = \int_{x_l}^{x_{l+1}} (\phi - \psi)dx \tag{5.13}$$

Método de Galerkin

Es uno de los más populares métodos y se basa en elegir

$$W_l = N_l \tag{5.14}$$

con lo cual el sistema a resolver está formado por:

$$K_{lm} = \int_{x_l}^{x_{l+1}} N_l N_m dx \qquad f_l = \int_{x_l}^{x_{l+1}} N_l(\phi - \psi)dx \tag{5.15}$$

Este método goza con la ventaja de que la matriz del sistema es simétrica Si elegimos como funciones de prueba y de peso aquellas que conforman la base de un desarrollo en series de Fourier se puede demostrar que el sistema queda reducido a una simple expresión para los coeficientes a_m ya que la matriz del sistema es diagonal. Esta característica tan particular se debe a que la base elegida es *ortogonal* con lo que $\int_\Omega N_l N_m d\Omega = 0 \;\; l \neq m$.

Otros pesos

En general existen muchas posibles elecciones de la función de peso. Entre las más conocidas aún no presentadas podemos mencionar el *método de los momentos* donde $W_l = x^{l-1}$ donde se requiere que no solo la integral del error sea nula sino algunos de sus momentos. Otro método del estilo de los presentados bajo el método de los residuos ponderados es el *método de los cuadrados mínimos*. Comúnmente definido como la minimización de un funcional formado como:

$$I(a_1, a_2, \ldots, a_M) = \int_\Omega (\phi - \hat{\phi})^2 d\Omega \tag{5.16}$$

la idea es hallar un extremo de dicho funcional mediante $\frac{\partial I}{\partial a_l} = 0$, que al introducirla en (5.16) produce que

$$\int_\Omega (\phi - \hat{\phi}) N_l d\Omega = 0 \tag{5.17}$$

habiendo usado el hecho que $\frac{\partial \hat{\phi}}{\partial z_l} = N_l$ a partir de (5.1) .
(5.17) equivale al método de Galerkin y es un caso particular del mismo.

5.1.3. Residuos ponderados para la resolución de ecuaciones diferenciales

En la sección anterior hemos visto como utilizar el método de los residuos ponderados para aproximar funciones conocidas analíticamente o aquellas en donde conocemos su evaluación en un conjunto discreto de puntos. En esos casos el residuo estaba asociado a la diferencia entre la función a aproximar y la aproximante. En esta sección trataremos el caso de la resolución de ecuaciones diferenciales, en donde el residuo viene dado por la diferencia entre un operador diferencial aplicado a la función a aproximar y el mismo operador aplicado a la aproximante. En el primer capítulo hemos hecho un repaso a los modelos físicos y matemáticos que gobiernan muchos de los problemas de interés en fluidodinámica computacional. Para comenzar con un caso simple tomemos un operador diferencial sencillo como la ecuación de Poisson,

Funciones de prueba que satisfacen las condiciones de contorno

En esta sección trataremos el caso de funciones aproximantes que satisfacen exactamente las condiciones de contorno a través de la elección apropiada de las funciones de prueba. Sea el problema de Poisson:

$$\begin{aligned} A(\phi) &= \mathcal{L}\phi + p = 0 \qquad \text{en } \Omega \\ \mathcal{L}\phi &= \frac{\partial}{\partial x}(\kappa\frac{\partial\phi}{\partial x}) + \frac{\partial}{\partial y}(\kappa\frac{\partial\phi}{\partial y}) \\ p &= Q \end{aligned} \tag{5.18}$$

donde κ puede ser la conductividad térmica de un material y Q el flujo de calor aportado por una fuente y en este caso ϕ sería la temperatura. En el caso lineal κ y Q son independientes de ϕ. En general un problema de valores de contorno como éste para estar bien planteado requiere definir las condiciones de frontera. Estas en general pueden ser escritas como otro operador diferencial, del tipo:

$$B(\phi) = \mathcal{M}\phi + r = 0 \qquad \text{sobre } \Gamma \tag{5.19}$$

En general existen diferentes tipos de condiciones de frontera. Las más conocidas son del tipo:

$$\begin{aligned} \mathcal{M}\phi &= \phi & r &= -\overline{\phi} & \text{sobre } \Gamma_\phi & \quad \text{DIRICHLET} \\ \mathcal{M}\phi &= -\kappa\frac{\partial\phi}{\partial n} & r &= -\overline{q} & \text{sobre } \Gamma_q & \quad \text{NEUMANN} \\ \mathcal{M}\phi &= -\kappa\frac{\partial\phi}{\partial n} + h\phi & r &= -h\overline{\phi} & \text{sobre } \Gamma_{q+\phi} & \quad \text{MIXTAS} \end{aligned} \tag{5.20}$$

Aproximando la solución mediante funciones del tipo (5.1) y eligiendo

$$\begin{aligned} \mathcal{M}\psi &= -r \\ \mathcal{M}N_m &= 0 \qquad\qquad \text{sobre } \Gamma \end{aligned} \tag{5.21}$$

entonces $\hat{\phi}$ automáticamente satisface las condiciones de borde chapV-19 para todos los valores de a_m. Aplicando el operador diferencial a la función aproximante (5.1) y asumiendo que las funciones de prueba y sus derivadas son continuas tenemos:

$$\begin{aligned} \phi \approx \hat{\phi} &= \psi + \sum_{m=1}^{M} a_m N_m \\ \frac{\partial\phi}{\partial x} \approx \frac{\partial\hat{\phi}}{\partial x} &= \frac{\partial\psi}{\partial x} + \sum_{m=1}^{M} a_m \frac{\partial N_m}{\partial x} \\ \frac{\partial^2\phi}{\partial x^2} \approx \frac{\partial^2\hat{\phi}}{\partial x^2} &= \frac{\partial^2\psi}{\partial x^2} + \sum_{m=1}^{M} a_m \frac{\partial^2 N_m}{\partial x^2} \end{aligned} \tag{5.22}$$

Aquí se requiere que hasta la segunda derivada sea continua. Cuando en las próximas secciones analicemos el método de los elementos finitos veremos como estas restricciones serán debilitadas. La forma en la cual

hemos construido la aproximación garantiza el cumplimiento de las condiciones de borde y entonces nos queda que $\hat{\phi}$ debe satisfacer solo la ecuación diferencial en el interior del dominio. Sustituyendo $\hat{\phi}$ en (5.18) :

$$R_\Omega = A(\hat{\phi}) = \mathcal{L}\hat{\phi} + p = \mathcal{L}\psi + \Big(\sum_{m=1}^{M} a_m \mathcal{L} N_m\Big) + p \tag{5.23}$$

obtenemos el residuo de la misma con $\mathcal{L}$ asumido un operador lineal. Una vez definido el residuo aplicamos el método de los residuos ponderados

$$\int_\Omega W_l R_\Omega d\Omega = \int_\Omega W_l \Big\{\mathcal{L}\psi + \Big(\sum_{m=1}^{M} a_m \mathcal{L} N_m\Big) + p\Big\} d\Omega = 0 \tag{5.24}$$

Esta ecuación contiene M incógnitas, entonces aplicando esta misma ecuación para $l = 1, 2, \ldots, M$ se obtiene un sistema de ecuaciones algebraicas que pueden ser escritas en forma compacta como:

$$\begin{aligned} \mathbf{Ka} &= \mathbf{f} \\ K_{lm} &= \int_\Omega W_l \mathcal{L} N_m d\Omega \qquad 1 \le l, m \le M \\ f_l &= -\int_\Omega W_l p d\Omega - \int_\Omega W_l \mathcal{L}\psi d\Omega \qquad 1 \le l \le M \end{aligned} \tag{5.25}$$

El procedimiento requiere calcular los coeficientes de las matriz y del miembro derecho y luego invertir el sistema para calcular los coeficientes con los cuales se obtiene la solución aproximada al operador diferencial de (5.18). Ya que las funciones de prueba elegidas para aproximar la solución tienen soporte global entonces la matriz de coeficientes será llena y no tendrá estructura de banda, típica de los métodos de diferencias finitas y elementos finitos. Además, en lo anterior nada se ha dicho acerca de la elección de las funciones de peso con lo cual uno puede aplicar todo lo anterior a las distintas alternativas mostradas en secciones anteriores.

A modo de ejemplo calcularemos la solución al siguiente problema de valores de contorno unidimensional:

Ejemplo 1D Hallar $\hat{\phi}$ solución aproximada de la siguiente ecuación diferencial

$$\begin{aligned} &\frac{d^2\phi}{dx^2} - \phi = 0 \\ &\phi(x=0) = 0 \\ &\phi(x=1) = 1 \end{aligned} \tag{5.26}$$

De acuerdo a las definiciones generales presentadas antes las condiciones de borde y las funciones de prueba elegidas son:

$$\begin{aligned} \mathcal{M}\phi = \phi \qquad & r = 0 \qquad \text{en } x = 0 \\ \mathcal{M}\phi = \phi \qquad & r = -1 \qquad \text{en } x = 1 \\ \psi &= x \\ N_m &= \sin(m\pi x) \end{aligned} \tag{5.27}$$

La elección de ψ y N_m es arbitraria, existen muchas otras posibles alternativas a estas pero aquí usaremos la base trigonométrica. Aplicando el método de los residuos ponderados y la definición de los coeficientes de la matriz y el vector del miembro derecho tenemos:

$$\begin{aligned} K_{lm} &= \int_0^1 W_l(1+m^2\pi^2)\sin(m\pi x)dx \\ f_l &= -\int_0^1 W_l x dx \end{aligned} \tag{5.28}$$

Si tomamos $M = 2$ dos términos en la expansión y si aplicamos el método de colocación puntual obtenemos la siguiente matriz:

$$\begin{aligned} K_{11} = (1+\pi^2)\sin(\pi/3) \qquad & K_{12} = (1+4\pi^2)\sin(2/3\pi) \\ K_{21} = (1+\pi^2)\sin(2/3\pi) \qquad & K_{22} = (1+4\pi^2)\sin(4/3\pi) \\ & f_1 = -1/3 \qquad f_2 = -2/3 \end{aligned} \tag{5.29}$$

mientras que si aplicamos Galerkin en virtud de la ortogonalidad de las funciones de prueba,

$$\begin{aligned} K_{lm} &= \int_0^1 (1+m^2\pi^2)\sin(m\pi x)\sin(l\pi x)dx = (1+m^2\pi^2)\int_0^1 \sin(m\pi x)\sin(l\pi x)dx = \\ &= (1+m^2\pi^2)\frac{1}{m\pi}\frac{m\pi x - \sin(2m\pi x)/2}{2}\Big|_0^1 = \tfrac{1}{2}(1+m^2\pi^2) \\ f_l &= -\int_0^1 x\sin(l\pi x)dx = -\frac{\sin(l\pi)-(l\pi)\cos(l\pi)}{(l\pi)^2} = \frac{(-1)^l}{(l\pi)} \end{aligned} \tag{5.30}$$

tenemos:

$$\begin{aligned} K_{11} = \tfrac{1}{2}(1+\pi^2) \qquad & K_{12} = 0 \\ K_{21} = 0 \qquad & K_{22} = \tfrac{1}{2}(1+4\pi^2) \\ f_1 = -\frac{1}{\pi} \qquad & f_2 = \frac{1}{2\pi} \end{aligned} \tag{5.31}$$

La solución numérica del sistema de ecuaciones resultante da:

$$\begin{aligned} a_1 = -0.05312 \qquad a_2 = 0.004754 \qquad & \text{COLOCACION PUNTUAL} \\ a_1 = -0.05857 \qquad a_2 = 0.007864 \qquad & \text{GALERKIN} \end{aligned} \tag{5.32}$$

La solución exacta a este problema es del tipo

$$\phi = \frac{1}{e-1/e}\left(e^x - e^{-x}\right) \tag{5.33}$$

La rutina Ej_2_1 contiene la resolución de este ejemplo.

La figura (5.3) muestra a la derecha la solucion exacta, la aproximada por Galerkin y la de colocación a la cual se le ha removido la función $\psi = x$ para poder notar mejor las diferencias. A la izquierda vemos una distribución puntual del error donde se alcanza a notar, para este ejemplo, una mejor aproximación obtenida mediante el método de Galerkin.

Ejemplo 2D La ecuación que gobierna la torsión elástica de barras prismáticas es:

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=-2G\theta \tag{5.34}$$

donde G es el módulo elástico de torsión, θ es el ángulo que se gira la sección y ϕ equivale a una función tensión que de acuerdo a la teoría es nula en todo el contorno. Detalles acerca de la forma que se obtiene esta ecuación pueden verse en libros sobre teoría de la elasticidad, como por ejemplo Timoschenko [Ti]. Esta ecuación tiene la estructura de una ecuación de Poisson y analogías con otros experimentos gobernados por la misma ecuación pueden hacerse. Por ejemplo la anterior también surguría si queremos resolver un problema de conducción del calor con una fuente aplicada en todo el volúmen del material asumiendo que en la dirección z la barra es infinita y la temperatura del contorno está fija a un valor de referencia.

Supongamos que en este ejemplo $G\theta=1$, con lo cual la ecuación a resolver se transforma en:

$$\begin{aligned}\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}&=-2 \qquad x\in[-3,3]\ ,y\in[-2,2]\\ \phi&=0 \qquad x=\pm3\ \ ,y=\pm2\end{aligned} \tag{5.35}$$

Aquí apelamos a la intuición. Siendo el problema simétrico respecto a los ejes x e y deberíamos elegir funciones de prueba que tengan esta propiedad y satisfagan las condiciones de contorno. Por ejemplo, si tomamos $\psi=0$ y usamos 3 términos una elección posible sería:

$$\begin{aligned}N_1&=\cos(\pi x/6)\cos(\pi y/4)\\ N_2&=\cos(3\pi x/6)\cos(\pi y/4)\\ N_3&=\cos(\pi x/6)\cos(3\pi y/4)\end{aligned} \tag{5.36}$$

La rutina Ej_2_2 muestra el aspecto que tienen estas tres funciones donde se alcanza a apreciar la paridad deseada.

De esta forma aplicando la aproximación (5.1), comparando (5.18) con (5.34) y usando la definición de la matriz y el vector derecho del sistema algebraico (5.9) que permite calcular los coeficientes a_m tenemos:

$$\begin{aligned}K_{lm}&=\int_{-3}^{3}\int_{-2}^{2}N_l\Big(\frac{\partial^2N_m}{\partial x^2}+\frac{\partial^2N_m}{\partial y^2}\Big)dydx\\ f_l&=-\int_{-3}^{3}\int_{-2}^{2}2N_l dydx\end{aligned} \tag{5.37}$$

Como vemos, ahora las integrales a calcular son bidimensionales por lo que debe estimarse con más detalle la forma de realizar el cálculo. Nosotros aquí solo deseamos mostrar la metodología a usar y en este caso estas integrales las realizaremos a mano. Cuando se pretende volcar estos conceptos en un programa para fines de cálculo intensivo deben adoptarse métodos más eficientes y generales para tal fin, como por ejemplo la integración numérica, tema que veremos más adelante cuando abordemos el estudio del método de los elementos finitos . Analizando (5.37) vemos que la derivada segunda de funciones tipo cosenos generan funciones cosenos, y considerando la ortogonalidad de estas bases las integrales en (5.37) se simplifican notablemente.

La rutina Ej_2_2 contiene el cálculo de este ejemplo donde se puede apreciar la forma de calcular la matriz (diagonal) y el miembro derecho del sistema. En este caso se ha empleado una resolucion analítica

de las integrales. La rutina mejorada Ej_2_2b muestra como puede emplearse una rutina de integración numérica con el fin de evitar tediosos calculos.

La figura (5.4) muestra en la parte superior la solución obtenida con la mencionada rutina en la cual se incluye el valor de la tensión máxima que es de 3.103, cercano al teórico de 2.96 $5\,\%$ de error.

Para terminar esta sección mencionamos que el método de los cuadrados mínimos aplicado a la resolucion de ecuaciones a derivadas parciales no es equivalente al método de los residuos ponderados de Galerkin. Para ver esto tomemos como antes el funcional definido como:

$$I(a_1, a_2, \ldots, a_M) = \int_\Omega R_\Omega^2 d\Omega = \int_\Omega \Big\{ \mathcal{L}\psi + \Big(\sum_{m=1}^{M} a_m \mathcal{L} N_m \Big) + p \Big\}^2 d\Omega$$

$$\begin{aligned} \frac{\partial I}{\partial a_l} &= 0 \qquad l = 1, 2, \ldots, M \\ \int_\Omega R_\Omega \frac{\partial R_\Omega}{\partial a_l} d\Omega &= 0 \qquad l = 1, 2, \ldots, M \end{aligned} \tag{5.38}$$

$$W_l = \frac{\partial R_\Omega}{\partial a_l} = \mathcal{L} N_l$$

o sea la función de peso que surge del método de los cuadrados mínimoses equivalente al operador diferencial del problema aplicado a las funciones de prueba. En algunas circumstancias este tipo de aproximación es deseable mientras que en algunos casos no.

Funciones de prueba que no satisfacen las condiciones de contorno

Hasta aquí hemos considerado aproximaciones elegidas de forma tal de satisfacer las condiciones de contorno. Esto muchas veces puede ser dificultoso y antes esto es preciso relajar tal requisito. Para poder elegir las funciones de prueba independientemente de las condiciones de contorno postulamos una expansión del tipo:

$$\phi \approx \hat{\phi} = \sum_{m=1}^{M} a_m N_m \tag{5.39}$$

que no satisface las condiciones de contorno. Entonces el residuo en el interior del dominio (R_Ω) es suplementado por otro en el borde (R_Γ):

$$\begin{aligned} R_\Omega &= A(\hat{\phi}) = \mathcal{L}\hat{\phi} + p \\ R_\Gamma &= B(\hat{\phi}) = \mathcal{M}\hat{\phi} + r \end{aligned} \tag{5.40}$$

En este caso el método de los residuos ponderados consiste en escribir:

$$\int_\Omega W_l R_\Omega d\Omega + \int_\Gamma \overline{W_l} R_\Gamma d\Gamma = 0 \tag{5.41}$$

con W_l y $\overline{W_l}$ elegidas en forma arbitraria e independiente. Reemplazando (5.40) en (5.41) llegamos a la definición del siguiente sistema de ecuaciones:

$$\begin{gathered}\mathbf{Ka}=\mathbf{f}\\ K_{lm}=\int_\Omega W_l\mathcal{L}N_m d\Omega+\int_\Gamma \overline{W_l}\mathcal{M}N_m d\Gamma \qquad 1\le l,m\le M\\ f_l=-\int_\Omega W_l p d\Omega-\int_\Gamma \overline{W_l} r d\Gamma \qquad 1\le l\le M\end{gathered} \tag{5.42}$$

Para ilustrar el procedimiento tomaremos el ejemplo 1D anteriormente presentado.

Ejemplo 1D (chapV-26) versión 2 Aquí resolveremos el mismo problema presentado en (5.26) con la diferencia que usaremos como funciones de prueba la base $N_m=x^{m-1}, m=1,2,\ldots$

En este caso simple la integral de borde en (5.41) se reduce a la evaluación del integrando en los dos puntos extremos de este dominio definido por el intervalo $[0,1]$. Entonces

$$\int_0^1 W_l R_\Omega dx+[\overline{W_l}R_\Gamma]_{x=0}+[\overline{W_l}R_\Gamma]_{x=1}=0 \tag{5.43}$$

En este ejemplo usaremos para el peso en el interior del dominio el método de Galerkin $W_l=N_l$ mientras que para el peso en el borde $\overline{W_l}=-N_l|_\Gamma$, entonces:

$$\int_0^1 N_l(\frac{\partial^2\hat\phi}{\partial x^2}-\hat\phi)dx-[N_l\hat\phi]_{x=0}-[N_l(\hat\phi-1)]_{x=1}=0 \tag{5.44}$$

Si usamos una expasión en tres términos se llega a la siguiente matriz:

$$\begin{gathered}\mathbf{K}=\begin{pmatrix}3 & 3/2 & -2/3\\ 3/2 & 4/3 & 1/4\\ 4/3 & 5/4 & 8/15\end{pmatrix}\\ \mathbf{f}=\begin{pmatrix}1\\1\\1\end{pmatrix}\end{gathered} \tag{5.45}$$

La solución a este problema es:

$$a_1=0.068 \qquad a_2=0.632 \qquad a_3=0.226 \tag{5.46}$$

Ejemplo 2D - versión 2 Usando como aproximación la siguiente

$$\hat\phi=(4-y^2)(a_1+a_2x^2+a_3y^2+a_4x^2y^2+a_5x^4)$$

es obvio ver que la misma satisface las condiciones de contorno en $y=\pm2$, mientras que la condición $x=\pm3$ debe incluirse en el cálculo.

Tomando (5.40) y (5.42) y reemplazando la anterior expresión vemos que:

$$\int_{-3}^3\int_{-2}^2 W_l\Big(\frac{\partial^2\hat\phi}{\partial x^2}+\frac{\partial^2\hat\phi}{\partial y^2}+2\Big)dydx+\int_{-2}^2\overline{W_l}\hat\phi|_{x=3}dy-\int_{-2}^2\overline{W_l}\hat\phi|_{x=-3}dy=0 \tag{5.47}$$

Usando $W_l = N_l$ y $\overline{W_l} = N_l|_\Gamma$ se llega a armar el sistema de ecuaciones. La rutina Ej_2_3 muestra una forma de resolver este problema apelando a las capacidades de cálculo simbólico de MatLab. Se recomienda editar y leer esta rutina para entender la metodología que es extensible a otros ejemplos utilizando diferentes funciones de base.

La figura (5.4) en la parte inferior muestra la solución numérica obtenida donde se alcanza a ver que la condición de contorno en $y = \pm 2$ se satisface exactamente pero aquella en $x = \pm 3$ se cumple solo aproximadamente. Esta es una de las diferencias entre las dos metodologías propuestas. Además la tensión máxima es un poco superior a la obtenida con la primera metodología alejándose un poco más del valor teórico. La figura chapV-4 en la parte inferior derecha muestra la variación de la tensión a lo largo del contorno $x = \pm 3$ en función de la coordenada y.

5.1.4. Condiciones de contorno naturales

Como hemos visto en la sección anterior es posible facilitar la selección de las funciones de prueba sin necesidad de que satisfagan las condiciones de contorno a expensas de un trabajo algebraico mayor y de una degradación en la calidad de la solución. El primer item está relacionado con la resolución de las integrales que permiten el cálculo de los coeficientes. En (5.42) vemos que además de las integrales en el interior tenemos la necesidad de evaluar integrales en el contorno del dominio, las cuales pueden contener operadores diferenciales que complican aún más la situación. Aquí veremos que existen ciertas condiciones de contorno las cuales surgen como las naturales al problema, en el sentido que permiten cierta cancelación de términos cuando el problema se expresa en su forma débil. Supongamos que aplicamos el método de los residuos ponderados al problema definido en (5.18) .

$$\int_\Omega W_l R_\Omega d\Omega = \int_\Omega W_l \Big(\mathcal{L}\hat{\phi} + p \Big) d\Omega \tag{5.48}$$

La *forma débil* de la anterior formulación integral se obtiene mediante la integración por partes que en su versión general puede escribirse como:

$$\int_\Omega W_l \mathcal{L}\hat{\phi} d\Omega = \int_\Omega \Big(\mathcal{C}W_l \Big) \Big(\mathcal{D}\hat{\phi} \Big) d\Omega + \int_\Gamma W_l \mathcal{E}\hat{\phi} d\Gamma \tag{5.49}$$

donde $\mathcal{C}, \mathcal{D}, \mathcal{E}$ son operadores diferenciales lineales involucrando órdenes de derivación inferiores a la del operadores $\mathcal{L}$. Usando (5.40) y (5.49) en (5.41) se llega a:

$$\int_\Omega \Big(\mathcal{C}W_l \Big) \Big(\mathcal{D}\hat{\phi} \Big) d\Omega + \int_\Gamma W_l \mathcal{E}\hat{\phi} d\Gamma + \int_\Gamma \overline{W_l} \mathcal{M}\hat{\phi} d\Gamma = -\Big(\int_\Omega W_l p d\Omega + \int_\Gamma \overline{W_l} r d\Gamma \Big) \tag{5.50}$$

donde lo que se pretende es anular las contribuciones al contorno del miembro izquierdo , o sea:

$$\int_\Gamma W_l \mathcal{E}\hat{\phi} + \overline{W_l} \mathcal{M}\hat{\phi} d\Gamma = 0 \tag{5.51}$$

Un ejemplo de condiciones de contorno natural que comúnmente se tiene en las aplicaciones es la especificación de un flujo impuesto en el contorno. Pensando en el problema térmico podemos imaginar que extraemos calor del contorno de una pieza con una magnitud $\overline{q}$ especificada. En esos casos la condición de contorno viene expresada como:

$$-\kappa\frac{\partial\phi}{\partial n}=\overline{q} \tag{5.52}$$

Si hechamos un vistazo a (5.40) vemos que $\mathcal{M}=\frac{\partial}{\partial n}$ y si lo que estamos resolviendo es la ecuación de conducción térmica en ese caso el operador $\mathcal{E}=\frac{\partial}{\partial n}=\mathcal{M}$, por lo cual $\overline{W_l}=-W_l$ satisface (5.51) simplificando (5.50) a:

$$\int_\Omega\Big(\mathcal{C}W_l\Big)\Big(\mathcal{D}\hat{\phi}\Big)d\Omega=-\Big(\int_\Omega W_l p d\Omega-\int_\Gamma W_l r d\Gamma\Big) \tag{5.53}$$

En lo anterior hemos utilizado el teorema de Green o su equivalente, la fórmula de integración por partes. Una versión un poco más detallada del mismo aplicada a un caso sencillo expresa que:

$$\int_\Omega\alpha\nabla\beta d\Omega=-\int_\Omega\nabla\alpha\beta d\Omega+\int_\Gamma\alpha\beta\mathbf{n}d\Gamma \tag{5.54}$$

con α y β funciones escalares. A pesar que (5.54) incluye solo derivadas de primer órden, la técnica es bien general como lo expresa chapV-46 e incluso se extiende al caso de funciones vectoriales. No obstante en la mayoría de las aplicaciones los operadores que se trabajan son de relativo bajo órden.

5.1.5. Métodos de solución del contorno

Hasta aquí los métodos empleados resolvían el problema en el interior del dominio pudiendo trabajar con funciones de prueba que satisfagan o no las condiciones de contorno . Si en lugar de elegir funciones de prueba que satisfagan las condiciones de contorno elegimos aquellas que satisfacen el operador diferencial en el interior del dominio el problema se reduce a resolver solo el residuo en el contorno. Este tipo de estrategia dió lugar al *método de los paneles* y al *método de los elementos de contorno*. De esta forma como las funciones de prueba satisfacen el operador diferencial en el interior entonces:

$$R_\Omega=A(\hat{\phi})=\sum_{m=1}^{M}a_m A(N_m)=0 \tag{5.55}$$

con lo cual la definición del método de los residuos ponderados (chapV-38) se reduce simplemente a:

$$\int_\Gamma\overline{W_l}R_\Gamma d\Gamma=0 \tag{5.56}$$

Un solo conjunto de funciones de prueba $\overline{W_l}$, definidas solamente sobre el borde del dominio deben definirse. Además como el problema se plantea sobre el contorno la dimensión espacial se reduce en una unidad con lo cual problemas en 3D se transforman en bidimensionales y aquellos en 2D en unidimensionales. Estas ventajas tiene su contracara en la dificultad de elegir funciones de prueba que satisfagan el operador diferencial en el interior del dominio. Este es el principal limitante de esta técnica que se muestra atractiva por todo lo que implica reducir la dimensión espacial. Una de las aplicaciones más divulgadas de esta técnica es la resolución de problemas gobernados por la ecuación de Laplace, por ejemplo: conducción del calor, flujo potencial, elasticidad lineal, y otros. La razón es que si pensamos en funciones analíticas de variable

compleja $z = x + iy$, estas satisfacen automáticamente la ecuación de Laplace. Supongamos una función analítica del tipo:

$$f(z) = u + iv \qquad u, v \in \mathbb{R} \tag{5.57}$$

luego,

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= f'' \\ \frac{\partial^2 f}{\partial y^2} &= i^2 f'' = -f'' \end{aligned}$$

sumando m.a.m. (5.58)

$$\begin{aligned} \nabla^2 f &= \nabla^2 u + i\nabla^2 v = 0 \\ &\Rightarrow \nabla^2 u = \nabla^2 v = 0 \end{aligned}$$

con $f' = \dfrac{\mathrm{d}f}{\mathrm{d}z}$

Por ejemplo tomando la función analítica

$$f(z) = z^n \tag{5.59}$$

Todas las funciones u y v que surgen de (5.59) satisfacen la ecuación de Laplace siendo todas ellas candidatas para integrar las funciones de prueba con las cuales armar una función aproximante que satisfaga este problema en particular en el interior del contorno. El siguiente ejemplo muestra una aplicacion del método de los elementos de contorno al problema de la torsión de una viga.

5.1.6. Sistema de ecuaciones diferenciales

El método de los residuos ponderados en cualquiera de sus versiones puede ser extendido para tratar el caso de sistemas de ecuaciones diferenciales . Un sistema de ecuaciones diferenciales surge cuando en el modelo matemático se pretende resolver campos vectoriales en una o varias dimensiones espaciales o cuando se pretenden acoplar varios campos escalares y/o vectoriales tanto en una como en varias dimensiones espaciales. En esos casos la solución a obtener viene representada por un vector

$$\boldsymbol{\phi} = \{\phi_1, \phi_2 \dots\} \tag{5.60}$$

que debe satisfacer ciertas ecuaciones diferenciales, una por cada componente del vector incógnita

$$\begin{aligned} A_1(\boldsymbol{\phi}) &= 0 \\ A_2(\boldsymbol{\phi}) &= 0 \end{aligned} \tag{5.61}$$

la cual en forma compacta puede escribirse como

$$\mathbf{A}(\boldsymbol{\phi}) = \begin{pmatrix} A_1(\boldsymbol{\phi}) \\ A_2(\boldsymbol{\phi}) \\ \vdots \end{pmatrix} = \mathbf{0} \quad \text{en } \Omega \tag{5.62}$$

Del mismo modo con las ecuaciones en el contorno

$$\mathbf{B}(\boldsymbol{\phi}) = \begin{pmatrix} B_1(\boldsymbol{\phi}) \\ B_2(\boldsymbol{\phi}) \\ \vdots \end{pmatrix} = \mathbf{0} \quad \text{en } \Gamma \tag{5.63}$$

Para cada componente del vector necesitamos definir una aproximación del tipo (5.1) , que escrita en forma compacta se expresa como:

$$\boldsymbol{\phi} \approx \hat{\boldsymbol{\phi}} = \boldsymbol{\psi} + \sum_{m=1}^{M} \mathbf{N}_m \, \mathbf{a}_m \tag{5.64}$$

donde $\mathbf{N}_m$ es una matriz diagonal donde en cada término de la diagonal se halla la función de prueba de cada componente del vector incógnita. Del mismo modo las funciones de peso, tanto para la integral sobre el interior del dominio como para la del contorno son también matrices diagonales. De esta forma la extensión del método de los residuos ponderados escalar aplicado al caso vectorial es directa.

$$\int_{\Omega} \mathbf{W}_l \mathbf{A}(\hat{\boldsymbol{\phi}}) d\Omega + \int_{\Gamma} \overline{\mathbf{W}_l} \mathbf{B}(\hat{\boldsymbol{\phi}}) d\Gamma = \mathbf{0} \tag{5.65}$$

La mayoría de los casos de interés tanto en la industria como en la ciencia involucran sistemas de ecuaciones diferenciales .

$$\begin{aligned} \boldsymbol{\phi} &= \{u, v\} \text{ Elasticidad lineal} \\ \boldsymbol{\phi} &= \{u, v, w, p\} \text{ Flujo incompresible viscoso (a)} \\ \boldsymbol{\phi} &= \{\psi, \omega\} \text{ Flujo incompresible viscoso (b)} \\ \boldsymbol{\phi} &= \{\rho, u, v, w, p\} \text{ Flujo compresible viscoso} \end{aligned} \tag{5.66}$$

Los problemas de elasticidad bidimensional están generalmente formulados en dos variables, los desplazamientos u, v según las componentes x e y respectivamente. Los problemas de flujo viscoso incompresible tridimensional vienen muchas veces expresados en término de las variables primitivas del problema, las tres componentes de la velocidad y la presión. No obstante en 2D muchos prefieren la formulación vorticidad ω , función de corriente ψ que desde el punto de vista computacional tiene una implementación más fácil y no requiere un tratamiento especial de la incompresibilidad como lo necesita la formulación en variables primitivas. El caso de flujo compresible 3D tiene un vector incógnita con 5 componentes. Esta es solo una breve descripción de algunos de los casos típicos donde se necesita resolver un sistemas de ecuaciones diferenciales . El agregado de modelos adicionales, por ej. turbulencia u otros, muchas veces también agrega componentes al vector incógnita. Por último en pos de reducir el órden de una ecuación diferencial se puede transformar la misma en un sistemas de ecuaciones diferenciales donde el orden se ha reducido completamente equivalente al original.

5.1.7. Problemas no lineales

Muchos problemas prácticos modelados físicamente producen tanto ecuaciones diferenciales como condiciones de contorno que son no lineales. Esta no linealidad se expresa porque existe una dependencia de los operadores, tanto el del interior como el del contorno, con la variable de estado o función incógnita. Dado que en todas las secciones anteriores hemos considerado la aplicación del método de los residuos ponderados al caso lineal se hace necesario hacer algunos comentarios respecto al caso no lineal. El método de los residuos ponderados es completamente aplicable al caso no lineal. Supongamos que queremos resolver un problema de conducción del calor donde la conductividad depende de la misma temperatura. La ecuación de gobierno puede escribirse como:

$$\begin{aligned} \frac{\partial}{\partial x}(\kappa(\phi)\frac{\partial \phi}{\partial x}) + \frac{\partial}{\partial y}(\kappa(\phi)\frac{\partial \phi}{\partial y}) + Q = 0 \qquad & \text{en } \Omega \\ \phi = \overline{\phi} \qquad & \text{en } \Gamma_\phi \\ \kappa(\phi)\frac{\partial \phi}{\partial n} = -\overline{q} \qquad & \text{en } \Gamma_q \end{aligned} \tag{5.67}$$

Planteando una aproximación del tipo (5.1) , introduciendo la misma en la formulación por residuos ponderados de (5.67) produce un sistema de ecuaciones algebraico no lineales del tipo:

$$\mathbf{K}(\mathbf{a})\mathbf{a} = \mathbf{f} \tag{5.68}$$

que puede resolverse iterativamente en la forma:

$$\mathbf{K}(\mathbf{a}^{n-1})\mathbf{a}^n = \mathbf{f}^{n-1} \tag{5.69}$$

Además del caso de la ecuación de conducción no lineal existen muchos otros problemas de interés a mencionar como el caso de la ec. de Bürgers, el caso de la ecuación de flujo viscoso incompresible expresada en una formulación $\psi - \omega$ donde la nolinealidad se halla en las condiciones de contorno. Obviamente las ecuaciones de flujo compresible e incompresible expresada en las variables primitivas o conservativas son también ejemplos claros de sistemas de ecuaciones diferenciales no lineales.

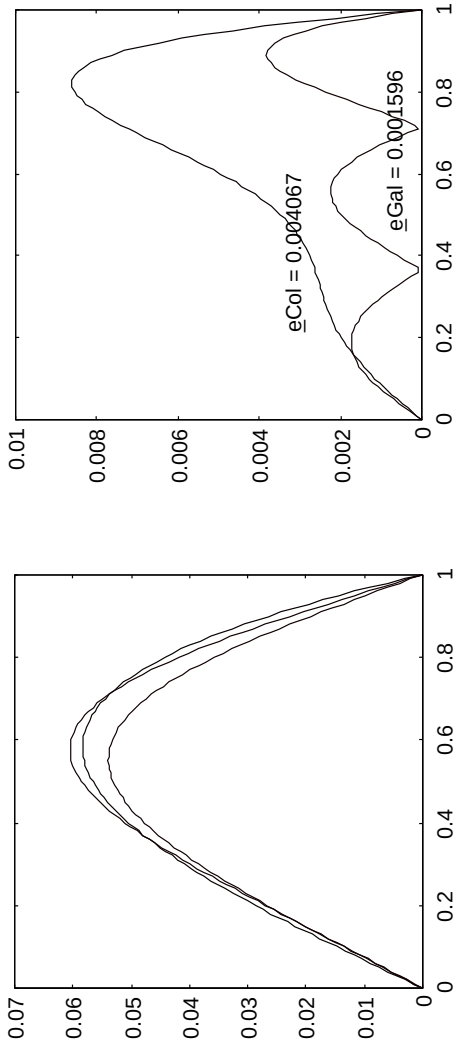
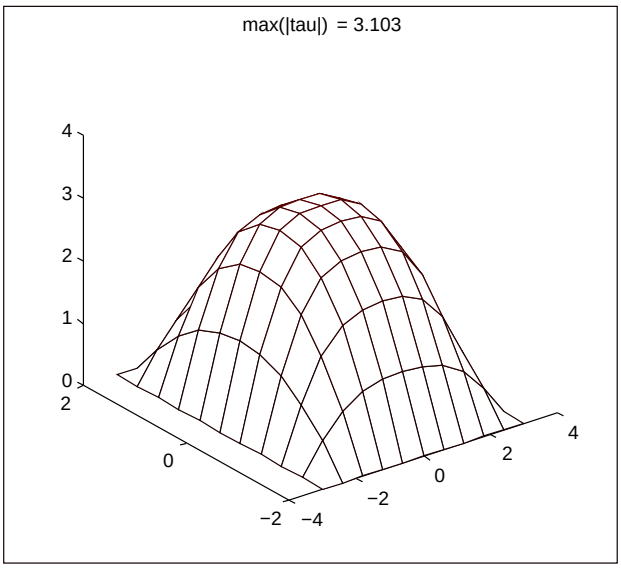
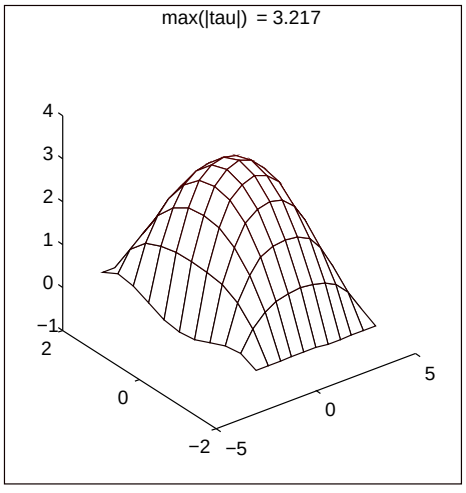


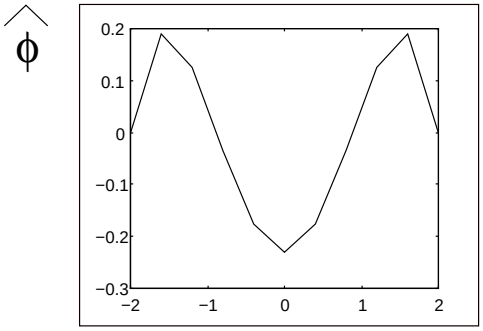
Figura 5.3: Solución aproximada a una ODE mediante residuos ponderados



Funciones de prueba que satisfacen
las condiciones de contorno



Funciones de prueba que no satisfacen
las condiciones de contorno



Solucion en x=3

Figura 5.4: Torsión de una barra prismática

5.1.8. Conclusiones

En esta primera parte de este capítulo que trata acerca de diferentes técnicas numéricas de discretización hemos presentado el caso del método de los residuos ponderados aplicado a la resolución de ecuaciones a derivadas parciales utilizando para aproximar un conjunto de funciones de prueba definidas globalmente, satisfaciendo o no las condiciones de contorno, de fácil extensión al caso de sistemas de ecuaciones diferenciales y al caso no lineal. No obstante, como se desprende de los ejemplos incluidos la elección de dichas funciones no es tarea fácil, incluso no es extensible al caso de geometrías arbitrarias si uno requiere que dichas funciones satisfagan las condiciones de contorno exactamente. Además, a medida que se aumenta el grado de la aproximación el condicionamiento del sistema lineal a resolver se vuelve crítico salvo que se usen funciones base con mayor grado de ortogonalidad. Esto se logra mediante el uso de polinomios de Legendre o de Chebyshev. En realidad estos son muy frecuentemente utilizados en el contexto de los *métodos espectrales* el cual en forma indirecta ha sido el tema de esta sección (5.1) . Este tema daría para una sección aparte pero por el momento diferimos un tratamiento más detallado para futuras versiones de estas notas. En las próximas secciones trataremos de relajar algunas de las limitaciones de esta técnica, en especial aquella relacionada con el tratamiento de dominios de forma arbitraria, presentando primero el método de los elementos finitos y posteriormente el método de los volúmenes finitos .

5.1.9. TP.chapV– Trabajo Práctico #2

1. Tome la rutina Ej_2_0.m y transfórmela para ser usada con funciones de prueba del tipo $N_m = \sin(m\pi x)$. Comience con $M = 2$ y refínelo para testear la convergencia de la aproximación.

2. Demuestre que la aproximación por residuos ponderados del tipo Galerkin, tomando como funciones de prueba la base $N_m = \sin(m\pi x/L_x)$ conduce a un sistema de ecuaciones con una matriz diagonal.

3. Un ensayo experimental sobre la deflección $u(x, y)$ de una placa cuadrada de lado unitario con todo su contorno empotrado dio como resultado los valores que se muestran en la figura.

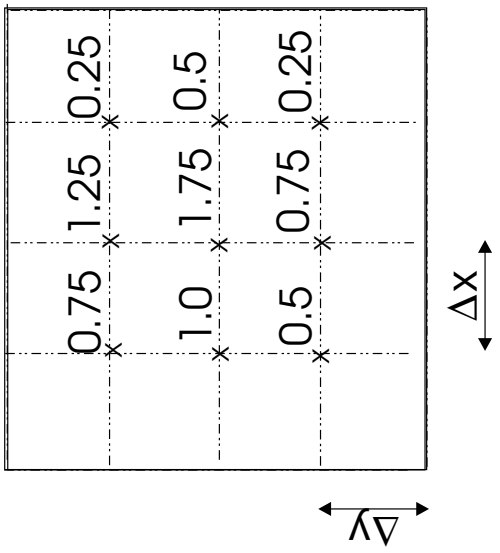


Figura 5.5: Ej. 3 Torsión de una barra

Aproximar la deflección mediante

$$\hat{u}(x, y) = \psi(x, y) + \sum_{\substack{l,m=1 \\ l+m\leq 4}}^{M} a_{lm} \sin(l\pi x) \sin(m\pi y) \tag{5.70}$$

y usando el método de los residuos ponderados estimar los coeficientes a_{lm}. *Ayuda: las integrales que aparecen del cálculo de los coeficientes pueden resolverse mediante integración numérica usando regla del trapecio bidimensional*

4. Mediante el uso de un adecuado conjunto de funciones de prueba aproxime la función $\phi = 1 + \sin(\pi x/2)$ en el rango $0 \leq x \leq 1$. Utilice colocación puntual, colocación por subdominios y el método de Galerkin e investigue numéricamente la convergencia de las sucesivas aproximaciones.

5. La rutina Ej_2_1 contiene la resolución del problema de valores de contorno presentado en las notas

teóricas:

$$\begin{aligned} &\frac{d^2\phi}{dx^2} - \phi = 0 \\ &\phi(x=0) = 0 \\ &\phi(x=1) = 1 \end{aligned} \tag{5.71}$$

El mismo se ha usado con $M = 2$ términos en la expansión. Pruebe de modificar la cantidad de términos y trace una curva donde se muestre la convergencia de cada uno de los métodos.

6. Resuelva el ejercicio anterior pero utilizando como conjunto de funciones de prueba la base $N_m = x^m(1-x)$. Construya una rutina en base a la Ej_2_1 para resolver este problema y muestre la convergencia de la misma. Saque conclusiones respecto a los obtenido en este ejercico y el anterior.

7. Utilizando como base la rutina Ej_2_2 realice las modificaciones necesarias para realizar un estudio de convergencia de la aproximación. Tenga en cuenta de mantener la simetría en la elección de las funciones de prueba y utilice las propiedades de ortogonalidad para calcular la matriz de coeficientes.

8. Utilice el método de los residuos ponderados aplicado al residuo en el dominio interior y agregado el proveniente del contorno para resolver el problema de la torsión de la barra definido en la teoría. Utilice una expansión del tipo

$$\hat{\phi} = (4 - y^2)(a_1 + a_2x^2 + a_3y^2 + a_4x^2y^2 + a_5x^4)$$

que satisface la condición de contorno en $y = \pm 2$ pero no satisface aquella en $x = \pm 3$. Elija $W_l = N_l$ y $\overline{W_l} = N_l|_\Gamma$ y obtenga el sistema de ecuaciones a resolver y la solución. Estudie la convergencia dl residuo al tomar diferente cantidad de términos en la expansión arriba presentada. *Sugerencia: Realice un programa del tipo Ej_2_2 para el mismo y para resolver las integrales use integración numérica*

9. **Condiciones de contorno naturales** Resolver el problema de conducción térmica estacionaria

$$\begin{aligned} \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} &= 0 \\ \phi &= 0 \qquad\qquad y = \pm 1(\Gamma_\phi) \\ \frac{\partial\phi}{\partial n} &= \cos(\pi y/2) \qquad x = \pm 1(\Gamma_q) \\ \kappa &= 1 \end{aligned} \tag{5.72}$$

usando como función aproximante

$$\hat{\phi} = (1 - y^2)(a_1 + a_2x^2 + a_3y^2 + a_4x^2y^2 + a_5x^4)$$

Utilice la rutina Ej_2_3 para este fin.

10. **Método de los elementos de contorno - Torsión de una viga.**

El problema de la torsión de una viga definido en la teoría

$$\begin{aligned} \frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}&=-2 \qquad \text{en } \Omega=[-3,3]\times[-2,2] \\ \phi&=0 \qquad\quad \text{en } \Gamma \end{aligned} \tag{5.73}$$

puede ser transformado a una ecuación de Laplace mediante la siguiente igualdad:

$$\phi=\theta-½(x^2+y^2)$$

Usando el método de los residuos ponderados aplicado solo al contorno y la siguiente base de funciones

$$\hat{\theta}=a_1+a_2(x^2-y^2)+a_3(x^4-6x^2y^2+y^4)$$

hallar la solución aproximada $\hat{\phi}$ y compararla con la obtenida por el método de los residuos ponderados aplicado al interior.

11. **Sistema de ecuaciones diferenciales**

El problema de conducción térmica

$$\begin{aligned} \frac{\partial}{\partial x}(\kappa\frac{\partial\phi}{\partial x})+\frac{\partial}{\partial y}(\frac{\partial\phi}{\partial y})+Q&=0 \qquad \text{en } \Omega \\ \phi(x=0)=0 \qquad q(x&=1)=0 \end{aligned} \tag{5.74}$$

puede descomponerse en un sistema de dos ecuaciones diferenciales de primer orden del tipo:

$$\begin{aligned} q+\kappa\frac{\mathrm{d}\phi}{\mathrm{d}x}&=0 \\ \frac{\mathrm{d}q}{\mathrm{d}x}-Q&=0 \end{aligned} \tag{5.75}$$

siendo el vector incógnita $\boldsymbol{\phi}=\begin{pmatrix} q \\ \phi \end{pmatrix}$. Usando la aproximación:

$$\begin{aligned} N_{m,1}&=x^{m-1}(1-x) \\ N_{m,2}&=x^m \end{aligned} \tag{5.76}$$

calcular la solución a este problema usando dos términos.

12. **Ejemplo : Problemas no lineales**

Resolver la ecuación no lineal

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(\kappa\frac{\mathrm{d}\phi}{\mathrm{d}x})&=-10x \\ \phi(x=0)&=0 \\ \phi(x=1)&=0 \\ \kappa&=1+0.1\phi \end{aligned} \tag{5.77}$$

usando como funciones de prueba la base $N_m=x^m(1-x)$ con dos términos mediante un método de los residuos ponderados por colocación puntual.

Capítulo 6

Método de los elementos finitos

6.1. Introducción

El método de los residuos ponderados presentado en el capítulo anterior sirvió como una teoría unificadora de una gran cantidad de métodos numéricos a la vez que en sí mismo puede concebirse como una técnica numérica en particular. Esa técnica comúnmente denominada *método espectral* goza de ciertas ventajas siendo su principal desventaja la de estar muy restringida a dominios con geometrías muy simples como zonas rectangulares, paralelepípidos u otras mapeables a las anteriores. Este no es el caso que más interesa a los ingenieros de las últimas décadas los cuales requieren herramientas computacionales de cálculo que permitan tratar dominios arbitrarios. Si bien en el procedimiento antes empleado uno planteaba la aproximación de forma tal de satisfacer las condiciones de contorno, hemos visto que existe la posibilidad de elegir funciones más generales que no satisfacen las condiciones de contorno y para las cuales un método de los residuos ponderados que incluya el residuo en el contorno debe plantearse. Esto provocaba la aparición de integrales adicionales sobre el contorno de difícil tratamiento analítico. En los métodos empleados en el capítulo anterior trabajamos en el espacio transformado en lugar que en el espacio físico y esto puede verse de inmediato si pensamos que el vector de incógnitas $\mathbf{a} = \{a_1, a_2, \ldots, a_M\}$ representan las amplitudes de diferentes componentes ondulatorias de la solución y no el valor de la incógnita del problema en cada posición de la malla. Una idea explorada en los últimos tiempos es la de trabajar con métodos espectrales pero en el dominio físico del problema. Entonces para poder aplicar todo lo anterior es necesario hacer una transformación al dominio de la frecuencia para luego volver al dominio físico con la solución del problema. Trabajar en el dominio físico del problema permite descomponer espacialmente el problema asi como el método espectral lo descompone en el espacio de las frecuencias. Esta descomposición del dominio en pedazos (elementos, volúmenes, celdas, etc) de tamaño finito le confiere el nombre al método. Aquí esta la *gran idea* en torno a estas técnicas las cuales no requieren demasiado trabajo para especificar las funciones de prueba ya que al ser de soporte compacto aproximan bien la mayoría de las funciones aún siendo de bajo orden. A su vez las integrales a calcular para obtener los coeficientes de la matriz y del vector miembro derecho son integrales sobre elementos que tienen una forma completamente mapeable a un cuadrado o cualquier otra figura geométrica simple (triángulos) lo cual permite su cálculo en forma muy sencilla, tanto analíticamente como mediante cuadratura numérica. En cuanto a las condiciones de contorno estas presentan un tratamiento mucho más simple debido a que las variables sobre las cuales se especifican coinciden con las variables a calcular. Diferente era el caso de los métodos espectrales en los cuales las condiciones de contorno se

especifican sobre las variables del problema siendo la incógnita las amplitudes de su descomposición espectral. Esto motivaba tener que elegir adecuadamente a priori las funciones de prueba antes de realizar el cálculo. No obstante esto, la selección de una técnica numérica no es una cosa obvia. Los métodos espectrales sirven y son insuperables para algunas aplicaciones donde el fenómeno físico requiere alto orden de precisión y la geometría no presenta dificultades, mientras que los métodos localizados son más aplicables a los casos donde no interesa entrar en demasiado detalle de la solución pero el dominio del problema es muy complicado.

Resumiendo, vimos que el método de los residuos ponderados consiste en aproximar la solución mediante

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M} a_m N_m \tag{6.1}$$

definida a lo largo de todo el dominio Ω y las integrales

$$\int_\Omega W_l R_\Omega d\Omega + \int_\Gamma \overline{W_l} R_\Gamma d\Gamma = 0 \tag{6.2}$$

se evalúan mediante una sola operación y sobre todo el dominio. La ida alternativa consiste en dividir la región Ω en un número de subdominios o elementos Ω^e sin solapamiento y que cubran todo el dominio, de forma que la aproximación $\hat{\phi}$ se construya *a pedazos* o *a trozos* sobre cada subdominio.

Expresando matemáticamente lo anterior,

$$\begin{aligned} \cap_e \Omega^e &= \emptyset \\ \cup_e \Omega^e &= \Omega \end{aligned} \tag{6.3}$$

De esta forma las integrales IV.38 se pueden calcular agregando la contribución a la integral proveniente de cada elemento,

$$\begin{aligned} \int_\Omega W_l R_\Omega d\Omega &= \sum_{e=1}^{E} \int_{\Omega^e} W_l R_\Omega d\Omega \\ \int_\Gamma \overline{W_l} R_\Gamma d\Gamma &= \sum_{e=1}^{E} \int_{\Gamma^e} \overline{W_l} R_\Gamma d\Gamma \end{aligned} \tag{6.4}$$

donde Γ^e es el borde del elemento que cae sobre algún borde del dominio Γ y es tal que $\cup_e \Gamma^e = \Gamma$. E denota la cantidad total de elementos en la partición. Si se usan elementos de forma simple y si la definición de las funciones de forma permite un cálculo de manera repetitiva entonces es posible tratar dominios con formas completamente arbitrarias con facilidad. Es de notar que el método de los residuos ponderados al estilo del presentado en el capítulo anterior equivale a hacer lo mismo pero sobre un solo elemento que coincide con el dominio Ω. La definición a trozos de la aproximación introduce ciertas discontinuidades en la solución o en alguna de sus derivadas. Cierto grado de discontinuidad es permisible y esto limita fuertemente la formulación a emplear. Por otro lado el hecho de que las funciones de prueba se elijan a trozos provoca un beneficio computacional importante respecto a la estructura de la matriz resultante. Un función a trozos en general tiene un soporte compacto, o sea esta no es nula solo en una región pequeña del dominio abarcando algunos pocos elementos del mismo. Esto produce matrices con estructura de banda lo cual tratada convenientemente puede reducir muchíismo el costo computacional de obtener una solución. Esto lo veremos en más detalle

más adelante al tratar el problema de la resolución numérica del sistema de ecuaciones. Nosotros ya hemos pasado por una situación similar al resolver el sistema de ecuaciones para calcular los coeficientes $\mathbf{a}$ en el capítulo anterior. Allí no hemos tenido mayor dificultad tanto porque los sistemas eran de un tamaño chico y además porque la matriz era completamente llena en cuyo caso recurrímos directamente a una especie de eliminación gaussiana. Ya veremos que esto no es admisible en especial en problema con gran número de incógnitas (sistemas de ecuaciones en 3D). Esta es otra de las grandes diferencias entre los métodos locales y aquellos globales, la resolución del sistema de ecuaciones.

6.2. Funciones de forma locales de soporte compacto

Para ilustrar lo anterior consideremos la aproximación de una función $\phi(x)$ en el espacio unidimensional definido por $\Omega = [0, L_x]$.

La división de Ω en $E(= M_n - 1)$ subregiones no solapadas se determina estableciendo un adecuado conjunto de puntos $\{x_l; l = 1, 2, \ldots, M_n\}$ en Ω con $x_l = 0$ y $x_{M_n} = L_x$ definiendo el elemento Ω^e como el intervalo $x_e \leq x \leq x_{e+1}$. Como vemos la función a trozos usada para aproximar la función ϕ es constante a trozos, constante dentro de cada elemento coincidiendo su valor con el de la función ϕ en el centro de cada elemento, lo cual equivale a hacer colocación en esos puntos, comúnmente llamados los nodos. En este ejemplo tan sencillo la numeración es obvia. Siendo la función aproximante constante a trozos la función de prueba que da orígen a la misma es una función que vale:

$$N^e = \begin{cases} 1 & x_e \leq x \leq x_{e+1} \\ 0 & x_e > x, x > x_{e+1} \end{cases} \tag{6.5}$$

A nivel global la función aproximante que resulta es:

$$N_e = \begin{cases} 1 & x_{e-1} \leq x \leq x_{e+1} \\ 0 & x_{e-1} > x, x > x_{e+1} \end{cases} \tag{6.6}$$

De esta forma la aproximación se escribe como:

$$\phi \approx \hat{\phi} = \sum_{m=1}^{M_n - 1} \phi_m N_m \qquad \in \Omega \tag{6.7}$$

donde ϕ_m es el valor de la función en cada nodo m de la malla (el centro del elemento) y reemplazaa a_m la amplitud de cada componente espectral en el método de los residuos ponderados presentado en el capítulo anterior. Esto ya fue comentado previamente y tiene la ventaja que en este caso los coeficientes a calcular tienen un significado físico más directo con el problema. La función ψ ha sido omitida por lo que la aproximante no coincidirá con el valor especificado en el borde aunque por un proceso de refinamiento se puede llegar tan cerca como se quiera a satisfacer los mismos. Sobre cualquier elemento e la aproximación global puede expresarse en términos del valor ϕ_e y de la función de forma del elemento N^e como:

$$\phi = \hat{\phi} = \phi_e N^e = \phi_e \qquad \text{sobre el elemento } e \tag{6.8}$$

Otro tipo de aproximación de frecuente uso y que mejora la anterior es la que surge de emplear funciones de forma lineales. Estas funciones de forma desde el punto de vista global asumen

$$N_i = \begin{cases} 1 & \text{en } x = x_i \\ 0 & \text{en } x = x_{i-1} \\ 0 & \text{en } x = x_{i+1} \\ 0 & \text{en el resto de } \Omega \end{cases} \tag{6.9}$$

con una variación lineal entre los valores mencionados. Esto puede construirse a nivel elemental si por cada elemento definimos dos funciones de forma, una por cada nodo del elemento. Ahora el concepto de nodo ya no coincide como antes con el centro del mismo sino que ahora se ubican en los extremos del intervalo que define al elemento. Estos son ahora los puntos de colocación. La figura 6.2 muestra graficamente lo que recien se acaba de mencionar.

La aproximación global para este caso es:

$$\phi \approx \hat{\phi} = \sum_{m=1}^{M_n} \phi_m N_m \qquad \in \Omega \tag{6.10}$$

donde a diferencia del caso anterior la suma se extiende a M_n puntos, los nodos de esta malla. Es de remarcar que esta aproximación satisfará automáticamente los valores en el contorno $x = 0, x = L_x$ sin necesidad de agregar una función ψ ya que ahora los puntos de colocación están justamente sobre los extremos de los elementos y para el caso del primer y último elemento estos coinciden con los extremos del dominio donde se agregan las condiciones de contorno . Sobre cualquier elemento e con nodos i, j la aproximación toma el valor:

$$\phi = \hat{\phi} = \phi_i N_i^e + \phi_j N_j^e = \qquad \text{sobre el elemento } e \tag{6.11}$$

siendo la variación en el interior del elemento lineal por ser lineales las funciones de prueba N_i, N_j.

Los dos conjutnos de funciones de prueba usados, aquellos constantes a trozos y los lineales a trozos forman un conjunto completo en el sentido que refinando la partici'on se obtienen soluciones con cada vez mayor precisión. A su vez estas funciones son generalizables a varias dimensiones como puede verse en la figura 6.3.

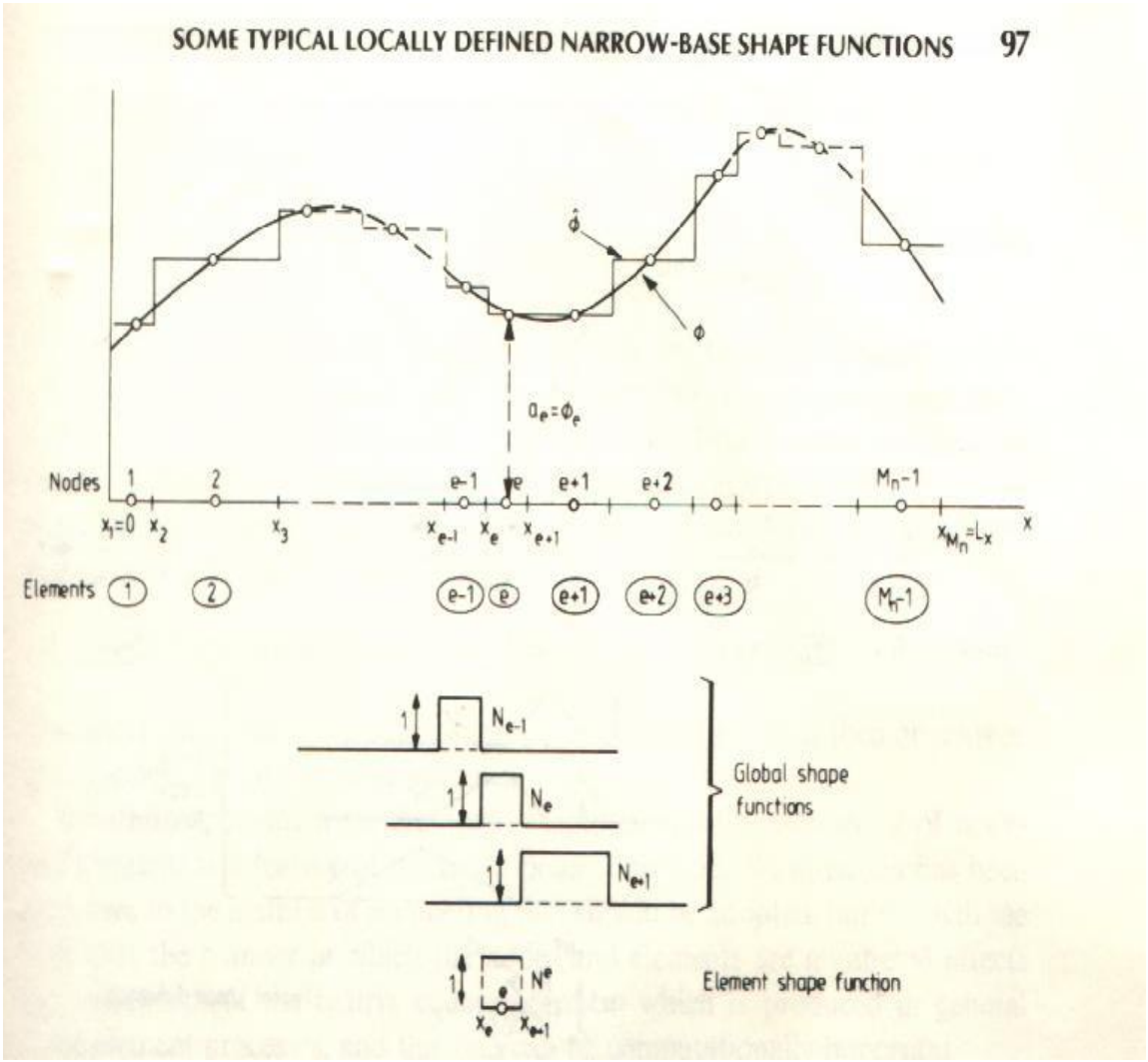


Figura 6.1: Aproximación por funciones a trozos

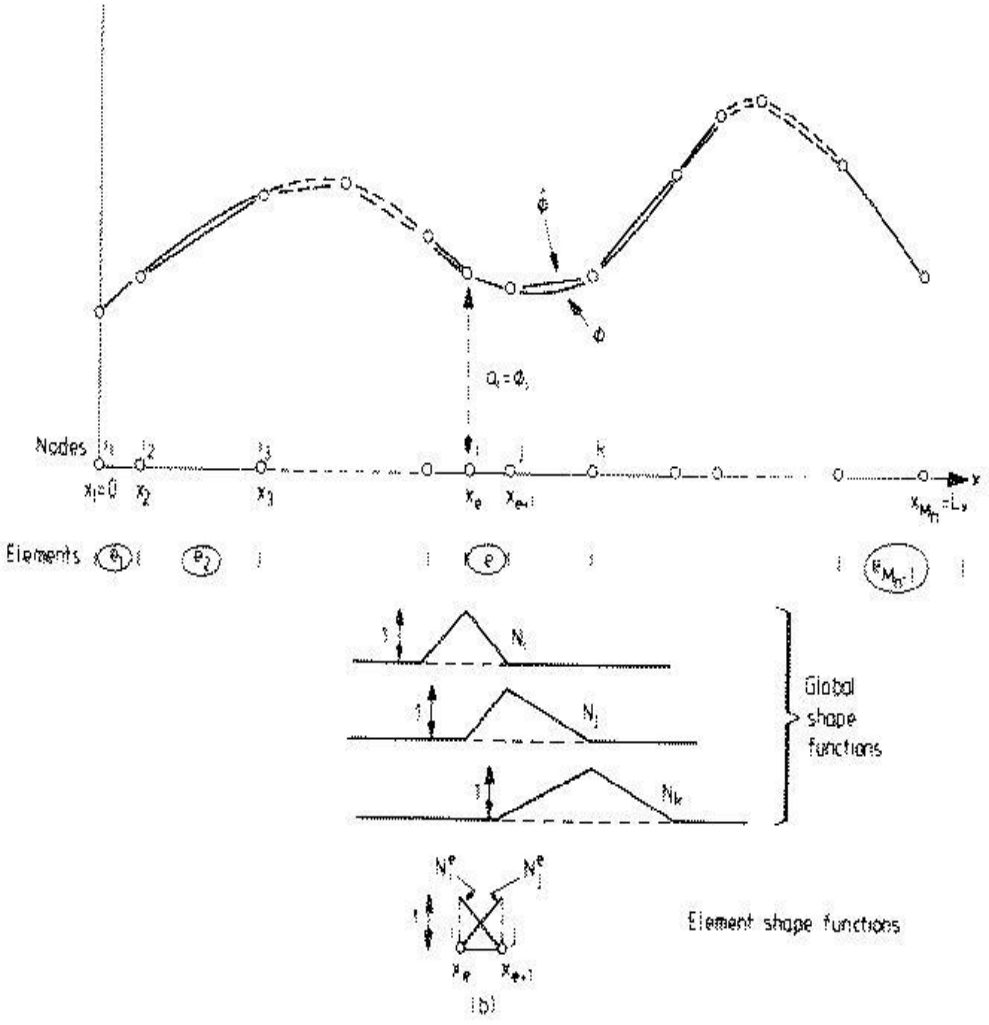


FIGURE 3.1. *(continued).*

Figura 6.2: Aproximación por funciones a trozos

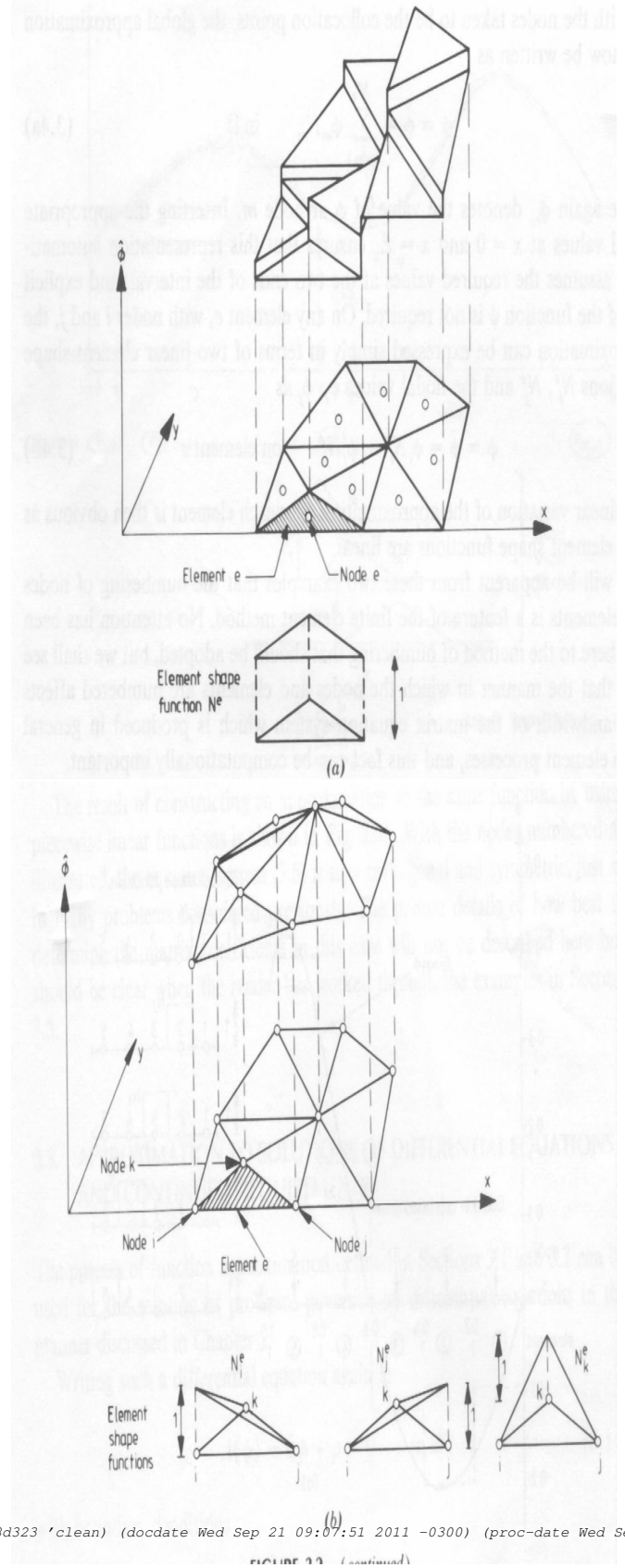


Figura 6.3: Funciones a trozos en varias dimensiones

A su vez es posible tanto aplicar un método de los residuos ponderados de colocación como uno tipo Galerkin. En el primer caso las integrales se pesan con distribuciones tipo *delta de dirac* en los nodos, ya sea en el centro del elemento como en sus extremos. En el caso de la formulación de Galerkin las funciones de peso coinciden con las de interpolación dando resultados distintos pero equivalentes. Al igual que en el capítulo anterior el método de los residuos ponderados puede aplicarse para aproximar una función como para resolver una ecuación diferencial. El primero es un caso particular del segundo donde el operador $\mathcal{L}$ es la identidad. Los detalles del procedimiento incluido en el método de los elementos finitos se verán en próximas secciones cuando resolvamos algunos ejemoplos en particular.

6.3. Aproximación a soluciones de ecuaciones diferenciales. Requisitos sobre la continuidad de las funciones de forma

El método de los residuos ponderados aplicado a la resolución de una ecuación diferencial del tipo:

$$A(\phi) = \mathcal{L}\phi + p = 0 \qquad \text{en } \Omega \tag{6.12}$$

con condiciones de contorno

$$B(\phi) = \mathcal{M}\phi + r = 0 \qquad \text{sobre } \Gamma \tag{6.13}$$

se escribe como

$$\int_\Omega W_l R_\Omega d\Omega + \int_\Gamma \overline{W_l} R_\Gamma d\Gamma = 0 \tag{6.14}$$

con

$$\begin{aligned} R_\Omega &= A(\hat{\phi}) = \mathcal{L}\hat{\phi} + p \\ R_\Gamma &= B(\hat{\phi}) = \mathcal{M}\hat{\phi} + r \end{aligned} \tag{6.15}$$

La idea de usar un método numérico basado en aproximar una solución mediante funciones de forma locales plantea el interrogante de si es posible su uso sabiendo que tanto la función como alguna de sus derivadas pueden ser discontinuas.

Para ilustrar la explicación observemos la figura 6.4 donde se presentan tres tipos de funciones, la de la izquierda es discontinua con derivada puntualmente no acotada, la del centro continua con derivada primera discontinua y derivada segunda puntualmente no acotada y aquella de la derecha que es continua y tiene primera derivada continua, segunda derivada discontinua y tercera derivada puntualmente no acotada. Pensemos que estas funciones pueden ser posibles candidatos a ser usados como funciones de forma para nuestro método numérico. El hecho que alguna de las derivadas presente una singularidad puede provocar problemas en el cómputo de las matrices por lo cual se debe evitar su uso. Para ello si los operadores involucrados en el cálculo de las matrices contienen derivadas de orden s entonces debemos garantizarnos que las funciones de forma tengan $s-1$ derivadas continuas o sea las funciones deben pertenecer a la clase C^{s-1}. Un ejemplo concreto lo tenemos si observamos de la sección anterior que la forma débil del problema de conducción térmica contiene a lo sumo primeras derivadas. En esa caso $s=1$ y necesitamos que las funciones de forma sean C^0, o sea funciones continuas, como aquella ubicada en el centro de la figura 6.4. Una observación a hacer es que no debe confundirse la continuidad de una función con el grado del polinomio

que la aproxima localmente. Por ejemplo una función de forma que usa polinomios de segundo órden en el interior de los elementos no necesariamente tiene mayor continuidad que una lineal. La continuidad depende de como se pegan las funciones elemento a elemento y no de como se interpola en su interior. Los elementos de la clase *Lagrange*, los más standard en método de los elementos finitos pertenecen a la clase C^0 mientras que aquellos del tipo *Hermite* son C^1. Mientras los primeros requieren que las funciones se peguen en forma continua entre los elementos los últimos son más estrictos y requieren que la primera derivada también sea continua, independientemente de lo que suceda adentro. La figura 6.5 pretende ilustrar una función de forma hermitiana para un elemento unidimensional de dos nodos.

Los requisitos de continuidad que estuvimos tratando también se aplican a las funciones de peso. En estos casos una sutileza debe remarcarse. Cuando hablamos del método de colocación hemos empleado la función *delta de Dirac* para tal fin. Esto solo es factible si se garantiza que el residuo sea finito.

6.4. Formulación débil y el método de Galerkin

En el capítulo anterior vimos como un término como el siguiente

$$\int_\Omega W_l \mathcal{L}\hat{\phi} d\Omega \tag{6.16}$$

con $\mathcal{L}$ un operador diferencial lineal puede ser reemplazado por otro como

$$\int_\Omega \left(\mathcal{C}W_l\right)\left(\mathcal{D}\hat{\phi}\right) d\Omega + \int_\Gamma W_l \mathcal{E}\hat{\phi} d\Gamma \tag{6.17}$$

donde los operadores lineales $\mathcal{C}, \mathcal{D}\mathcal{E}$ tienen un orden de diferenciación estrictamente inferior al de $\mathcal{L}$. Tal reformulación es muy ventajosa cuando se usan funciones de forma definidas localmente debido a que esto disminuye el grado de continuidad a ser demandado sobre las mismas. Uno de los operadores que frecuentemente aparecen es el de segundo orden que cuando se lo somete a la debilitación (6.16,6.17) da orígen a dos operadores de primer orden, uno aplicado a la función de peso y el otro a la interpolación. Esto pone de manifiesto que la elección de la misma clase de funciones para W como para N, base del método de Galerkin, posee muchas ventajas. Además de equiparar el grado de continuidad queda garantizada la simetría del operador.

6.5. Aspectos computacionales del método de los elementos finitos

En esta sección tomaremos algunos ejemplos muy simples que servirán para explicar la metodología empleada en el método de los elementos finitos . Estos ejemplos consisten de dominios unidimensionales, con una partición muy gruesa (pocos elementos empleados) y con funciones de prueba sencillas. No obstante esto el procedimiento es bien general y puede extenderse tanto a mallas muy finas como al caso multidimensional y con funciones de alto orden.

6.5.1. Ejemplo 1

Este primer ejemplo consiste en resolver:

$$\begin{aligned} &\frac{d^2\phi}{dx^2} - \phi = 0 \qquad 0 \leq x \leq 1 \\ &\phi(x=0) = 0 \\ &\phi(x=1) = 1 \end{aligned} \tag{6.18}$$

Paso 1: Elección de la funciones de forma y formulación del problema Teniendo en cuenta las consideraciones acerca del grado de continuidad de las funciones de forma relativa al orden de diferenciación del operador involucrado en este ejemplo vemos de inmediato que funciones de la clase C^0 son suficientes para resolver este problema. Esto permite usar funciones lineales a trozos aunque si se desea también puede aumentarse el grado del polinomio interpolante, como por ejemplo usar elementos cuadráticos o de mayor orden. No obstante esto involucra en general un mayor costo computacional muchas veces innecesario para la precisión requerida. Entonces la aproximación se escribe como:

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M+1} \phi_m N_m \qquad 0 \leq x \leq 1 \tag{6.19}$$

Al establecer el método de los residuos ponderados sobre este problema tenemos

$$\int_0^1 W_l \Big(\frac{d^2\hat{\phi}}{dx^2} - \hat{\phi} \Big) dx + \Big[\overline{W_l} R_\Gamma \Big]_0^1 = 0 \tag{6.20}$$

donde el último término representa la contribución del residuo en el borde porque la aproximación elegida no necesariamente satisface las condiciones de borde. No obstante este término puede omitirse si uno elije funciones de peso que se anulen sobre la parte del contorno donde se prescriben condiciones tipo Dirichlet. Entonces si lo anulamos y si debilitamos la integral sobre el interior del dominio para poder usar funciones con menores requisitos de continuidad llegamos a:

$$-\int_0^1 \Big(\frac{\mathrm{d}W_l}{\mathrm{d}x} \frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x} + W_l \hat{\phi} \Big) dx + \Big[W_l \frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x} \Big]_0^1 = 0 \qquad l = 1, 2, \ldots, M, M+1 \tag{6.21}$$

Usando el método de Galerkin $W_l = N_l$ surge obviamente la necesidad de emplear funciones clase C^0.

Paso 2: discretizacion de las variables independientes generacion de la malla Esto consiste en subdividir el dominio en elementos tales que

$$\begin{aligned} \cap_e \Omega^e &= \emptyset \\ \cup_e \Omega^e &= \Omega \end{aligned} \tag{6.22}$$

En este caso simple la partición es trivial, el dominio es un segmento de recta y la partición consiste en dividir este segmento en intervalos que satisfagan (6.22). La situación unidimensional es muy particular y simple ya que (6.22) se puede alcanzar en forma exacta. En varias dimensiones esto se obtiene en general

en forma aproximada. Para simplificar los pasos siguientes dividamos el dominio $\Omega = [0, 1]$ en 3 elementos (4 nodos) como se muestra en la figura 6.6.

De esta forma quedan definidas las coordenadas de los nodos

$$x = \{x_1, x_2, x_3, x_4\} = \{0, L/3, 2L/3L\} \tag{6.23}$$

y las conectividades de los elementos (los nodos asociados a cada elemento)

$$IX = \begin{pmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{pmatrix} \tag{6.24}$$

En lo que siguen definimos el tamaño del elemento como $h^e = |x_j - x_i|$ con i, j los nodos que lo definen.

A continuación presentamos una transformación de coordenadas que facilita mucho la tarea a nivel computacional. Esta se define como:

$$x_k = f(\xi_k) \qquad\qquad k = 1, \ldots ndm \tag{6.25}$$

donde x_k son las coordenadas del elemento en el dominio real del problema y ξ_k son las correspondientes en un elemento denominado *master*. En una dimensión este elemento máster se define como:

$$\hat{\Omega}^e = \{\xi | \xi \in [-1, 1]\} \tag{6.26}$$

La idea es transformar el elemento real en otro muy simple, por ejemplo pensemos simplemente en una cuadrángulo en 2D completamente distorsionado y su mapeo a un cuadrado con sus aristas paralelas a los ejes coordenados. Realizar cálculos de derivadas e integrales en el primero suele ser mucho más trabajoso que en el elemento de forma simple. Si bien este tema presentado en el caso unidimensional parece de poca importancia en varias dimensiones este mapeo es clave para poder facilitar los cálculos. Este tema será presentando con más detalle más adelante en el contexto del método de los elementos finitos en varias dimensiones.

Paso 3: Definición de las funciones de forma Este paso es una consecuencia de los dos anteriores. Habiendo elegido el tipo de funciones de forma necesarios para satisfacer los requerimentos de continuidad que plantea el operador diferencial a resolver y habiendo realizado la discretización de las variables independientes surge de su combinación la definición de las funciones de forma. Estas se pueden escribir como:

$$\begin{aligned} N_i = N_i^e &= \frac{\chi}{h^e} \\ N_j = N_j^e &= \frac{h^e - \chi}{h^e} \\ \text{con } \chi &= x - x_i \end{aligned} \tag{6.27}$$

donde se asume que $x_i < x_j$ y tiene un valor unitario en el nodo cayendo linealmente hasta ser nula en el otro nodo del mismo elemento.

Esta definición es muy ad-hoc para el caso unidimensional y no usa la idea de mapeo al elemento master definida anteriormente. La siguiente si presenta esa idea y la incluimos porque será la que en definitiva usaremos al momento de calcular las integrales.

$$\begin{aligned} N_i &= N_i^e = 1/2(1-\xi) \\ N_j &= N_j^e = 1/2(1+\xi) \end{aligned} \tag{6.28}$$

Esta definición como vemos satisface la definición de la función, en el nodo i, $\xi = -1 \Rightarrow N_i = 1$ y $N_j = 0$ mientras que en el nodo j, $\xi = 1 \Rightarrow N_i = 0$ y $N_j = 1$. Una definición equivalente a la anterior que suele unificar la anterior y es válida para los dos nodos es la siguiente:

$$\begin{aligned} N_l = N_l^e &= 1/2(1+\xi_l\xi) \\ \xi_l &= \begin{cases} -1 & ; l = 1 \\ +1 & ; l = 2 \end{cases} \end{aligned} \tag{6.29}$$

donde l representa la numeración local de los nodos, a nivel del elemento. Esta tiene la ventaja que permite operar algebraicamente con más flexilibidad.

Ensamblándola sobre todo el conjunto de elementos produce una función que tiene la forma de un sombrero como fue mostrada en la figura 6.2.

Paso 4: Cálculo de la matriz del sistema y del miembro derecho Como hemos visto en el primer paso (6.21) la formulación del problema contiene el cálculo de varias integrales, una por cada nodo de la malla. Reemplazando en ella la aproximación (6.19), la definición de la función de peso que para este ejemplo coincide con la de interpolación (Galerkin) y la definición de las funciones de forma (6.27), entonces el cálculo de estas integrales se puede escribir como:

$$\begin{aligned} K_{lm} &= \int_0^1 \Big(\frac{\mathrm{d}N_l}{\mathrm{d}x}\frac{\mathrm{d}N_m}{\mathrm{d}x} + N_l N_m \Big) dx \qquad 1 \le l, m \le M+1 \\ f_l &= \Big[N_l \frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x} \Big]_0^1 \qquad\qquad 1 \le l \le M+1 \end{aligned} \tag{6.30}$$

Estas integrales definidas sobre todo el dominio $\Omega = [0,1]$ pueden descomponerse aditivamente por una de las propiedades de la integración en varias integrales, una por cada elemento. Entonces, calcular las integrales a nivel del elemento y luego ensamblarlas o agregarlas cada una de sus contribuciones es equivalente a integrar sobre todo el dominio. Esto es así porque las funciones que aproximan a la solución fueron definidas elemento a elemento. Por lo tanto lo dicho equivale a:

$$K_{lm} = \sum_{e=1}^{E} \Upsilon^{lm}_{l'm'} K^e_{l'm'}$$

$$\begin{aligned} K^e_{l'm'} &= \int_0^{h^e} \left[\frac{\mathrm{d}N^e_{l'}}{\mathrm{d}x}\frac{\mathrm{d}N_{m'}}{\mathrm{d}x} + N_{l'}(x)N_{m'}(x)\right]dx = \\ &= \int_{-1}^{1} \left[\frac{\mathrm{d}N^e_{l'}}{\mathrm{d}\xi}\frac{\mathrm{d}\xi}{\mathrm{d}x}\frac{\mathrm{d}N_{m'}}{\mathrm{d}\xi}\frac{\mathrm{d}\xi}{\mathrm{d}x} + N_{l'}(\xi)N_{m'}(\xi)\right]\frac{\mathrm{d}x}{\mathrm{d}\xi}d\xi = \\ &= \int_{-1}^{1} \left[\left(\tfrac{1}{2}\xi_{l'}\frac{2}{h^e}\right)\left(\tfrac{1}{2}\xi_{m'}\frac{2}{h^e}\right) + \left(\tfrac{1}{2}(1+\xi_{l'}\xi)\right)\left(\tfrac{1}{2}(1+\xi_{m'}\xi)\right)\right]\frac{h^e}{2}d\xi = \\ &= \int_{-1}^{1} \left[\left(\tfrac{1}{2}\xi_{l'}\frac{2}{h^e}\right)\left(\tfrac{1}{2}\xi_{m'}\frac{2}{h^e}\right) + \left(\tfrac{1}{2}(1+\xi_{l'}\xi)\right)\left(\tfrac{1}{2}(1+\xi_{m'}\xi)\right)\right]\frac{h^e}{2}d\xi = \\ &= \frac{1}{h^e}\xi_{l'}\xi_{m'} + \frac{h^e}{8}(2 + \frac{2}{3}\xi_{l'}\xi_{m'}) \end{aligned} \tag{6.31}$$

Reemplazando por sus valores

$$\xi_{l'} = \begin{cases} -1 & l' = 1 \\ +1 & l' = 2 \end{cases} \quad \mathbf{K}^e = \begin{pmatrix} \frac{1}{h^e} + \frac{h^e}{3} & -\frac{1}{h^e} + \frac{h^e}{6} \\ -\frac{1}{h^e} + \frac{h^e}{6} & \frac{1}{h^e} + \frac{h^e}{3} \end{pmatrix} \tag{6.32}$$

Esta expresión contiene dos juegos de índices, unos sin primas que representan la numeración global de la matriz y está asociado a la numeración global de la malla y otros índices primados que tienen la numeración local dentro del elemento. Hay una relación entre ambos y viene expresada por la tabla de conectividades definida antes (6.24). Para el el elemento é-simo, fila e del arreglo IX se satisface que

con lo cual el álgebra anterior se pudo compactar bastante. Nosotros para simplificar la notación y expresar este cambio de numeración en la expresión (6.31) usamos $\Upsilon^{lm}_{l'm'}$.

Esta relación entre numeraciones es la que permite el ensamble de contribuciones elementales en globales. Esto significa que

$$\begin{aligned} \mathbf{K} &= \mathcal{A}_{e=1}^{E}\mathbf{K}^e \\ K_{l,m} &= \sum_{e=1}^{E}\sum_{l',m'=1}^{2} K^e_{l',m'} \end{aligned} \tag{6.33}$$

con $\mathcal{A}$ el operador de ensamblaje. Esto se ilustra en la figura 6.6.

Si bien todo el cálculo ha sido llevado a cabo manualmente esto tiene un alto grado de automatización debido a lo simple que resultan las funciones de forma y sus derivadas en el elemento master.

Paso 5: Resolución del sistema de ecuaciones Con el miembro derecho el procedimiento es similar y finalmente se alcanza el siguiente sistema de ecuaciones lineales a resolver:

$$\mathbf{K}\boldsymbol{\phi} = \mathbf{f}$$

$$\begin{pmatrix} \frac{1}{h^e}+\frac{h^e}{3} & -\frac{1}{h^e}+\frac{h^e}{6} & 0 & 0 \\ -\frac{1}{h^e}+\frac{h^e}{6} & 2\left(\frac{1}{h^e}+\frac{h^e}{3}\right) & -\frac{1}{h^e}+\frac{h^e}{6} & 0 \\ 0 & -\frac{1}{h^e}+\frac{h^e}{6} & 2\left(\frac{1}{h^e}+\frac{h^e}{3}\right) & -\frac{1}{h^e}+\frac{h^e}{6} \\ 0 & 0 & -\frac{1}{h^e}+\frac{h^e}{6} & \frac{1}{h^e}+\frac{h^e}{3} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix} = \begin{pmatrix} -\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}\Big|_{x=0} \\ 0 \\ 0 \\ \frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}\Big|_{x=1} \end{pmatrix} \tag{6.34}$$

Dado que las condiciones de contorno definidas inicialmente implicaban imponer el valor de $\phi(x=0)=0$ y $\phi(x=1)=1$, esto implica primero reemplazar dichos valores en el vector incógnita y pasar para el miembro derecho todos aquellos términos que dichos valores generan para luego remover las filas y las columnas correspondientes a estas variables impuestas.

Por lo tanto el sistema (6.34) se transforman en otro más reducido:

$$\begin{pmatrix} 2\left(\frac{1}{h^e}+\frac{h^e}{3}\right) & -\frac{1}{h^e}+\frac{h^e}{6} \\ -\frac{1}{h^e}+\frac{h^e}{6} & 2\left(\frac{1}{h^e}+\frac{h^e}{3}\right) \end{pmatrix} \begin{pmatrix} \phi_2 \\ \phi_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -(-\frac{1}{h^e}+\frac{h^e}{6}) \end{pmatrix} \tag{6.35}$$

La resolución de este sistema es inmediata obteniendo los valores del arreglo $\boldsymbol{\phi}$, solución a nuestro problema. Con ellos es posible estimar las derivadas en los extremos reemplazando simplemente el vector solución en (6.34) y despejando el valor de $\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}\Big|_{x=0,1}$.

Comentarios finales Para terminar con este ejemplo haremos dos comentarios, uno acerca del cálculo de las integrales elementales y otro acerca de la resolución del sistema algebraico. Respecto al primero es de destacar que si bien aquí hemos recurrido a la integración analítica en general se trabaja utilizando integración numérica con lo cual solo se necesita evaluar el integrando en determinados puntos del dominio y sumar las contribuciones, como es el caso de integración por cuadratura numérica. Con esto es posible extender el tratamiento a operadores de diferente tipo, incluso con coeficientes variables dentro del elemento, etc. El tema de la integración numérica será abordado más adelante. Respecto a la resolución del sistema se debe evaluar el método a seguir según el tipo de problema a resolver, lineal o no lineal, estacionario o transiente, 2D o 3D y fundamentalmente teniendo en cuenta los recursos computacionales disponibles. Este tema que hoy en dia es un area en si misma será tratada en un próximo capítulo.

6.5.2. Ejemplo 2

Este ejemplo es similar al anterior salvo en el tipo de condiciones de contorno impuesta. En este caso las mismas son: $\phi|_{x=0}=0$ y $\frac{\mathrm{d}\phi}{\mathrm{d}x}|_{x=1}=1$

La formulación del método de los residuos ponderados para este problema es la siguiente:

$$\int_0^1 W_l\Big(\frac{d^2\hat{\phi}}{dx^2}-\hat{\phi}\Big)dx + \Big[\overline{W_l}\Big(\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}-1\Big)\Big]_{x=1} - \Big[\overline{W_l}\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}\Big]_{x=1} = 0 \tag{6.36}$$

Ya que la integral sobre el dominio no ha cambiado y los términos de contorno no influyen sobre la matriz del sistema, esta última queda igual que en la del ejemplo anterior. Por el cambio en el tipo de derivada solo

se modifica el miembro derecho con lo que el sistema a resolver se transforma en uno idéntico a (6.34) con un miembro derecho

$$\left(-\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}\Big|_{x=0} \quad 0 \quad 0 \quad 1\right)^T \tag{6.37}$$

Lo restante es todo similar a lo ya visto.

6.5.3. Ejemplo 3

Este ejemplo muestra como tratar un sistema de ecuaciones diferenciales e incluso para darle cierta generalidad usaremos una formulación mixta.

El problema a resolver es el de conducción del calor con fuente (ec. de Poisson) reemplazado por una formulación mixta del tipo:

$$\begin{aligned} \kappa\frac{\mathrm{d}\phi}{\mathrm{d}x} + q &= 0 \\ \frac{\mathrm{d}q}{\mathrm{d}x} - Q &= 0 \end{aligned} \tag{6.38}$$

Escribiendo para cada variable incógnita su aproximación en la forma:

$$\begin{aligned} q \approx \hat{q} &= \sum_{m=1}^{M_q} q_m N_{m,1} \\ \phi \approx \hat{\phi} &= \sum_{m=1}^{M_\phi} \phi_m N_{m,2} \end{aligned} \tag{6.39}$$

El método de los residuos ponderados aplicado a la anterior se puede escribir como:

$$\begin{aligned} \int_0^1 \kappa\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x}N_{l,1}dx + \int_0^1 \hat{q}N_{l,1}dx = 0 \qquad & l = 1,2,\ldots,M_q \\ \int_0^1 \frac{\mathrm{d}\hat{q}}{\mathrm{d}x}N_{l,2}dx = \int_0^1 QN_{l,2}dx \qquad & l = 1,2,\ldots,M_\phi \end{aligned} \tag{6.40}$$

Aquí hay varias alternativas para elegir las funciones de forma. La primera que vamos a adoptar es aquella en la que ambas variables se aproximan con funciones lineales a trozos con los pesos coincidentes entre si y coincidente a las funciones de interpolación. O sea $N_{l,1} = N_{l,2} = N_l$ Por cada elemento hay 4 incógnitas, 2 nodos con 2 grados de libertad por nodo.

Las expresión para el cálculo de las matrices elementales ahora contienen a su vez una matriz, o sea que al agregar la contribución de cada uno de los nodos del elemento contra todos los otros nodos del mismo elemento (en 1D, 2×2) se llega a:

$$\mathbf{K}^e_{l',m'} = \begin{pmatrix} \int_{-1}^1 N_{l',1}\kappa\frac{\mathrm{d}N_{m',1}}{\mathrm{d}\xi}d\xi & \int_{-1}^1 N_{l',1}N_{m',2}\frac{h^e}{2}d\xi \\ 0 & \int_{-1}^1 N_{l',2}\frac{\mathrm{d}N_{m',2}}{\mathrm{d}\xi}d\xi \end{pmatrix} \tag{6.41}$$

lo cual produce la siguiente matriz elemental

$$\mathbf{K}^e = \begin{pmatrix} \int_{-1}^{1} N_{1,1}N_{1,2}\frac{h^e}{2}d\xi & \int_{-1}^{1} N_{1,1}\kappa\frac{\mathrm{d}N_{1,1}}{\mathrm{d}\xi}d\xi & \int_{-1}^{1} N_{1,1}N_{2,2}\frac{h^e}{2}d\xi & \int_{-1}^{1} N_{1,1}\kappa\frac{\mathrm{d}N_{2,1}}{\mathrm{d}\xi}d\xi \\ \int_{-1}^{1} N_{1,2}\frac{\mathrm{d}N_{1,2}}{\mathrm{d}\xi}d\xi & 0 & \int_{-1}^{1} N_{1,2}\frac{\mathrm{d}N_{2,2}}{\mathrm{d}\xi}d\xi & 0 \\ \int_{-1}^{1} N_{2,1}N_{1,2}\frac{h^e}{2}d\xi & \int_{-1}^{1} N_{2,1}\kappa\frac{\mathrm{d}N_{1,1}}{\mathrm{d}\xi}d\xi & \int_{-1}^{1} N_{2,1}N_{2,2}\frac{h^e}{2}d\xi & \int_{-1}^{1} N_{2,1}\kappa\frac{\mathrm{d}N_{2,1}}{\mathrm{d}\xi}d\xi \\ \int_{-1}^{1} N_{2,2}\frac{\mathrm{d}N_{1,2}}{\mathrm{d}\xi}d\xi & 0 & \int_{-1}^{1} N_{2,2}\frac{\mathrm{d}N_{2,2}}{\mathrm{d}\xi}d\xi & 0 \end{pmatrix} \tag{6.42}$$

Lo que sigue es simplemente calcular las integrales y resolver el sistema.

Otra alternativa para (6.40) es debilitar la segunda ecuación y cambiarla de signo:

$$\int_0^1 \hat{q}\frac{\mathrm{d}N_{l,2}}{\mathrm{d}x}dx - \left[\hat{q}N_{l,2}\right]_0^1 = -\int_0^1 QN_{l,2}dx \qquad l = 1, 2, \ldots, M_\phi \tag{6.43}$$

y por haber disminuido el orden de diferenciación sobre q podemos usar funciones de forma constantes a trozo tanto para interpolar esta función como para pesar la misma ($N_{l,1} = N_{m,1}$ constantes a trozos). De esta forma en cada elemento la cantidad de grados de libertad se reduce de 4 a 3, los dos valores de ϕ en los nodos y el valor de q en el centro del elemento. La contribución elemental se escribe como:

$$\mathbf{K}^e\boldsymbol{\phi}^e = \int_{\Omega^e} \begin{pmatrix} 0 & \frac{\mathrm{d}N_i}{\mathrm{d}x} & 0 \\ \frac{\mathrm{d}N_i}{\mathrm{d}x} & 1 & \frac{\mathrm{d}N_j}{\mathrm{d}x} \\ 0 & \frac{\mathrm{d}N_j}{\mathrm{d}x} & 0 \end{pmatrix} \begin{pmatrix} \phi_i \\ q_e \\ \phi_j \end{pmatrix} dx \tag{6.44}$$

6.6. Interpolación de mayor orden en 1D

Tanto en el capítulo anterior como en las secciones anteriores del prsente hemos introducido el concepto de aproximación mediante el uso de funciones de prueba con cierta suavidad aplicada a todo el dominio *métodos globales* o aquellas definidas elemento a elemento *métodos locales* con interpolación constante o lineal por elemento. Esta última alternativa condujo a la forma más simple de definir el método de los elementos finitos . La motivación de mejorar la aproximación a la solución del problema conduce a varias clases de refinamiento, entre ellas las más usuales son:

1.- usar una interpolación de bajo orden y refinar la malla *tipo h*,

2.- usar una malla fija y una interpolación de mayor orden *tipo p*,

3.- una combinación de las dos anteriores *tipo h-p*.

Desde el punto de vista práctico la mejor solución es la alcanzar la mayor precisión al menor costo posible. Si bien la elección es muy dependiente del problema es cierto que en muchas aplicaciones refinar en el polinomio de aproximación es el camino óptimo. No obstante, en ciertas aplicaciones la definición de las funciones de interpolación está muy restringida por la formulación del problema. Ejemplos de este caso lo tenemos en las formulaciones aplicadas a la resolución de flujo compresible e incompresible. El tratamiento de la compresibilidad pone una fuerte restricción en la elección de los espacios funcionales. No todas las combinaciones posibles para aproximar la velocidad y la presión son permisibles, existiendo un criterio a satisfacer (*critero de inf-sup*) para los espacios funcionales. Por otro lado la advección dominante requiere

la estabilización numérica del problema que se hace cada vez más complicada de definir a medida crece el orden de los polinomios interpolantes. Estas son las principales causas por las cuales en el tratamiento numérico de modelos físicos más complicados se prefiere el uso de funciones de interpolación de bajo orden y se refina sobre el tamaño de la malla. De todos modos nosotros aquí presentaremos algunos detalles acerca de la interpolación de mayor orden restringida al caso unidimensional, aunque la extensión a muchas dimensiones es tan directa como la del caso lineal. Dado un grado polinomial, existen muchas formas para generar funciones de forma que tengan ese orden de aproximación. La más standard es la de utilizar una base de Lagrange en la cual cada grado polinomial diferente se alcanza con un juego de bases diferentes. No obstante existen otros métodos que son aditivos en el sentido que dada una función de forma de grado n se puede generar aquella de orden $n+1$ usando la de grado n y agregandole alguna función adicional. Este método es denominado *formas jerárquicas* de aproximación y son computacionalmente muy eficientes aunque su uso no ha sido muy difundido aún.

6.6.1. Grado de las funciones de prueba y velocidad de convergencia

La discusión acerca de la estimación del error y la convergencia de una aproximación en general es algo complicado de tratar y anteriormente solo mencionamos que la convergencia se alcanzará si la aproximación es *completa*. Esta forma vaga de definirla se debía a la generalidad de la elección de las funciones de prueba. Si nos limitamos a una interpolación polinomial el tratamiento se simplifica mucho.

Consideremos un dominio Ω subdividido en elementos Ω^e de tamaño h y tomemos funciones de prueba interpoladas mediante un polinomio completo de grado p. Es claro que si la solución exacta del problema es un polinomio de grado menor o igual a p la solución numérica será exacta (sin tener en cuenta errores de redondeo) independientemente del peso utilizado. Obviamente que la solución en general no será polinomial, no obstante, si no contiene singularidades (en la función o en alguna de sus derivadas) puede ser bien aproximada por una expansión en series de Taylor.

$$\phi(\Delta x) = \phi|_O + \Delta x \frac{\mathrm{d}\phi}{\mathrm{d}x}|_O + \frac{(\Delta x)^2}{2}\frac{d^2\phi}{dx^2}|_O + \dots \tag{6.45}$$

donde la serie de Taylor se ha tomado alrededor del punto O. Entonces al usar funciones de prueba polinomiales de grado p estamos aproximando exactamente los primeros $p+1$ término en (6.45) y el error de la aproximación se vuelve $O(h^{p+1}$. Es de notar que la aproximación a la primera derivada será $O(h^p)$ y aquella correspondiente a la derivada d-ésima será $O(h^{p+1-d})$

Volviendo a la formulación general del método de los residuos ponderados o a aquella del método de los elementos finitos vimos que estaban involucrados dos operadores diferenciales $\mathcal{L}$ y $\mathcal{M}$. Estos operadores contienen derivadas y supongamos que la máxima derivada sea d, es claro que el menor orden para la aproximación de la solución debe ser tal que la representación del operador diferencial de la solución sea $O(h)$ por razones de completitud. Entonces se requiere que:

$$p+1-d \geq 1 \Rightarrow p-d \geq 0 \tag{6.46}$$

Este requisito de completitud pone de manifiesto la utilidad de la formulación débil presentada anteriormente. Habímos visto oportunamente que al debilitar la formulación disminuían los requisitos de continuidad de la aproximación. Ahora se pone de manifiesta una segunda ventaja de la debilitación, reducir el orden de diferenciación del operador diferencial disminuye el orden de la mínima interpolación necesaria para garantizar convergencia en malla. Como ejemplo a esto podemos mencionar el caso del operador involucrado en

un problema de conducción del calor. Este contiene como máximo segundas derivadas, al debilitarlo usando el método de Galerkin las mayores derivadas son de primer orden con lo cual se requieren funciones C^0 para satisfacer continuidad entre elementos y como mínimo funciones lineales para obtener convergencia en malla.

6.6.2. Funciones de forma de alto orden standard de la clase C^0

Sea el conjunto de elementos unidimensionales como el ilustrado en la figura 6.7 y tomamos un elemento e de la malla con sus nodos extremos numerados como 0 y 1. La aproximación lineal mostrada en la parte superior de la figura asegura que la aproximación es lineal sobre todo el elemento y que los parámetros asociados a los nodos corresponden a los valores nodales en los mismos nodos. Además de esta forma se garantiza la continuidad C^0. Usando la idea de mapeo al elemento de referencia las funciones de forma se escriben como:

$$N_0^e = \tfrac{1}{2}(1-\xi) \qquad\qquad N_1^e = \tfrac{1}{2}(1+\xi) \tag{6.47}$$

La extensión al caso cuadrático se logra definiendo un nodo intermedio y tomando ahora tres funciones, una por cada nodo, con la condición de ser cuadrática dentro del elemento y valer uno en el nodo asociado y cero en los otros dos, o sea:

$$\begin{gathered} N_l^e(\xi) \in \mathcal{P}_2(\xi) \\ N_l^e(\xi) = \begin{cases} 1 & \xi = \xi_l \\ 0 & \xi = \xi_m \end{cases} \quad l,m = 0,1,2 \qquad \text{y } m \neq l \\ \xi_0 = -1 \qquad \xi_1 = 0 \qquad \xi_2 = 1 \end{gathered} \tag{6.48}$$

Extendiendo lo anterior la caso cúbico y luego generalizando es posible definir cada una de estas funciones resolviendo un sistema lineal para los coeficientes de cada función de forma asociada con cada nodo. En general, si el grado del polinomio es p, entonces habrá $p+1$ nodos en el elemento y

$$N_l^e = \alpha_0 + \alpha_1\xi + \alpha_2\xi^2 + \cdots + \alpha_p\xi^p \tag{6.49}$$

$$\begin{aligned} \xi = \xi_0 \qquad & N_l^e = \alpha_0 + \alpha_1\xi_0 + \alpha_2\xi_0^2 + \cdots + \alpha_p\xi_0^p \\ \xi = \xi_1 \qquad & N_l^e = \alpha_0 + \alpha_1\xi_1 + \alpha_2\xi_1^2 + \cdots + \alpha_p\xi_1^p \\ & \vdots \qquad\qquad \vdots \\ \xi = \xi_l \qquad & N_l^e = \alpha_0 + \alpha_1\xi_l + \alpha_2\xi_l^2 + \cdots + \alpha_p\xi_l^p \\ & \vdots \qquad\qquad \vdots \\ \xi = \xi_p \qquad & N_l^e = \alpha_0 + \alpha_1\xi_p + \alpha_2\xi_p^2 + \cdots + \alpha_p\xi_p^p \end{aligned} \tag{6.50}$$

De esta forma surgen las funciones de prueba para el caso cuadrático:

$$N_0^e = \tfrac{1}{2}\xi(1-\xi) \qquad N_1^e = (1+\xi)(1-\xi) \qquad N_2^e = \tfrac{1}{2}\xi(1+\xi) \tag{6.51}$$

y para el caso cúbico

$$N_0^e = -\frac{9}{16}(\xi+\frac{1}{3})(\xi-\frac{1}{3})(\xi-1) \qquad N_1^e = \frac{27}{16}(\xi+1)(\xi-\frac{1}{3})(\xi-1)$$
$$N_2^e = -\frac{27}{16}(\xi+1)(\xi+\frac{1}{3})(\xi-1) \qquad N_3^e = \frac{9}{16}(\xi+\frac{1}{3})(\xi-\frac{1}{3})(\xi+1) \tag{6.52}$$

6.7. Problemas con advección dominante - Método de Petrov-Galerkin

Como ya hemos visto al introducir los método de los residuos ponderados estos métodos se basan en definir funciones de peso diferentes a las elegidas para interpolar la solución, lo cual da orígen a una gran variedad de posibilidades. Lo que se pretende aquí es introducir este tema que inmediatamente encontrará su utilidad cuando se pretenda resolver problemas dominados por advección, naturalmente hallados en problemas de mecánica de fluidos.

Nuestro interés por este tema surge de la gran popularidad que ha tomado este tipo de formulaciones en el área de la mecánica de fluidos, especialmente el método denominado SUPG. Si bien no es nuestro interés aquí hacer historia sobre este tema podemos mencionar que la idea del método SUPG fue tratar de buscar el efecto producido por las ya conocidas y efectivas técnicas de *upwinding* para evitar oscilaciones numéricas producidas cuando en las ecuaciones de transporte dominan los términos convectivos sobre los restantes. La siguiente es un ejemplo típico de una ecuación de transporte donde el miembro izquierdo representa la convección y es proporcional a la velocidad con que se mueve el fluido mientras que el miembro derecho es el término difusivo, si ϕ es la temperatura equivale a la conducción del calor.

$$u\frac{\mathrm{d}\phi}{\mathrm{d}x} = \kappa\frac{d^2\phi}{dx^2} \tag{6.53}$$

Entonces adimensionalizando la ecuación anterior surge el número de Peclet,

$$Pe = \frac{|u|L}{\kappa} \tag{6.54}$$

relación entre la convección y la difusión, con L una dimensión característica del problema. Si barremos el valor de Peclet desde $0 \rightarrow \infty$ vemos que la ecuación cambia de tipo, de ser una elíptica pasa a ser hiperbólica y este cambio depende sobre la competencia entre los términos convectivos y los difusivos. Al resolver el problema por diferencias finitas centradas aparecen oscilaciones cuando el Peclet de la grilla supera un valor próximo a la unidad. La técnica del upwindind surgió en el área de las diferencias finitas como intento de evitar las mencionadas oscilaciones y fue planteada sobre la base de aplicar una aproximación en diferencias decentrada a la derivada primera. El decentraje debía hacerse aguas arriba considerando la orientación del flujo y esto pudo explicarse desde muchos puntos de vista. Uno de los más importantes es aquel ligado al concepto de error de truncamiento explicándose la aparición de las oscilaciones del hecho que al truncar la aproximación centrada esta introduce una especie de difusión numérica negativa que compite con la física. Existe un cierto valor del número de Peclet para el cual la difusión numérica supera a la física y de acuerdo a argumentos termodinámicos se viola el segundo principio y el problema pasa a estar mal planteado. Para ver mejor esto recurrimos nuevamente a la ecuación de advección-difusión en 1D y la discretizamos por diferencias finitas centradas, lo cual es completamente equivalente a haberlo hecho por método de los elementos finitos . El esquema resultante es:

$$\begin{aligned} a\frac{\phi_{i+1}-\phi_{i-1}}{2\Delta x} &= k\frac{\phi_{i+1}-2\phi_i+\phi_{i-1}}{\Delta x^2} \\ Pe(\phi_{i+1}-\phi_{i-1}) &= \phi_{i+1}-2\phi_i+\phi_{i-1} \end{aligned} \tag{6.55}$$

con $Pe = a\Delta x 2k$ es el número de Peclet del elemento y aquí hemos asumido coeficientes constantes y malla uniforme. La solución exacta a este problema es del tipo:

$$\phi_i = \xi^i \tag{6.56}$$

que reemplazada en la ecuación produce la siguiente ecuación algebraica de segundo grado:

$$\xi^2(Pe-1)+2\xi-(Pe+1)=0 \tag{6.57}$$

la cual produce las siguientes raíces:

$$\xi_{1,2} = \begin{cases} 1 \\ \frac{1+Pe}{1-Pe} \end{cases} \tag{6.58}$$

Vemos que si $Pe = 1$ la ecuación degenera en una de primer grado con la solución $\xi = 1$ o sea $\phi =$ constante. Si $Pe < 1$ las dos raíces son positivas lo cual genera soluciones positivas, combinaciones de la solución constante y otra del tipo $\phi = \frac{1+Pe}{1-Pe}^i$. El problema surge cuando $Pe > 1$ ya que en este caso una de las raíces es negativa y genera soluciones del tipo $\phi = (-1)^i \frac{1+Pe}{|1-Pe|}^i$. Esta solución contiene una oscilación numérica que puede arreglarse refinando el elemento siempre que exista algo de difusión en el problema, o sea que el $Pe \neq \infty$. Extrapolando al caso de mecánica de fluidos esta situación se presenta en el caso de flujo viscoso (Navier-Stokes) cuando el Reynolds del elemento supera un valor crítico, del órden de la unidad. Una situación extrema ocurre en el caso de los modelos invíscidos. Allí la difusión física es nula y por más que refinemos el $Pe \to \infty$, generando las raices $\xi = \pm 1$, lo cual implica una oscilación irremediable. Usando diferencias finitas descentradas aguas arriba equivale a que la raiz negativa introducida por el término cuadrático, o sea el nodo aguas abajo, sea removida. Es por ello que es usual aproximar la primera derivada con una diferencia hacia atrás, de forma de que ahora el esquema es:

$$\begin{aligned} a\frac{\phi_i-\phi_{i-1}}{\Delta x} &= k\frac{\phi_{i+1}-2\phi_i+\phi_{i-1}}{\Delta x^2} \\ 2Pe(\phi_i-\phi_{i-1}) &= \phi_{i+1}-2\phi_i+\phi_{i-1} \end{aligned} \tag{6.59}$$

El caso de $Pe \to \infty$ transforma la ecuación de segundo grado en otra de primer grado, lo cual genera una única solución (constante) lo cual tiene sentido físico ya que el operador diferencial también cambia de tipo (órden) cuando sucede esto.

El *upwind* puede verse como el agregado de viscosidad artificial o difusión artificial en pos de contrarestar la difusión negativa que introduce la discretización. Controlar esta difusión agregada artificialmente es importante ya que si nos excedemos la solución es sobredifusiva y estamos resolviendo un problema con mayor difusión que la real, y por el lado contrario si no introducimos la suficiente aparecerán oscilaciones no físicas. Para ver esto se puede partir de la ecuación discreta centrada (6.55) y restarle la recién obtenida (6.59), lo cual pone en evidencia el término introducido artificialmente:

$$\frac{a}{2\Delta x}\Big(\phi_{i+1} - 2\phi_i + \phi_{i-1}\Big) = \frac{a\Delta x}{2}\Big(\frac{\phi_{i+1} - 2\phi_i + \phi_{i-1}}{\Delta x^2}\Big) = \tag{6.60}$$

con $\frac{a\Delta x}{2}$ funcionando como una especie de difusión artificial.

Lo anterior se lo conoce como la técnica de *full upwind*. Se sabe que esto es correcto solo cuando $Pe \to \infty$ y que cuando Pe asume valores próximos al valor crítico la difusión introducida debe ser corregida para evitar soluciones muy suavizadas. Para ello se ha recurrido al caso unidimensional lineal el cual tiene solución exacta y se ha demostrado que la forma de corregir es introducir una función denominada *mágica* del tipo:

$$\psi(Pe) = coth(Pe) - \frac{1}{Pe} \tag{6.61}$$

En el contexto del método de los elementos finitos el mismo efecto puede lograrse de varias formas, por ejemplo mediante el uso del método de los residuos ponderados Petrov-Galerkin. Tomando la ecuación (6.53).

$$\int_\Omega W_l u \frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x} - \frac{\mathrm{d}W_l}{\mathrm{d}x}\kappa\frac{\mathrm{d}\hat{\phi}}{\mathrm{d}x} = 0 \tag{6.62}$$

y definiendo a la función de peso como:

$$W_l = N_l + \tau u \frac{\mathrm{d}N_l}{\mathrm{d}x} \tag{6.63}$$

donde el primer término reproduce el método de Galerkin mientras que el segundo es una perturbación cuyo efecto en la matriz es tal que al ser aplicado a un término proporcional a $\frac{\mathrm{d}\phi}{\mathrm{d}x}$ produce un término similar a uno difusivo. El parámetro τ debe ajustarse en función de la cantidad de perturbación (difusión artificial) a agregar. De todos modos existen muchos trabajos que dan cuenta de la buena confiabilidad de la definición:

$$\tau = \psi(Pe)\frac{1}{||\frac{\mathrm{d}\xi}{\mathrm{d}x}u||} \tag{6.64}$$

donde $||\frac{\mathrm{d}\xi}{\mathrm{d}x}u||$ equivale al vector velocidad transformado al elemento máster y la función $\psi(Pe)$ es la llamada función mágica definida más arriba.

Un aspecto de importancia es la continuidad de las funciones de peso. Según la definición $N_l \in C^0$, luego $W_l \in C^{-1}$ con lo cual se plantea una dificultad matemática que ha sido muy bien estudiada. Sin entrar en detalles las conclusiones han sido que el problema se suscita en los bordes entre elementos y que la formulación variacional que surge del planteamiento de la forma débil del método de los residuos ponderados arroja la satisfacción de las ecuaciones en el interior de los elementos y de los flujos en los bordes.

Su extensión al caso multidimensional dió orígen a los métodos denominados *Streamline diffusion* ya que introducen la difusión según las líneas de corriente. Estos métodos han mostrado ser muy robustos y eficientes a la hora de resolver problemas de flujo de fluidos no obstante, a diferencia del caso unidimensional, ahora no se cuenta con una solución que permita controlar la cantidad de difusión a agregar y esta debe ser introducida acorde a criterios ad-hoc.

La extensión al caso de sistemas de ecuaciones, tal como ocurre con los modelos de Navier-Stokes y Euler requieren un estudio un poco más detallado del tema y será abordado en futuros capítulos.

6.8. El caso multidimensional

6.8.1. Introducción

En esta sección por su simplicidad presentaremos el caso 2D siendo directa la extensión a 3D. En cuanto a las funciones de interpolación existe mucha bibliografía en la cual se puede hallar las expresiones de las mismas tanto para funciones de diferente grado de continuidad como de diferente orden polinomial. Nosotros aquí solo trataremos el caso de funciones de prueba de clase C^0 multilineales sobre elementos de forma triangular o cuadrangular. Dado que en el caso multidimensional se presentan situaciones geométricas completamente arbitrarias lo mejor para ordenar el tratamiento del tema es introducir las funciones de forma en el elemento master y luego mencionar los detalles del mapeo que transforma las coordenadas reales en las del elemento de referencia.

6.8.2. Elemento triangular

El triángulo es una forma particularmente muy útil porque permite representar con bastante precisión dominios de forma arbitraria. Para un triángulo típico e con nodos numerados en sentido antihorario i, j, k y ubicados en los vértices del mismo como se muestra en la figura 6.3 buscamos una funciones de forma elemental $N_i^e(x, y)$ tal que,

$$\begin{aligned} N_i^e(x,y) &= 1 \qquad \text{en } x = x_i, y = y_i \\ N_i^e(x,y) &= o \qquad \text{en } x = x_j, x_k, y = y_j, y_k \\ \text{con } N_i(x,y) &\text{ continua sobre } \cup_{e\in\mathcal{S}_i^e} \Gamma_e \\ \text{con } N_i(x,y) &\neq 0 \cup_{e\in\mathcal{S}_i^e} \Omega_e \end{aligned} \tag{6.65}$$

con $\mathcal{S}_i^e$ el conjunto de elementos que contienen al nodo i.

El mapeo de cualquier triángulo al elemento master puede verse en la figura 6.8, donde los nodos i, j, k se posicionan en $1, 2, 3$ respectivamente.

Las funciones de interpolación lineales que satisfacen lo anterior son:

$$\begin{aligned} N_1(\xi,\eta) &= 1 - \xi - \eta \\ N_2(\xi,\eta) &= \xi \\ N_3(\xi,\eta) &= \eta \end{aligned} \tag{6.66}$$

Como se puede apreciar geométricamente con 3 puntos definimos exactamente un plano y por lo tanto el gradiente de cualquier función ϕ definida como combinación lineal de sus tres valores nodales será constante dentro del elemento. Calcularemos primero el mapeo.

$$\begin{aligned} x(\xi,\eta) &= \sum_{k=1}^{3} x_k N_k(\xi,\eta) \\ &= x_1(1-\xi-\eta) + x_2\xi + x_3\eta \\ &= x_1 + \xi(x_2-x_1) + \eta(x_3-x_1) \\ y(\xi,\eta) &= \sum_{k=1}^{3} y_k N_k(\xi,\eta) \\ &= y_1(1-\xi-\eta) + y_2\xi + y_3\eta \\ &= y_1 + \xi(y_2-y_1) + \eta(y_3-y_1) \end{aligned} \tag{6.67}$$

De lo cual surge que dado un par $(\bar{\xi},\bar{\eta})$ podemos calcular inmediatamente su correspondiente $(\bar{x},\bar{y})$ y viceversa. Retomando lo dicho antes con la aproximación de la solución tenemos

$$u^e(x,y) = \sum_{k=1}^{3} u_k N_k^e(x(\xi,\eta), y(\xi,\eta)) = \sum_{k=1}^{3} u_k N_k^e(\xi,\eta) \tag{6.68}$$

Entonces

$$\begin{aligned} u^e &= u_1 + \xi(u_2-u_1) + \eta(u_3-u_1) \\ \nabla u^e &= \Big(\frac{\partial u^e}{\partial x}, \frac{\partial u^e}{\partial y}\Big) = \Big(\frac{\partial u^e}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u^e}{\partial \eta}\frac{\partial \eta}{\partial x}, \frac{\partial u^e}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial u^e}{\partial \eta}\frac{\partial \eta}{\partial y}\Big) \end{aligned} \tag{6.69}$$

donde

$$\begin{aligned} \frac{\partial \phi}{\partial \xi} &= \phi_2 - \phi_1 = \text{ constante} \\ \frac{\partial \phi}{\partial \eta} &= \phi_3 - \phi_1 = \text{ constante} \end{aligned} \tag{6.70}$$

con $\phi = x, y, u$

Lo mismo vale para los elementos del jacobiano y su inversa

$$J = \begin{pmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \eta}{\partial x} \\ \frac{\partial \xi}{\partial y} & \frac{\partial \eta}{\partial y} \end{pmatrix} \quad J^{-1} = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{pmatrix} \tag{6.71}$$

$$J = \frac{1}{|J|}\begin{pmatrix} (y_3-y_1) & -(y_2-y_1) \\ -(x_3-x_1) & (x_2-x_1) \end{pmatrix} \quad J^{-1} = \begin{pmatrix} (x_2-x_1) & (y_2-y_1) \\ (x_3-x_1) & (y_3-y_1) \end{pmatrix} \tag{6.72}$$

con el determinante del jacobiano expresado como

$$|J| = (x_2-x_1)(y_3-y_1) - (y_2-y_1)(x_3-x_1) \tag{6.73}$$

Con toda esta información es posible reemplazar en las integrales que surgen al aplicar el método de los residuos ponderados y obtener tanto la matriz del sistema como el vector miembro derecho. La ventaja del uso de elementos triangulares radica en que al ser los gradientes constantes permiten simples reglas de

integración. No obstante más adelante veremos que en pos dee generalizar el tratamiento introduciremos la integración numérica como medio para evaluar las mismas. Una vez que las contribuciones elementales han sido computadas se deben agregar a la matriz global. Este procedimiento puede visualizarse en la figura 6.9 donde se muestra un ejemplo 2D discretizado mediante elementos triangulares lineales. La forma en que se hace la numeración de la malla influye notablemente en la posición de los elementos no nulos de la malla y esto influye directamente sobre la definición del ancho de banda de la matriz, un parámetro que afecta el costo de resolver el sistema en forma drástica. Visualizando el elemento $e = 1$ si por un momento alteramos la numeración intercambiando aquella del nodo 4 con la del nodo 17, entonces la matriz será llena y el ancho de banda coincide con la dimensión completa de la matriz.

6.8.3. Elemento cuadrangular

Los elementos cuadrangulares si bien poseen menos posibilidades de representar con precisión un dominio de forma arbitraria tienen como ventaja su buena performance en aplicaciones de mecánica de fluidos. En el caso de elementos cuadrangulares la transformación al elemento máster se ilustra en la figura 6.10.

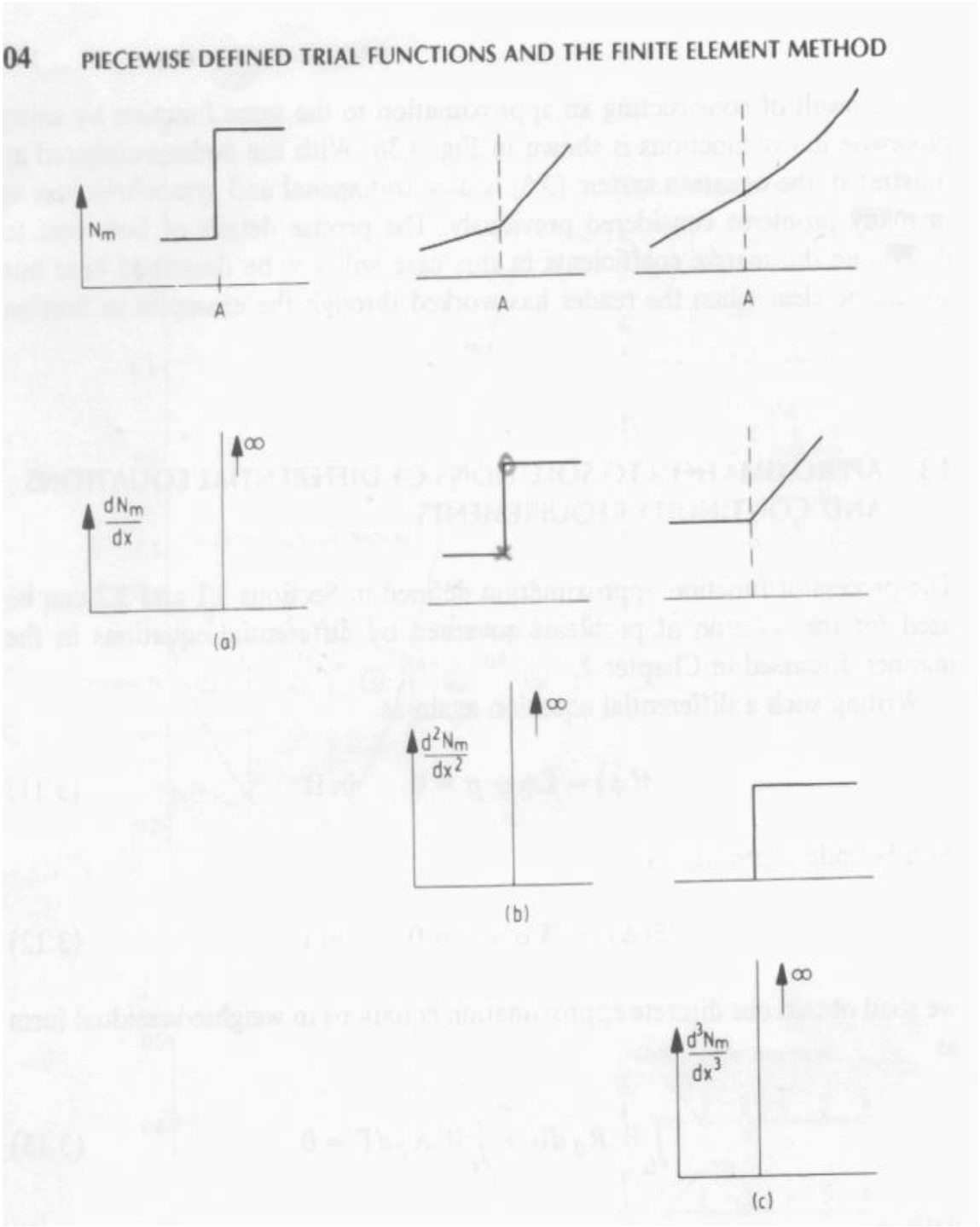


Figura 6.4: Continuidad de las funciones de forma

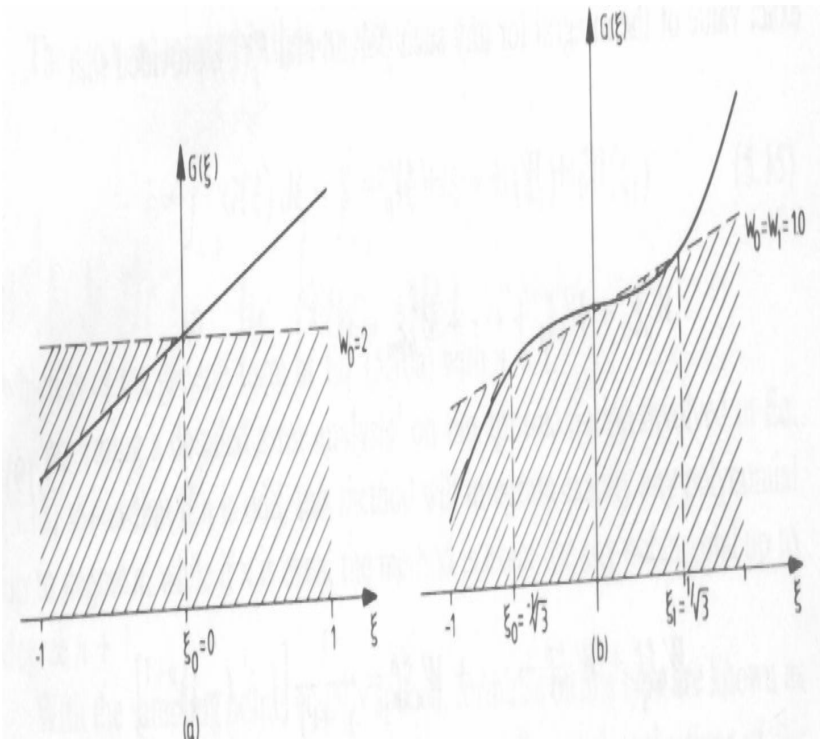


Figura 6.5: Funciones de forma de mayor orden

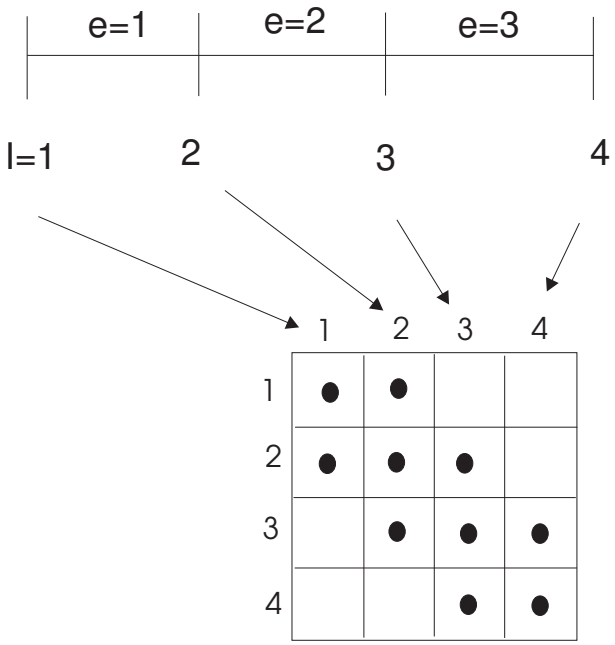


Figura 6.6: Malla del ejemplo 1

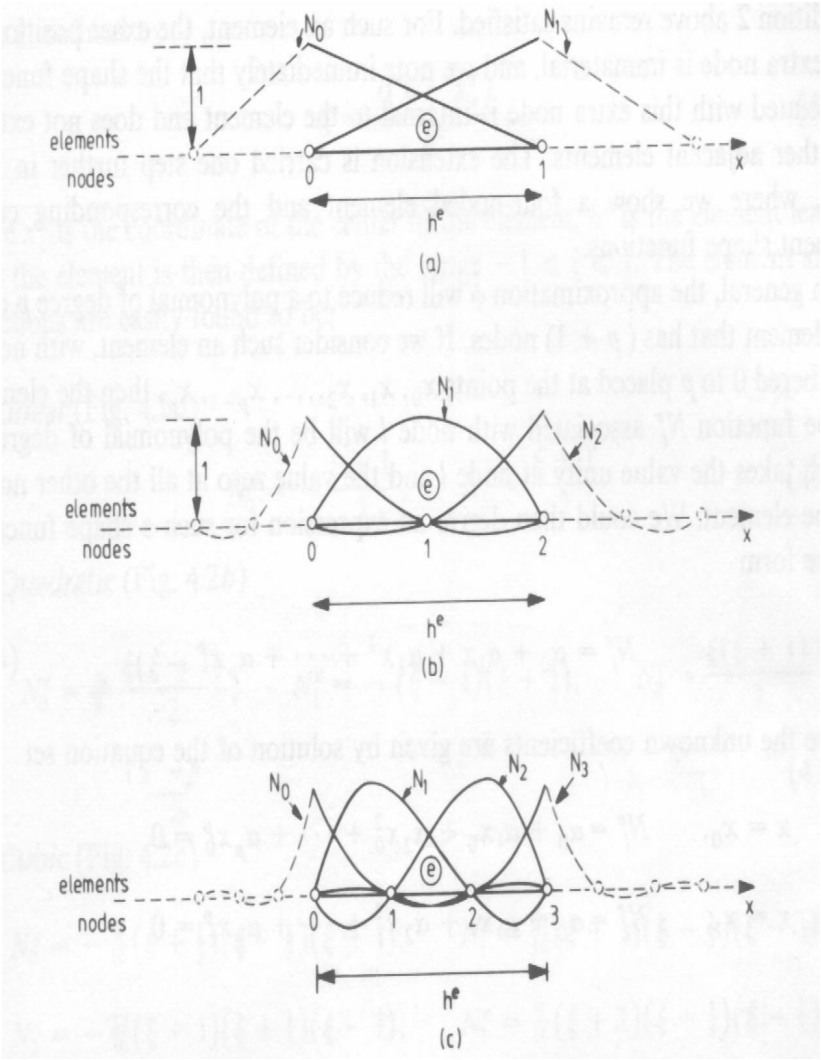


Figura 6.7: Aproximaciones de mayor orden

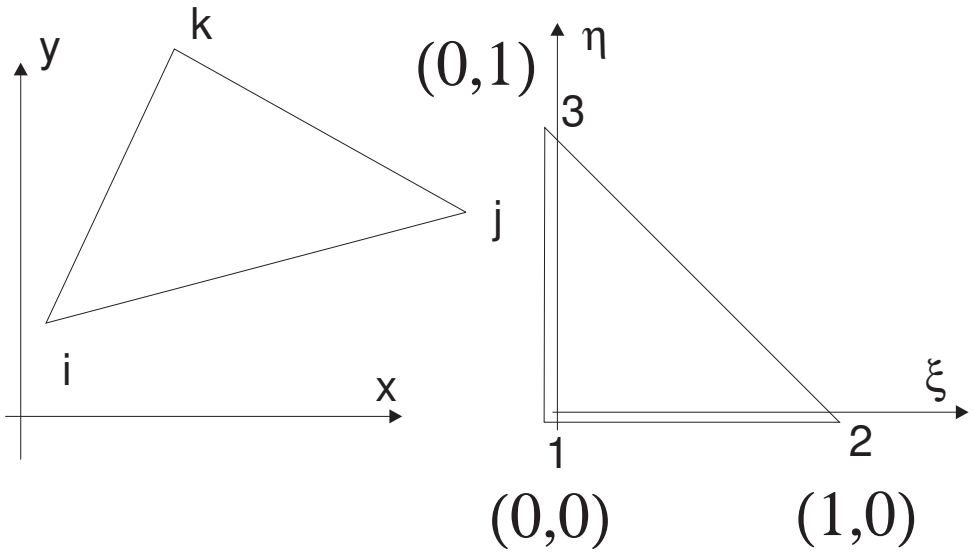


Figura 6.8: Mapeo para elementos triangulares

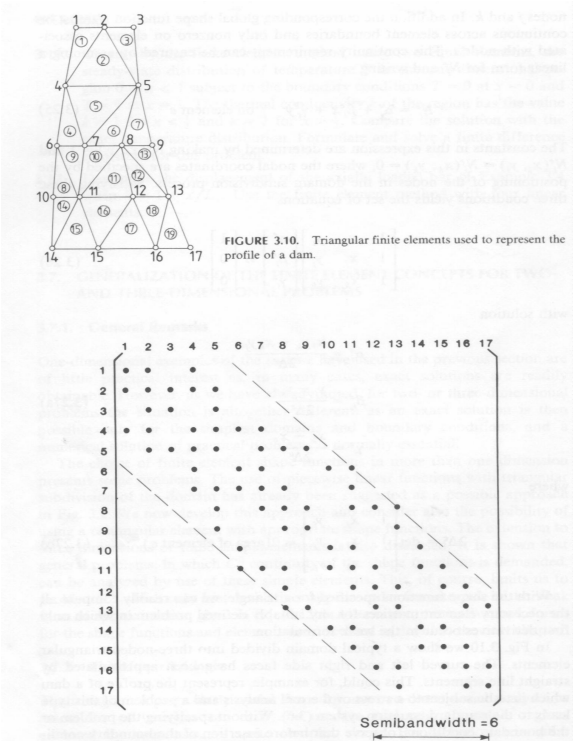


FIGURE 3.10. Triangular finite elements used to represent the profile of a dam.

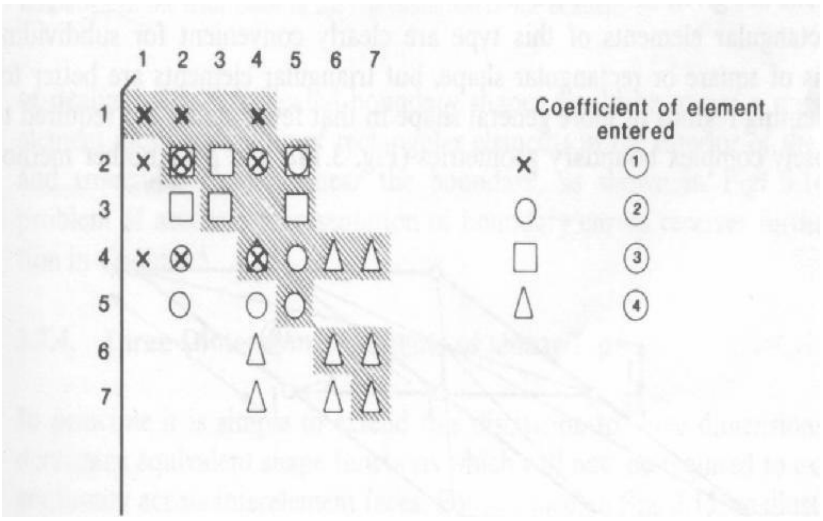


Figura 6.9: Ensamble para mallas triangulares

Las funciones de interpolación se calculan como el producto cartesiano de las funciones de prueba unidimensionales (funciones bilineales) y pueden escribirse como:

$$\begin{aligned} N_i &= \frac{1}{4}(1+\xi_i\xi)(1+\eta_i\eta) \\ \xi_i &= \Big\{ -1\,;+1\,;+1\,;-1 \Big\} \\ \eta_i &= \Big\{ -1\,;-1\,;+1\,;+1 \Big\} \end{aligned} \tag{6.74}$$

Esto origina un polinomio incompleto de segundo grado. De la misma forma que en el caso triangular el mapeo se logra interpolando cada coordenada según:

$$\begin{aligned} & x(\xi,\eta) = \sum_{k=1}^{4} x_k N_k(\xi,\eta) = \\ & = \frac{1}{4}\Big\{ x_1(1-\xi)(1-\eta) + x_2(1+\xi)(1-\eta) + x_3(1+\xi)(1+\eta) + x_4(1-\xi)(1+\eta) \Big\} = \\ & = \frac{x_1+x_2+x_3+x_4}{4} + \xi\Big[\frac{(x_2+x_3)-(x_1+x_4)}{4}\Big] + \eta\Big[\frac{(x_3+x_4)-(x_1+x_2)}{4}\Big] + \\ & + \xi\eta\Big[\frac{(x_1+x_3)-(x_2+x_4)}{4}\Big] \end{aligned} \tag{6.75}$$

Entonces dado $(\bar{\xi},\bar{\eta})$ podemos calcular inmediatamente su correspondiente $(\bar{x},\bar{y})$ pero no podemos plantear la correspondencia inversa ya que el sistema es incompletamente cuadrático.

En estos casos ni los gradientes ni los jacobianos son constantes por elemento sino que dependen de la posición. Esto agrega complicación algebraica al cálculo de las integrales y es aquí donde más rédito se obtiene de emplear integración numérica.

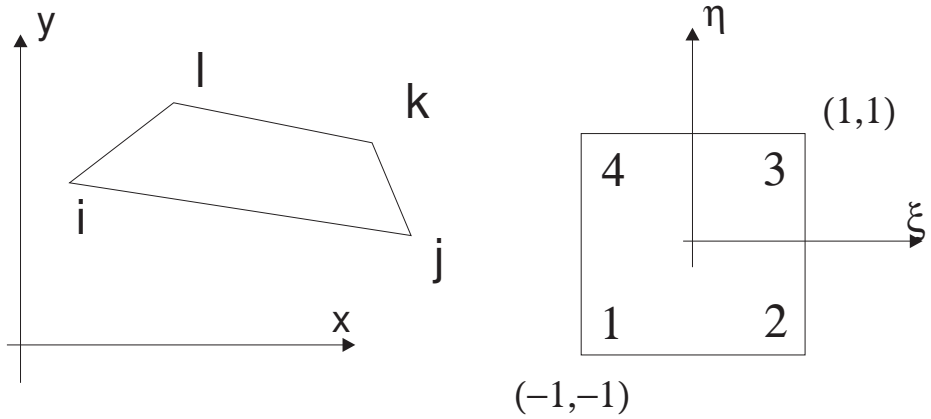


Figura 6.10: Mapeo elemento cuadrangular

6.8.4. Transformación de coordenadas

Tanto los métodos globales como los locales de alto órden requieren realizar integrales sobre porciones extendidas del dominio, incluso sobre la totalidad del mismo, los cuales difícilmente pueden ser representados por elementos de forma simple. La complejidad de los dominios de cálculo hace necesaria una transformación de coordenadas de forma de poder llevar un dominio arbitrario $\Omega(x, y)$ a otro mucho más simple $\hat{\Omega}(\xi, \eta)$ donde realizar las operaciones tales como la integración para el cálculo del sistema algebraico a resolver.

Por transformación de coordenadas o *mapeo* del inglés *mapping* entendemos una transformación unívoca entre (ξ, η) y (x, y). Para entrar en tema consideremos la familiar transformación de coordenadas cilíndricas polares a cartesianas, donde

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta) \\ \xi &= r \\ \eta &= \theta \end{aligned} \tag{6.76}$$

En la figura **??** vemos detalles bien conocidos de esta transformación. Supongamos que un mapeo en general se describa mediante funciones

$$\begin{aligned} x &= f_1(\xi, \eta) \\ y &= f_2(\xi, \eta) \\ z &= f_3(\xi, \eta) \end{aligned} \tag{}$$

$$x_k = f_k(\xi_j) \qquad j, k = 1, \ldots, ndm$$

con ndm el número de dimensiones del dominio o número de variables independientes espaciales. Una vez que se conocen las coordenadas en el dominio físico de cada uno de los elementos de la malla $(x^e, y^e) \in \Omega^e \subset \Omega$ y elegido el tipo de mapeo a realizar $f_k(\xi_j)$ se pueden escribir las funciones de forma y sus derivadas sobre el elemento de referencia *master* $(\hat{\Omega})$ de tal forma de poder realizar las integraciones propias del método sobre este dominio sencillo apelando a técnicas standard de cálculo numérico como la integración por cuadratura gaussiana u otras. Los requisitos de continuidad de las funciones en los contornos entre elementos se garantiza eligiendo apropiadamente las funciones de forma. En las aplicaciones standard del método de los elementos finitos se usan funciones de forma del tipo C^0, o sea continuas con todas sus derivadas discontinuas. No obstante eligiendo apropiadamente estas funciones se podrían obtener aproximaciones con mayor suavidad. Cuando presentamos las diferentes técnicas numéricas aplicadas a los modelos vimos que existe la necesidad de derivar las funciones de forma respecto a las variables independientes, tanto las coordenadas espaciales en el dominio global como el tiempo. Con respecto a las derivadas especiales es necesario aplicar la regla de la cadena:

$$\begin{aligned}
\frac{\partial N_l^e}{\partial x} &= \frac{\partial N_l^e}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial N_l^e}{\partial \eta}\frac{\partial \eta}{\partial x} \\
\frac{\partial N_l^e}{\partial y} &= \frac{\partial N_l^e}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial N_l^e}{\partial \eta}\frac{\partial \eta}{\partial y} \\
\nabla N_l^e &= \mathbf{J}^{-1}\nabla_\xi N_l^e \\
\nabla_\xi N_l^e &= \begin{pmatrix} \frac{\partial N_l^e}{\partial \xi} \\ \frac{\partial N_l^e}{\partial \eta} \end{pmatrix} \\
\mathbf{J} &= \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{pmatrix}
\end{aligned} \tag{6.77}$$

donde $\mathbf{J}$ es el *jacobiano* de la transformación, una medida de la deformación que se requiere para llevar un diferencial de elemento en el dominio local a otro en el global. Como se alcanza a ver para que la transformación exista se requiere que el jacobiano sea inversible. Para ver como se transforma el area tomamos:

$$dxdy = |\mathbf{J}|d\xi d\eta \tag{6.78}$$

De esta forma tenemos todos los elementos necesarios para realizar cualquier integración que surge de la aplicación del método de los elementos finitos . Supongamos que tomamos la siguiente integral proveniente de un problema de conducción térmica sobre un dominio mapeable en un cuadrado:

$$\begin{aligned}
I &= \int_{\Omega^e} \kappa \nabla N_l^e \cdot \nabla N_m^e dxdy \\
&= \int_{-1}^{1}\int_{-1}^{1} \kappa(\mathbf{J}^{-1}\nabla_\xi N_l^e)\cdot(\mathbf{J}^{-1}\nabla_\xi N_m^e)|\mathbf{J}|d\xi d\eta
\end{aligned} \tag{6.79}$$

Mapeo paramétrico

Una forma útil de definir un mapeo de entre muchas posibles alternativas es utilizando la misma clase de funciones utilizadas para interpolar la o las variables dependientes de un problema ϕ. En este caso cada una de las coordenadas espaciales pueden tratarse como si fueran una función más a aproximar:

$$\begin{aligned}
x &= \sum_{l=1}^{M} N_l^e(\xi,\eta)x_l \\
y &= \sum_{l=1}^{M} N_l^e(\xi,\eta)y_l
\end{aligned} \tag{6.80}$$

Si elegimos las mismas funciones de forma para las variables dependientes e independientes, entonces la aproximación se denomina *isoparamétrica*. Ya que las funciones de forma son continuas el mapeo paramétrico será continuo. De todas formas es posible aproximar a distinto orden las variables independientes y las dependientes aunque no es lo usual. Con el mapeo se simplifica mucho el análisis y la programación de los esquemas numéricos.

6.8.5. Integración numérica

El cálculo de las contribuciones elementales tanto de la matriz como del vector miembro derecho requiere resolver integrales sobre el dominio ocupado por cada elemento con un integrando que contiene al operador diferencial discreto pesado con alguna función de prueba. Tanto la forma del elemento en principio arbitraria como el grado de las funciones de forma pueden complicar demasiado la evaluación analítica del problema haciendo este procedimiento completamente dependiente de la aplicación. Para ganar en generalidad e incluso para facilitar la tarea se recurre a integrar numéricamente. Esto consiste en reemplazar la integral por una suma donde cada término de la misma es el producto del integrando evaluado en cada punto de muestreo pesado con un valor correspondiente a cada uno de esos mismos puntos. Por ejemplo en 1D

$$I = \int_{-1}^{1} G(\xi) d\xi = \sum_{pg}^{Npg} G(\xi_{pg}) w_{pg} \tag{6.81}$$

donde ξ_{pg}, w_{pg} son las coordenadas de cada punto de muestreo y su correspondiente peso. La forma en que se deducen las coordenadas y los pesos en los puntos de muestreo diferencian a un método de otro. Sin entrar en detalles técnicos acerca de este tema el cual es ampliamente tratado en cursos regulares de cálculo numérico podemos decir que en el contexto de método de los elementos finitos es muy usual el método de la cuadratura gaussiana. Este procedimiento define la posición de los puntos de muestreo en el interior del dominio de forma de aproximar exactamente la integral de una función aproximada por un polinomio de grado $\leq p$. Sea la función

$$F_p(\xi) = \alpha_0 + \alpha_1 \xi + \cdots + \alpha_p \xi^p + \tag{6.82}$$

cuya integral evaluada numéricamente mediante (6.81) nos da:

$$\begin{aligned} I = \int_{-1}^{1} F_p(\xi) d\xi = \sum_{pg}^{Npg} F_p(\xi_{pg}) w_{pg} = \\ w_0(\alpha_0 + \alpha_1 \xi_0 + \cdots + \alpha_p \xi_0^p) + w_1(\alpha_0 + \alpha_1 \xi_1 + \cdots + \alpha_p \xi_1^p) + \cdots + \\ w_{Npg}(\alpha_0 + \alpha_1 \xi_{Npg} + \cdots + \alpha_p \xi_{Npg}^p) \end{aligned} \tag{6.83}$$

Comparando la integral exacta

$$I_{ex} = 2\alpha_0 + \frac{2\alpha_2}{3} + \ldots \frac{\alpha_p}{p+1}[1 - (-1)^{p+1}] \tag{6.84}$$

con la numérica (6.83) se establece un sistema de $p+1$ ecuaciones con $2(Npg+1)$ incógnitas y la solución solo es posible cuando

$$p + 1 = 2(Npg + 1) \tag{6.85}$$

y ya que Npg es un entero p será siempre un número impar.

Comparando este método de cuadratura de Gauss-Legendre con el método de Newton-Cotes vemos que este permite con 3 evaluaciones aproximar un integrando de 5to orden mientras que el de Newton-Cotes con 3 evaluaciones solo permite aproximar exactamente una función de tercer orden. Las coordenadas de los puntos de muestreo y sus pesos correspondientes tanto para el caso 1D como para el multidimensional pueden consultarse en la abundante bibliografía del tema.

$n+1$	p
1	1
2	3
3	5
4	4

Cuadro 6.1: Precisión de la integración por Puntos de Gauss

6.9. Problemas dependientes del tiempo

En esta clase de problemas el tiempo aparece como una variable independiente adicional y lo que se busca es la evolución temporal de la solución que a su vez es variable en el espacio. Estos problemas, denominados *de valores iniciales* ocurren muy frecuentemente en las aplicaciones, en problemas de difusión transiente, en propagación de ondas, en problemas de inestabilidad de flujos, etc. En el caso estacionario la discretización de las variables independientes (en este caso las espaciales) conducía a un sistema de ecuaciones algebraicas. Ahora, discretizando en una primera etapa las variables espaciales se alcanza un sistema de ecuaciones diferenciales ordinarias que puede ser resuelto tal como está aplicando técnicas ad-hoc para tal fin. Este método se lo conoce como de *semidiscretización parcial*. La otra alternativa es en una segunda etapa discretizar la variable independiente tiempo mediante alguna técnica, método de los elementos finitos o método de las diferencias finitas , conduciendo nuevamente a un sistemas de ecuaciones algebraicas .

6.9.1. Discretización parcial

Si bien este método fue presentado como para desacoplar el tratamiento de las variables espaciales respecto a la temporal también puede ser aplicado al caso estacionario cuando queremos discretizar solo algunas de las variables espaciales y no todas. Supongamos que $\phi = \phi(x, y, z)$ sea una función a encontrar dependiente de las tres coordenadas espaciales. Podemos aproximar esta función de la forma habitual

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M} a_m(y) N_m(x, z) \tag{6.86}$$

Reemplazando la anterior en el problema diferencial produce que todas las derivadas respecto a y permanecerán tal cual y las restantes derivadas asumirán su versión discreta, llegando finalmente al sistema de ecuaciones diferenciales ordinarias siguiente:

$$\mathbf{K}\mathbf{a} + \mathbf{C}\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}y} + \cdots = \mathbf{f} \tag{6.87}$$

con el orden de la ecuación diferencial ordinaria determinado por el máximo orden de derivación respecto a y. Este tipo de solución prueba ser útil cuando el dominio es prismático en y, o sea no depende de y.

En general podemos decir que si el problema físico viene gobernado por la ecuación diferencial

$$\mathcal{L}\phi + p - \alpha\frac{\partial \phi}{\partial t} - \beta\frac{\partial^2 \phi}{\partial t^2} = 0 \qquad \text{en } \Omega \tag{6.88}$$

entonces si la aproximación planteada es del tipo

$$\phi \approx \hat{\phi} = \psi + \sum_{m=1}^{M} a_m(t) N_m(x, y, z) \tag{6.89}$$

entonces el sistemas de ecuaciones diferenciales ordinarias resultante se puede escribir como:

$$\mathbf{M}\frac{d^2\mathbf{a}}{dt^2} + \mathbf{C}\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t} + \mathbf{K}\mathbf{a} = \mathbf{f}$$

$$\begin{aligned} M_{lm} &= \int_{\overline{\Omega}} \beta W_l N_m d\Omega \\ C_{lm} &= \int_{\overline{\Omega}} \alpha W_l N_m d\Omega \\ K_{lm} &= -\int_{\overline{\Omega}} W_l \mathcal{L} N_m d\Omega \\ f_l &= \int_{\overline{\Omega}} \Big(p + \mathcal{L}\psi - \alpha\frac{\partial\psi}{\partial t} - \beta\frac{\partial^2\psi}{\partial t^2}\Big) W_l d\Omega \end{aligned} \tag{6.90}$$

Adecuadas condiciones de borde sobre $\overline{\Gamma}$ para todo instante t junto con valores iniciales para $\mathbf{a}(t=0)$ y para $\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t}(t=0)$ si $\beta \neq 0$ deben suministrarse para un buen planteamiento del problema.

Ejemplos típicos de las ecuaciones anteriores son:

$$\begin{aligned} \frac{\partial^2\phi}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = 0 \qquad & \textbf{propagación de ondas} \\ \kappa\Big(\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2}\Big) + Q - \rho c\frac{\partial^2\phi}{\partial t^2} = 0 \qquad & \textbf{ecuación del calor lineal} \end{aligned} \tag{6.91}$$

Posponemos para capítulos posteriores la resolución de algunos problemas no estacionarios.

6.9.2. Discretización espacio-temporal por elementos finitos

La resolución del sistema de ecuaciones diferenciales ordinarias que fue presentado en la sección anterior requiere de utilizar algoritmos numéricos que permitan integrar en el tiempo a las mismas. La otra alternativa posible es la de discretizar el operador temporal de alguna forma para convertir el sistema de ecuaciones diferenciales ordinarias en un sistemas de ecuaciones algebraicas que puede resolverse de la misma forma que hemos visto para problemas en estado estacionario. La discretización de la variable independiente tiempo se puede efectuar de muchas formas, entre ellas la más usual es recurrir a esquemas simples en diferencias finitas. Esto consiste en discretizar la o las derivadas temporales mediante esquemas en diferencias y este tema será tratado con mayor grado de detalle más adelante. Por el momento solo ejemplificamos como se podría efectuar esta discretización para el caso de una ecuación como la (6.90).

$$\mathbf{M}\frac{d^2\mathbf{a}}{dt^2} + \mathbf{C}\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t} + \mathbf{K}\mathbf{a} = \mathbf{f}$$

$$\mathbf{M}\frac{\mathbf{a}^{n+1} - 2\mathbf{a}^n + \mathbf{a}^{n-1}}{(\Delta t)^2} + \mathbf{C}\frac{\mathbf{a}^{n+1} - \mathbf{a}^{n-1}}{\Delta t} + \mathbf{K}\mathbf{a}^n = \mathbf{f} \tag{6.92}$$

donde vemos que existen tres niveles de tiempo $t^{n-1} = t^n - \Delta t = (n-1)\Delta t$, $t^n = n\Delta t$ y $t^{n+1} = t^n + \Delta t = (n+1)\Delta t$ y además hemos empleado operadores en diferencias centradas. La forma de llevar a cabo el cálculo da oria gen a dos clases de métodos, los *explícitos* y los *implícitos*. En los primeros no se requiere invertir ninguna matriz y el valor de la solución en un instante de tiempo se puede despejar directamente a partir de aquellos valores de la solución evaluada en tiempos anteriores. En el caso implícito esto no es factible y se requiere invertir un sistema. Este tema será profundizado más adelante.

La segunda alternativa que exploraremos con más detalle a continuación es la de usar funciones de prueba dependientes del tiempo. El hecho de que no sea esta una forma general de trabajo se debe a que:

1.- problemas con tiempos característicos largos involucran un excesivo volúmen de cálculo,

2.- las matrices resultan en general no simétricas aun usando el método de Galerkin,

3.- la simple topología que existe en el dominio temporal ofrece poco atractivo para usar discretización irregulares en el tiempo.

En los últimos tiempos ha habido un gran interés por el uso de formulaciones espacio-temporales para tratar problemas con dominios variables en el tiempo.

Por su mayor grado de aplicación en el area de la mecánica de fluidos en la sección que sigue trataremos exclusivamente el caso de ecuaciones de primer orden.

Ecuaciones de primer orden

Tomando el caso de $\beta = 0$ en (6.88) llegamos al siguiente sistema de ecuaciones diferenciales ordinarias

$$\mathbf{C}\frac{\mathrm{d}\mathbf{a}}{\mathrm{d}t} + \mathbf{K}\mathbf{a} = \mathbf{f} \tag{6.93}$$

Las condiciones iniciales implican conocer el valor de $\phi(t = 0)$ lo cual en forma discreta equivale a conocer $\mathbf{a}(t = 0) = \mathbf{a}^0$. De acuerdo a (6.89) vemos que el problema espacial y el temporal están desacoplados por lo que una posterior aproximación, ahora exclusivamente en el tiempo, para la incógnita $\mathbf{a}$ se puede plantear. Entonces,

$$\mathbf{a} \approx \hat{\mathbf{a}} = \sum_{m=1}^{\infty} \mathbf{a}^m N_m \tag{6.94}$$

con $\mathbf{a}^m = \mathbf{a}(t = t_m)$. Consideremos un elemento n el tiempo con sus nodos extremos $t = t_n$ y $t = t_{n+1}$ como se muestra en la figura 6.8.

Por lo tanto

$$\begin{aligned} N_n^n &= 1-T \qquad & N_{n+1}^n &= T \\ \frac{\mathrm{d}N_n^n}{\mathrm{d}t} &= \frac{-1}{\Delta t_n} \qquad & \frac{\mathrm{d}N_{n+1}^n}{\mathrm{d}t} &= \frac{1}{\Delta t_n} \\ T &= \frac{t-t_n}{\Delta t_n} \qquad & \Delta t_n &= t_{n+1}-t_n \end{aligned} \tag{6.95}$$

Aplicando el método de los residuos ponderados a (6.93) llegamos a:

$$\int_0^\infty \Big(\mathbf{C}\frac{\mathrm{d}\hat{\mathbf{a}}}{\mathrm{d}t} + \mathbf{K}\hat{\mathbf{a}} - \mathbf{f}(t)\Big) W_n dt = 0 \qquad n = 0,1,2,\ldots \tag{6.96}$$

Si las funciones de peso elegidas satisfacen que

$$W_n = 0 \ , \qquad t < t_n \qquad \text{y } t > t_{n+1} \tag{6.97}$$

(6.96) se escribe como:

$$\int_{t_n}^{t_{n+1}} \Big(\mathbf{C}\frac{\mathrm{d}\hat{\mathbf{a}}}{\mathrm{d}t} + \mathbf{K}\hat{\mathbf{a}} - \mathbf{f}(t)\Big) W_n dt = 0 \qquad n = 0,1,2,\ldots \tag{6.98}$$

mientras que (6.94) se escribe como:

$$\hat{\mathbf{a}} = \mathbf{a}^n N_n^n + \mathbf{a}^{n+1} N_{n+1}^n \tag{6.99}$$

la cual reemplazada en (6.98) produce

$$\int_0^1 \Big[\frac{\mathbf{C}}{\Delta t_n}(-\mathbf{a}^n + \mathbf{a}^{n+1}) + \mathbf{K}\{\mathbf{a}^n(1-T) + \mathbf{a}^{n+1}T\} - \mathbf{f}(t_n + \Delta t_n T)\Big] W_n dT = 0 \qquad n = 0,1,2,\ldots \tag{6.100}$$

Estas funciones de peso equivalen a funciones constantes a trozos por elemento.
Si las matrices $\mathbf{C}$ y $\mathbf{K}$ fueran independientes del tiempo entonces:

$$\begin{aligned} \{\frac{\mathbf{C}}{\Delta t_n}\int_0^1 W_n dT + \mathbf{K}\int_0^1 T W_n dT\}\mathbf{a}^{n+1} + \{-\frac{\mathbf{C}}{\Delta t_n}\int_0^1 W_n dT + \mathbf{K}\int_0^1 (1-T) W_n dT\}\mathbf{a}^n = \\ = \int_0^1 \mathbf{f}(t_n + \Delta t_n T) W_n dT \end{aligned} \tag{6.101}$$

(6.101) es válida para cada elemento n de la discretización temporal o sea que permite para cada n obtener los valores de la incógnita $\mathbf{a}^1, \mathbf{a}^2, \mathbf{a}^3 \ldots$ comenzando con el valor $\mathbf{a}^0$. Este tipo de esquema de marcha temporal se lo denomina de *dos niveles* ya que cada cálculo involucra solo dos instantes de tiempo, el actual y el anterior. Existen extensiones a esto que involucran varios niveles de tiempo pero no serán abordados por el momento.

Para generalizar el tratamiento la expresión (6.101) puede reescribirse para ser usada con cualuquier función de peso de la siguiente forma:

$$\Big(\frac{\mathbf{C}}{\Delta t_n}+\gamma_n\mathbf{K}\Big)\mathbf{a}^{n+1}+\Big(-\frac{\mathbf{C}}{\Delta t_n}+(1-\gamma_n)\mathbf{K}\Big)\mathbf{a}^{n}=\overline{\mathbf{f}}^{n}$$
$$\gamma_n=\int_0^1 W_n T dT\Big/\int_0^1 W_n dT \tag{6.102}$$
$$\overline{\mathbf{f}}^{n}=\int_0^1 \mathbf{f}(t_n+\Delta t_n T)W_n dT\Big/\int_0^1 W_n dT$$

y ya que la función $\mathbf{f}$ en general varían suavemente en el tiempo es algunas veces conveniente interpolarla mediante:

$$\mathbf{f}(t_n+T\Delta t_n)=\mathbf{f}^n N_n^n(T)+\mathbf{f}^{n+1}N_{n+1}^n(T)\ ,\qquad 0\le T\le 1 \tag{6.103}$$

Reemplazando (6.103) en (6.102) llegamos a:

$$\overline{\mathbf{f}}^{n}=(1-\gamma_n)\mathbf{f}^n+\gamma_n\mathbf{f}^{n+1} \tag{6.104}$$

con lo cual (6.102) se escribe como:

$$\Big(\frac{\mathbf{C}}{\Delta t_n}+\gamma_n\mathbf{K}\Big)\mathbf{a}^{n+1}+\Big(-\frac{\mathbf{C}}{\Delta t_n}+(1-\gamma_n)\mathbf{K}\Big)\mathbf{a}^{n}=(1-\gamma_n)\mathbf{f}^n+\gamma_n\mathbf{f}^{n+1} \tag{6.105}$$

Si $\mathbf{f}$ no fuera una función suave entonces (6.102) debe ser evaluado exactamente.

Algunos esquemas en particular

Colocación Tomemos el caso del método de los residuos ponderados utilizando como funcón de peso la colocación puntual. En este caso lo de puntual debe reinterpretarse como su equivalente temporal, o sea evaluada no en un punto en el espacio sino en un instante de tiempo. Si $W_n=\delta(T-\theta),\ n=0,1,2,\ldots,$ entonces de (6.102) surge que

$$\gamma_n=\theta \tag{6.106}$$

la cual reemplazada en (6.105) nos da el bien conocido *método theta* que se escribe como:

$$\Big(\frac{\mathbf{C}}{\Delta t_n}+\theta\mathbf{K}\Big)\mathbf{a}^{n+1}+\Big(-\frac{\mathbf{C}}{\Delta t_n}+(1-\theta)\mathbf{K}\Big)\mathbf{a}^{n}=(1-\theta)\mathbf{f}^n+\theta\mathbf{f}^{n+1} \tag{6.107}$$

Este método sirve como un marco teórico general ya que de él se derivan muchos de los esquemas frecuentemente usado en la práctica. Algunos de los ejemplos más citados son:

1.- *Esquema Forward Euler,* $\theta=0$,

2.- *Esquema Backward Euler,* $\theta=1$,

3.- *Crank-Nicolson,* $\theta=1/2$

Galerkin Si en lugar de usar colocación aplicamos el método de Galerkin $(W_n = N_n)$ y si tomamos funciones de prueba de la clase C^0 satisfaciendo que:

$$N_n = 0 \qquad \forall t \leq t_{n-1} \quad \text{y } t \geq t_{n+1} \tag{6.108}$$

al reemplazar en (6.98) se obtiene el siguiente esquema:

$$\Big(\frac{\mathbf{C}}{2\Delta t} + \frac{1}{6}\mathbf{K}\Big)\mathbf{a}^{n+1} + \frac{2}{3}\mathbf{K}\mathbf{a}^{n} + \Big(-\frac{\mathbf{C}}{2\Delta t} + \frac{1}{6}\mathbf{K}\Big)\mathbf{a}^{n-1} = \frac{1}{6}\mathbf{f}^{n+1} + \frac{2}{3}\mathbf{f}^{n} + \frac{1}{6}\mathbf{f}^{n-1} \tag{6.109}$$

que es un esquema de tres niveles de tiempo. De esta forma hemos visto que el tratamiento empleado sobre la discretización espacial puede ser extendido a la temporal en forma directa donde los esquemas multipasos tienen su correspondencia con la aproximación de mayor orden en el tiempo. Detalles acerca de la forma de resolver la semidiscretización asi como la discretización completa se posponen para un próximo capítulo.

6.10. El método de los elementos finitos aplicado a las leyes de conservación

En esta sección presentamos una aplicación del método de los elementos finitos a un sistema de leyes de conservación, típica en la modelización matemática de problemas de mecánica de fluidos. Hasta aquí habíamos considerado casi exclusivamente el caso escalar y lineal. Las leyes de conservación tal como fueron presentadas al comienzo forman un sistema de ecuaciones no lineales. Su tratamiento no difiere a lo visto para el caso escalar y tampoco respecto a lo dicho sobre sistemas en el caso de la aplicación del método de los residuos ponderados con aproximaciones globales. Por lo tanto y por no abundar en detalles pasamos directamente a la aplicación del método de los elementos finitos a las ecuaciones de conservación.

Consideremos las leyes de conservación de la forma:

$$\frac{\partial \mathbf{U}}{\partial t} + \nabla \cdot \mathbf{F} = \mathbf{Q} \tag{6.110}$$

siendo $\mathbf{F}$ el vector de flujos. Supongamos por un momento que solo incluimos los flujos convectivos dentro de éste y que aplicamos las siguientes condiciones iniciales sobre el dominio Ω y sobre el contorno $\Gamma = \Gamma_0 \cup \Gamma_1$:

$$\begin{aligned} \mathbf{U}(\mathbf{x}, 0) &= \mathbf{U}_0(\mathbf{x}) \qquad t = 0 \quad \mathbf{x} \in \Omega \\ \mathbf{U}(\mathbf{x}, t) &= \mathbf{U}_1(\mathbf{x}) \qquad t \geq 0 \quad \mathbf{x} \in \Gamma_0 \\ \mathbf{F} \cdot \mathbf{n} \equiv \mathbf{F}_n &= \mathbf{g} \qquad t \geq 0 \quad \mathbf{x} \in \Gamma_1 \end{aligned} \tag{6.111}$$

Definiendo una forma débil con $\mathbf{W} = \mathbf{0}$ sobre Γ_0 llegamos a

$$\int_\Omega \mathbf{W}\frac{\partial \mathbf{U}}{\partial t} d\Omega + \int_\Omega \mathbf{W}(\nabla \cdot \mathbf{F}) d\Omega = \int_\Omega \mathbf{W}\mathbf{Q} \tag{6.112}$$

que seguido a una integración por partes conduce a

$$\int_\Omega \mathbf{W}\frac{\partial \mathbf{U}}{\partial t} d\Omega - \int_\Omega (\mathbf{F} \cdot \nabla)\mathbf{W} d\Omega + \int_\Gamma \mathbf{W}\mathbf{F} \cdot d\Gamma = \int_\Omega \mathbf{W}\mathbf{Q} \tag{6.113}$$

Interpolando la solución

$$\mathbf{U} = \sum_m \mathbf{U}_m(t)\mathbf{N}_m(\mathbf{x}) \tag{6.114}$$

y debido a que el flujo es generalmente una función no lineal de $\mathbf{U}$ es preferible una representación separada para los flujos

$$\mathbf{F} = \sum_m \mathbf{F}_m\mathbf{N}_m(\mathbf{x}) \tag{6.115}$$

Usando el método de Galerkin $(\mathbf{W} = \mathbf{N})$, la ecuación para el nodo j es:

$$\sum_m \frac{\mathrm{d}\mathbf{U}_m}{\mathrm{d}t}\int_{\Omega_j}\mathbf{N}_m\mathbf{N}_j d\Omega - \sum_m \mathbf{F}_m\int_{\Omega_j}\mathbf{N}_m\cdot\nabla\mathbf{N}_j d\Omega + \int_{\Gamma_1}\mathbf{g}\mathbf{N}_j d\Gamma = \int_{\Omega_j}\mathbf{N}_j\mathbf{Q} \tag{6.116}$$

donde Ω_j es el subdominio formado por todos los elementos que contienen al nodo j y la suma sobre m cubre todos los nodos de Ω_j.

La matriz de masa se define como:

$$\mathbf{M}_{mj} = \int_{\Omega_j}\mathbf{N}_m\mathbf{N}_j d\Omega \tag{6.117}$$

mientas que la de rigidez es

$$\mathbf{K}_{mj} = \int_{\Omega_j}\mathbf{N}_m\cdot\nabla\mathbf{N}_j d\Omega \tag{6.118}$$

que como se alcanza a ver es no simétrica.

Lo anterior conduce a:

$$\sum_m \mathbf{M}_{mj}\frac{\mathrm{d}\mathbf{U}_m}{\mathrm{d}t} - \sum_m \mathbf{F}_m\cdot\mathbf{K}_{mj} = \int_{\Omega_j}\mathbf{N}_j\mathbf{Q} - \int_{\Gamma_1}\mathbf{N}_j\mathbf{g}d\Gamma \tag{6.119}$$

Si tuviéramos difusión física entonces la discretización se lleva a cabo conforme a lo ya visto anteriormente y queda solo un comentario acerca del tratamiento de la matriz de masa. En diferencias finitas es usual una aproximación a $\frac{\mathrm{d}\mathbf{U}_j}{\mathrm{d}t}$ que conduce a una matriz diagonal mientras que en elementos finitos usando un método tipo Galerkin la misma no es diagonal *masa consistente*. En las aplicaciones habituales los sistemas a resolver son enormes en tamaño por lo que es casi imprescindible muchas veces recurrir a métodos de resolución explícitos. Estos requieren que la matriz asociada al miembro izquierdo de la ecuación algebraica sea diagonal por lo que en el caso del método de los elementos finitos lo que se suele hacer es *concentrar* los términos en la diagonal, técnica conocida con el nombre de *lumping*,

$$\mathbf{M}_{mj} = (\sum_m \mathbf{M}_{mj})\delta_{mj} \tag{6.120}$$

6.11. TP.VI- Trabajo Práctico

método de los elementos finitos en 1D

1. Dada una malla unidimensional formada por 5 nodos equiespaciados y numerados consecutivamente de izquierda a derecha, con el extremo izquierdo en $z = 0$ y el derecho en $x = 1$. Calcular la matriz

$$\begin{aligned} \mathbf{K} &= \{K_{ij}\} \\ K_{ij} &= \int_0^1 N_i N_j dx \\ i, j &= 1, \ldots, 5 \end{aligned} \tag{6.121}$$

usando

a.- funciones de prueba lineales a trozo definidas como

$$N_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad , \quad i, j \text{ nodos de la malla}$$

variando linealmente dentro de cada elemento.

b.- usando funciones del tipo $N_i = x^i$. Compare con la matriz obtenida en (a)

c.- Repita (a) pero usando la siguiente numeración:

x	i
0.00	1
0.25	5
0.50	3
0.75	4
1.00	2

(6.122)

Compare con (a)

2. Evalúe que tipo de funciones de forma utilizar para resolver las ecuaciones diferenciales que abajo se muestran si se adopta el método de Galerkin. Fundamentar la selección hecha.

$$\begin{aligned} &\text{(a)} \quad \frac{d^2\phi}{dx^2} + \phi = 0 \\ &\text{(b)} \quad \begin{cases} EI\frac{d^2\phi}{dx^2} + M = 0 \\ \frac{d^2M}{dx^2} + k\phi = -w \end{cases} \\ &\text{(c)} \quad EI\frac{d^4\phi}{dx^4} - k\phi = -w \end{aligned} \tag{6.123}$$

3. Resolver la ecuacón $\frac{d^2\phi}{dx^2} + \phi = 0$ con $\phi(x=0) = 1$ y $\phi(x=1) = 0$ usando el método de Galerkin con 4 elementos lineales de igual tamaño. Tome la ecuación del nodo central y evalúe el orden de precisión del esquema. *Ayuda: Expandir usando series de Taylor la ecuación del nodo*

4. (Uso del *FemCode* en 1D)

   Usando el programa *FemCode* resolver la ecuación

$$\begin{aligned} &\frac{d^2\phi}{dx^2} - \phi = 1 \qquad 0 \leq x \leq 1 \\ &\phi(x=0) = 0 \\ &\phi(x=1) = 1 \end{aligned} \tag{6.124}$$

   usando 4,8,16,32 y 64 elementos lineales.

   a.- Calcule la solución exacta de este problema.

   b.- Trazar la curva de error vs tamaño de elemento en forma logarítmica y estimar el orden de convergencia de la aproximación.

   c.- Compare lo obtenido con lo teórico.

5. (Uso del *FemCode* en 1D)

   Usando el *FemCode* resolver el problema unidimensional

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x}(\kappa\frac{\mathrm{d}\phi}{\mathrm{d}x}) + \phi = 1 \\ \phi(x=0) = 0 \\ \phi(x=1) = 1 \end{aligned} \tag{6.125}$$

   utilizando elementos cuadráticos y cúbicos. Hacer un estudio del orden de convergencia de cada uno de ellos.

6. (Mapeo)

   Un elemento de 4 nodos se muestra en la figura siguiente. Construya un mapeo isoparamétrico entre este y un elemento bilineal cuadrado sobre $-1 \leq \xi, \eta \leq 1$.

7. (Mapeo)

   Encuentre el mapeo isoparamétrico a un elemento serendípido de 8 nodos del elemento de la figura, con el borde curvo descripto por la función $y^4 = 5x$.

8. (Integración numérica)

   En el cómputo por elementos finitos es muy usual evaluar integrales del tipo
   $\int_{\Omega^e} \frac{\partial N_i}{\partial x_k}\frac{\partial N_j}{\partial x_l} dx_k dx_l$
   y $\int_{\Omega^e} N_i N_j dx_k dx_l$.
   Determine el número de puntos de Gauss que se necesitan tomar si requerimos que ambas integrales se evalúen exactamente usando elementos :

(a) triangulares

(b) cuadrangulares de 4 nodos

(c) cuadrangulares de 9 nodos

9. (Uso del *FemCode* en 2D)

   Resolver para una geometría plana anular con $r_i = 0.1$ y $r_e = 1$ un problema de conducción térmica con conductividad constante $\kappa = 1$ imponiendo la temperatura sobre la pared circular interior a un valor nulo. Sobre la pared circular exterior la misma está aislada salvo en la porción $0 < \theta < \pi/2$ donde el flujo térmico alcanza un valor unitario. Determinar el campo de temperaturas resultante $T = T(r, theta)$ y mostrar las isotermas.

10. (Upwind)

    Resolver la ecuación de transporte

$$\begin{aligned} \mathbf{v} \cdot \nabla T &= \nabla \cdot (\kappa \nabla T) \\ T(x = 0, y) &= 0 \\ T(x = 1, y) &= 1 \\ \kappa &= 10^{-3} \\ (x, y) &\in [0, 1] \times [0, 1] \end{aligned} \tag{6.126}$$

    usando el programa *FemCode* para una malla bidimensional compuesta por 10×10 elementos tomando:

    a.- $\mathbf{v} = [10^{-3}, 0]$ sin upwind,

    b.- $\mathbf{v} = [1, 0]$ sin upwind,

    c.- $\mathbf{v} = [1, 0]$ con upwind,

    Sacar algunas conclusiones de acuerdo a los resultados obtenidos.

11. Repita el ejemplo anterior usando una velocidad no alineada con la malla. Tome lo mismo que antes con una velocidad orientada a 30 grados respecto al eje horizontal.

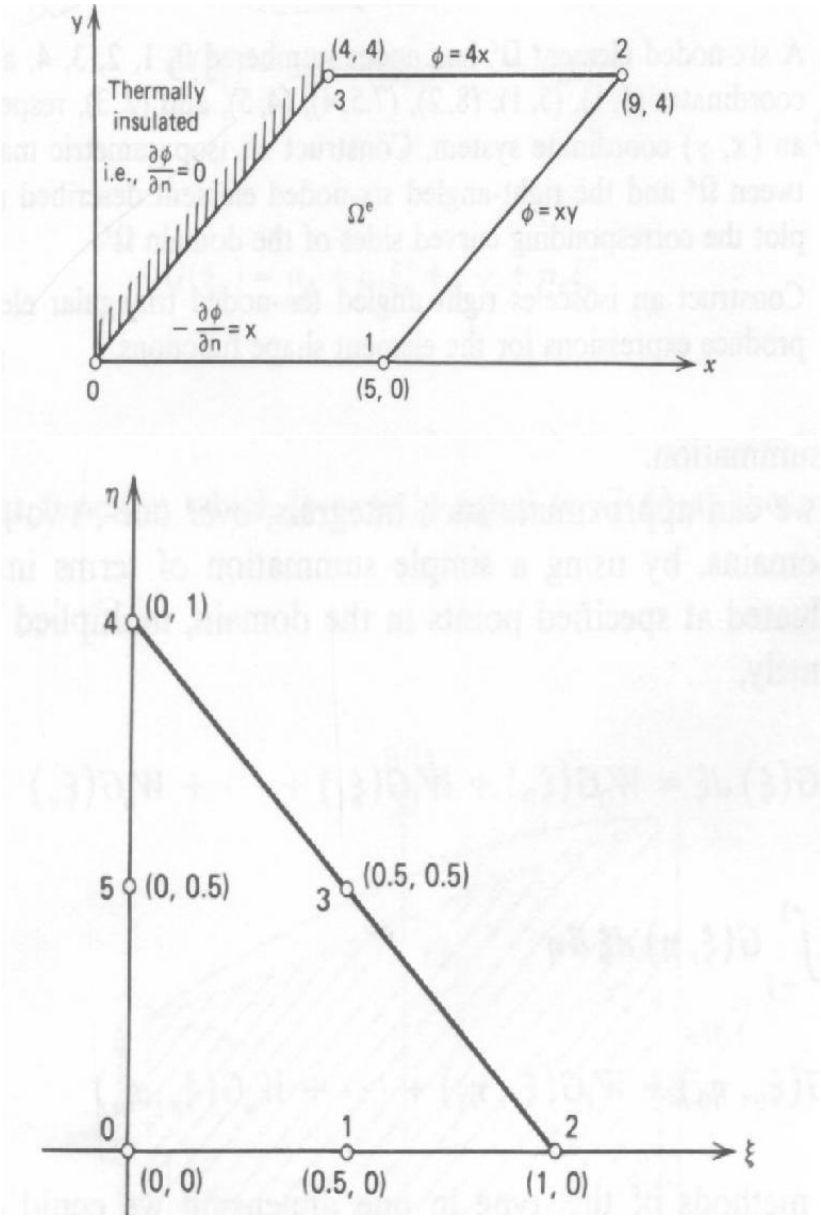


Figura 6.11: Mapeo isoparamétrico

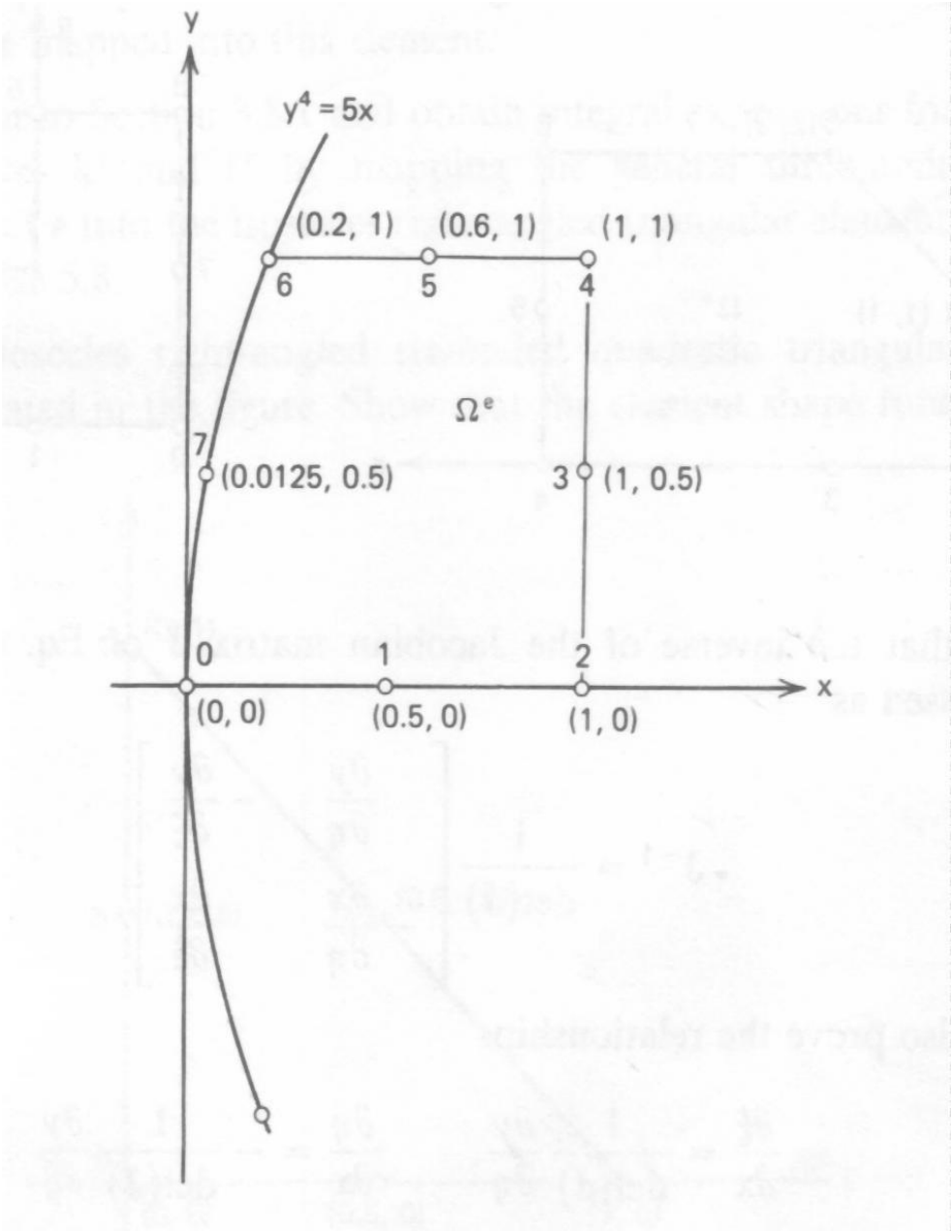


Figura 6.12: Mapeo isoparamétrico

Capítulo 7

Método de los volúmenes finitos

7.1. Introducción

Continuando con la presentación de los diferentes métodos que surgen a partir del método de los residuos ponderados ahora trataremos el caso del método de los volúmenes finitos . A modo de resúmen de lo visto antes la aplicación del método de los residuos ponderados a las ecuaciones de conservación fue escrita como:

$$\int_{\Omega} \mathbf{W} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \int_{\Omega} \mathbf{W} \nabla \cdot \mathbf{F} d\Omega = \int_{\Omega} \mathbf{W}\mathbf{Q} \tag{7.1}$$

Los métodos de colocación, tanto el caso puntual como el de subdominios, usan la ecuación residual sin integración parcial sobre la función de peso quedando el problema en forma diferencial. Si a cada nodo j del dominio le asignamos un subdominio Ω_j y si tomamos una función de peso definida como:

$$\begin{aligned} W_j(\mathbf{x}) &= 0 \qquad \mathbf{x} \notin \Omega_j \\ W_j(\mathbf{x}) &= 1 \qquad \mathbf{x} \in \Omega_j \end{aligned} \tag{7.2}$$

que aplicada a (7.1) nos da:

$$\int_{\Omega_j} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \int_{\Omega_j} \nabla \cdot \mathbf{F} d\Omega = \int_{\Omega_j} \mathbf{Q} d\Omega \tag{7.3}$$

que equivale a la misma ley de conservación aplicada a cada subdominio. La idea detrás del método de los volúmenes finitos es discretizar cada integral siendo esta la principal diferencia del método respecto a muchos otros que llevan el problema a una formulación diferencial. Para poder arribar a la forma más conveniente del método de los volúmenes finitos debemos aplicar a (7.3) el teorema de Gauss-Green o de la divergencia y transformar la integral de volúmen en otra de superficie,

$$\int_{\Omega_j} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \oint_{\Gamma_j} \mathbf{F} \cdot d\mathbf{S} = \int_{\Omega_j} \mathbf{Q} d\Omega \tag{7.4}$$

(7.4) es la ecuación básica del método de los volúmenes finitos que tiene como una de sus principales ventajas la de trabajar con el término de los flujos sobre el contorno del dominio, con lo cual si el costo computacional es dominado por esta operación la reducción del mismo puede ser notable. A partir de (7.4)

se necesita discretizar las integrales de alguna forma y lograr el sistema discreto final a resolver. Ya que el método es planteado sobre la forma integral de las leyes de conservación es de notar que al satisfacer las mismas sobre cada subdominio implica satisfacerlas sobre el dominio global. Por ejemplo, si planteamos las leyes de conservación sobre los 3 dominios de la figura 7.1 llegamos a:

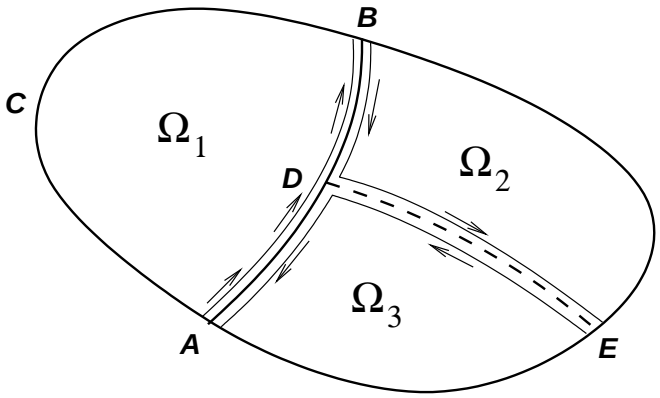


Figura 7.1: Leyes de conservación para subdominios del dominio Ω

$$\begin{aligned}
\int_{\Omega_1} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \oint_{ABCA} \mathbf{F} \cdot d\mathbf{S} &= \int_{\Omega_1} \mathbf{Q} d\Omega \\
\int_{\Omega_2} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \oint_{DEBD} \mathbf{F} \cdot d\mathbf{S} &= \int_{\Omega_2} \mathbf{Q} d\Omega \\
\int_{\Omega_3} \frac{\partial \mathbf{U}}{\partial t} d\Omega + \oint_{AEDA} \mathbf{F} \cdot d\mathbf{S} &= \int_{\Omega_3} \mathbf{Q} d\Omega
\end{aligned} \tag{7.5}$$

que sumados producen el mismo resultado que si lo hubiéramos aplicado a todo el dominio. Esto se explica ya que dos subdominios vecinos por una cara o arista comparten los términos de flujo, con la diferencia que debido a la orientación de la normal exterior a cada uno, los mismos se deben balancear,

$$\int_{ED} \mathbf{F} \cdot d\mathbf{S} = -\int_{DE} \mathbf{F} \cdot d\mathbf{S} \tag{7.6}$$

Esta propiedad debe ser satisfecha si se requiere que el esquema sea *conservativo*, caso contrario pueden aparecer contribuciones internas produciendo esquemas no conservativos.

Como para ilustrar la diferencia entre un esquema conservativo y otro que no lo es planteamos el caso de una ley de conservación en un dominio unidimensional que solo cuenta con un término de flujo convectivo,

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = q \tag{7.7}$$

En la figura 7.2 se muestra la discretización espacial adoptada.

$$\frac{\partial u_i}{\partial t} + \frac{f_{i+1/2} - f_{i-1/2}}{\Delta x} = q_i \tag{7.8}$$

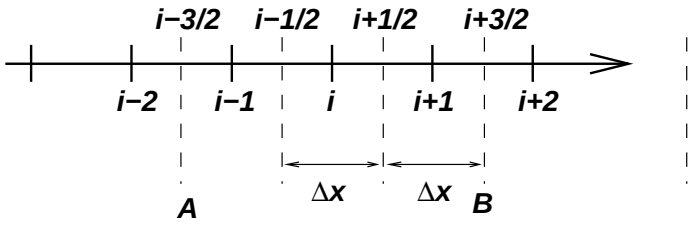


Figura 7.2: Grilla unidimensional

Este esquema sencillo, similar a aquel obtenido por diferencias finitas equivale a haber tomado como volúmen la longitud de cada segmento que representa cada celda, con el valor de la variable en cada nodo de la grilla, la fuente evaluada también en cada nodo de la misma y los flujos evaluados en los puntos medios entre los nodos. Lo anterior tambien es equivalente a haber tomado los flujos en los nodos y tanto la variable como la fuente en el centro de cada celda.

Si aplicamos (7.8) para los nodos $i+1$ e $i-1$ y sumamos llegamos a:

$$\frac{\partial}{\partial t}\frac{(u_i+u_{i+1}+u_{i-1})}{3}+\frac{f_{i+3/2}-f_{i-3/2}}{3\Delta x}=\frac{1}{3}(q_{i+1}+q_i+q_{i-1}) \tag{7.9}$$

Esto es posible ya que los flujos en los puntos intermedios se van cancelando por una propiedad denominada *telescópica*. Este esquema es *conservativo*. Veamos para este mismo problema como llegamos a un esquema *no conservativo*. Supongamos que $f=f(u)$ como sucede en el flujo convectivo de la ecuación del momento lineal, entonces:

$$\frac{\partial f}{\partial x}=(\frac{\partial f}{\partial u})\frac{\partial u}{\partial x}=a(u)\frac{\partial u}{\partial x} \tag{7.10}$$

donde $a(u)$ es conocido como el jacobiano del flujo convectivo. Pensando nuevamente en la ecuación de momento podemos tomar el caso de $f=u^2/2$ donde en este caso $a(u)=u$, entonces la forma matemáticamente equivalente a (7.7) expresada en términos de este jacobiano es

$$\frac{\partial u}{\partial t}+a(u)\frac{\partial u}{\partial x}=q \tag{7.11}$$

Discretizando la (7.11)

$$\begin{aligned}\frac{\partial u_i}{\partial t}+a_i\frac{u_{i+1/2}-u_{i-1/2}}{\Delta x}&=q_i\\ a_i&=\frac{a_{i+1/2}+a_{i-1/2}}{2}\end{aligned} \tag{7.12}$$

Obteniendo similares expresiones para los nodos $i-1$ e $i+1$ y sumando llegamos a:

$$\begin{aligned}&\frac{\partial}{\partial t}\frac{(u_i+u_{i+1}+u_{i-1})}{3}+(a_{i+3/2}+a_{i-3/2})\frac{u_{i+3/2}-u_{i-3/2}}{6\Delta x}-\frac{1}{3}(q_{i+1}+q_i+q_{i-1})=\\ &-(a_{i+1/2}-a_{i-3/2})\frac{u_{i+3/2}-u_{i-1/2}}{6\Delta x}+(a_{i+3/2}-a_{i-1/2})\frac{u_{i+1/2}-u_{i-3/2}}{6\Delta x}\end{aligned} \tag{7.13}$$

Si hubiéramos discretizado (7.11) sobre un solo subdominio AB en lugar de agregar las contribuciones de 3 subdominios equivalentes hubiéramos obtenido solo el miembro izquierdo de (7.13) , con lo cual el miembro derecho equivale a una fuente interna numérica que no representa la física del problema. Haciendo un análisis numérico sobre el esquema no conservativo mediante expansión en series de Taylor se puede ver que la importancia de estas fuentes internas es similar al error de truncamiento con lo cual en principio parecería ser despreciable. No obstante, lo anterior no es válido en el caso de existir fuertes gradientes en la solución tal el caso de flujo transónico con ondas de choque, bastante citado en la bibliografía.

La expresión formal de una discretización conservativa a (7.7) puede escribirse como:

$$\frac{\partial u_i}{\partial t} + \frac{f^*_{i+1/2} - f^*_{i-1/2}}{\Delta x} = q_i \tag{7.14}$$

donde f^* es llamado el *flujo numérico* y es una función de los valores de u en los puntos vecinos,

$$f^*_{i+1/2} = f^*(u_{i+k}, \ldots, u_{i-k+1}) \tag{7.15}$$

La consistencia de (7.14) requiere que cuando la solución u es constante el flujo numérico sea igual al del continuo,

$$f^*(u, \ldots, u) = f(u) \tag{7.16}$$

Una consecuencia interesante de lo anterior la establece el siguiente teorema:

Teorema de Lax y Wendroff (1960) *Si la solución u_i de (7.14) converge acotadamente en casi todo punto a alguna función $u(x,t)$ cuando $\Delta x, \Delta t \to 0$, luego $u(x,t)$ es una solución débil de (7.7) .*

7.2. Formulación del método de los volúmenes finitos

Hemos visto que las leyes de conservación escritas en forma integral y aplicadas a un volúmen discreto Ω_j pueden escribirse como en (7.4) . La forma discreta de esta se escribe como:

$$\frac{\partial}{\partial t}(\mathbf{U}_j \Omega_j) + \sum_{\text{lados}} (\mathbf{F} \cdot \mathbf{S}) = \mathbf{Q}_j \Omega_j \tag{7.17}$$

donde la suma de los flujos se refiere a todos los contornos externos de la celda de control Ω_j. La figura 7.3 muestra en su parte superior un ejemplo de grilla aplicable a volúmenes finitos. Tomando la celda 1 identificada por los índice (i,j) entonces $\mathbf{U}_j = U_{ij}, \Omega_j =$ area $(ABCD)$ y los términos de flujo se obtienen como suma sobre los 4 lados AB, BC, CD, DA. Asimismo en la figura inferior Ω_j está representada por el area sombreada formada por los triángulos que tienen como nodo común al nodo j y los flujos se obtienen sumando sobre los lados $12, 23, 34, 45, 56, 61$. Esta es la formulación general del método de los volúmenes finitos , y el usuario tiene que definir para un volúmen Ω_j seleccionado cómo estimará el volúmen y las caras de la celda y cómo aproximará los flujos sobre estas caras. Esto en diferencias finitas equivale a elegir la aproximación en diferencias para las derivadas.

Para definir una formulación conservativa se requiere:

1.- $\sum_j \Omega_j = \Omega$,

2.- $\cap\Omega_j \neq \emptyset$, pueden solaparse pero solo si los contornos internos que surgen del solapamiento son comunes entre dos celdas,

3.- los flujos en las superficies de las celdas deben calcularse con independencia de a cual celda le corresponde.

(2) significa que todos los contornos de las celdas deben pertenecer a lo sumo a dos celdas y solo aquellos que están en el contorno exterior del dominio pueden no satisfacer este requisito.

La condición (3) garantiza la conservatividad.

A continuación y tomando como referencia la figura 7.3 mencionaremos algunas diferencias entre el método de los volúmenes finitos con el método de las diferencias finitas y el método de los elementos finitos .

1.- Las coordenadas del nodo j que es la precisa ubicación de la variable $\mathbf{U}$ dentro de la celda Ω_j no aparece explicitamente. Consecuentemente $\mathbf{U}_j$ no está asociada con ningún punto fijo del dominio y puede considerarse como un *valor promedio* de la variable de flujo $\mathbf{U}$ sobre la celda. Esta interpretación surge inmediatamente inspeccionando la figura superior de la gráfica 7.3.

2.- Las coordenadas de la malla aparecen solamente para definir el volúmen de la celda y las areas de las caras, por ejemplo en la misma figura superior las coordenadas A, B, C, D son suficientes.

3.- En problemas estacionarios sin fuentes el único término que permanece es el de la suma de los flujos sobre las caras con lo cual el método puede programarse para barrer estas caras e ir descargando los flujos sobre las dos celdas al mismo tiempo considerando la diferencia en signo.

4.- El método de los volúmenes finitos permite una fácil introducción de las condiciones de contorno, especialmente aquellas que vienen expresadas en término de los flujos que pueden ser directamente impuestos en el respectivo término.

7.2.1. Mallas y volúmenes de control

En cuanto a las mallas que puede manipular el método de los volúmenes finitos este cuenta con la misma flexibilidad que el método de los elementos finitos pero restringido a elementos con lados rectas o caras planas. Entre las mallas posibles podemos hacer una división entre:

1.- *mallas estructuradas*, son aquellas en las cuales cualquier nodo puede ser ubicable en término de dos índices (i, j) en 2D o tres índices (i, j, k) en 3D. Estas mallas son muy comúnmente denominadas mallas de diferencias finitas. (Ver figuras 7.3, 7.4.)

2.- *mallas no estructuradas*, son aquellas comúnmente empleadas por el método de los elementos finitos y donde no es posible encontrar una expresión de 2 o 3 índices que permita ubicar un nodo. (Ver figuras 7.5.)

Una vez que la malla se ha construido el usuario debe decidir entre 2 opciones propias del método de los volúmenes finitos :

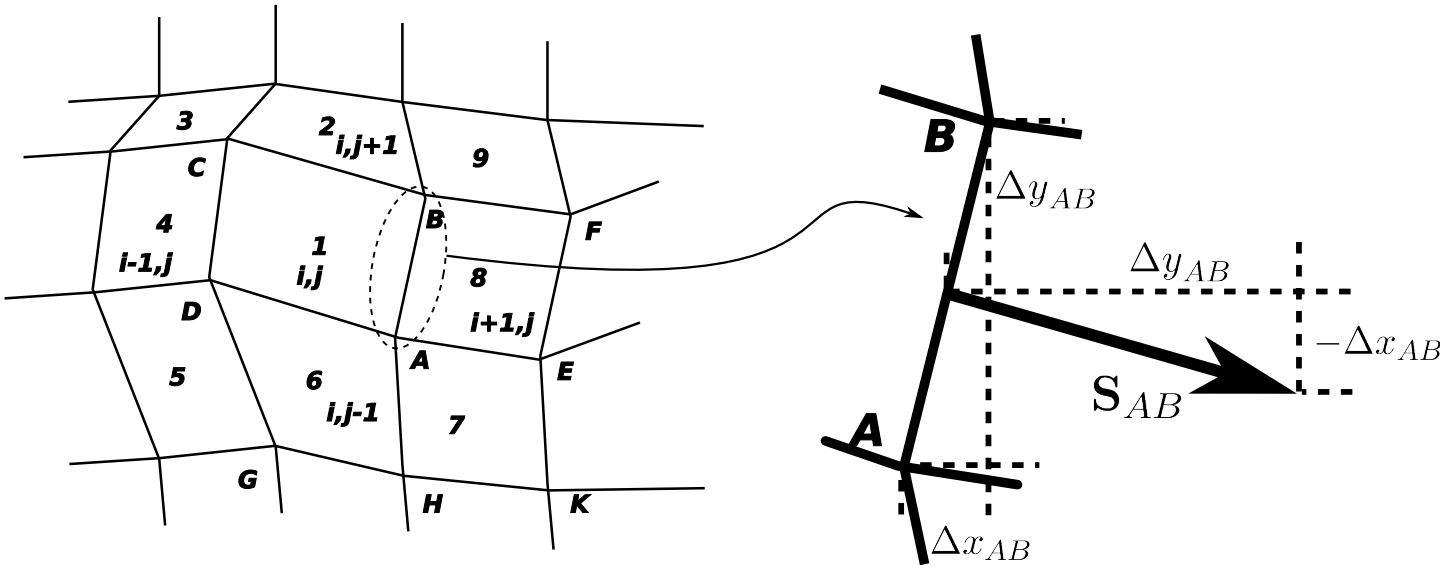


Figura 7.3: Grillas bidimensionales de volúmenes finitos. Malla estructurada. Formulación centrada en las celdas.

1.- *centrado en las celdas*, las variables de flujo representan un promedio de los valores de la misma sobre la celda y están asociadas a la celda. (ver figura 7.3 (a) y (c))

2.- *centrado en los vértices*, en este caso las variables de flujo representan los valores en los vértices o los puntos de la malla. (ver figura 7.3 (b) y (d))

7.3. El método de los volúmenes finitos en 2D

Cconsideremos la ecuación (7.4) aplicada al volúmen de control ABCD de la figura 7.3 (a),

$$\frac{\partial}{\partial t}\int_{\Omega_j}\mathbf{U}d\Omega + \oint_{ABCD}(\mathbf{f}dy - \mathbf{g}dx) = \int_{\Omega_j}\mathbf{Q}d\Omega \tag{7.18}$$

donde la derivada temporal pudo ser extraida de la integral siempre que consideremos el caso de dominio y malla fija. f y g representan las dos componentes cartesianas del vector flujo $\mathbf{F}$. Al discretizar estas integrales llegamos a la siguiente expresión:

$$\begin{aligned}\frac{\partial}{\partial t}(\mathbf{U}\Omega)_{ij} + \sum_{ABCD}[\mathbf{f}_{AB}(y_B - y_A) - \mathbf{g}_{AB}(x_B - x_A)] &= (\mathbf{Q}\Omega)_{ij}\\ \Omega_{ABCD} = \tfrac{1}{2}|\mathbf{x}_{AC} \wedge \mathbf{x}_{BD}|&\end{aligned} \tag{7.19}$$

7.3.1. Evaluación de los flujos convectivos

La evaluación de los flujos f_{AB} y g_{AB} depende del esquema elegido tambien como de la localización de las variables de flujo con respecto a la malla. Como es habitual en métodos numéricos en mecánica

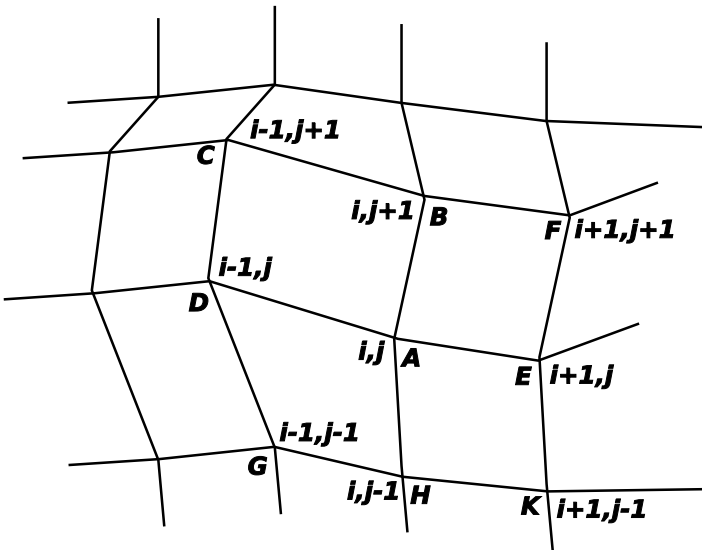


Figura 7.4: Grillas bidimensionales de volúmenes finitos. Malla estructurada. Formulación centrada en los vértices.

de fluidos podemos distinguir entre *esquemas centrados* y aquellos que tienen *upwinding*. Los primeros requieren estimaciones locales del flujo mientras que los últimos determinan los flujos acorde con la dirección de propagación de las componentes ondulatorias. Ahora exploraremos las distintas posibilidades.

Esquema central - centrado en la celda

1.- Promedio de flujos

$$\begin{aligned} f_{AB} &= \tfrac{1}{2}(f_{ij} + f_{i+1,j}) \\ f_{ij} &= f(U_{ij}) \end{aligned} \tag{7.20}$$

2.- El flujo de los promedios

$$f_{AB} = f\Big(\frac{U_{ij} + U_{i+1,j}}{2}\Big) \tag{7.21}$$

3.- Otro promedio de flujos

$$f_{AB} = \tfrac{1}{2}(f_A + f_B) \tag{7.22}$$

con

$$\begin{aligned} f_A &= f(U_A) \\ U_A &= \frac{1}{4}(U_{ij} + U_{i+1,j} + U_{i+1,j-1} + U_{i,j-1}) \end{aligned} \tag{7.23}$$

$$f_A = \frac{1}{4}(f_{ij} + f_{i+1,j} + f_{i+1,j-1} + f_{i,j-1}) \tag{7.24}$$

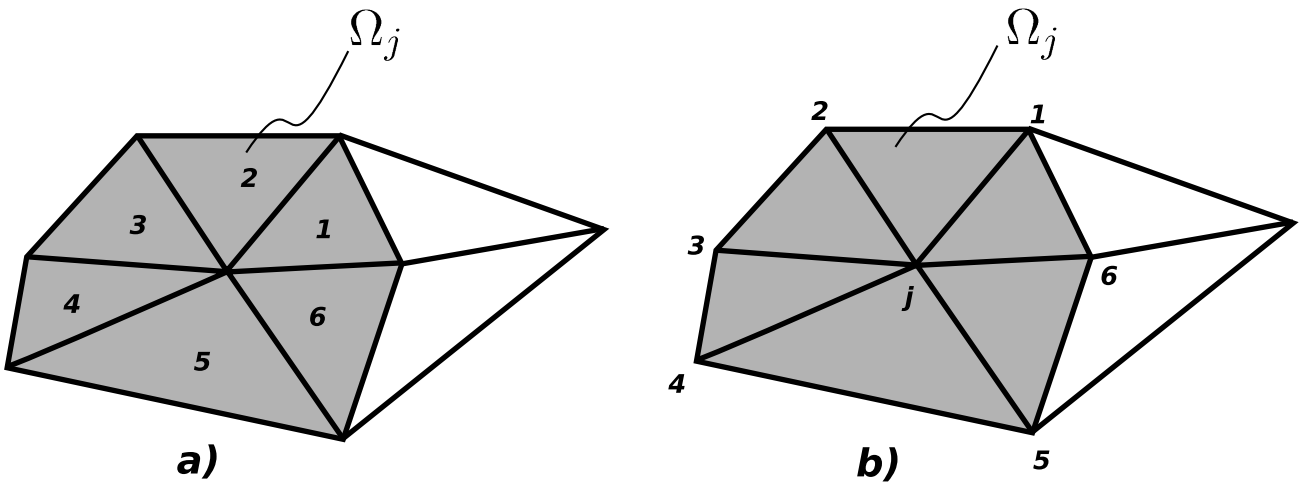


Figura 7.5: Grillas bidimensionales de volúmenes finitos. Malla no estructurada. a) Formulación centrada en las celdas. b) Formulación centrada en los vértices.

Tanto (7.21) como (7.22) son aproximaciones directas al flujo f_{AB}. En especial esta última equivale a integrar el flujo sobre cada cara usando la regla del trapecio $\int_{AB} f dy = (f_A + f_B)(y_B - y_A)/2$. Sumando sobre todas las caras llegamos a:

$$\begin{aligned} &\oint_{ABCD} \mathbf{F} \cdot d\mathbf{S} = \\ &= \tfrac{1}{2}[(\mathbf{f}_A - \mathbf{f}_C)\Delta y_{DB} + (\mathbf{f}_B - \mathbf{f}_D)\Delta y_{AC} - (\mathbf{g}_A - \mathbf{g}_C)\Delta x_{DB} - (\mathbf{g}_B - \mathbf{g}_D)\Delta x_{AC}] \end{aligned} \tag{7.25}$$

lo cual equivale al método de Galerkin sobre elementos triangulares o cuadrangulares.

Para el método de los volúmenes finitos centrado en las celdas usando esquemas upwind el flujo convectivo es evaluado como una función de la dirección de propagación de la asociada velocidad convectiva. Si pensamos en el caso escalar bidimensional esto último puede expresarse como:

$$\begin{aligned} \mathbf{A}(U) = \frac{\partial \mathbf{F}}{\partial U} = a(U)\mathbf{i} + b(U)\mathbf{j} \\ a(U) = \frac{\partial f}{\partial U} \qquad b(U) = \frac{\partial g}{\partial U} \end{aligned} \tag{7.26}$$

Considerando la figura 7.3 en su parte superior (centrado en las celdas), se podría definir:

$$\begin{aligned} (\mathbf{F} \cdot \mathbf{S})_{AB} = (\mathbf{F} \cdot \mathbf{S})_{ij} \qquad &\text{si } (\mathbf{A} \cdot \mathbf{S})_{AB} > 0 \\ (\mathbf{F} \cdot \mathbf{S})_{AB} = (\mathbf{F} \cdot \mathbf{S})_{i+1,j} \qquad &\text{si } (\mathbf{A} \cdot \mathbf{S})_{AB} < 0 \end{aligned} \tag{7.27}$$

mientras que para el caso de centrado en los vértices (figura (b)) tenemos:

$$\begin{aligned} (\mathbf{F} \cdot \mathbf{S})_{AB} = (\mathbf{F} \cdot \mathbf{S})_{CD} \qquad &\text{si } (\mathbf{A} \cdot \mathbf{S})_{AB} > 0 \\ (\mathbf{F} \cdot \mathbf{S})_{AB} = (\mathbf{F} \cdot \mathbf{S})_{EF} \qquad &\text{si } (\mathbf{A} \cdot \mathbf{S})_{AB} < 0 \end{aligned} \tag{7.28}$$

La desventaja de esta técnica está en que se aumenta bastante el soporte de información en forma innecesaria ya que están involucrados los vértices $(i-2, j)$ e $(i, j-2)$.

Esquema central sobre una malla cartesiana

Si bien el método de los volúmenes finitos es tan general como el método de los elementos finitos y puede ser aplicado a mallas completamente no estructuradas en esta sección trataremos el caso de una grilla uniforme y demostraremos que el mismo es completamente equivalente a un esquema obtenido por el método de las diferencias finitas . Si miramos la malla superior de la figura 7.3 y si por un momento la rectificamos un poco de forma de alinear los lados con los ejes cartesianos y planteamos las integrales de contorno sobre la curva ABCD encontramos:

$$
\begin{aligned}
&\text{LADO AB}\\
&\Delta y_{AB} = y_{i+1/2,j+1/2} - y_{i+1/2,j-1/2} = \Delta y\\
&\Delta x_{AB} = x_{i+1/2,j+1/2} - x_{i+1/2,j-1/2} = 0\\
&\text{LADO BC}\\
&\Delta y_{BC} = y_{i-1/2,j+1/2} - y_{i+1/2,j+1/2} = 0\\
&\Delta x_{BC} = x_{i-1/2,j+1/2} - x_{i+1/2,j+1/2} = -\Delta x\\
&\text{LADO CD}\\
&\Delta y_{CD} = y_{i-1/2,j-1/2} - y_{i-1/2,j+1/2} = -\Delta y\\
&\Delta x_{CD} = x_{i-1/2,j-1/2} - x_{i-1/2,j+1/2} = 0\\
&\text{LADO DA}\\
&\Delta y_{DA} = y_{i+1/2,j-1/2} - y_{i-1/2,j-1/2} = 0\\
&\Delta x_{DA} = x_{i+1/2,j-1/2} - x_{i-1/2,j-1/2} = \Delta x
\end{aligned}
\tag{7.29}
$$

$$
\begin{aligned}
\Omega_{ij} &= \Delta x \Delta y\\
f_{AB} &= f_{i+1/2,j}\\
f_{BC} &= f_{i,j+1/2}\\
f_{CD} &= f_{i-1/2,j}\\
f_{DA} &= f_{i,j-1/2}
\end{aligned}
$$

Reemplazando lo anterior en (7.19) llegamos a

$$
\Delta x \Delta y \frac{\partial U_{ij}}{\partial t} + (f_{i+1/2,j} - f_{i-1/2,j})\Delta y + (g_{i,j+1/2} - g_{i,j-1/2})\Delta x = Q_{ij}\Delta x \Delta y
\tag{7.30}
$$

$$
\frac{\partial U_{ij}}{\partial t} + \frac{(f_{i+1/2,j} - f_{i-1/2,j})}{\Delta x} + \frac{(g_{i,j+1/2} - g_{i,j-1/2})}{\Delta y} = Q_{ij}
$$

Ahora tenemos que especificar como definir los flujos en los centros de los lados $f_{i\pm 1/2, j\pm 1/2}$

Caso (a) Aplicando (7.20) a un esquema centrado en las celdas

$$\frac{\partial U_{ij}}{\partial t} + \frac{(f_{i+1,j} - f_{i-1,j})}{2\Delta x} + \frac{(g_{i,j+1} - g_{i,j-1})}{2\Delta y} = Q_{ij} \tag{7.31}$$

Caso (b) Aplicando (7.24) a un esquema centrado en las celdas

$$\begin{aligned}\frac{\partial U_{ij}}{\partial t} + \frac{1}{4}\Big[2\frac{(f_{i+1,j} - f_{i-1,j})}{2\Delta x} + \frac{(f_{i+1,j+1} - f_{i-1,j+1})}{2\Delta x} + \frac{(f_{i+1,j-1} - f_{i-1,j-1})}{2\Delta x}\Big] + \\ \frac{1}{4}\Big[2\frac{(g_{i,j+1} - g_{i,j-1})}{2\Delta y} + \frac{(g_{i+1,j+1} - g_{i+1,j-1})}{2\Delta y} + \frac{(g_{i-1,j+1} - g_{i-1,j-1})}{2\Delta y}\Big] = Q_{ij}\end{aligned} \tag{7.32}$$

Comentarios:

1.- Se puede demostrar que estos esquemas son de segundo orden de precisión,

2.- no dependen de los valores de los flujos f_{ij}, g_{ij},

3.- (7.31) puede generar oscilaciones debido al desacoplamiento de las ecuaciones correspondientes a $(i + j)$ par de aquellas donde $(i + j)$ es impar. (7.32) no contiene este inconveniente.

Idéntico tratamiento puede hacerse con el caso de formulaciones centradas en los vértices pero esto es dejado como ejercicio práctico.

Esquemas upwind en mallas cartesianas

Supongamos la ecuación de advección pura lineal en 2D,

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} + b\frac{\partial U}{\partial y} = 0 \qquad a, b > 0 \tag{7.33}$$

sobre una malla similar a la representada en la parte superior de la figura 7.3 pero rectificada de forma de lograr alinearla con los ejes coordenados cartesianos. Si $f = aU$ y $g = bU$ entonces,

$$\begin{aligned}(\mathbf{F}\cdot\mathbf{S})_{AB} &= f_{ij}\Delta y = aU_{ij}\Delta y \\ (\mathbf{F}\cdot\mathbf{S})_{CD} &= -f_{i-1,j}\Delta y = -aU_{i-1,j}\Delta y \\ (\mathbf{F}\cdot\mathbf{S})_{BC} &= g_{ij}\Delta x = bU_{ij}\Delta x \\ (\mathbf{F}\cdot\mathbf{S})_{DA} &= -g_{ij,j-1}\Delta x = -bU_{i,j-1}\Delta x\end{aligned} \tag{7.34}$$

que reemplazada en la (7.19) aplicada a este problema nos da:

$$\frac{\partial U_{ij}}{\partial t} + \frac{a}{\Delta x}(U_{ij} - U_{i-1,j}) + \frac{b}{\Delta y}(U_{ij} - U_{i,j-1}) = 0 \tag{7.35}$$

Mallas no uniformes

Garantizar la precisión de segundo orden de algunos esquemas cuando las mallas involucradas son no uniformes no es tarea fácil. No entraremos en detalles aquí por ser un tema demasiado específico pero hacer promedios sobre mallas no uniformes es muy dependiente del problema, de la discretización usada, etc. Por el momento nos conformamos con lo ya visto y debemos cuidar que las mallas no tengan fuertes gradientes, o sea que sean suaves. Esto hace que la generación de mallas para el método de los volúmenes finitos sea bastante restrictiva.

7.3.2. Fórmulas generales de integración

A diferencia de los problemas convectivos donde los flujos son funciones homogéneas de primer orden en las variables de estado, los problemas con difusión contienen en general flujos difusivos que son dependientes tanto de la variable a resolver como de su gradiente. En esto casos se hace necesario promediar derivadas de las variables en lugar de las variables en si misma. Un ejemplo típico de esto son las ecuaciones de Navier-Stokes que contienen $\mathbf{F}_d = \mathbf{F}_d(\mathbf{U}, \nabla\mathbf{U})$. Una forma bastante general de resolver esto es apelando al teorema de la divergencia y a la integración por partes. Supongamos que tenemos integrales de flujo y la queremos aproximar por el valor promedio del integrando multiplicado por la medida del dominio de integración. Entonces, aplicando el teorema de la divergencia,

$$\begin{aligned}
\int_\Omega \nabla U d\Omega = \int_\Omega (\frac{\partial U}{\partial x}\mathbf{i} + \frac{\partial U}{\partial y}\mathbf{j}) d\Omega = \oint_S U d\mathbf{S} \\
\Big(\overline{\frac{\partial U}{\partial x}}\Big)_\Omega = \frac{1}{\Omega}\int_\Omega \frac{\partial U}{\partial x} d\Omega = \frac{1}{\Omega}\oint_S U\mathbf{i}\cdot d\mathbf{S} \\
\Big(\overline{\frac{\partial U}{\partial y}}\Big)_\Omega = \frac{1}{\Omega}\int_\Omega \frac{\partial U}{\partial y} d\Omega = \frac{1}{\Omega}\oint_S U\mathbf{j}\cdot d\mathbf{S} \\
d\mathbf{S} = dy\mathbf{i} - dx\mathbf{j} \\
\Big(\overline{\frac{\partial U}{\partial x}}\Big)_\Omega = \frac{1}{\Omega}\oint_S U dy = -\frac{1}{\Omega}\oint_S y dU \\
\Big(\overline{\frac{\partial U}{\partial y}}\Big)_\Omega = -\frac{1}{\Omega}\oint_S U dx = \frac{1}{\Omega}\oint_S x dU
\end{aligned} \tag{7.36}$$

Aplicando la regla de integración del trapecio llegamos a:

$$\begin{aligned}
\Big(\overline{\frac{\partial U}{\partial x}}\Big)_\Omega = \frac{1}{\Omega}\int_\Omega \frac{\partial U}{\partial x} d\Omega &= \frac{1}{2\Omega}\sum_l (U_l + U_{l+1})(y_{l+1} - y_l) \\
&= \frac{-1}{2\Omega}\sum_l (y_l + y_{l+1})(U_{l+1} - U_l) \\
&= \frac{1}{2\Omega}\sum_l U_l (y_{l+1} - y_{l-1}) \\
&= \frac{-1}{2\Omega}\sum_l y_l (U_{l+1} - U_{l-1})
\end{aligned} \tag{7.37}$$

$$\begin{aligned}\left(\overline{\frac{\partial U}{\partial y}}\right)_\Omega = \frac{1}{\Omega}\int_\Omega \frac{\partial U}{\partial y}d\Omega &= \frac{-1}{2\Omega}\sum_l (U_l + U_{l+1})(x_{l+1} - x_l) \\ &= \frac{1}{2\Omega}\sum_l (x_l + x_{l+1})(U_{l+1} - U_l) \\ &= \frac{-1}{2\Omega}\sum_l U_l(x_{l+1} - x_{l-1}) \\ &= \frac{1}{2\Omega}\sum_l x_l(U_{l+1} - U_{l-1})\end{aligned} \tag{7.38}$$

donde la suma se extiende a todos los vértices desde 1 hasta 6 que rodean al nodo j en la figura 7.3 con $U_0 = U_6$ y $U_7 = U_1$.

La misma expresión se puede aplicar para el cálculo del area de las celdas,

$$\begin{aligned}\Omega = \frac{1}{2}\sum_l (x_l + x_{l+1})(y_{l+1} - y_l) \\ = \frac{-1}{2}\sum_l (y_l + y_{l+1})(x_{l+1} - x_l) \\ = \frac{1}{2}\sum_l x_l(y_{l+1} - y_{l-1}) \\ = \frac{-1}{2}\sum_l y_l(x_{l+1} - x_{l-1})\end{aligned} \tag{7.39}$$

Si las celdas fueras cuadrangulares se puede hallar una forma interesante y particular de la anterior tomando la tercera de las ecuaciones (7.37,7.38,7.39) y agrupando por valores en los nodos opuestos. Esto queda para la práctica.

Ejemplo: Ecuación de difusión en 2D

Tomemos la ecuación

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}(\kappa\frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(\kappa\frac{\partial U}{\partial y}) = 0 \tag{7.40}$$

considerando κ constante las componentes del flujo son:

$$\begin{aligned} f &= \kappa\frac{\partial U}{\partial x} \\ g &= \kappa\frac{\partial U}{\partial y}\end{aligned} \tag{7.41}$$

Considerando (7.19) llegamos a una ecuación para cada celda, siendo la de (i, j) escrita como:

$$\left(\frac{\partial U}{\partial t}\right)_{ij}\Delta x\Delta y + (f_{AB} - f_{CD})\Delta y + (g_{BC} - g_{DA})\Delta x = 0 \tag{7.42}$$

El flujo f_A se toma como valor promedio entre las celdas 1678 y de manera similar con f_B sobre las celdas 1892 con lo cual se escriben como:

$$\begin{aligned} f_A &= \kappa(\frac{\partial U}{\partial x})_A = \frac{\kappa}{2\Delta x}(U_{i+1,j} + U_{i+1,j-1} - U_{i,j} - U_{i,j-1}) \\ f_B &= \kappa(\frac{\partial U}{\partial x})_B = \frac{\kappa}{2\Delta x}(U_{i+1,j} + U_{i+1,j+1} - U_{i,j} - U_{i,j+1}) \end{aligned} \tag{7.43}$$

Tomando $f_{AB} = \frac{1}{2}(f_A + f_B)$, multiplicando por Δy y haciendo lo mismo con las restantes 3 caras se obtiene una expresión sencilla si consideramos $\Delta x = \Delta y$,

$$\frac{\partial U_{ij}}{\partial t} + \frac{\kappa}{4(\Delta x)^2}\Big[U_{i+1,j+1} + U_{i+1,j-1} + U_{i-1,j+1} + U_{i-1,j-1} - 4U_{ij}\Big] = 0 \tag{7.44}$$

Si en su lugar usamos

$$f_{AB} = \kappa\Big(\frac{\partial U}{\partial x}\Big)_{AB} = \frac{\kappa}{\Delta x}(U_{i+1,j} - U_{ij}) \tag{7.45}$$

esta conduce a la standard forma del operador de Laplace en el método de las diferencias finitas

$$\frac{\partial U_{ij}}{\partial t} + \frac{\kappa}{(\Delta x)^2}\Big[U_{i+1,j} + U_{i,j-1} + U_{i-1,j} + U_{i,j+1} - 4U_{ij}\Big] = 0 \tag{7.46}$$

7.4. El método de los volúmenes finitos en 3D

En el caso 3D los volúmenes finitos más comúnmente utilizados son los tetraedros y los hexaedros. Con los primeros se tiene la ventaja de que son muy generales ya que cualquier dominio tridimensional puede ser mallado con tetraedros, mientras que con hexaedros el problema no es tan sencillo existiendo algunas restricciones. No obstante en lo que sigue pensaremos en volúmenes independientemente de su forma geométrica para aplicar algunos conceptos particulares al caso 3D.

7.4.1. Evaluación del area de las caras de la celda

Una importante propiedad del vector area asociado con cada celda se puede derivar a través del teorema de la divergencia. Por ejemplo si en (7.36) usamos $U = 1$ entonces

$$\oint_S d\mathbf{S} = \int_\Omega \nabla 1 d\Omega = 0 \tag{7.47}$$

mostrando que el vector superficie de una dada cara contenida en una superficie cerrada S y orientado en la dirección saliente

$$\mathbf{S}_{face} = \int_{face} d\mathbf{S} \tag{7.48}$$

depende solamente del propio contorno de la cara. Tomemos la figura 7.6 donde se muestra el vector superficie exterior para algunas de las caras del hexaedro. Si por ejemplo tomamos la cara $ABCD$ vemos que entre las muchas alternativas para evaluar el vector area podemos citar a:

1.- *usando las diagonales*

$$\mathbf{S}_{ABCD} = \tfrac{1}{2}(\mathbf{x}_{AC} \wedge \mathbf{x}_{BD}) \tag{7.49}$$

2.- *usando las aristas*

$$\mathbf{S}_{ABCD} = \tfrac{1}{2}[(\mathbf{x}_{AB} \wedge \mathbf{x}_{BC}) + (\mathbf{x}_{CD} \wedge \mathbf{x}_{DA})] \tag{7.50}$$

No obstante ambas expresiones conducen a idénticos resultados aun en el caso general en el que la cara no sea coplanar, siendo la (7.49) la más económica desde el punto de vista de la cantidad de operaciones involucradas.

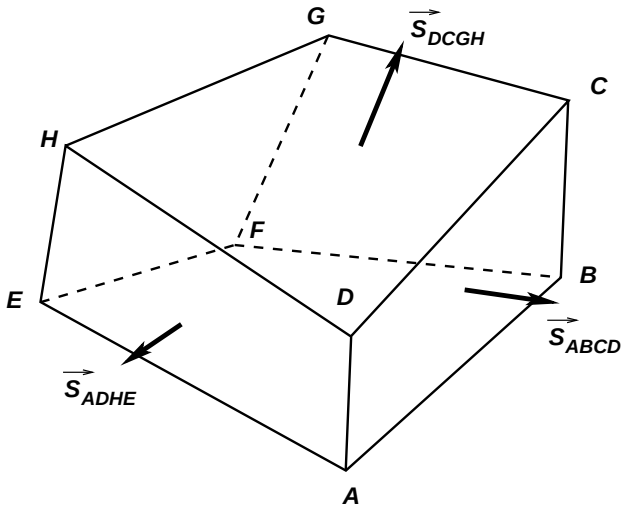


Figura 7.6: Volumen finito hexaédrico en 3D

7.4.2. Evaluación del volúmen de la celda de control

En la figura 7.7 vemos una forma de calcular el volúmen de un hexaedro mediante división en tetraedros o pirámides. Para calcular el volúmen de un tetraedro podemos apelar a las siguientes identidades:

$$\begin{aligned} \int_\Omega \nabla \cdot \mathbf{a} d\Omega &= \oint_S \mathbf{a} \cdot d\mathbf{S} \\ \text{en 2D tomando } \mathbf{a} = \mathbf{x} &\Rightarrow \nabla \cdot \mathbf{x} = 2 \\ 2\Omega = \oint_S \mathbf{x} \cdot d\mathbf{S} &= \oint_S (xdy - ydx) \\ \text{en 3D tomando } \mathbf{a} = \mathbf{x} &\Rightarrow \nabla \cdot \mathbf{x} = 3 \\ 3\Omega_{PABC} = \oint_{PABC} \mathbf{x} \cdot d\mathbf{S} &= \sum_{faces} \mathbf{x} \cdot \mathbf{S}_{faces} \end{aligned} \tag{7.51}$$

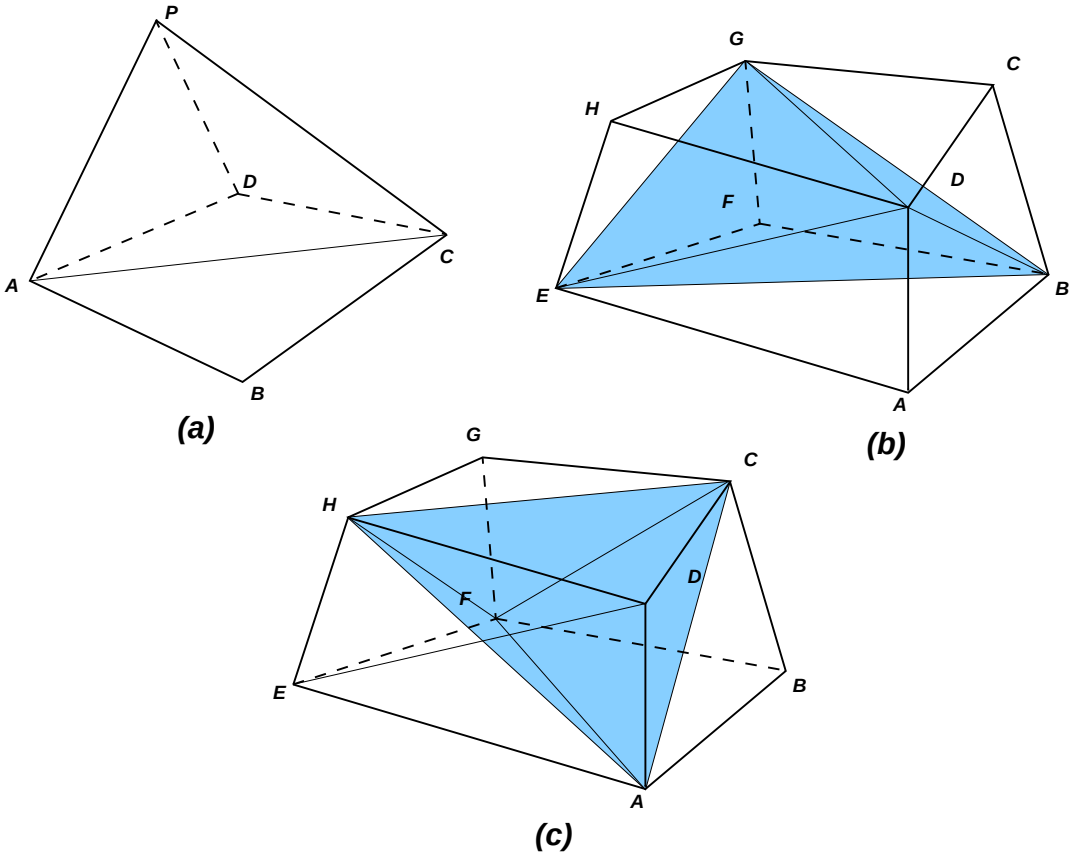


Figura 7.7: Calculo del volúmen de una celda en 3D

La última expresión en (7.51) es equivalente a

$$\Omega_{PABC} = \frac{1}{3}\mathbf{x}_{(P)} \cdot \mathbf{S}_{ABC} \tag{7.52}$$

con $\mathbf{x}_{(P)}$ el vector posición respecto al punto P. Este punto es arbitrario pero por comodidad conviene que coincida con uno de los vértices del tetraedro. Ya que el punto P pertenece a todas las caras excepto a la cara ABC solo con ella el producto escalar es no nulo. Si recurrimos a la expresión (7.49) para evaluar el area de las caras de la celda y la reemplazamos en (7.52) hallamos:

$$\Omega_{PABC} = \frac{1}{6}\mathbf{x}_{(PA)} \cdot (\mathbf{x}_{AB} \wedge \mathbf{x}_{BC}) \tag{7.53}$$

que puede escribirse también como un determinante :

$$\Omega_{PABC} = \frac{1}{6}\begin{vmatrix} x_P & y_P & z_P & 1 \\ x_A & y_A & z_A & 1 \\ x_B & y_B & z_B & 1 \\ x_C & y_C & z_C & 1 \end{vmatrix} \tag{7.54}$$

Se debe tener cuidado con la numeración dada al tetraedro ya que de esta dependerá el signo de algunas contribuciones. Una regla general podría ser utilizar la regla de la mano derecha partiendo de la cara y tomando el cuarto nodo en el sentido que indica el pulgar. Otra recomendación importante es que cuando partimos una malla de hexaedros en tetraedros se debe tomar la precaución de que las caras hexaédricas comunes a dos celdas tengan diagonal coincidente. Estas diagonales surgen al hacer la partición.

En el caso de una partición del hexaedro en pirámides esta puede dividirse en dos tetraedros y el cálculo del volúmen y las areas se reduce a lo visto anteriormente.

Como caso de interés se puede demostrar que si alguna de las caras de la celda hexa edrica no es coplanar entonces las dos únicas particiones del hexaedro en 5 tetraedros no producen el mismo valor de volúmen. Si tomamos un promedio de los dos valores obtenidos este equivale a hacer una transformación isoparamétrica trilineal al estilo método de los elementos finitos e integrarla con $2 \times 2 \times 2$ puntos de Gauss.

7.5. TP.VII.- Trabajo Práctico

método de los volúmenes finitos en 1D

- Muestre que la ley de conservación integral sobre un dominio unidimensional $a \leq x \leq b$ aplicada a

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} &= 0 \\ f(a) &= f(b) \end{aligned} \tag{7.55}$$

  se reduce a que

$$\int_a^b u dx$$

  es constante en el tiempo. Aplique esta condición a la variable espacial discretizada con una distribución arbitraria de puntos en la malla y muestre que la anterior condición se reduce a:

$$\begin{aligned} \tfrac{1}{2} \sum_i \Delta u_i (x_{i+1} - x_{i-1}) &= 0 \\ \Delta u_i = u_i^{n+1} - u_i^n &= \left(\frac{\partial u_i}{\partial t} \right) \Delta t \end{aligned} \tag{7.56}$$

  Ayuda: Aplique la regla del trapecio para evaluar la integral $\frac{\partial}{\partial t} \int_a^b u dx = 0$ *y reescriba la suma aislando los términos en* u_i*.*

- Escriba un programa en MatLab para resolver la ecuación de advección difusión 1D no estacionaria utilizando una formulación

- 1.- centrada en la celda,

- 2.- centrada en los vértices

  Tenga en cuenta la necesidad de ingresar una condición inicial, condiciones de contorno del tipo Dirichlet y Neumann y como datos del problema físico la velocidad de transporte y el coeficiente de difusión.

método de los volúmenes finitos en 2D

- Hemos visto algunas formas de evaluar los flujos convectivos como las expresiones (7.20) y (7.22), entre otras. Su aplicación al caso del método de los volúmenes finitos centrado en las celdas produjo las ecuaciones (7.31) y (7.32) respectivamente. Obtenga las expresiones equivalentes al método de los volúmenes finitos centrado en los vértices.

- Tome una celda cuadrangular y partiendo de las terceras expresiones de (7.37,7.38,7.39) obtenga una expresión para el promedio de las derivadas y otra para el volúmen de la celda tratando de que la expresión final quede expresada como diferencia entre valores opuestos por cada diagonal.

- Considere la ecuación de difusión bidimensional

$$\frac{\partial U}{\partial t} + \frac{\partial}{\partial x}(\kappa \frac{\partial U}{\partial x}) + \frac{\partial}{\partial y}(\kappa \frac{\partial U}{\partial y}) = 0$$

y escriba la formulación centrada en la celda para una malla estructurada considerando que los coeficientes de difusión son variables con la posicón y se definen en los vértices de la celda, siendo el valor en el centro de cada arista aquel que surge como promedio de los valores en los vértices extremos de cada arista.

- De la figura 7.8 surgen algunas alternativas para definir las celdas en 2D para el método de los volúmenes finitos . Tome la mostrada a la izquierda ((a) Mc Donald-1971), aplique la definición del método de los volúmenes finitos (7.17) a la celda $ACDEGH$

$$\frac{\partial}{\partial t}(\mathbf{U}_j \Omega_j) + \sum_{\text{lados}} (\mathbf{F} \cdot \mathbf{S}) = \mathbf{Q}_j \Omega_j$$

y derive el esquema para el nodo (i, j) usando como definición de flujos por cara aquella definida como el promedio de los flujos en los puntos extremos, es decir, $f_{AC} = 1/2(f_A + f_C)$ y compare con la expresión (7.57)

$$\begin{aligned} &\frac{\partial U_{ij}}{\partial t} + \frac{1}{4}\Big[2\frac{(f_{i+1,j} - f_{i-1,j})}{2\Delta x} + \frac{(f_{i+1,j+1} - f_{i-1,j+1})}{2\Delta x} + \frac{(f_{i+1,j-1} - f_{i-1,j-1})}{2\Delta x}\Big] + \\ &\frac{1}{4}\Big[2\frac{(g_{i,j+1} - f_{i,j-1})}{2\Delta y} + \frac{(g_{i+1,j+1} - f_{i+1,j-1})}{2\Delta y} + \frac{(g_{i-1,j+1} - f_{i-1,j-1})}{2\Delta y}\Big] = Q_{ij} \end{aligned} \tag{7.57}$$

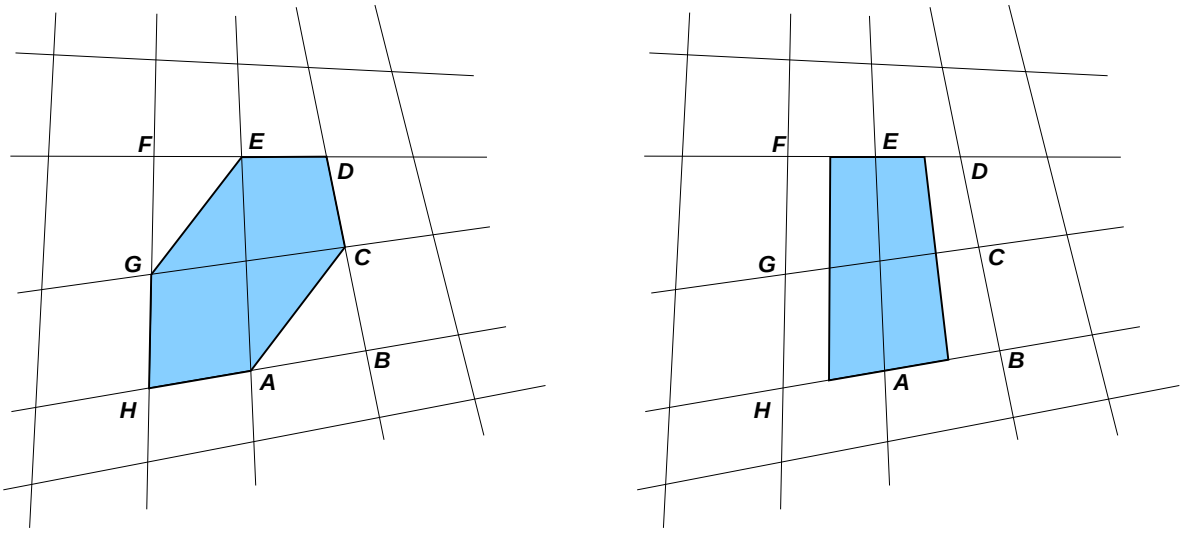


Figura 7.8: Volúmenes de control centrado en los vértices

- **método de los volúmenes finitos en 3D** Mostrar que las expresiones

$$\mathbf{S}_{ABCD} = \tfrac{1}{2}[(\mathbf{x}_{AB} \wedge \mathbf{x}_{BC}) + (\mathbf{x}_{CD} \wedge \mathbf{x}_{DA})]$$

o

$$\mathbf{S}_{ABCD} = \frac{1}{4}[(\mathbf{x}_{AB} + \mathbf{x}_{DC}) \wedge (\mathbf{x}_{BC} + \mathbf{x}_{AD})]$$

utilizadas para evaluar el area de las caras de la celda de control en 3D son equivalente a la expresión:

$$\mathbf{S}_{ABCD} = \tfrac{1}{2}(\mathbf{x}_{AC} \wedge \mathbf{x}_{BD})$$

- Tome un hexaedro con al menos una cara no coplanar. Realice las dos únicas posibles particiones en 5 tetraedros cada una y calcule el volúmen de cada partición utilizando la expresión del determinante:

$$\Omega_{1234} = \frac{1}{6}\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix} \tag{7.58}$$

  (a) Qué valores de volúmenes obtuvo ?,

  (b) son coincidentes ?.

  (c) Tome su promedio.

  (d) Demostrar que el promedio coincide con el volúmen obtenido mediante el cálculo por el método de los elementos finitos usando una transformación isoparamétrica trilineal integrada con $2 \times 2 \times 2$ puntos de Gauss.

Capítulo 8

Análisis de esquemas numéricos

8.1. Introducción

De acuerdo a lo establecido en los capítulos precedentes vimos que la discretización de las variables independientes y de las ecuaciones conducen a un esquema numérico particularmente dependiente del método usado. Un esquema numérico en general se lo representa mediante un *stencil* o *estrella* del operador discreto que permite vincular la solución entre los nodos de la malla. Si bien en el continuo el dominio de dependencia e influencia de un operador abarca regiones extensas del dominio espacial en el caso discreto este se restringe a una zona más acotada y que dependerá del orden y tipo de aproximación utilizado. Uno de los pasos que anteceden a la aplicación práctica de los esquemas numéricos es llevar a cabo un análisis del mismo para tener una idea de lo que puede esperarse del mismo. El análisis más clásico incluye tópicos como *consistencia, precisión, estabilidad y convergencia*.

Existe una gran variedad de métodos de análisis desde aquellos que estudian el comportamiento del operador discreto en un medio infinito, otros que incluyen el tratamiento del contorno y algunos otros que dan cuenta de la influencia de las no linealidades. Antes de pasar a tratar los más importantes métodos de análisis introduciremos al lector en los conceptos básicos de los 4 tópicos antes mencionados.

8.2. Definiciones básicas

Consideraremos una de las ecuaciones más representativas de los modelos matemáticos que aparecen en mecánica de fluidos, la ecuación hiperbólica de advección pura no estacionaria, que en su versión unidimensional se escribe como:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{8.1}$$

donde a es la velocidad de propagación de la onda cuya amplitud es u. Reescribimos la anterior usando subíndices para referirnos a las derivadas, entoces:

$$u_t + au_x = 0 \tag{8.2}$$

Consideremos un problema de valores iniciales de frontera, entonces para poder resolver la (8.2) necesitamos especificar :

$$\begin{array}{lll} t = 0 & u(x,0) = f(x) & 0 \leq x \leq L \\ x = 0 & u(0,t) = g(t) & t \geq 0 \end{array} \tag{8.3}$$

Supongamos que aplicamos un esquema en diferencias finitas centradas de segundo órden para la primera derivada espacial o uno de volúmenes finitos equivalente, entonces el esquema semidiscreto nos queda

$$(u_t)_i = -\frac{a}{2\Delta x}(u_{i+1} - u_{i-1}) \tag{8.4}$$

Este es comúnmente referenciado como el *método de las líneas*.

El próximo paso es definir una discretización para el tiempo. Esto implica el reemplazo de la derivada temporal por una forma discreta y además involucra tomar la decisión de elegir el nivel de tiempo en el cual evaluar el lado derecho.

- *Esquema explícito* Eligiendo una forma en diferencias hacia adelante, el esquema más simple que se obtendría sería evaluar el lado derecho en el tiempo anterior. Este esquema se denomina *forward Euler* y se escribe como:

  $$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) \tag{8.5}$$

  Surge por inspección directa que la solución se obtiene despejando directamente

  $$u_i^{n+1} = f(u_{i-1}^n, u_i^n, u_{i+1}^n) \tag{8.6}$$

- *Esquema implícito*

  Si en cambio evaluamos el lado derecho en el tiempo posterior y si usamos diferencias hacia atrás para la forma discreta de la derivada temporal tenemos el esquema *backward Euler*:

  $$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{2\Delta x}(u_{i+1}^{n+1} - u_{i-1}^{n+1}) \tag{8.7}$$

  Como vemos aquí la solución se obtiene invirtiendo un sistema de ecuaciones de la forma

  $$f(u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^{n+1}) = u_i^n \tag{8.8}$$

  donde la matriz que representa el sistema es tridiagonal debido a que la estrella es centrada y fue generada mediante la combinación de tres nodos.

Estos esquemas producen aproximaciones con una precisión de segundo órden en el espacio y primer órden en el tiempo. Usando esquemas de primer órden en el espacio surgen los siguientes dos esquemas:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{\Delta x}(u_i^n - u_{i-1}^n) \tag{8.9}$$

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{\Delta x}(u_i^{n+1} - u_{i-1}^{n+1}) \tag{8.10}$$

Analizando lo que ocurre con los esquemas (8.5-8.10) se puede comprobar que los dos primeros son incondicionalmente inestables mientras que el tercero es condicionalmente estables y el cuarto es incondicionalmente estable. Esto significa que los dos primeros producen un error que crece en amplitud con el paso de tiempo (divergen) mientras que en el tercer esquema debemos elegir una relación entre el paso de tiempo y el de la malla acorde a un criterio de estabilidad. Como veremos más adelante en problemas de advección pura el parámetro que gobierna la estabilidad de un esquema numérico es el número de *Courant*. Este se define como:

$$\sigma = \frac{a\Delta t}{\Delta x} \tag{8.11}$$

El cuarto esquema (8.10) converge sin restricciones pero tiene una precisión baja (primer órden).

- **Problema:** Usando Matlab implementar los 4 esquemas (8.5-8.10) para una función $f(x)$

$$f(x) = \begin{cases} 0 & ; x < -0.2 \\ 1 + 1/2 x & ; -0.2 \leq x \leq 0 \\ 1 - 1/2 x & ; 0 \leq x \leq 0.2 \\ 0 & ; x > 0.2 \end{cases} \tag{8.12}$$

  con $a = 1, \Delta x = 1, g(x = -5) = 0, L = 10$. Tome distintos valores del número de Courant, desde $\sigma = 0.1$ hasta valores superiores a $\sigma = 1$.

Este ejemplo simple muestra la importancia del análisis antes de realizar cualquier experimento numérico. Podemos formularnos las siguientes preguntas a la hora de analizar un esquema:

- Qué condiciones imponer para una aproximación aceptable ?
- Cómo predecir los límites de estabilidad del esquema ?
- Cómo obtener información cuantitativa de la precisión ?

Para responder estas preguntas recurrimos a los conceptos de:

consistencia Esta define una relación entre la ecuación diferencial y la formulación discreta.

estabilidad Establece una relación entre la solución calculada y la solución exacta de las ecuaciones discretas

convergencia Conecta la solución calculada con la solución exacta de la ecuación diferencial.

En la figura 8.1 vemos un cuadro sinóptico con la relación entre los operadores, sus soluciones y los tópicos de análisis. A continuación entraremos más en detalle en ellos.

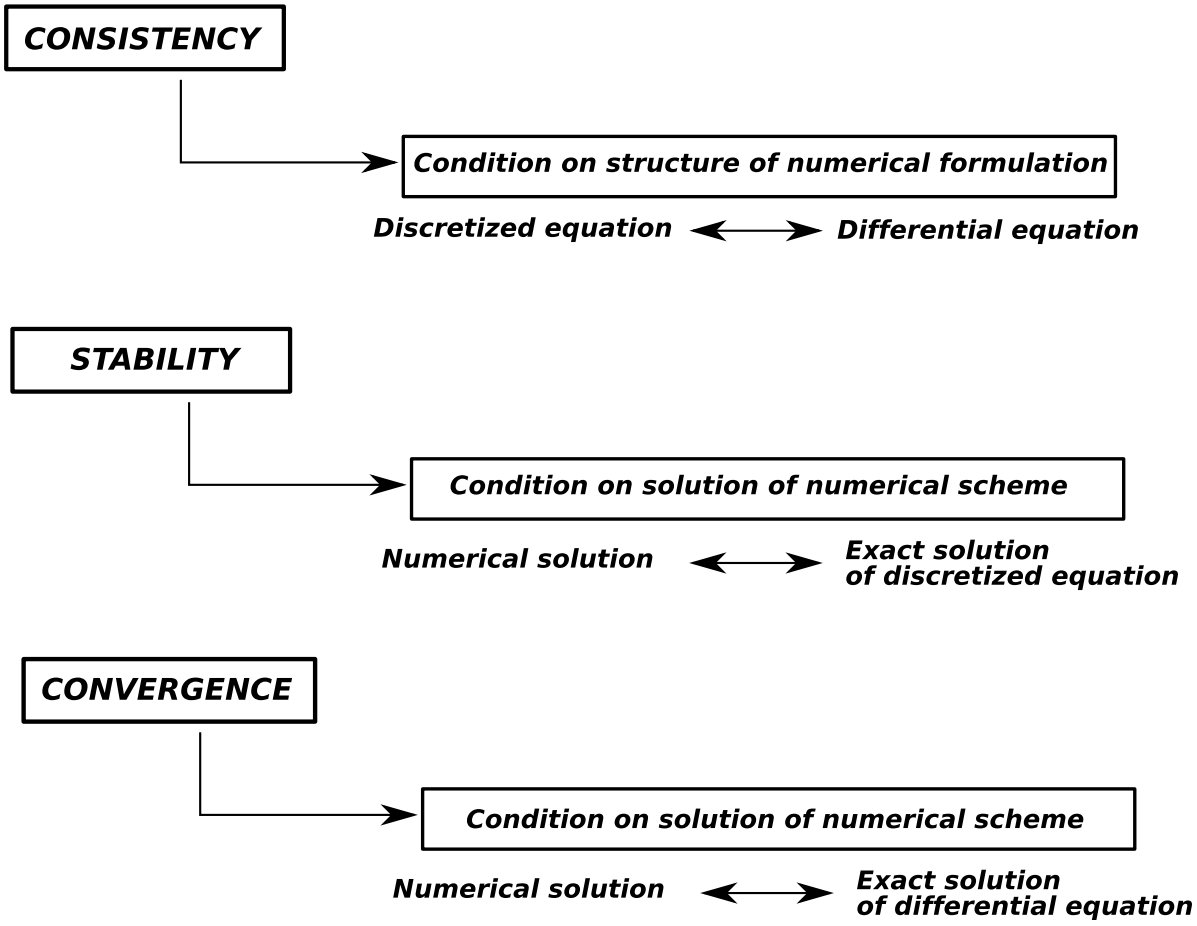


Figura 8.1: Definición de consistencia - convergencia - estabilidad

8.3. Consistencia

Un esquema es consistente si la ecuación discreta tiende al operador diferencial cuando todos los incrementos de las variables independientes tienden a cero. Esto se puede expresar como:

$$L^h u \to Lu \qquad \text{para} \quad \Delta x_j, \Delta t \to 0 \tag{8.13}$$

con Δx_j representando los pasos de la malla en todas las direcciones. Como vemos ambos operadores se aplican a la solución exacta del problema.

Apliquemos la definición al caso (8.2). Para ello desarrollemos en series de Taylor alrededor del punto u_i^n todas los valores nodales que están incluidos en el esquema numérico (8.5).

$$u_i^{n+1} = u_i^n + \Delta t\Big(u_t\Big)_i^n + \frac{\Delta t^2}{2}\Big(u_{tt}\Big)_i^n + \dots$$

$$u_{i+1}^n = u_i^n + \Delta x\Big(u_x\Big)_i^n + \frac{\Delta x^2}{2}\Big(u_{xx}\Big)_i^n + \frac{\Delta x^3}{6}\Big(u_{xxx}\Big)_i^n + \dots \tag{8.14}$$

$$u_{i-1}^n = u_i^n - \Delta x\Big(u_x\Big)_i^n + \frac{\Delta x^2}{2}\Big(u_{xx}\Big)_i^n - \frac{\Delta x^3}{6}\Big(u_{xxx}\Big)_i^n + \dots$$

Para los esquemas de primer órden como el aplicado a la derivada temporal hemos cortado el desarrollo en el término $O(\Delta t^2)$ mientras que en las aproximaciones espaciales centradas, que son de segundo órden hemos extendido el desarrollo un término más. Reemplazando (8.14) en el esquema (8.5) y usando la definición (8.13) vemos que

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} + \frac{a}{2\Delta x}(u_{i+1}^n - u_{i-1}^n) - (u_t + au_x)_i^n = \frac{\Delta t}{2}\Big(u_{tt}\Big)_i^n + \frac{\Delta x^2}{6}a\Big(u_{xxx}\Big)_i^n + O(\Delta t^2, \Delta t^4) \tag{8.15}$$

Se ve claramente que si $\Delta x\ , \Delta t\ \to 0$ el segundo miembro se anula y el esquema es, por la definición, *consistente*.

La *precisión* del esquema puede verse en los términos que lideran el segundo miembro. Este esquema es de primer órden en el tiempo y de segundo órden en el espacio. No obstante la precisión global podría alterarse si alguna relación entre Δx y Δt se establece, por ejemplo si se fija que $\frac{\Delta t}{\Delta x}$ sea constante, entonces será globalmente de primer órden, mientras que si se fija la relación $\frac{\Delta t}{\Delta x^2}$ constante será globalmente de segundo órden.

Otra forma interesante de interpretar la consistencia es la siguiente:

Supongamos que $\bar{u}_i^n$ es la solución exacta del esquema numérico, o sea las $\bar{u}_i^n$ son tales que satisfacen exactamente (8.5). Al reemplazarlas en el operador diferencial arrojan la siguiente identidad:

$$\Big(\bar{u}_t + a\bar{u}_x\Big)_i^n = -\frac{\Delta t}{2}\Big(u_{tt}\Big)_i^n - \frac{\Delta x^2}{6}a\Big(u_{xxx}\Big)_i^n + O(\Delta t^2, \Delta t^4) \tag{8.16}$$

mostrando que la solución exacta de una formulación discreta no satisface exactamente la ecuación diferencial para valores finitos del paso de tiempo y el incremento de la malla. No obstante la solución exacta de la ecuación en diferencias satisface una ecuación diferencial equivalente, llamada *modificada*, que difiere de la original en el error de truncamiento, representado por los términos del lado derecho. En este ejemplo

$$\epsilon_T = -\frac{\Delta t}{2}\Big(u_{tt}\Big)_i^n - \frac{\Delta x^2}{6}\Big(u_{xxx}\Big)_i^n + O(\Delta t^2, \Delta t^4) \tag{8.17}$$

que puede ser expresado en forma equivalente aplicando la ecuación diferencial sobre él para eliminar la derivada temporal,

$$\begin{aligned}\Big(u_t\Big)_i^n &= -a\Big(u_x\Big)_i^n + O(\Delta t, \Delta x^2)\\ \Big(u_{tt}\Big)_i^n &= -a\Big(u_{xt}\Big)_i^n + O(\Delta t, \Delta x^2)\\ &= -a^2\Big(u_{xx}\Big)_i^n + O(\Delta t, \Delta x^2)\end{aligned} \tag{8.18}$$

con lo cual el error de truncamiento puede ser escrito como

$$\epsilon_T = -\frac{\Delta t}{2}a^2\Big(u_{xx}\Big)_i^n - a\frac{\Delta x^2}{6}\Big(u_{xxx}\Big)_i^n + O(\Delta t^2, \Delta x^2) \tag{8.19}$$

lo cual equivale finalmente a resolver una ecuación diferencial modificada

$$u_t + au_x = -\frac{\Delta t}{2}a^2u_{xx} + O(\Delta t^2, \Delta x^2) \tag{8.20}$$

Esto muestra porqué el esquema es inestable. Tal como mencionáramos anteriormente el esquema numérico (8.5) genera un error de truncamiento equivalente a una difusión negativa, con un coeficiente $-\frac{\Delta t}{2}a^2$

Finalmente la consistencia y el órden de precisión de un esquema lo dictamina el error de truncamiento, en el primer caso tendiendo a cero con los incrementos espaciales y temporales y en el segundo caso a través de las potencias con las cuales se va a cero.

Concluyendo, si

$$\epsilon_T = O(\Delta t^q, \Delta x^p) \tag{8.21}$$

- la *consistencia* exige $p > 0\ ,\ q > 0$
- la *precisión* está asociada con los valores de $p\ ,\ q$

8.4. Estabilidad

La definición de estabilidad ha sufrido bastantes cambios desde que fue por primera vez introducida por O'Brien en 1950. Sin entrar en detalles históricos saltamos hasta 1956 cuando Lax y Richtmyer introdujeron el concepto de estabilidad basado en la evolución temporal de la solución. Posteriormente, en 1967, Ritchmyer y Morton desarrollaron esta idea que se basa en definir en forma matricial la influencia del operador discreto. Asi como un operador diferencial acepta una descomposición espectral sobre un espacio de dimensión infinita, el operador discreto acepta una equivalente pero en un espacio de dimensión finita, o sea representable por matrices. Supongamos que expresamos todas las incógnitas en cada punto del espacio y del tiempo (variables nodales) en un arreglo, que al tiempo $n\Delta t$ se puede expresar como:

$$U^n = \begin{pmatrix} u_1^n \\ \vdots \\ u_{i-1}^n \\ u_i^n \\ u_{i+1}^n \\ \vdots \end{pmatrix} \tag{8.22}$$

El esquema numérico puede ser escrito en forma de operador $C : \mathbb{R}^N \to \mathbb{R}^N$ como:

$$U^{n+1} = C \cdot U^n \tag{8.23}$$

con $C = C(\Delta x, \Delta t)$

A modo de ejemplo tomemos el esquema numérico (8.5) y veamos que forma toma el operador. Una ecuación típica de aquel esquema es

$$u_i^{n+1} = u_i^n - \sigma(u_{i+1}^n - u_{i-1}^n)/2$$

$$CU^n = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \sigma/2 & 1 & -\sigma/2 & \dots & \dots & \dots \\ \dots & \dots & \sigma/2 & 1 & -\sigma/2 & \dots & \dots \\ \dots & \dots & \dots & \sigma/2 & 1 & -\sigma/2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix} \begin{pmatrix} \vdots \\ u_{i-1}^n \\ u_i^n \\ u_{i+1}^n \\ \vdots \end{pmatrix} \tag{8.24}$$

Si tomamos en cambio el esquema (8.7,8.8) por su carácter de implícito debemos previamente definir un operador B sobre U^{n+1} para luego generar el operador C sobre U^n.

$$u_i^{n+1} = u_i^n - \sigma(u_{i+1}^{n+1} - u_{i-1}^{n+1})/2$$

$$BU^{n+1} = \begin{pmatrix} \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & -\sigma/2 & 1 & +\sigma/2 & \dots & \dots & \dots \\ \dots & \dots & -\sigma/2 & 1 & +\sigma/2 & \dots & \dots \\ \dots & \dots & \dots & -\sigma/2 & 1 & +\sigma/2 & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \end{pmatrix} \begin{pmatrix} \vdots \\ u_{i-1}^{n+1} \\ u_i^{n+1} \\ u_{i+1}^{n+1} \\ \vdots \end{pmatrix} \tag{8.25}$$

- **Problema:** Calcular el operador discreto C del esquema (8.9) y el operador discreto B del esquema (8.10).

La condición de estabilidad se establece del siguiente modo: Dada una solución inicial U^0 a tiempo $t = 0$ la acción repetida de C sobre ella produce que

$$U^n = C^n \cdot U^0$$

Para que todas las soluciones U^n permanezcan acotadas y el esquema definido por C sea estable el operador C tiene que ser uniformemente acotado. Esto es, existe una constante K tal que

$$||C^n|| < K \quad \text{para} \quad \begin{cases} 0 < \Delta t < \tau \\ 0 \leq n\Delta t \leq T \end{cases} \tag{8.26}$$

para valores fijos de τ, T y $\forall n$. A continuación presentamos el método de *Von Neumann*, uno de los métodos más conocidos para el análisis de estabilidad.

8.5. El método de Von Neumann

Este método es bien popular y simple de aplicar, pero como tal tiene algunas limitaciones relacionadas con: • válido solamente para operadores lineales

- válido solamente para coeficientes constantes
- válido solamente para problemas periódicos

Como vimos en la figura 8.1 el análisis de estabilidad se basa en la relación que existe entre la solución computada u_i^n y la solución exacta de la ecuación discretizada $\bar{u}_i^n$. Entre ellas existe una diferencia que viene dada por los errores de redondeo y los errores en los valores iniciales. Por lo tanto

$$u_i^n = \bar{u}_i^n + \epsilon_i^n \tag{8.27}$$

Aplicando el esquema (8.5) sobre la solución exacta de la ecuación discreta vemos que por (8.27)

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{\Delta t} + \frac{\epsilon_i^{n+1} - \epsilon_i^n}{\Delta t} = -\frac{a}{2\Delta x}(\bar{u}_{i+1}^n - \bar{u}_{i-1}^n) - \frac{a}{2\Delta x}(\epsilon_{i+1}^n - \epsilon_{i-1}^n) \tag{8.28}$$

Ya que $\bar{u}_i^n$ satisface exactamente la ecuación (8.5) entonces nos queda una ecuación para el error

$$\frac{\epsilon_i^{n+1} - \epsilon_i^n}{\Delta t} = -\frac{a}{2\Delta x}(\epsilon_{i+1}^n - \epsilon_{i-1}^n) \tag{8.29}$$

que es idéntica al esquema original. Por lo tanto, según una visión lineal los errores evolucionan con el tiempo en la misma forma que la solución. Del mismo modo podemos plantear algo equivalente aplicando una visión de operadores, como presentamos en (8.23), llegando a que

$$e^{n+1} = Ce^n \tag{8.30}$$

con e^n un vector equivalente al definido en (8.22) pero conteniendo a los errores en lugar de los valores nodales. Si las condiciones de borde son periódicas entonces podemos descomponer al error en series de Fourier para la variable espacial en cada paso de tiempo. Ya que el dominio tiene longitud finita la representación de Fourier será discreta y sumada sobre un conjunto finito de armónicas.

Sea un dominio de longitud L, la representación compleja de Fourier refleja la región $(0, L)$ en $(-L, L)$ con un rango de frecuencias y longitudes de onda ($k = 2\pi/\lambda$) como el siguiente:

$$\begin{aligned} k_{\min} &= \pi/L \qquad & \lambda_{\max} &= 2L \\ k_{\max} &= \pi/\Delta x \qquad & \lambda_{\min} &= 2\Delta x \end{aligned} \tag{8.31}$$

La figura 8.2 muestra los dos modos extremos, uno que cubre todo el dominio $2L$(ondas largas) y el otro que abarca la menor región en la cual se puede representar una onda, $2\Delta x$.

Por lo tanto estableciendo que $\Delta x = L/N$, con N el número de nodos en una grilla definida como el conjunto de puntos de coordenadas $x_i = i\Delta x$ podemos representar todas las armónicas visibles por la grilla como

$$k_j = jk_{\min} = j\frac{\pi}{L} = j\frac{\pi}{N\Delta x} \qquad j = 0, \ldots, N \tag{8.32}$$

La descomposición en series de Fourier del error es:

$$\epsilon_i^n = \sum_{j=-N}^{N} E_j^n e^{Ik_j \cdot i\Delta x} = \sum_{j=-N}^{N} E_j^n e^{Iij\pi/N} \tag{8.33}$$

donde $I = \sqrt{-1}$ y E_j^n es la amplitud de la j-ésima armónica.
Definimos la fase como

$$\phi = k_j \cdot \Delta x = \frac{j\pi}{N} \tag{8.34}$$

que cubre el dominio $(-\pi, \pi)$ en pasos de π/N. La región alrededor de $\phi = 0$ corresponden a las bajas frecuencias mientras que las cercanas a $\phi = \pi$ están asociadas a las altas frecuencias. Por la linealidad del operador no solamente el error satisface (8.30) sino cada armónica.

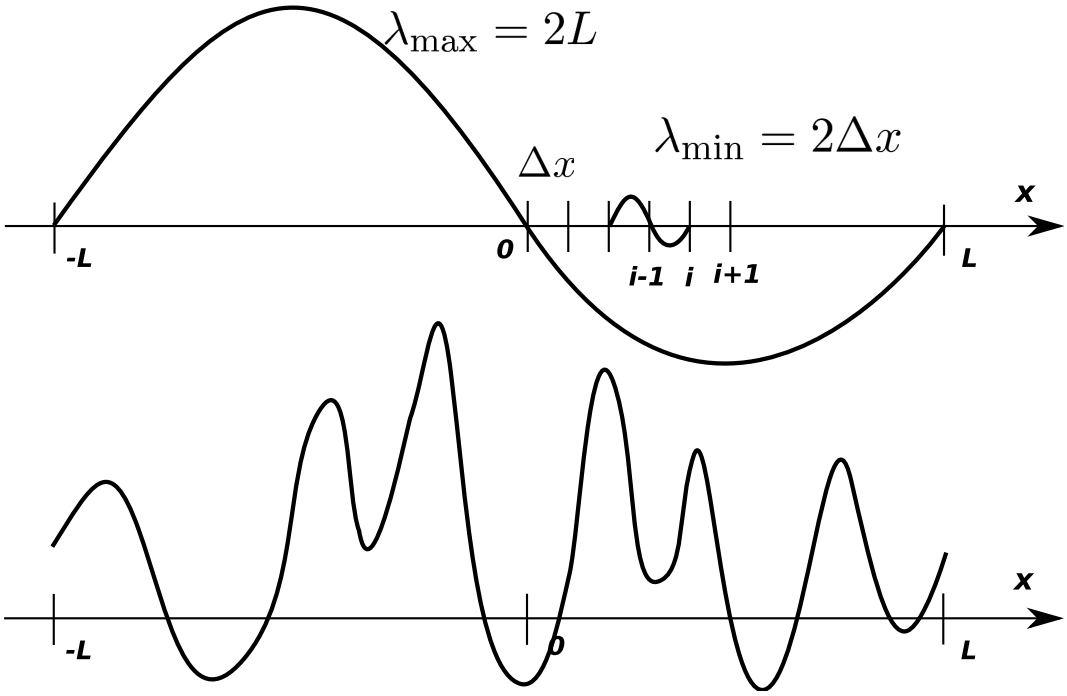


Figura 8.2: Representación de Fourier del error numérico

8.5.1. Factor de amplificación

El hecho que cada armónica satisfaga (8.30) implica que existe un desacoplamiento de los modos y cada uno de ellos puede tratarse por separado. Consideremos una de ellas, $E_j^n e^{Ii\phi}$, introduciéndola en el esquema (8.29), introduciendo el número de Courant σ y sacando factor común $e^{Ii\phi}$ llegamos a una ecuación del tipo:

$$(E^{n+1} - E^n) + \sigma/2E^n(e^{I\phi} - e^{-I\phi}) = 0 \tag{8.35}$$

La condición de estabilidad (8.26) se satisface si las amplitudes del error no crecen con el tiempo, o sea si

$$|G| = \left|\frac{E^{n+1}}{E^n}\right| \leq 1 \qquad \forall\phi \tag{8.36}$$

La cantidad $G(\Delta t, \Delta x, k) = \frac{E^{n+1}}{E^n}$ se la define como el factor de amplificación.

Para el esquema presentado llegamos a que el factor de amplificación es

$$G = 1 - I\sigma\sin(\phi) \tag{8.37}$$

lo cual indica que su módulo será siempre mayor que uno y por lo tanto *incondicionalmente inestable*.

Habíamos visto anteriormente sin demostración evidente que el esquema (8.9) era *condicionalmente estable*. Tratemos de justificarlo con las herramientas recién presentadas. Para ello tomemos una armónica como la anterior, ingresémosla en el esquema (8.9)

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{\Delta x}(u_i^n - u_{i-1}^n) \tag{8.9}$$

aplicado al error y saquemos factor común $E^n e^{Ii\phi}$, entonces arribamos a la siguiente expresión para el factor de amplificación:

$$G = 1 - \sigma + \sigma e^{-I\phi} = 1 - 2\sigma\sin^2(\phi/2) - I\sigma\sin(\phi) \tag{8.38}$$

Esto se representa gráficamente por un círculo centrado en $1-\sigma$ de radio σ, como lo muestra la figura 8.3.

La versión geométrica de la condición de estabilidad (8.26) es que el factor de amplificación de todas las armónicas deben ubicarse dentro de un círculo centrado en el orígen de radio unitario. En el caso del esquema (8.9) esto se cumple solo si

$$0 < \sigma \leq 1 \tag{8.39}$$

La condición (8.39) es muy conocida y denominada *condición de Courant-Friedrichs-Lewy (CFL)*

Otra forma de ver esta condición es mediante teoría de características. Si trazamos las características a partir de un punto de la malla se debe elegir la relación entre el paso espacial y el temporal de forma tal de satisfacer que *el dominio de dependencia de la ecuación diferencial debe estar contenido en el dominio de dependencia de las ecuaciones discretizadas*. La figura 8.4 muestra lo que recién comentamos. Es la discretización empleada capaz de capturar el dominio de dependencia de las características del problema ?

Finalmente si tomamos un esquema como el (8.10)

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = -\frac{a}{\Delta x}(u_i^{n+1} - u_{i-1}^{n+1}) \tag{8.10}$$

vemos que este tiene una *estabilidad incondicional*. Esto se puede demostrar siguiendo los mismos pasos que en los casos anteriores, llegando finalmente a que el factor de amplificación tiene una expresión como:

$$G = \frac{1}{1 + I\sigma\sin(\phi)} \tag{8.40}$$

lo cual implica que su módulo será siempre menor que la unidad.

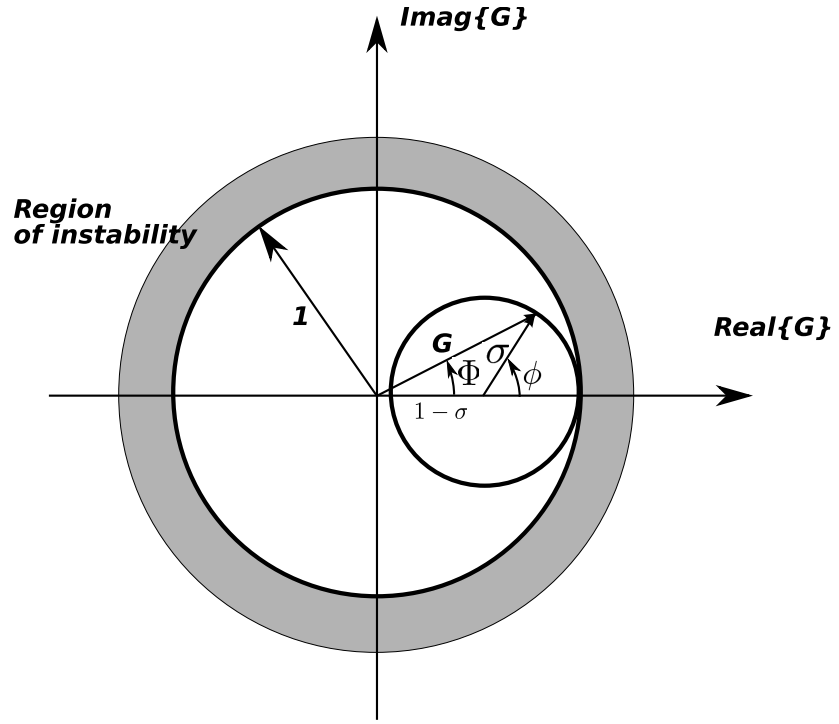


Figura 8.3: Factor de amplificación para un esquema con upwind.

8.5.2. Extensión al caso de sistema de ecuaciones

Consideremos que un esquema numérico es construido en dos pasos:

- (1) Aplicación de la discretización espacial

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = Su_i + q_i \tag{8.41}$$

  donde el término q_i contiene eventuales fuentes y las contribuciones del contorno. La representación matricial de lo anterior se transforma en

$$\frac{\mathrm{d}U}{\mathrm{d}t} = SU + Q \tag{8.42}$$

- (2) Aplicación de un esquema de integración temporal.

$$u_i^{n+1} = Cu_i^n + \bar{q}_i \tag{8.43}$$

  que en forma matricial se escribe como

$$U^{n+1} = CU^n + \bar{Q} \tag{8.44}$$

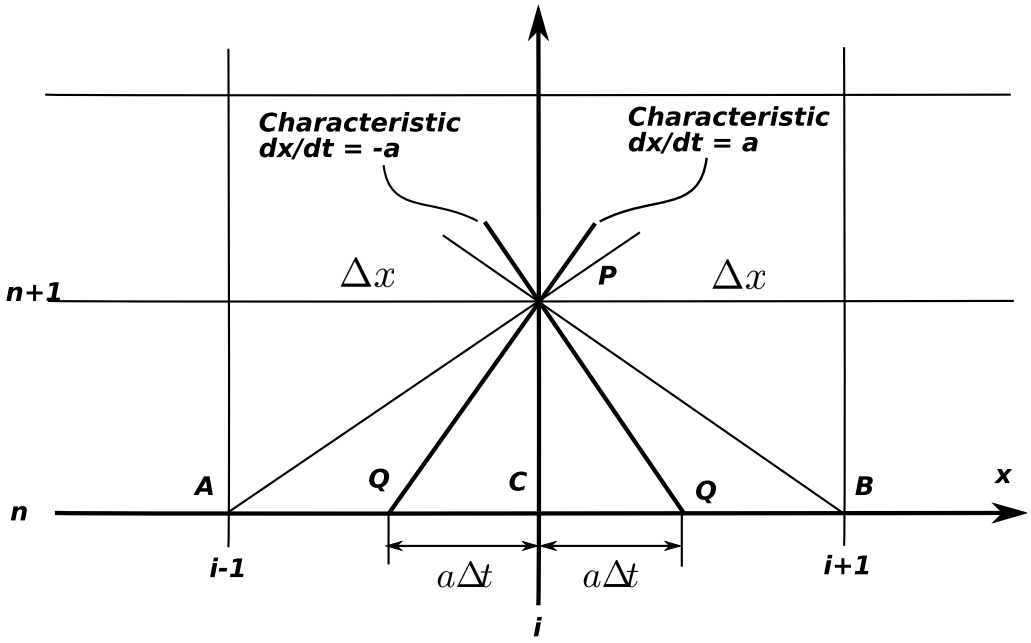


Figura 8.4: Interpretación geométrica de la condición CFL

En estos casos $C = 1 + \Delta t S$. Esto corresponde al caso de un esquema de dos niveles explícitos. El caso implícito adopta la forma

$$B_1 U^{n+1} = B_0 U^n \tag{8.45}$$

donde el operador discreto se define como $C = B_1^{-1} B_0$.

En el caso del esquema de integración temporal del tipo Forward Euler el operador $C = I + \Delta t S$.

En el caso de varias ecuaciones el factor de amplificación se transforma en la matriz de amplificación. Existe una relación entre el concepto de matriz de amplificación G y el operador C. Mientras el primero habita en el espacio de las frecuencias el segundo lo hace en el dominio espacial. En general $G(\phi)$ o $G(k)$ se puede ver como el equivalente de Fourier discreto del operador de discretización C, con una relación como:

$$G(\phi) = C(e^{I\phi}) \tag{8.46}$$

La anterior nos dice que la matriz de amplificación se puede obtener reemplazando el argumento de la matriz C, que en general está asociado con el operador de desplazamientos E por $e^{I\phi}$. Si recordamos que ϕ está definido desde $(-\pi, \pi)$ en pasos de π/N, entonces existe una equivalencia entre E y $e^{I\phi}$. El primero desplaza en la malla y el segundo en el espacio de las frecuencias.

Habiendo establecido una forma de calcular la matriz de amplificación queda por ver como establecer un criterio de estabilidad general. El concepto de módulo del factor de amplificación cambia por el de norma de la matriz de amplificación. Definiendo ésta como:

$$||G|| = \max_{u \neq 0} \frac{|G \cdot u|}{|u|} \tag{8.47}$$

y considerando el hecho que

$$||G|| \geq \rho(G) \tag{8.48}$$

donde $\rho(G) = \max_{j=1,p} |\lambda_j|$ es el radio espectral de la matriz G de $p \times p$ y λ sus autovalores. Entonces una condición suficiente de estabilidad es que

$$\rho(G) \leq 1 + O(\Delta t) \qquad \Delta t \text{ finito} \quad , \quad \forall \phi \in (-\pi, \pi) \tag{8.49}$$

La posibilidad que el radio espectral sea mayor que uno surge en considerar algunos problemas donde la solución exacta crece exponencialmente, no obstante una condición estricta del tipo

$$\rho(G) \leq 1 \tag{8.50}$$

puede ser necesaria para evitar que algunos modos numéricos crezcan más rápidamente que sus correspondientes modos físicos. Por ejemplo cuando la matriz de amplificación tiene autovalores de valor 1 con multiplicidad mayor que uno. Estas condiciones suficientes de estabilidad (b) son válidas también para el caso de matrices normales donde G conmuta con su hermitiana conjugada y el radio espectral coincide con la norma.

Propiedades de la matriz de amplificación

1. Si la matriz de amplificación puede expresarse como un polinomio de la matriz A, $G = P(A)$, luego el teorema del mapeo espectral es válido y establece que:

$$\lambda(G) = P(\lambda(A)) \tag{8.51}$$

   Por ejemplo, si $G = 1 - I\alpha A + \beta A^2$, entonces

$$\lambda(G) = 1 - I\alpha\lambda(A) + \beta(\lambda(A))^2$$

2. Si G puede expresarse como una función de varias matrices que conmutan entre si, esto es:

$$G = P(A, B) \qquad \text{con} \quad AB = BA$$

   entonces, A, B tienen el mismo conjunto de autovectores y

$$\lambda(G) = P(\lambda(A), \lambda(B))$$

   Esta condición deja de ser válida cuando las matrices no conmutan como es el caso de las ecuaciones de movimiento fluido en varias dimensiones.

- *Ejemplo: Ecuaciones de shallow water*

  Las ecuaciones de *shallow water (aguas poco profundas)* son un importante modelo para tratar la hidrodinámica de zonas costeras, algunos rios y lagos donde la profundidad es mucho menor que las restantes dos direcciones coordenadas. El modelo se puede escribir como:

$$\begin{aligned} &\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\cdot\nabla\mathbf{U} = \mathbf{0} \\ &\mathbf{U} = \{h\ ; u\ ; v\} \\ \mathbf{A}_x = \begin{pmatrix} u & h & 0 \\ g & u & 0 \\ 0 & 0 & u \end{pmatrix} \quad &; \quad \mathbf{A}_y = \begin{pmatrix} v & 0 & h \\ 0 & v & 0 \\ g & 0 & v \end{pmatrix} \end{aligned} \tag{8.52}$$

  El modelo unidimensional linealizado alrededor de un estado de referencia $\mathbf{U}_0 = \{h_0\ ; v_0\}$, expresado en forma diferencial se puede escribir como:

$$\begin{aligned} &\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = \mathbf{0} \\ &\mathbf{U} = \{h\ ; v\} \\ \mathbf{A}_y = \begin{pmatrix} v_0 & h_0 \\ g & v_0 \end{pmatrix} \end{aligned} \tag{8.53}$$

  Un esquema espacialmente centrado se escribe como:

$$\begin{aligned} \frac{\mathrm{d}\mathbf{U}_i}{\mathrm{d}t} &= -\mathbf{A}\frac{\mathbf{U}_{i+1} - \mathbf{U}_{i-1}}{2\Delta x} \\ \mathbf{U}_i &= \mathbf{E}_\phi e^{Ii\phi} \\ \mathbf{U}_{i+1} &= \mathbf{E}_\phi e^{I(i+1)}\phi = \mathbf{E}_\phi e^{Ii\phi}e^{I\phi} = \mathbf{U}_i e^{I\phi} \\ \mathbf{U}_{i-1} &= \mathbf{E}_\phi e^{I(i-1)\phi} = \mathbf{E}_\phi e^{Ii\phi}e^{-I\phi} = \mathbf{U}_i e^{-I\phi} \end{aligned} \tag{8.54}$$

  En cuanto a la discretización temporal primero consideraremos el caso de una integración temporal siguiendo el esquema de Euler y luego tomaremos el caso del esquema de Lax-Friedrichs.

- *Esquema Euler*

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \mathbf{A}\Delta t\frac{\Delta - \Delta^{-1}}{2\Delta x}\mathbf{U}_i^n = \Big(\mathbf{I} - \mathbf{A}\Delta t\frac{\Delta - \Delta^{-1}}{2\Delta x}\Big)\mathbf{U}_i^n = \mathbf{C}\mathbf{U}_i^n \tag{8.55}$$

  donde Δ representa el operador de corrimientos espaciales y si llevamos a cabo el ensamblaje de la matriz C esta se expresa como:

$$\mathbf{C} = \begin{pmatrix} \ddots & & & & \\ \frac{1}{2}(\frac{\mathbf{A}\Delta t}{\Delta x}) & \mathbf{I} & \frac{1}{2}(-\frac{\mathbf{A}\Delta t}{\Delta x}) & & \\ & \frac{1}{2}(\frac{\mathbf{A}\Delta t}{\Delta x}) & \mathbf{I} & \frac{1}{2}(-\frac{\mathbf{A}\Delta t}{\Delta x}) & \\ & & \frac{1}{2}(\frac{\mathbf{A}\Delta t}{\Delta x}) & \mathbf{I} & \frac{1}{2}(-\frac{\mathbf{A}\Delta t}{\Delta x}) \\ & & & & \ddots \end{pmatrix} \tag{8.56}$$

Por otro lado podemos arribar a la matriz de amplificación reemplazando (8.54) en (8.55),

$$\begin{aligned} \mathbf{E}_i^{n+1} &= \Big(\mathbf{I} - \frac{\mathbf{A}\Delta t}{2\Delta x}(e^{I\phi} - e^{-I\phi})\Big)\mathbf{E}_i^n \\ \mathbf{G}(\phi) &= \mathbf{I} - \frac{\mathbf{A}\Delta t}{2\Delta x}(e^{I\phi} - e^{-I\phi}) \end{aligned} \tag{8.57}$$

Comparando (8.55) y (8.57) vemos que

$$\mathbf{G}(\phi) = \mathbf{C}(\Delta = e^{I\phi}) \tag{8.58}$$

Aplicando la primera propiedad de la matriz de amplificación tenemos que

$$\begin{aligned} \lambda(\mathbf{G}(\phi)) &= P(\lambda(\mathbf{A})) \\ \mathbf{G}(\phi) &= \mathbf{I} - \frac{\mathbf{A}\Delta t}{2\Delta x}(e^{I\phi} - e^{-I\phi}) \\ \lambda(\mathbf{G}(\phi)) &= 1 - \frac{\lambda(\mathbf{A})\Delta t}{2\Delta x}(e^{I\phi} - e^{-I\phi}) \end{aligned} \tag{8.59}$$

o sea los autovalores de la matriz $\mathbf{A}$ se transforman conforme a 8.59)

Los autovalores de la matriz $\mathbf{A}$ se calculan en forma directa como:

$$\begin{aligned} |\mathbf{A} - \lambda\mathbf{I}| &= 0 \\ \begin{vmatrix} v_0 - \lambda & h_0 \\ g & v_0 - \lambda \end{vmatrix} &= 0 \\ \lambda(\mathbf{A})_{1,2} &= v_0 \pm \sqrt{gh_0} \end{aligned} \tag{8.60}$$

Por lo tanto los autovalores de la matriz de amplificación serán:

$$\lambda(\mathbf{G}) = 1 - \frac{(v_0 \pm \sqrt{gh_0})\Delta t \sin(\phi)}{\Delta x} I \tag{8.61}$$

Este esquema es a simple vista inestable. En el caso en que los autovalores fueran de dificultosa evaluación podemos recurrir a una resolución numérica.

- *Esquema Lax-Friedrichs*

  En 1954 Lax sugirió una forma de estabilizar el esquema anterior conservando la discretización espacial centrada. Este esquema bien popular en estos dias se puede escribir como:

$$\mathbf{U}_i^{n+1} = \tfrac{1}{2}(\mathbf{U}_{i+1}^n + \mathbf{U}_{i-1}^n) - \frac{\mathbf{A}\Delta t}{2\Delta x}(\mathbf{U}_{i+1}^n - \mathbf{U}_{i-1}^n) \tag{8.62}$$

  Realizando un procedimiento similar al caso del integrador Euler hacia adelante llegamos a que existe una condición tipo CFL aquí del tipo,

$$(|v_0| + \sqrt{gh_0})\frac{\Delta t}{\Delta x} \leq 1 \tag{8.63}$$

8.5.3. Análisis espectral del error numérico

Habímos visto que una forma de caracterizar la evolución temporal del error en un esquema numérico era mediante la definición de la matriz de amplificación o en el caso de modelos escalares simplemente del factor de amplificación. Surgió que dicha evolución estaba regida por una expresión del tipo:

$$v^{n+1} = G(\phi)v^n = [G(\phi)]^n v^1 \tag{8.64}$$

Descomponiendo temporalmente a la solución numérica en armónicas podemos ver que

$$v^n = \hat{v}e^{-I\omega n\Delta t} \tag{8.65}$$

con $\omega = \omega(k)$ una función compleja del número de onda que representa la dispersión numérica. Relacionando (8.64) con (8.65) vemos que existe una relación directa entre el factor o matriz de amplificación y esta dispersión numérica,

$$G = e^{-I\omega\Delta t} \tag{8.66}$$

Entonces, la representación de la solución numérica en término de armónicas simples se puede escribir como:

$$u_i^n = \hat{v}e^{-I\omega n\Delta t}e^{Ii\phi} \tag{8.67}$$

Del mismo modo la solución exacta acepta una descomposición similar obteniendo:

$$\tilde{u}_i^n = \hat{v}e^{-I\tilde{\omega} n\Delta t}e^{Ii\phi} \tag{8.68}$$

con su correspondiente función de amplificación exacta

$$\tilde{G} = e^{-I\tilde{\omega}\Delta t} \tag{8.69}$$

Para poder estudiar el espectro de los errores numéricos dividiremos la dispersion ω en su parte real y su parte imaginaria,

$$\omega = \xi + I\eta \tag{8.70}$$

de forma tal que si reemplazamos (8.70) en (8.69) tenemos

$$\begin{aligned} \tilde{G} = e^{+\eta\Delta t}e^{-I\xi\Delta t} &= \parallel G \parallel e^{-I\Phi} \\ \parallel G \parallel &= e^{+\eta\Delta t} \\ \Phi &= \xi\Delta t \end{aligned} \tag{8.71}$$

Del mismo modo con la solución exacta podemos obtener

$$\begin{aligned} \parallel \tilde{G} \parallel &= e^{+\tilde{\eta}\Delta t} \\ \tilde{\Phi} &= \tilde{\xi}\Delta t \end{aligned} \tag{8.72}$$

Si comparamos ambos módulos y ambas fases entre si llegamos a la definición de dos parámetros de error muy importantes:

$$\begin{aligned} \epsilon_D &= \frac{|G|}{e^{\tilde{\eta}\Delta t}} && \text{ERROR POR DISIPACION} \\ \epsilon_\phi &= \Phi - \tilde{\Phi} && \text{ERROR POR DISPERSION} \end{aligned} \tag{8.73}$$

La definición del error de dispersión dada en (8.73) es adecuada en problemas parabólicos sin términos convectivos $(\tilde{\Phi} = 0)$, caso contrario esta puede cambiarse por

$$\epsilon_\phi = \Phi/\tilde{\Phi} \tag{8.74}$$

mejor adecuada en aquellos problemas dominados por convección.

Análisis de error en problemas parabólicos

Tomemos la ecuación del calor discretizada en el espacio en forma centrada y en el tiempo con un esquema explícito. El esquema se puede escribir como:

$$u_i^{n+1} = u_i^n + \frac{\alpha\Delta t}{\Delta x^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) \tag{8.75}$$

El factor de amplificación se obtiene mediante el procedimento presentado oportunamente y este se expresa como:

$$\begin{aligned} G &= 1 - 4\beta\sin^2(\phi/2) \\ \beta &= \frac{\alpha\Delta t}{\Delta x^2} \\ \alpha &\geq 0 \qquad\qquad \beta \leq 1/2 \end{aligned} \tag{8.76}$$

El factor de dispersión del continuo se obtiene reemplazando la representación espectral de la solución exacta en el operador diferencial,

$$\tilde{\omega} = -\alpha k^2 I = -\beta\phi^2/\Delta t I \tag{8.77}$$

y el error de disipación se obtiene como:

$$\epsilon_D = \frac{1 - 4\beta \sin^2(\phi/2)}{e^{-\beta\phi^2}} \tag{8.78}$$

que si lo expandimos en series de potencias en ϕ se obtiene después de algo de algebra

$$\begin{aligned} \epsilon_D &\simeq 1 - \frac{\beta^2\phi^4}{2} + \frac{\beta\phi^4}{12} + \dots \\ &\simeq 1 - \frac{\alpha^2 k^4 \Delta t^2}{2} + \frac{\alpha k^4 \Delta t \Delta x^2}{12} \end{aligned} \tag{8.79}$$

Este esquema produce errores numéricos bajos para modos de baja frecuencia pero para aquellos de alta frecuencia estos pueden ser inaceptables, especialmente para aquellos $\beta \leq 1/2$ próximos a la cota superior. No obstante para $\beta = 1/6$ los dos últimos términos se cancelan y este esquema presenta un alto orden $O(\Delta t^2, \Delta x^4)$. Además ya que G es real no existe error en phase por lo tanto no presenta dispersión numérica.

Análisis de error en problemas hiperbólicos

Tomemos como ejemplo típico de una ecuación hiperbólica la ecuación de advección pura:

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{8.80}$$

Físicamente el transporte por convección se puede ejemplificar colocando inicialmente en el dominio una onda y viendo que esta viaja en una determinada dirección guiada por el campo de velocidades sin modificarse, o sea sin amortiguarse ni dispersarse. Si como hicimos para el caso parabólico reemplazamos una armónica del tipo $u = e^{-I\tilde{\omega}t}e^{Ikx}$ en el operador diferencial encontramos que:

$$\begin{aligned} \tilde{\omega} = ka &= \tilde{\xi} \\ \tilde{\eta} &= 0 \end{aligned} \tag{8.81}$$

Como vemos a diferencia del caso parabólico aquí el coeficiente de dispersión numérica es un número real lo que le da un carácter ondulatorio a la solución y sin amortiguamiento. Entonces, de acuerdo a la definición (8.73) el error por disipación y por dispersión vienen dados por:

$$\begin{aligned} \epsilon_D &= \frac{|G|}{e^{\tilde{\eta}\Delta t}} = |G| \\ \epsilon_\phi = \Phi/\tilde{\Phi} &= \Phi/(ka\Delta t) = \Phi/(\sigma\phi) \end{aligned} \tag{8.82}$$

Entonces, $\parallel G \parallel$ es la amortiguación numérica del esquema mientras que ϵ_ϕ es el error por dispersión. Si $\epsilon_\phi > 1$ la solución numérica viaja más rápido que la exacta mientras que si es $\epsilon_\phi < 1$ viaja más lento.

- *Esquema explícito con upwind*

  A modo de ilustración tomemos el caso de un esquema explícito con upwind para resolver la ecuación de advección pura no estacionaria (8.80). Este esquema ya lo hemos tratado a la hora de calcular el factor de amplificación y habíamos obtenido una expresión para el mismo (8.38)

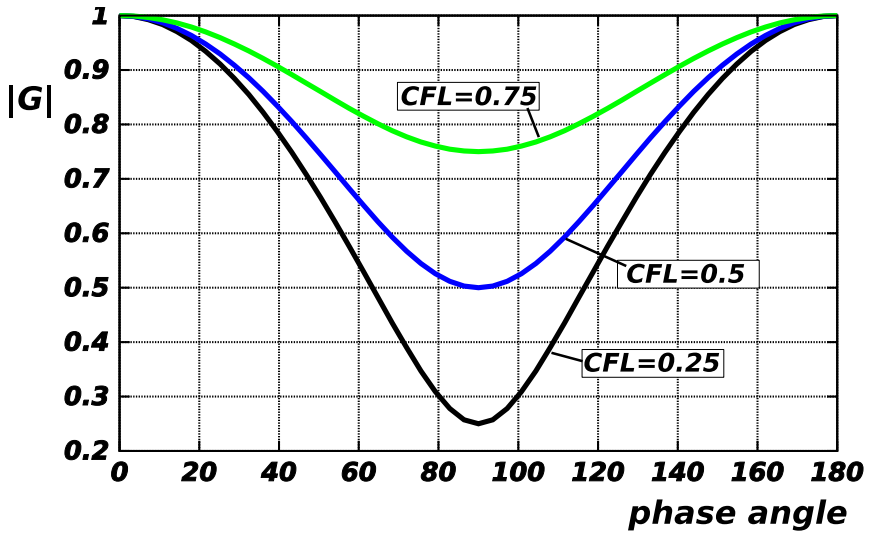


Figura 8.5: Error de disipación del método de Lax-Friedrichs para la ec. de advección pura.

$$\| G \| = [(1 - \sigma + \sigma\cos(\phi))^2 + \sigma^2 \sin^2(\phi)]^{1/2} = [1 - 4\sigma(1-\sigma)\sin^2(\phi/2)]^{1/2} \tag{8.83}$$

O sea que de acuerdo a (8.82) el amortiguamiento numérico lo podemos apreciar graficando la expresión (8.83). La figura 8.5 muestra el error de disipación para un esquema explícito de primer orden estabilizado espacialmente mediante la introducción de *upwind*. (FIXME:= la figura que se referencia es para Lax-Friedrichs).

El error de dispersión se calcula a partir de la relación entre las velocidades de fase numérica y la exacta. La primera se calcula mediante su definición (8.66) y finalmente el error de dispersión se escribe como:

$$\Phi = \Delta t \,\mathrm{Re}\,\{\omega\} = \Delta t \tan^{-1}\Big(-\frac{\mathrm{Im}\,\{G\}}{\mathrm{Re}\,\{G\}}\Big) \tag{8.84}$$

$$\epsilon_\phi = \Phi/\tilde{\Phi} = \frac{\tan^{-1}\Big[\sigma\sin(\phi)/(1-\sigma+\sigma\cos(\phi))\Big]}{\sigma\phi}$$

La figura 8.6 (FIXME:= la figura referenciada es para Lax-Friedrichs) muestra como es el error de dispersión para diferentes números de Courant, fundamentalmente para $\sigma = 1/2$ que da dispersión nula y para $\sigma = 1/4$ y $3/4$ que son casos representativos de ondas numéricas que viajan más lentas y más rápido que las exactas respectivamente.

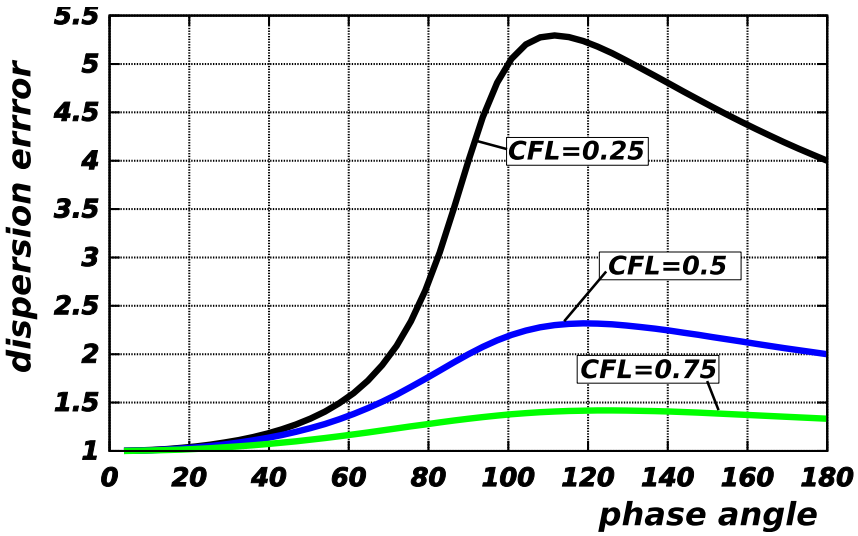


Figura 8.6: Error de dispersión del método de Lax-Friedrichs para la ec. de advección pura.

8.5.4. Extensión a esquemas de tres niveles

Una forma muy simple de ver un esquema de tres niveles es mediante la introducción de una ecuación adicional. De esta forma el problema queda definido mediante dos ecuaciones de 2 niveles y es completamente análogo al caso de tratar con sistema de ecuaciones. De esta forma la matriz de amplificación que surge requiere un cálculo de autovalores para poder identificar los errores de disipación y dispersión. Dejamos como ejemplos dos casos interesantes.

- *El esquema leapfrog para la ecuación de convección*

$$\begin{aligned} u_i^{n+1} &= z_i^n - \sigma(u_{i+1}^n - u_{i-1}^n) \\ z_i^{n+1} &= u_i^n \end{aligned} \tag{8.85}$$

- *El esquema Du Fort-Frankel para la ecuación de difusión*

$$\begin{aligned} (1-2\beta)u_i^{n+1} &= (1-2\beta)u_i^{n-1} + 2\beta(u_{i+1}^n - u_{i-1}^n) \\ \beta &= \alpha\Delta t/\Delta x^2 \qquad \text{Número de Fourier} \end{aligned} \tag{8.86}$$

8.5.5. El concepto de velocidad de grupo

Un paquete de ondas contiene en general armónicas con varias frecuencias. En ese caso la velocidad de grupo representa la velocidad a la cual la energía de la onda se propaga. En el caso de una onda unidimensional la definición matemática de la velocidad de grupo de la solución exacta es:

$$\tilde{v}_G(k) = \frac{\mathrm{d}\tilde{\omega}}{\mathrm{d}k} \tag{8.87}$$

Su contraparte numérica la definimos como

$$v_G(k) = \frac{\mathrm{d}\xi}{\mathrm{d}k} = Re\{\frac{\mathrm{d}\omega}{\mathrm{d}k}\} \tag{8.88}$$

Un ejemplo lo tenemos con el esquema *leapfrog* que tiene una relación de dispersión numérica expresada como:

$$\sin(\omega\Delta t) = \sigma \sin(\phi) \tag{8.89}$$

y aplicando (8.88) se llega a que:

$$v_G(k) = \frac{a\cos(\phi)}{\cos(\omega\Delta t)} = \frac{a\cos(\phi)}{(1-\sigma^2\sin^2(\phi))^{1/2}} \tag{8.90}$$

Para bajas frecuencias la velocidad de grupo es similar a la velocidad de transporte a pero a altas frecuencias $\phi \approx \pi$ esta tiende a $-a$ o sea en contrafase con la exacta.

8.5.6. Análisis de Von Neumann multidimensional

Para extender lo visto en las secciones anteriores al caso multidimensional introducimos el vector número de onda $\mathbf{k}$, de forma que una solución descompuesta en armónicas puede escribirse como:

$$u(\mathbf{x},t) \sim \hat{v}e^{-I\omega t}e^{I\mathbf{k}\cdot\mathbf{x}} \tag{8.91}$$

El producto escalar en su versión discretizada se escribe como:

$$(\mathbf{k}\cdot\mathbf{x})_{i,j,k} = k_x i\Delta x + k_y j\Delta y + k_z k\Delta z = i\phi_x + j\phi_y + k\phi_z \tag{8.92}$$

donde cada número de fase en cada dirección espacial barre el rango $[-\pi,\pi]$. El resto del análisis permanece idéntico al caso unidimensional. Para ejemplificar tomemos la ecuación del calor, un ejemplo de una ecuación del tipo parabólica.

$$\frac{\partial u}{\partial t} = \alpha\Big(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\Big) \tag{8.93}$$

Si usamos un esquema centrado en las variables espaciales y explícito de primer órden en el tiempo la versión discreta se escribe como:

$$u_{i,j}^{n+1} - u_{i,j}^n = \alpha\Delta t\Big[\frac{u_{i+1,j}^n - 2u_{i,j}^n + u_{i-1,j}^n}{\Delta x^2} + \frac{u_{i,j+1}^n - 2u_{i,j}^n + u_{i,j-1}^n}{\Delta y^2}\Big] \tag{8.94}$$

Una descomposición de Fourier discreta se define por:

$$u_{i,j}^n = \sum_{k_x,k_y} v^n e^{Ik_x i\Delta x}e^{Ik_y j\Delta y} \tag{8.95}$$

Definiendo la matriz de amplificación en forma similar al caso unidimensional $v^{n+1} = Gv^n$ e introduciéndola junto con (8.95) en (8.94) se llega a:

$$G - 1 = \beta[(e^{I\phi_x} + e^{-I\phi_x} - 2) + (\frac{\Delta x}{\Delta y})^2(e^{I\phi_y} + e^{-I\phi_y} - 2)] \tag{8.96}$$

con β el número de Fourier antes definido. Otra vez, haciendo que el factor de amplificación se mantenga en módulo menor que la unidad se llega a dos condiciones similares al caso 1D apenas más restrictivas,

$$\begin{aligned} \alpha &> 0 \\ \beta(1 + (\frac{\Delta x}{\Delta y})^2) &\leq 1/2 \end{aligned} \tag{8.97}$$

8.6. Convergencia

La convergencia puede ser establecida siguiendo a Richtmyer y Morton de la siguiente forma:

$$\lim_{\Delta t \to 0\,,\, n \to \infty} ||[C(\Delta t)]^n U^0 - \tilde{U}(t)|| = 0 \tag{8.98}$$

para $n\Delta t$ fijo.

Siguiendo el *teorema de equivalencia de Lax* podemos decir que:

para un problema de valores iniciales bien planteado y una discretización consistente, la estabilidad es condición necesaria y suficiente para la convergencia

Con esto concluimos diciendo que los dos pasos necesarios del análisis son:

- (1) La consistencia conduce a la determinación del órden de precisión del esquema y su error de truncación
- (2) la estabilidad nos brinda información detallada de la distribución en frecuencias del error como función del contenido en frecuencia de la solución calculada.

De esta forma la convergencia queda definida sin necesitar análisis.

8.7. TP. Trabajo Práctico

1. Escriba un programa en MatLab que grafique el factor de amplificación para la ecuación de advección-difusión discretizada en el espacio en forma central y en el tiempo usando un esquema explícito *Forward Euler* para diferentes números de Peclet. Considere al menos lo que sucede cuando el Peclet de la malla es:

   a.- $Pe^h << 1$
   b.- $Pe^h < 1$
   c.- $Pe^h = 1$
   d.- $Pe^h > 1$
   e.- $Pe^h >> 1$

2. Obtener la matriz de amplificación para el modelo unidimensional linealizado de *shallow water* usando un esquema centrado espacialmente y un integrador temporal tipo Lax-Friedrichs,

$$u_i^{n+1} = ½(u_{i+1}^n + u_{i-1}^n) - \frac{\sigma}{2}(u_{i+1}^n - u_{i-1}^n)$$

3. Calcule la matriz de amplificación para la ecuación de las ondas unidimensional de segundo orden:

$$\frac{\partial^2 w}{\partial t^2} - a^2 \frac{\partial^2 w}{\partial x^2} = 0$$

   usando el siguiente esquema de discretización basado en una descomposición del operador de segundo orden en 2 ecuaciones de primer orden:

$$\begin{aligned} v_i^{n+1} - v_i^n &= \frac{a\Delta t}{\Delta x}(w_{i+1}^n - w_i^n) \\ w_i^{n+1} - w_i^n &= \frac{a\Delta t}{\Delta x}(v_i^{n+1} - w_{i-1}^{n+1}) \end{aligned} \tag{8.99}$$

   a.- Encuentre la condición de estabilidad de este esquema.
   b.- Es este esquema adecuado para resolver una ecuación elíptica del tipo

$$\frac{\partial^2 w}{\partial t^2} + |a^2| \frac{\partial^2 w}{\partial x^2} = 0$$

4. Calcule el error por disipación y por dispersión para la ecuación de advección pura discretizada espacialmente en forma centrada y temporalmente mediante un esquema del tipo Lax-Friedrichs.

5. Obtenga el esquema numérico de Lax-Wendroff. Este se puede escribir partiendo de una expansión en series de Taylor como la siguiente:

$$u_i^{n+1} = u_i^n + \Delta t \frac{\partial u}{\partial t}_i + \frac{1}{2}\Delta t^2 \frac{\partial^2 u}{\partial t^2}_i + O(\Delta t^3)$$

reemplazando la ecuación diferencial $\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$ y una versión que surge de derivar esta última nuevamente respecto al tiempo e intercambiar índices de derivaciones.

A continuación calcule el error por disipación y por dispersión para la ecuación de advección pura discretizada espacialmente en forma centrada y temporalmente mediante el esquema de Lax-Wendroff.

6. Programe los siguientes esquemas aplicados a la ecuación de advección pura:

   a.- explícito con upwind,

   b.- centrado y Lax-Friedrichs,

   c.- centrado y Lax-Wendroff,

   y ensayar su comportamiento para las siguientes condiciones iniciales:

   1.- una solución constante pero discontinua en el centro del dominio,

   2.- una onda senoidal con fase $\phi = \pi/10$,

   3.- una onda senoidal con fase $\phi = \pi/5$.

   Analice el comportamiento bajo diferentes números de Courant (σ) y saque conclusiones.

7. El esquema *leapfrog* aplicado a la ecuación de advección no estacionaria $u_t + au_x = 0$ se puede escribir como:

$$u_i^{n+1} = u_i^{n-1} - \sigma(u_{i+1}^n - u_{i-1}^n) \tag{8.100}$$

   Programe un esquema de este tipo e inicialice la solución con:

$$u(x, t=0) = e^{-\alpha x^2} \sin(2\pi k x)$$

$$\begin{aligned} \phi &= k\Delta x = \pi/4 \\ \sigma &= 0.4 \\ \Delta x &= 1/80 \\ a &= 1 \end{aligned} \tag{8.101}$$

   Estime la velocidad de grupo y analice la solución numérica obtenida comparándola con la teórica.

8. Analice el esquema *leapfrog* en combinación con una discretización espacial con upwind para resolver la ecuación de advección pura no estacionaria en 1D. Calcule la matriz de amplificación y muestre que el esquema es inestable.

9. Aplique una discretización espacial con *full upwind* en ambas direcciones a la ecuación de convección pura no estacionaria en 2D

$$u_t + au_x + bu_y = 0$$

usando un esquema explícito de primer order para integrarla en el tiempo. Realice un análisis de estabilidad de von Neumann y compare la condición de estabilidad con aquella del caso 1D.

10. Escriba un programa para analizar la estabilidad de von Neumann del esquema de Lax-Wendroff aplicado a una discretización espacial centrada en 2D para resolver la ecuación de advección pura. Grafique el factor de amplificación en función de la fase para diferentes números de Courant. Compare con el caso 1D.

Capítulo 9

Métodos iterativos para la resolución de ecuaciones lineales

9.1. Conceptos básicos de métodos iterativos estacionarios

9.1.1. Notación y repaso

Denotamos a un sistema lineal como

$$Ax = b \tag{9.1}$$

con A no-singular de $N \times N$ y $b \in R^N$. La solución del sistema la denotamos como $x^* = A^{-1}b \in R^N$, mientras que x representa una solución potencial.

Los métodos iterativos se basan en encontrar una secuencia $\{x_k\}_{k=0}^{\infty}$ tal que

$$\lim_{k\to\infty} x_k = x^* \tag{9.2}$$

Debemos asegurar la convergencia del método iterativo y si es posible determinar la *tasa de convergencia*, es decir como se comporta el error $\|x - x_k\|$ para $k \to \infty$.

Normas inducidas

Dada una norma para vectores $\| \, . \, \|$, denotamos por $\|A\|$ la norma de A inducida por $\| \, . \, \|$, definida por

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{9.3}$$

Las normas inducidas tienen la importante propiedad de que

$$\|Ax\| \leq \|A\| \, \|x\| \tag{9.4}$$

y también:

$$\|AB\| \leq \|A\| \, \|B\| \tag{9.5}$$

ya que

$$\begin{aligned} \|AB\| &= \max_{\|x\|=1} \|ABx\| && (9.6)\\ &\leq \max_{\|x\|=1} \|A\| \, \|Bx\| && (9.7)\\ &= \|A\| \max_{\|x\|=1} \|Bx\| && (9.8)\\ &= \|A\| \, \|B\| && (9.9) \end{aligned}$$

Obviamente $\|I\| = 1$ y $\|0\| = 0$.

Diferentes normas utilizadas son

- $\|A\|_2 = \sqrt{\text{maximo autovalor de}(A^T A)}$
- $\|A\|_\infty = \max_i \left(\sum_j |a_{ij}| \right)$
- $\|A\|_1 = \max_j \left(\sum_i |a_{ij}| \right)$

Norma inducida L_2 de una matriz. Sea $B = A^T A$ entonces B es simétrica y definida positiva. Sean $\{v_j, \lambda_j\}_{j=1}^N$ los autovalores de B, $v_j^T v_k = \delta_{jk}$, entonces

$$\|A\|_2^2 = \max_{x \neq 0} \frac{x^T B x}{x^T x} \tag{9.10}$$

Si $x = \sum_j \alpha_j v_j$ entonces

$$Bx = \sum_j \alpha_j \lambda_j v_j \tag{9.11}$$

y

$$\begin{aligned} x^T B x &= (\sum \alpha_k v_k^T)(\sum \alpha_j \lambda_j v_j) && (9.12)\\ &= \sum_{jk} \alpha_k \, \alpha_j \, \lambda_j \, v_k^T \, v_j && (9.13)\\ &= \sum_j \alpha_j^2 \lambda_j && (9.14) \end{aligned}$$

Por otra parte,

$$x^T x = \sum_j \alpha_j^2 \tag{9.15}$$

y

$$\|A\|_2^2 = \max_{a \in \mathbb{R}^N} \frac{\sum_j \alpha_j^2 \lambda_j}{\sum_j \alpha_j^2} = \max_j \lambda_j \tag{9.16}$$

Además, si A es simétrica, entonces es diagonalizable, con autovalores λ'_j, entonces $A^T A = A^2$ y

$$\text{autovalores de } A^2 = (\lambda'_j)^2 \tag{9.17}$$

de manera que

$$\|A\|_2 = \max |\text{autovalores de } A| \tag{9.18}$$

Norma infinito inducida de una matriz.
Por definición

$$\|x\|_\infty = \max |(x)_i| \tag{9.19}$$

y entonces, la norma inducida es

$$\begin{aligned} \|Ax\|_\infty &= \max_i |\sum_j a_{ij} x_j| \leq \max_i \sum_j |a_{ij}|\,|x_j| && (9.20)\\ &\leq \max_i \left\{ \sum_j |a_{ij}|\ \max_k |x_k| \right\} && (9.21)\\ &= \max_i \left(\sum_j |a_{ij}| \right)\ \|x\|_\infty && (9.22) \end{aligned}$$

por lo tanto

$$\|A\|_\infty \leq \max_i \left(\sum_j |a_{ij}| \right). \tag{9.23}$$

Tomemos v tal que

$$(v)_j = \operatorname{sign} a_{ij}, \quad \text{con } \ i = \operatorname*{argmax}_k \sum_j |a_{kj}| \tag{9.24}$$

entonces, es obvio que

$$\|v\|_\infty = 1 \tag{9.25}$$

y

$$\begin{aligned} |(Av)_k| &= \left| \sum a_{kj}\, v_j \right| \ \leq \sum |a_{kj}|\,|v_j| && (9.26)\\ &\leq \sum_j |a_{kj}| < \sum_j |a_{ij}| && (9.27) \end{aligned}$$

Para $k = i$ se satisface que

$$\begin{aligned} |(Av)_i| &= \left| \sum a_{ij}\, v_j \right| && (9.28)\\ &= \left| \sum a_{ij}\ \operatorname{sign} a_{ij} \right| && (9.29)\\ &= \sum |a_{ij}|. && (9.30)\\ & && (9.31) \end{aligned}$$

Por lo tanto,

$$\|Av\|_\infty = \frac{\|Av\|_\infty}{\|v\|_\infty} = \max_i \sum_j |a_{ij}| \tag{9.32}$$

Norma inducida 1 de una matriz.
Por definición,

$$\|x\|_1 = \sum_i |(x)_i| \tag{9.33}$$

y entonces,

$$\begin{aligned}
\|Ax\|_1 \quad &= \sum_i |(Ax)_i| = \sum_i |\sum_j a_{ij}\, x_j| && (9.34)\\
&\leq \sum_i \sum_j |a_{ij}|\, |x_j| = \sum_j \left(\sum_i |a_{ij}| \right)\, |x_j| && (9.35)\\
&\leq \sum_j \left[\max_k \sum_i |a_{ik}| \right]\, |x_j| && (9.36)\\
&= \left(\max_k \sum_i |a_{ik}| \right)\, \|x\|_1 && (9.37)
\end{aligned}$$

y entonces

$$\|A\|_1 \leq \max_k \sum_i |a_{ik}| \tag{9.38}$$

Sea v tal que $(v)_i = \delta_{ij}$ con $j = \text{argmax}_k \sum_i |a_{ik}|$, entonces

$$\|v\|_1 = \sum_i \delta_{ij} = 1 \tag{9.39}$$

y como

$$(Av)_i = a_{ij} \tag{9.40}$$

entonces

$$\|Av\|_1 = \sum_i |a_{ij}| = \max_k \sum_i |a_{ik}| \tag{9.41}$$

y por definición de norma inducida

$$\|A\|_1 \geq \frac{\|Av\|_1}{\|v\|_1} = \max_k \sum_i |a_{ik}| \tag{9.42}$$

y entonces, de (9.38) y (9.42) se deduce que

$$\|A\|_1 = \max_k \sum_i |a_{ik}| \tag{9.43}$$

Normas no inducidas. Es fácil demostrar que existen normas que no son inducidas, es decir que satisfacen

$$\begin{aligned}
\|A\| > 0, \ \text{ si } A \neq 0 && (9.44)\\
\|\alpha A\| = |\alpha|\, \|A\| && (9.45)\\
\|A + B\| \leq \|A\| + \|B\| && (9.46)
\end{aligned}$$

pero no provienen de ninguna norma para vectores. Por ejemplo, si definimos la norma $\|\ .\ \|_*$ como

$$\|A\|_* = c\,\|A\|_2 \tag{9.47}$$

con $c > 0, c \neq 1$, entonces es claro que es una norma pero $\|I\|_* = c \neq 1$ y por lo tanto no es inducida.

Número de condición

El número de condición de una matriz no-singular en la norma $\|\ .\ \|$ es

$$\kappa(A) = \|A\| \left\|A^{-1}\right\| \tag{9.48}$$

Podemos ver que

$$1 = \|I\| = \left\|AA^{-1}\right\| \leq \|A\| \left\|A^{-1}\right\| = \kappa(A) \tag{9.49}$$

Si A es singular tomamos como que $\kappa(A) = \infty$.

Criterios de convergencia de los métodos iterativos

Desearíamos *"detener"* un método iterativo cuando $\|e_k\| = \|x - x_k\| < \mathrm{tol}$, pero como no conocemos x^* esto es imposible en los casos prácticos. Pero sí podemos calcular el residuo de la ecuación para la iteración k como

$$r_k = b - Ax_k \tag{9.50}$$

Si $\|r_k\|$ es suficientemente pequeño esperamos que $\|e_k\|$ sea pequeño y podemos poner como criterio de detención

$$\frac{\|r_k\|}{\|r_0\|} \leq \mathrm{tol} \tag{9.51}$$

El siguiente lema nos permite relacionar ambos criterios de detención

Lema 1.1.1. Dados $b, x, x_0 \in \mathbb{R}^N$, A no-singular y $x^* = A^{-1}b$, entonces

$$\frac{\|e\|}{\|e_0\|} \leq \kappa(A)\frac{\|r\|}{\|r_0\|} \tag{9.52}$$

Demostración.

$$r = b - Ax = Ax^* - Ax = A(x - x^*) = -Ae \tag{9.53}$$

entonces

$$\|e\| = \left\|A^{-1}Ae\right\| \leq \left\|A^{-1}\right\| \|Ae\| = \left\|A^{-1}\right\| \|r\| \tag{9.54}$$

y

$$\|r_0\| \leq \|A\|\,\|e_0\| \tag{9.55}$$

Por lo tanto

$$\frac{\|e\|}{\|e_0\|} \leq \frac{\left\|A^{-1}\right\| \|r\|}{\|A^{-1}\|\,\|r_0\|} = \kappa(A)\,\frac{\|r\|}{\|r_0\|}\ \square. \tag{9.56}$$

La división por $\|e_0\|$ y $\|r_0\|$ es para *adimensionalizar* el problema. Por ejemlo, si consideramos un problema térmico, entonces x puede representar temperaturas nodales y por lo tanto tendrá dimensiones de temperatura ($^\circ$C), el miembro derecho q tendrá dimensiones de potencia (Watts) y los coeficientes $[A_ij] = \mathrm{W}/^\circ\mathrm{C}$,

pero lo importante es que el residuo tendrá las mismas dimensiones que b es decir potencia (Watts). Ahora si hacemos un cambio de unidades tomando como unidad de potencia a $\mathrm{cal/s}$ entonces el criterio de parada $\|r\| < \mathrm{tol}$ es completamente diferente, mientras que $\|r\| / \|r_0\|$ es el mismo en los dos sistemas de unidades. Por otra parte, este criterio depende de la solución inicial x_0 lo cual puede llevar a iterar demasiado en el caso en que partimos de una buena solución ($\|r_0\|$ muy pequeño). Otra posibilidad es

$$\frac{\|r_k\|}{\|b\|} < \mathrm{tol} \tag{9.57}$$

Ambos coinciden si $x_0 = 0$.

9.1.2. El lema de Banach

La forma más directa de obtener (y posteriormente analizar) un método iterativo es reescribir (9.1) como un problema de iteración de punto fijo. Una forma de hacer esto es reescribir la ecuación como

$$x = (I - A)x + b \tag{9.58}$$

lo cual induce el siguiente método iterativo de Richardson

$$x_{k+1} = (I - A)x_k + b \tag{9.59}$$

El análisis de tales secuencias recursivas es más general y vamos a estudiar la convergencia de relaciones recursivas generales de la forma

$$x_{k+1} = Mx_k + c \tag{9.60}$$

donde M es la llamada *matriz de iteración* o *matriz de amplificación*. Estos métodos son llamados métodos iterativos estacionarios porque la transición de x_k a x_{k+1} no depende de la historia anterior:

Métodos estacionarios: $x_{k+1} = f(x_k)$

Métodos no estacionarios: $x_{k+1} = f(x_k, x_{k-1}, x_{k-2}, \ldots)$

Los métodos de Krylov que discutiremos más adelante *no son métodos estacionarios*. Es interesante también ver que el esquema de Richardson puede ponerse de la forma

$$x_{k+1} = x_k + (b - Ax_k) = x_k + r_k \tag{9.61}$$

Nótese que r_k actúa como una pequeña corrección a la iteración k para obtener la siguiente $k+1$. Si estamos suficientemente cerca de la solución x^* entonces r_k será pequeño y la corrección será pequeña.

Aplicando recursivamente (9.60) obtenemos una expresión general para la iteración x_k. Asumamos por simplicidad que $x_0 = 0$, entonces

$$\begin{aligned}
x_1 &= c && (9.62)\\
x_2 &= Mc + c = (M + I)\, c && (9.63)\\
x_3 &= M(M + I)\, c + c = (M^2 + M + I)\, c && (9.64)\\
&\vdots \quad \vdots && (9.65)\\
x_k &= \left(\sum_{j=0}^{k-1} M^j \right) c && (9.66)
\end{aligned}$$

Es claro que la convergencia del método está ligada a la convergencia de la suma de M^j.

Lemma 1.2.1. Si $M \in \mathbb{R}^{N\times N}$ satisface $\|M\| < 1$ entonces $I - M$ es no-singular y

$$\left\|(I-M)^{-1}\right\| \leq \frac{1}{1-\|M\|} \tag{9.67}$$

Demostración. Veremos que la serie converge a $(1-M)^{-1}$. Sea $S_k = \sum_{l=0}^{k} M^l$. Mostraremos que es una secuencia de Cauchy

$$\begin{aligned}
\|S_k - S_m\| &= \left\| \sum_{l=k+1}^{m} M^l \right\| && (9.68)\\
&\leq \sum_{l=k+1}^{m} \left\| M^l \right\| && (9.69)\\
&\leq \sum_{l=k+1}^{m} \|M\|^l && (9.70)\\
&= \|M\|^{k+1} \left(\frac{1-\|M\|^{m-k}}{1-\|M\|} \right) && (9.71)\\
&\to 0 && (9.72)
\end{aligned}$$

para $m, k \to \infty$. Entonces $S_k \to S$ para algún S. Pero entonces tomando el límite en la relación de recurrencia

$$MS_k + I = S_{k+1} \tag{9.73}$$

obtenemos

$$MS + I = S \tag{9.74}$$

Por lo tanto

$$(I-M)S = I \tag{9.75}$$

de donde $I - M$ es no-singular y $S = (I-M)^{-1}$. Por otra parte

$$\left\|(I-M)^{-1}\right\| \leq \sum_{l=0}^{\infty} \|M\|^l = (1-\|M\|)^{-1} \,\square. \tag{9.76}$$

Matrices diagonalizables bajo la norma 2. En el caso en que M es diagonalizable y consideramos la norma $\|M\|_2$ es más fácil de visualizar las implicancias del lema. Sean S no-singular y Λ diagonal las matrices que dan la descomposición diagonal de M

$$\begin{aligned}
M &= S^{-1}\Lambda S && (9.77)\\
(I-M) &= S^{-1}(I-\Lambda)S && (9.78)
\end{aligned}$$

Como $\|M\| = \max_j |\lambda_j| < 1$ esto quiere decir que todos los autovalores λ_j de M, que son reales ya que M es simétrica, están estrictamente contenidos en el intervalo $-1 < \lambda_j < 1$. Por otra parte

$$\begin{aligned} \left\|(I-M)^{-1}\right\|_2 &= \max_j \left|\frac{1}{1-\lambda_j}\right| && (9.79)\\ &= \frac{1}{\min|1-\lambda_j|} && (9.80)\\ &= \frac{1}{1-\max\lambda_j} && (9.81)\\ &\leq \frac{1}{1-\max|\lambda_j|} = \frac{1}{1-\|M\|_2} && (9.82) \end{aligned}$$

Corolario 1.2.1. Si $\|M\| < 1$ entonces la iteración (9.60) converge a $x = (I-M)^{-1} c$ para cualquier punto inicial x_0.

Demostración. Si $x = x_0$ ya está demostrado. Si $x \neq x_0$ haciendo el cambio de variables $x'_k = x_k - x_0$ llegamos al esquema recursivo

$$\begin{aligned} x_{k+1} - x_0 &= M(x_k - x_0) + c - (I-M)x_0 && (9.83)\\ x'_{k+1} &= Mx'_k + c' && (9.84) \end{aligned}$$

El cual converge y converge a $(I-M)^{-1} c'$, por lo tanto x_k converge a

$$\begin{aligned} (I-M)^{-1} c' + x_0 &= (I-M)^{-1}[c-(I-M)x_0] + x_0 && (9.85)\\ &= (I-M)^{-1} c \,\square\,. && (9.86) \end{aligned}$$

Una consecuencia del corolario es que la iteración de Richardson (1.6) converge si $\|I-A\| < 1$. A veces podemos precondicionar el sistema de ecuaciones multiplicando ambos miembros de (9.1) por una matriz B

$$BA\,x = Bb \tag{9.87}$$

de manera que la convergencia del método iterativo es mejorada. En el contexto de la iteración de Richardson las matrices B tales que permiten aplicar el Lema de Banach y su corolario se llaman *inversas aproximadas*

Definición 1.2.1. B es una inversa aproximada de A si $\|I-BA\| < 1$.

El siguiente teorema es llamado comúnmente *Lema de Banach*.

Teorema 1.2.1. Si A y $B \in \mathbb{R}^{N\times N}$ y B es una inversa paroximada de A, entonces A y B son no singulares y

$$\left\|A^{-1}\right\| \leq \frac{\|B\|}{1-\|I-BA\|}, \quad \left\|B^{-1}\right\| \leq \frac{\|A\|}{1-\|I-BA\|}, \tag{9.88}$$

y

$$\left\|A^{-1}-B\right\| \leq \frac{\|B\|\,\|I-BA\|}{1-\|I-BA\|}, \quad \left\|A-B^{-1}\right\| \leq \frac{\|A\|\,\|I-BA\|}{1-\|I-BA\|}, \tag{9.89}$$

Demostración. Sea $M = I - BA$, entonces $I - M = BA$ es no singular y por lo tanto B y A son no-singulares y

$$\left\|(BA)^{-1}\right\| = \left\|A^{-1}B^{-1}\right\| \leq \frac{1}{1-\|I-BA\|} \tag{9.90}$$

y

$$\left\|A^{-1}\right\| \leq \left\|A^{-1}B^{-1}\right\| \|B\| \leq \frac{\|B\|}{1-\|I-BA\|}. \tag{9.91}$$

Por otra parte,

$$\left\|A^{-1}-B\right\| = \left\|(I-BA)A^{-1}\right\| \leq \frac{\|B\|\,\|I-BA\|}{1-\|I-BA\|} \tag{9.92}$$

La demostración de las otras desigualdades es similar. □

Notar que hay una cierta simetría en el rol jugado por A y B. De hecho deberíamos definir inversa aproximada *por derecha* y *por izquierda* y B sería la inversa aproximada por derecha de A.

La iteración de Richardson, precondicionada aproximadamente, tiene la forma

$$x_{k+1} = (I-BA)\,x_k + Bb \tag{9.93}$$

Si $\|I-BA\| \ll 1$ las iteraciones convergen rápido y además, (por el Lema 1.2.1) las decisiones basadas en el residuo precondicionado $\|B(b-A\,x)\|$ reflejarán muy bien el error cometido.

9.1.3. Radio espectral

El análisis de §9.1.2 relacionó la convergencia del esquema iterativo (9.60) a la norma de la matriz M. Sin embargo la norma de M puede ser pequeña en alguna norma y grande en otras. De aquí que la performance de la iteración no sea completamente descripta por $\|M\|$. El concepto de *radio espectral* da una descripción completa. Sea $\sigma(A)$ el conjunto de autovalores de A.

Definición 1.3.1. El radio espectral de A es

$$\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda| = \lim_{n\to\infty} \|A^n\|^{1/n} \tag{9.94}$$

El radio espectral de A es independiente de la norma particular de M. De hecho

$$\rho(A) \leq \|A\| \tag{9.95}$$

ya que, si $Av = \lambda_{\max} v$, entonces

$$\|A\| \geq \frac{\|Av\|}{\|v\|} = |\lambda_{\max}| = \rho(A) \tag{9.96}$$

Siempre que $\|\,.\,\|$ sea una norma inducida. En realidad puede demostrarse algo así como que $\rho(A)$ es el ínfimo de todas las normas de A. Esto es el enunciado del siguiente teorema (que no demostraremos aquí).

Teorema 1.3.1. Sea $A \in \mathbb{R}^{N\times N}$. Para cada $\epsilon > 0$ existe una norma inducida $\|\,.\,\|$ tal que $\rho(A) > \|A\| - \epsilon$.

Puede verse que, si $\rho(M) \geq 1$ entonces existen x_0 y c tales que (9.60) diverge. Efectivamente, sea v el autovalor tal que $Mv = \lambda v$ y $|\lambda| = \rho(M) \geq 1$. Tomemos $x_0 = v$, $c = 0$, entonces

$$x_k = M^k\,x_0 = \lambda^k\,x_0 \tag{9.97}$$

es claro que no converge.

Teorema 1.3.2. Sea $M \in \mathbb{R}^{N\times N}$. La iteración (9.60) converge para todos $x_0, c \in \mathbb{R}^{N\times N}$ sy y solo si $\rho(M) < 1$.

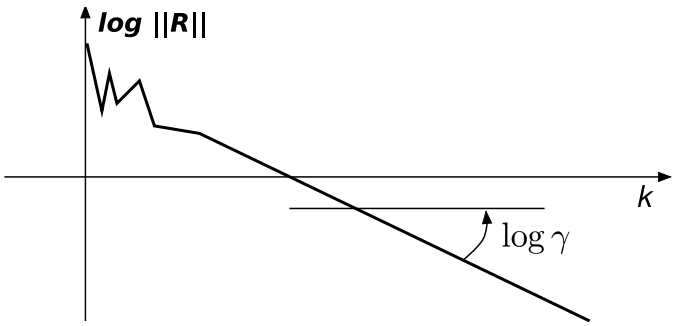


Figura 9.1: Historia de convergencia típica

Demostración. Si $\rho(M) > 1$ entonces en cualquier norma tal que $\|M\| > 1$ la iteración no converge por el párrafo anterior. Por otra parte, si $\rho(M) < 1$ entonces, tomando $\epsilon = (1 - \rho(M))/2$, existe una norma tal que

$$\|M\| < \rho(M) + \epsilon = \frac{1}{2}(1 + \rho(M)) < 1 \tag{9.98}$$

y por lo tanto converge. □

Tasa de convergencia de métodos estacionarios. Para un esquema convergente de la forma

$$x_{k+1} = (I - BA)\, x_k + Bb \tag{9.99}$$

podemos estimar la tasa de convergencia como,

$$\begin{aligned} x_k - x^* &= (I - BA)\, x_{k-1} + BA\, x^* - x^* && (9.100)\\ &= (I - BA)\, (x_{k-1} - x^*) && (9.101) \end{aligned}$$

y por lo tanto

$$\begin{aligned} \|x_k - x^*\| &\leq \|I - BA\|\ \|x_{k-1} - x^*\| && (9.102)\\ &\leq \|I - BA\|^k\ \|x_0 - x^*\| && (9.103)\\ & && (9.104) \end{aligned}$$

Es usual visualizar la convergencia del método graficando $\|r_k\|$ versus k. Como $\|r_k\|_2$ puede reducirse en varios órdenes de magnitud durante la iteración, es usual usuar ejes logarítmicos para $\|r_k\|_2$. Por otra parte (9.103) no sólo es una estimación de la tasa de convergencia sino que, de hecho, muchas veces el residuo de la ecuación termina comportándose de esta forma, después de un cierto transitorio inicial (ver figura) es decir

$$\|r_k\| \sim \gamma^k\, \|r_0\| \tag{9.105}$$

Este tipo de comportamiento se refleja en una recta de pendiente $\log \gamma$ en el gráfico. Un índice de la velocidad de convergencia es el número de iteraciones n necesario para bajar el residuo un factor 10,

$$\begin{aligned} \|r_{k+n}\| &= 1/10\ \|r_k\| && (9.106)\\ \gamma^{k+n}\, \|r_0\| &= 1/10\, \gamma^k\, \|r_0\| && (9.107)\\ \log 10 = -n \log \gamma, \quad n &= \frac{\log 10}{\log(1/\gamma)} && (9.108) \end{aligned}$$

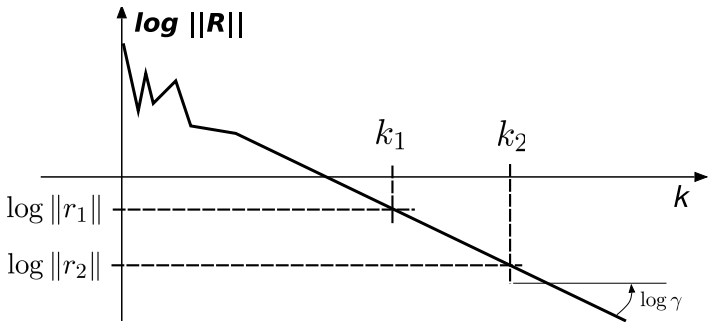


Figura 9.2: Estimación de la tasa de convergencia

para problemas muy mal condicionados

$$\gamma = 1 - \epsilon, \quad \epsilon \ll 1 \tag{9.109}$$

y entonces,

$$\log(1/\gamma) \sim \epsilon, \quad n = \frac{\log 10}{\epsilon} \tag{9.110}$$

A veces podemos calcular la tasa de convergencia "experimentalmente" conociendo el valor del residuo $\|r_k\|$ para dos iteraciones k_1, k_2. Asumiendo que en el intervalo $k_1 \leq k \leq k_2$ la tasa de convergencia es constante como en (9.105) (ver figura 9.2), entonces

$$\|r_2\| = \gamma^{k_2 - k_1} \|r_1\| \tag{9.111}$$

de donde

$$\log \gamma = \frac{\log(\|r_2\| \,/\, \|r_1\|)}{k_2 - k_1} \tag{9.112}$$

y entonces,

$$n = \frac{(\log 10)\,(k_2 - k_1)}{\log(\|r_1\| \,/\, \|r_2\|)} = \frac{(k_2 - k_1)}{\log_{10}(\|r_1\| \,/\, \|r_2\|)} \tag{9.113}$$

9.1.4. Saturación del error debido a los errores de redondeo.

A medida que vamos iterando el residuo va bajando en magnitud. Cuando el valor del residuo pasa por debajo de cierto valor umbral dependiendo de la precisión de la máquina ($\sim 10^{-15}$ en Octave, Fortran (`real *8`) o C (tipo `double`)) se produce, debido a errores de redondeo, un efecto de *saturación* (ver figura 9.3). Es decir, al momento de calcular (9.50), es claro que incluso reemplazando x_k por la solución exacta x^* el residuo (calculado en una máquina de precisión finita) dara un valor no nulo del orden de la precisión de la máquina. Por supuesto este umbral de saturación es relativo a la norma de cada uno de los términos intervinientes y además para sistemas mal condicionados, el umbral se alcanza antes por un factor $\kappa(A)$, es decir que

$$\|r\|_{\text{sat}} \approx 10^{-15} \times \kappa(A) \times \|b\| \tag{9.114}$$

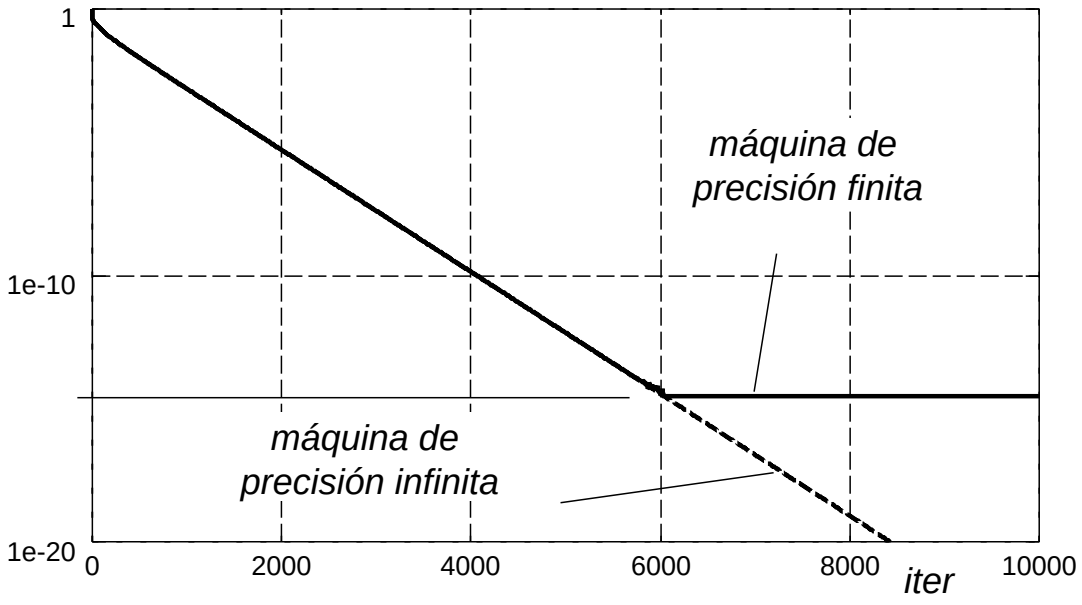


Figura 9.3: Saturación del error por errores de redondeo.

9.1.5. Métodos iterativos estacionarios clásicos

Hay otras formas alternativas a (9.58) de llevar $Ax = b$ a un problema de punto fijo. Los métodos como Jacobi, Gauss-Seidel y relajaciones sucesivas se basan en descomposiciones de A (*"splittings"*) de la forma,

$$A = A_1 + A_2 \tag{9.115}$$

con A_1 no-singular y fácil de factorizar. El nuevo problema de punto fijo es

$$x = A_1^{-1}\left(b - A_2 x\right) \tag{9.116}$$

El análisis del método se basa en estimar el radio espectral de $M = -A_1^{-1}A_2$.

Iteración de Jacobi. Corresponde a tomar

$$A_1 \quad = D = \mathrm{diag}(A) \tag{9.117}$$
$$A_2 \quad = L + U = A - D \tag{9.118}$$

donde L, U, D son las partes triangular inferior, superior y diagonal de $A = L + U + D$. El esquema iterativo es

$$(x_{k+1})_i = a_{ii}^{-1}\left(b_i - \sum_{j\neq i} a_{ij}(x_k)_j\right) \tag{9.119}$$

Notar que A_1 es diagonal y, por lo tanto, trivial de invertir. La matriz de iteración correspondiente es

$$M_{\mathrm{Jac}} = -D^{-1}\left(L + U\right) \tag{9.120}$$

Teorema 1.4.1. Sea $A \in \mathbb{R}^{N\times N}$ y asumamos que para cualquier $1 \leq i \leq N$

$$0 < \sum_{j\neq i} |a_{ij}| < |a_{ii}| \tag{9.121}$$

entonces A es no-singular y Jacobi converge.

Demostración. Veamos que

$$\sum_{j=1}^{N} |m_{ij}| = \frac{\sum_{j\neq i} |a_{ij}|}{|a_{ii}|} < 1 \tag{9.122}$$

Entonces,

$$\|M_{\text{Jac}}\|_\infty = \max_i \sum_j |m_{ij}| < 1 \tag{9.123}$$

Por lo tanto la iteración converge a la solución de

$$x = Mx + D^{-1}b \tag{9.124}$$
$$x = (I - D^{-1}A)x + D^{-1}b \tag{9.125}$$

entonces $Ax = b$ y $I - M = D^{-1}A$ es no-singular y A es no-singular.

Iteración de Gauss-Seidel. En este esquema se reemplaza la solución aproximada con el nuevo valor tan pronto como éste es calculado,

$$(x_{k+1})_i = a_{ii}^{-1}\left(b_i - \sum_{j<i} a_{ij}\,(x_{k+1})_j - \sum_{j>i} a_{ij}\,(x_k)_j\right) \tag{9.126}$$

que puede escribirse como

$$(D + L)\,x_{k+1} = b - U\,x_k \tag{9.127}$$

El split correspondiente es

$$A_1 = D + L, \quad A_2 = U \tag{9.128}$$

y la matriz de iteración

$$M_{\text{GS}} = -(D + L)^{-1}U \tag{9.129}$$

A_1 es triangular inferior y por lo tanto A_1^{-1} es fácil de calcular. Notar también que a diferencia de Jacobi, depende del ordenamiento de las incógnitas. También podemo hacer un "backward Gauss-Seidel" con el splitting

$$A_1 = D + U, A_2 = L \tag{9.130}$$

La iteración es

$$(D + U)\,x_{k+1} = (b - Lx_k) \tag{9.131}$$

y la matriz de iteración

$$M_{\text{BGS}} = -(D + U)^{-1}L \tag{9.132}$$

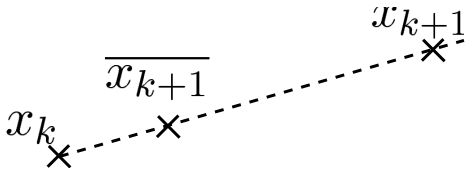


Figura 9.4: Aceleración de la convergencia por sobre-relajación

Gauss-Seidel simétrico. Podemos combinar alternando una iteración de forward GS con una de backward GS poniendo

$$(D+L)\,x_{k+1/2} = b - Ux_k \tag{9.133}$$
$$(D+U)\,x_{k+1} = b - Lx_{k+1/2} \tag{9.134}$$

La matriz de iteración es

$$M_{\mathrm{SGS}} = M_{\mathrm{BGS}}\,M_{\mathrm{GS}} = (D+U)^{-1}\,L\,(D+L)^{-1}\,U \tag{9.135}$$

Si A es simétrica, entonces $U = L^T$ y

$$M_{\mathrm{SGS}} = (D+L^T)^{-1}\,L\,(D+L)^{-1}\,L^T \tag{9.136}$$

Queremos escribir estos esquemas como una iteración de Richardson precondicionada, es decir que queremos encontrar B tal que $M = I - BA$ y usar B como inversa aproximada. Para la iteración de Jacobi,

$$B_{\mathrm{Jac}} = D^{-1} \tag{9.137}$$

y para Gauss-Seidel simétrico

$$B_{\mathrm{SGS}} = (D+L^T)^{-1}\,D\,(D+L)^{-1} \tag{9.138}$$

Verificación. Debemos demostrar que $M_{\mathrm{SGS}} = I - B_{\mathrm{SGS}}\,A$. Efectivamente,

$$\begin{aligned}
I - B_{\mathrm{SGS}}\,A &= I - (D+L^T)^{-1}\,D\,(D+L)^{-1}\,(D+L+L^T) && (9.139)\\
&= I - (D+L^T)^{-1}\,D\,(I+(D+L)^{-1}L^T) && (9.140)\\
&= (D+L^T)^{-1}\,[D+L^T-D-D(D+L)^{-1}L^T] && (9.141)\\
&= (D+L^T)^{-1}\,[I-D(D+L)^{-1}]\,L^T && (9.142)\\
&= (D+L^T)^{-1}\,[(D+L)-D]\,(D+L)^{-1}\,L^T && (9.143)\\
&= (D+L^T)^{-1}\,L\,(D+L)^{-1}\,L^T && (9.144)\\
&= M_{\mathrm{SGS}} && (9.145)
\end{aligned}$$

Sobrerelajación. Consideremos iteración simple de Richardson

$$x_{k+1} = x_k + (b - Ax_k) \tag{9.146}$$

Si vemos que x_k se acerca monótonamente y lentamente a x^* podemos pensar en "acelerarlo" (ver figura 9.4) diciendo que el método iterativo en realidad nos predice un estado intermedio $\overline{x_{k+1}}$, y buscamos el vector de iteración x_{k+1} sobre la recta que une x_k con $\overline{x_{k+1}}$

$$\overline{x_{k+1}} = x_k + (b - Ax_k) \tag{9.147}$$
$$x_{k+1} = x_k + \omega(\overline{x_{k+1}} - x_k) \tag{9.148}$$

Veremos más adelante que el valor óptimo de ω está relacionado con la distribución de autovalores de la matriz de iteración en el plano complejo. Mientras tanto podemos ver intuitivamente que

- $\omega = 1$ deja el esquema inalterado
- $\omega > 1$ tiende a acelerar la convergencia si el esquema converge lenta y monótonamente
- $\omega < 1$ tiende a desacelerar la convergencia si el esquema se hace inestable.

Esquema iterativo de Richardson con relajación para matrices spd. Aplicando el método de relajación a la iteración básica de Richardson obtenemos el esquema

$$x_{k+1} = x_k + \omega r_k \tag{9.149}$$

que puede reescribirse como

$$x_{k+1} = (I - \omega A)\, x_k + \omega b \tag{9.150}$$

de manera que la matriz de iteración es

$$M_{\mathrm{SR}} = I - \omega A \tag{9.151}$$

Asumiendo que A es *simétrica y definida positiva (spd)*, el espectro de M_{SR} esta dado por

$$\sigma(M_{\mathrm{SR}}) = 1 - \omega\sigma(A) \tag{9.152}$$

pero como A es spd, los autovalores $\lambda \in \sigma(A)$ son reales y positivos. Asumamos que están ordenados $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. También denotaremos $\lambda_{\min} = \lambda_N$, $\lambda_{\max} = \lambda_1$. Los autovalores de M_{SR} se comportan en función de ω como se observa en la figura 9.5. Para un dado ω todos los autovalores de M_{SR} están comprendidos entre los autovalores correspondientes a $\lambda_{\min}$ y $\lambda_{\max}$ (la región rayada de la figura). Para ω suficientemente chico todos los autovalores se concentran cerca de la unidad. Para un cierto valor ω_{crit} el autovalor correspondiente a $\lambda_{\max}$ se pasa de -1 con lo cual la iteración no converge. El valor de ω_{crit} está dado entonces por

$$1 - \omega_{\mathrm{crit}}\lambda_{\max} = -1, \quad \omega_{\mathrm{crit}} = \frac{2}{\lambda_{\max}} \tag{9.153}$$

La tasa de convergencia está dada por $\gamma = \max_j |1 - \omega\lambda_j|$, y queremos buscar el ω que da la mejor tasa de convergencia, es decir el mínimo γ. Como los $1 - \omega\lambda_j$ están comprendidos en el intervalo $1 - \omega\lambda_{\min}, 1 - \omega\lambda_{\max}$, se cumple que

$$\gamma = \max(|1 - \omega\lambda_{\min}|, |1 - \omega\lambda_{\max}|) \tag{9.154}$$

Ahora bien, para un cierto intervalo de valores $0 < \omega < \omega_{\mathrm{opt}}$ el máximo corresponde a $1 - \omega\lambda_{\min}$ y γ decrece al crecer ω, mientras que para $\omega_{\mathrm{opt}} < \omega < \omega_{\mathrm{crit}}$ el máximo está dado por $-1 + \omega\lambda_{\max}$ y crece con ω. Para $\omega = \omega_{\mathrm{opt}}$ ambos valores coinciden y entonces

$$1 - \omega_{\mathrm{opt}}\lambda_{\min} = -1 + \omega_{\mathrm{opt}}\lambda_{\max} \tag{9.155}$$

de donde

$$\omega_{\rm opt} = \frac{2}{\lambda_{\max} + \lambda_{\min}} \tag{9.156}$$

Además es fácil ver que γ es mínimo para $\omega = \omega_{\rm opt}$, de ahí el nombre de *coeficiente de relajación óptimo*.

Podemos ver también que, para números de condición muy altos el valor óptimo se encuentra muy cerca del valor crítico,

$$\omega_{\rm opt} = \frac{1}{1 + \kappa(A)^{-1}}\, \omega_{\rm crit} \sim \omega_{\rm crit} \tag{9.157}$$

La tasa de convergencia que se obtiene para $\omega_{\rm opt}$ es de

$$n_{\rm opt} = \frac{\log 10}{\log(1/\gamma_{\rm opt})} \tag{9.158}$$

$$= \frac{\log 10}{\log[1 - 2\lambda_{\min}/(\lambda_{\max} + \lambda_{\min})]} \tag{9.159}$$

$$= \frac{\log 10}{\log[(\kappa + 1)/(\kappa - 1)]} \tag{9.160}$$

que para sistemas mal condicionados ($\kappa \gg 1$) es

$$n_{\rm opt} = \frac{\kappa \log 10}{2} \sim 1.15\, \kappa \tag{9.161}$$

Método de relajaciones sucesivas. La combinación de Gauss-Seidel puede mejorarse dramáticamente con sobre-relajación para una cierta elección apropiada del parámetro de relajación. Este método es muy popular y se llama SSOR por *"Successive Standard Over-Relaxation"*. Partiendo de (9.127) y reescribiéndolo como

$$D\, \overline{x_{k+1}} + L\, x_{k+1} = b - U\, x_k \tag{9.162}$$

y combinando con la sobre-relajación estándar (9.148)

$$\overline{x_{k+1}} = \omega^{-1}(x_{k+1} - (1 - \omega)\, x_k) \tag{9.163}$$

llegamos a

$$D[x_{k+1} - (1 - \omega)\, x_k] + \omega L x_{k+1} = \omega\, (b - U\, x_k) \tag{9.164}$$

$$(D + \omega\, L)x_{k+1} = [(1 - \omega)\, D - \omega U]x_k + \omega b \tag{9.165}$$

de manera que la matriz de iteración es

$$M_{\rm SOR} = (D + \omega\, L)^{-1}\, [(1 - \omega)\, D - \omega\, U] \tag{9.166}$$

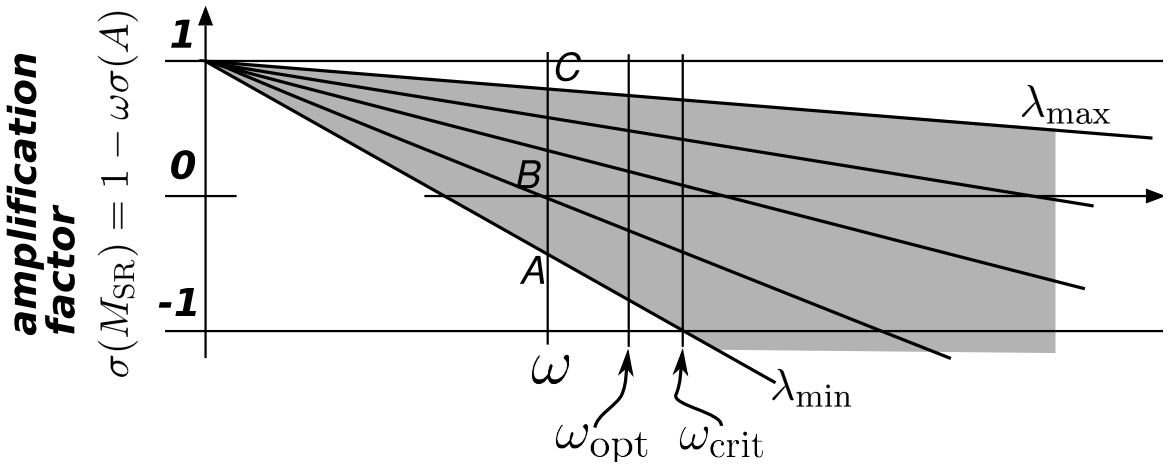


Figura 9.5: Comportamiento de los autovalores de la matriz de iteración en función del parámetro de relajación

9.2. Método de Gradientes Conjugados

9.2.1. Métodos de Krylov y propiedad de minimización

Los métodos de Krylov (a diferencia de los métodos estacionarios que vimos hasta ahora), no tienen una matriz de iteración. Los que describiremos más en profundidad son Gradientes Conjugados y GMRES. Ambos se basan en que la k-ésima iteración minimiza alguna medida del error en el espacio afín

$$x_0 + \mathcal{K}_k \tag{9.167}$$

donde x_0 es la iteración inicial y el subespacio de Krylov $\mathcal{K}_k$ es

$$\mathcal{K}_k = \mathrm{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{k-1} r_0\}, \quad k \geq 1 \tag{9.168}$$

Nota: El papel de x_0 y b puede intercambiarse, lo importante es r_0 (esto vale prácticamente para todos los métodos iterativos). De hecho, el esquema es invariante ante translaciones del origen, es decir las iteraciones para los dos sistemas siguientes

$$Ax = b, \quad \text{inicializando en } x_0 \tag{9.169}$$
$$Ax' = b', \quad \text{inicializando en } x'_0 \tag{9.170}$$

con

$$x' = x - \bar{x} \tag{9.171}$$
$$b' = b + A\bar{x} \tag{9.172}$$
$$x'_0 = x_0 - \bar{x} \tag{9.173}$$

son equivalentes (es decir, $x'_k = x_k - \bar{x}$), ya que

$$\begin{aligned} r'_0 &= b' - Ax'_0 && (9.174)\\ &= b + A\bar{x} - Ax_0 - A\bar{x} && (9.175)\\ &= b - Ax_0 && (9.176)\\ &= r_0 && (9.177) \end{aligned}$$

de manera que el espacio de Krylov es el mismo. Entonces podemos, o bien eliminar b haciendo $\bar{x} = x^*$ o eliminar x_0 haciendo $\bar{x} = x_0$.

El residuo es $r = b - Ax$, de manera que $\{r_k\}$ para $k \geq 0$ denota la secuencia de residuos $r_k = b - Ax_k$. Como antes, $x^* = A^{-1}b$, es la solución del sistema. El método de GC es en realidad un método directo, en el sentido de que llega a la solución en un número finito de pasos, de hecho veremos más adelante que converge en N (o menos) iteraciones, es decir que

$$x_k = x^*, \text{ para un cierto } k \text{ con } k \leq N \tag{9.178}$$

GC es un método que sirve en principio sólo para matrices (spd). Recordemos que A es simétrica si $A = A^T$ y definida positiva si

$$x^T Ax > 0, \text{ para todo } x \neq 0 \tag{9.179}$$

Luego veremos que podemos extender cualquier sistema (no simétrico ni definido positivo) a un sistema spd. Como A es spd. podemos definir una norma como

$$\|x\|_A = \sqrt{x^T Ax} \tag{9.180}$$

es la llamada *"norma A"* o *"norma energía"* ya que en muchos problemas prácticos el escalar resultante ($\|x\|_A^2$) representa la energía contenida en el campo representado por el vector x.

El esquema a seguir será

- Descripción formal del método y consecuencias de la propiedad de minimización
- Criterio de terminación, performance, precondicionamiento
- Implementación

La k-ésima iteración de GC x_k minimiza el funcional cuadrático

$$\phi(x) = (1/2)x^T Ax - x^T b \tag{9.181}$$

en $x_0 + \mathcal{K}_k$.

Notemos que si hacemos el mínimo sobre todo $\mathbb{R}^N$, entonces si

$$\tilde{x} = \underset{x \in \mathbb{R}^N}{\operatorname{argmin}}\, \phi(x), \tag{9.182}$$

vale que

$$\nabla\phi(\tilde{x}) = A\tilde{x} - b = 0, \quad \text{implica } \tilde{x} = x^* \tag{9.183}$$

Lema 2.1.1. Sea $S \in \mathbb{R}^N$, si x_k minimiza ϕ sobre S, entonces también minimiza $\|x^* - x\|_A = \|r\|_{A^{-1}}$ sobre S.

Demostración. Notemos que

$$\|x - x^*\|_A^2 = (x - x^*)^T A (x - x^*) \tag{9.184}$$
$$= x^T A x - x^{*T} A x - x^T A x^* + x^{*T} A x^* \tag{9.185}$$

pero $A = A^T$, entonces $x^{*T} A x = x^T A^T x^* = x^T A x^*$ y $Ax^* = b$, de manera que

$$\|x - x^*\|_A^2 = x^T Ax - 2x^{*T} b + x^{*T} A x^* \tag{9.186}$$
$$= 2\phi(x) + x^{*T} A x^* \tag{9.187}$$

Pero $x^{*T} A x^*$=cte de manera que x_k minimiza $\|x - x^*\|_A$. Sea $e = x - x^*$, entonces,

$$\|e\|_A^2 = e^T A e \tag{9.188}$$
$$= [A(x - x^*)]^T A^{-1} [A(x - x^*)] \tag{9.189}$$
$$= (b - Ax)^T A^{-1} (b - Ax) \tag{9.190}$$
$$= \|b - Ax\|_{A^{-1}}^2 \ \square. \tag{9.191}$$

Usaremos este lema para el caso $S = x_0 + \mathcal{K}_k$.

9.2.2. Consecuencias de la propiedad de minimización.

El lema 2.1.1 implica que, como x_k minimiza ϕ sobre $x_0 + \mathcal{K}_k$,

$$\|x^* - x_k\|_A \leq \|x^* - w\|_A \tag{9.192}$$

para todo $w \in x_0 + \mathcal{K}_k$. Como $w \in x_0 + \mathcal{K}_k$ puede ser escrito como

$$w = \sum_{j=0}^{k-1} \gamma_j A^j r_0 + x_0 \tag{9.193}$$

para ciertos coeficientes $\{\gamma_j\}$, podemos expresar $x^* - w$ como

$$x^* - w = x^* - x_0 - \sum_{j=0}^{k-1} \gamma_j A^j r_0 \tag{9.194}$$

Pero

$$r_0 = b - Ax_0 = A(x^* - x_0) \tag{9.195}$$

entonces

$$x^* - w = (x^* - x_0) - \sum_{j=0}^{k-1} \gamma_j A^{j+1} (x^* - x_0) \tag{9.196}$$
$$= p(A) (x^* - x_0) \tag{9.197}$$

donde el polinomio

$$p(z) = 1 - \sum_{j=0}^{k-1} \gamma_j \, z^{j+1} \tag{9.198}$$

tiene grado k y satisface $p(0) = 1$. Entonces,

$$\|x^* - x_k\|_A = \min_{p \in \mathcal{P}_k, p(0)=1} \|p(A)\,(x^* - x_0)\|_A \tag{9.199}$$

$\mathcal{P}_k$ denota el conjunto de los polinomios de grado k. El teorema espectral para matrices spd afirma que existe una base de autovectores $\{u_i\}_{i=1}^N$ con autovalores $\{\lambda_i\}$

$$A\,u_i = \lambda_i\,u_i \tag{9.200}$$

con u_i, λ_i reales, $\lambda_i > 0$ y los u_i ortogonales entre sí, es decir que

$$u_i^T u_j = \delta_{ij} \tag{9.201}$$

Además formamos las matrices

$$U = \left[\begin{array}{cccc} u_1 & u_2 & \ldots & u_N \end{array}\right] \tag{9.202}$$

$$\Lambda = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\} \tag{9.203}$$

La descomposición diagonal de A es

$$A = U\,\Lambda\,U^T \tag{9.204}$$

Además

$$A^j = (U\Lambda U^T)\,(U\Lambda U^T)\,\ldots\,(U\Lambda U^T) \tag{9.205}$$

$$= U\Lambda^j U^T \tag{9.206}$$

y

$$p(A) = U\,p(\Lambda)\,U^T \tag{9.207}$$

Definamos

$$A^{1/2} = U\,\Lambda^{1/2}\,U^T \tag{9.208}$$

y notemos que

$$\|x\|_A^2 = x^T\,A\,x = \left\|A^{1/2}\,x\right\|_2^2 \tag{9.209}$$

Entonces, para todo $x \in \mathbb{R}^N$

$$\|p(A)\,x\|_A = \left\|A^{1/2}\,p(A)\,x\right\|_2 \tag{9.210}$$

$$= \left\|p(A)\,A^{1/2}\,x\right\|_2 \tag{9.211}$$

$$\leq \|p(A)\|_2 \left\|A^{1/2}\,x\right\|_2 \tag{9.212}$$

$$\|p(A)\|_2\ \|x\|_A \tag{9.213}$$

y volviendo a (9.199)

$$\|x_k - x^*\|_A \leq \|x_0 - x^*\|_A \min_{p\in\mathcal{P}_k, p(0)=1} \left\{ \max_{z\in\sigma(A)} |p(z)| \right\} \tag{9.214}$$

donde $\sigma(A)$ es el conjunto de autovalores de A.

Corolario 2.2.1. Sea A spd y $\{x_k\}$ las iteraciones de GC. Sea $\bar{p}_k$ cualquier polinomio de grado k tal que $\bar{p}_k(0) = 1$, entonces

$$\frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \leq \max_{z\in\sigma(A)} |\bar{p}_k(z)| \tag{9.215}$$

Los polinomios que satisfacen $\bar{p}_k(0) = 1$ se llaman polinomios residuales.

Definición 2.2.1. El conjunto de polinomios residuales de orden k es

$$\mathcal{P}_k = \{p \ / \ p \text{ es un polinomio de grado } k \text{ y } p(0) = 1\} \tag{9.216}$$

La forma de estimar la convergencia de GC es construir secuencias de polinomios residuales basados en la información de como están distribuidos los autovalores de A (es decir de $\sigma(A)$). Un primer resultado es ver que GC es un método directo

Teorema 2.2.1. Sea A spd. Entonces GC converge antes de las N iteraciones.

Demostración. Sea $\{\lambda_i\}_{i=1}^N$ los autovalores de A. Tomemos como polinomio residual

$$\bar{p}(z) = \prod_{i=1}^{N} (\lambda_i - z)/\lambda_i \tag{9.217}$$

Pero $\bar{p} \in \mathcal{P}_N$ ya que tiene grado N y $\bar{p}(0) = 1$. Entonces, de la estimación de error (9.215)

$$\|x_N - x^*\|_A \leq \|x_0 - x^*\|_A \max_{z\in\sigma(A)} |\bar{p}(z)| = 0 \tag{9.218}$$

ya que, por construcción, $\bar{p}(z) = 0$ para todos los $z \in \sigma(A)$. □.

Sin embargo, desde el punto de vista práctico esto no es tan bueno como suena. Veremos que, bajo ciertas condiciones, la convergencia puede ser muy lenta y desde el punto de vista práctica haya que esperar hasta N iteraciones para que el residuo baje de la tolerancia aceptable.

Es decir, si evaluamos a GC como un método iterativo, entonces debemos poder evaluar la tasa de convergencia para $k < N$ (la pendiente de la curva de convergencia).

Teorema 2.2.2. Sea A spd con autovalores $\{u_i\}_{i=1}^N$. Sea b una combinación lineal de k de los autovectores de A. Por simplicidad asumiremos que los autovectores de A están ordenados de manera que estos autovectores son los primeros k

$$b = \sum_{l=1}^{k} \gamma_l \, u_l \tag{9.219}$$

Entonces la iteración de GC para $Ax = b$ con $x_0 = 0$ termina en, a lo sumo, k iteraciones

Demostración. Por el teorema espectral

$$x^* = \sum_{l=1}^{k} (\gamma_l/\lambda_l) \, u_l \tag{9.220}$$

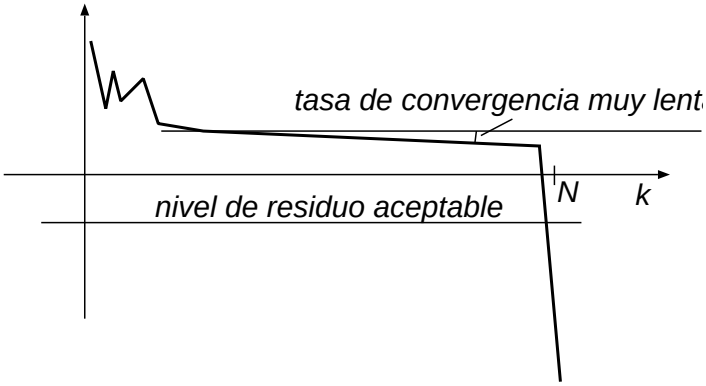


Figura 9.6: GC con tasa de convergencia muy lenta.

ya que

$$Ax^* = \sum_{l=1}^{k} (\gamma_l/\lambda_l) A u_l \tag{9.221}$$

$$= \sum_{l=1}^{k} (\gamma_l/\lambda_l) \lambda_l \, u_l \tag{9.222}$$

$$= b \tag{9.223}$$

Usamos el polinomio residual

$$\bar{p}(z) = \prod_{l=1}^{k} (\lambda_l - z)/\lambda_l \tag{9.224}$$

y puede verse que $\bar{p} \in \mathcal{P}_k$ y $\bar{p}(\lambda_l) = 0$ para $1 \leq l \leq k$ y

$$\bar{p}(A)\, x^* = \sum_{l=1}^{k} \bar{p}(\lambda_l)\, (\gamma_l/\lambda_l)\, u_l = 0 \tag{9.225}$$

De manera que, por (9.199) y $x_0 = 0$ se deduce que

$$\|x^* - x_k\|_A \leq \|\bar{p}(A)\, x^*\|_A = 0 \; \square. \tag{9.226}$$

Teorema 2.2.3. Sea A spd y supongamos que hay exactamente $k \leq N$ autovalores distintos de A. Entonces GC termina en, a lo sumo, k iteraciones

Demostración. Usar el polinomio residual

$$\bar{p}_k(z) = \prod_{l=1}^{k} (\lambda_l - z)/\lambda_l \; \square. \tag{9.227}$$

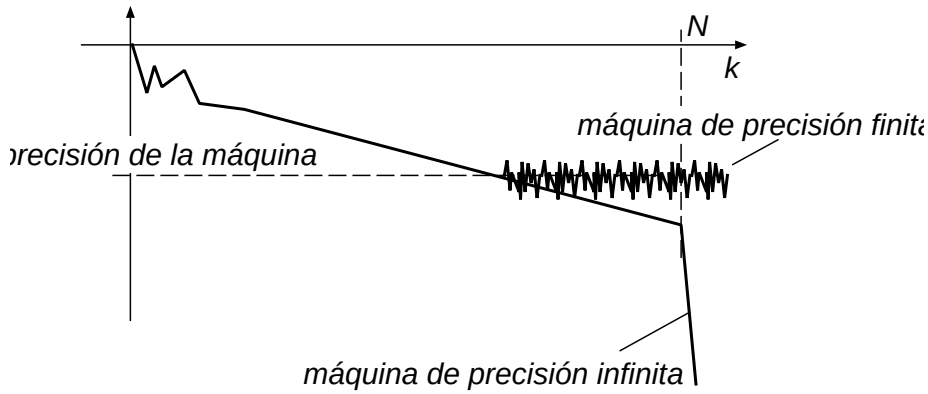


Figura 9.7: Comportamiento del residuo en máquinas de precisión finita.

9.2.3. Criterio de detención del proceso iterativo.

En una máquina de precisión infinita debemos ver que el residuo desciende bruscamente a cero en la iteración N (puede ser antes bajo ciertas condiciones, como hemos visto en algunos casos especiales), pero en la práctica no se itera GC hasta encontrar la solución exacta sino hasta que cierto criterio sobre el residuo (por ejemplo $\|r_k\| < 10^{-6}$) es alcanzado. De hecho, debido a errores de redondeo, ocurre que no se puede bajar de un cierto nivel de residuo relacionado con la precisión de la máquina (10^{-15} en Fortran doble precisión y también en Octave) y el número de operaciones necesarias para calcular el residuo. Entonces, si ese nivel de residuo está por encima del valor del residuo que obtendríamos en la iteración $N-1$ (ver figura 9.7), no veremos el efecto de la convergencia en N iteraciones, sino que GC se comporta como un método iterativo, de manera que lo importante es la tasa de convergencia media del método.

En la práctica se usa usualmente como criterio de detención del proceso iterativo

$$\|b - Ax_k\|_2 \leq \eta \, \|b\|_2 \tag{9.228}$$

Sin embargo, las estimaciones de error basadas en la propiedad de minimización están expresadas en la norma energía del error

$$\frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \leq \max_{z \in \sigma(A)} |\bar{p}_k(z)| \tag{9.229}$$

El siguiente lema relaciona la norma euclídea del error con la norma energía del error

Lema 2.3.1. Sea A spd, con autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$. Entonces para todo $z \in \mathbb{R}^N$

$$\left\| A^{1/2} z \right\|_2 = \|z\|_A \tag{9.230}$$

y

$$\lambda_N^{1/2} \, \|z\|_A \leq \|z\|_2 \leq \lambda_1^{1/2} \, \|z\|_A \tag{9.231}$$

Demostración.

$$\|z\|_A^2 = z^T A Z = (A^{1/2} z)^T (A^{1/2} z) = \left\| A^{1/2} z \right\|_2^2 \tag{9.232}$$

Sean $\{u_i, \lambda_i\}$ los autovectores, autovalores de A, con $u_i^T u_j = \delta_{ij}$. Entonces,

$$z = \sum_{i=1}^{N} (u_i^T z)\, u_i \tag{9.233}$$

$$Az = \sum_{i=1}^{N} \lambda_i\, (u_i^T z)\, u_i \tag{9.234}$$

Entonces,

$$\lambda_N \left\| A^{1/2} z \right\|_2^2 = \lambda_N \sum_{i=1}^{N} \lambda_i\, (u_i^T z)^2 \tag{9.235}$$

$$\leq \sum_{i=1}^{N} \lambda_i^2\, (u_i^T z)^2 = \|Az\|_2^2 \tag{9.236}$$

$$\leq \lambda_1 \sum_{i=1}^{N} \lambda_i\, (u_i^T z)^2 = \lambda_1 \left\| A^{1/2} z \right\|_2^2 \ \square. \tag{9.237}$$

Lema 2.3.2.

$$\frac{\|b - Ax_k\|_2}{\|b\|_2} \leq \frac{\sqrt{\kappa_2(A)}\ \|r_0\|_2}{\|b\|_2} \frac{\|x_k - x^*\|_A}{\|x_0 - x^*\|_A} \tag{9.238}$$

Recordemos que en la norma L_2 para matrices spd. vale que

$$\|A\|_2 = \text{máx autovalor de } A = \lambda_1 \tag{9.239}$$

$$\left\|A^{-1}\right\|_2 = \text{mín autovalor de } A = \lambda_N \tag{9.240}$$

$$\kappa_2(A) = \lambda_1/\lambda_N \tag{9.241}$$

Usando (9.230) y (9.231)

$$\frac{\|b - Ax_k\|_2}{\|b - Ax_0\|_2} = \frac{\|A\,(x^* - x_k)\|_2}{\|A\,(x^* - x_0)\|_2} \leq \frac{\lambda_1^{1/2}\ \|x^* - x_k\|_A}{\lambda_N^{1/2}\ \|x^* - x_0\|_A} \tag{9.242}$$

Consideremos un ejemplo simple

$$x_0 = 0, \ \ \lambda_1 = 11, \ \ \lambda_N = 9, \tag{9.243}$$

Por lo tanto $k_2 = 1.22$ (relativamente pequeño). Tomemos el polinomio (ver figura 9.8)

$$\bar{p}(z) = (10 - z)^k / 10^k \in \mathcal{P}_k \tag{9.244}$$

Como vemos en la figura todos los polinomios se anulan en $z = 10$ la mitad del intervalo donde se encuentran los autovalores. Esta región está sombreada en la figura. El máximo valor de $\bar{p}_k(z)$ sobre el intervalo se alcanza en los extremos del intervalo

$$\max_{9 \leq z \leq 11} |\bar{p}_k(z)| = |\bar{p}_k(9)| = |\bar{p}_k(11)| = 10^{-k} \tag{9.245}$$

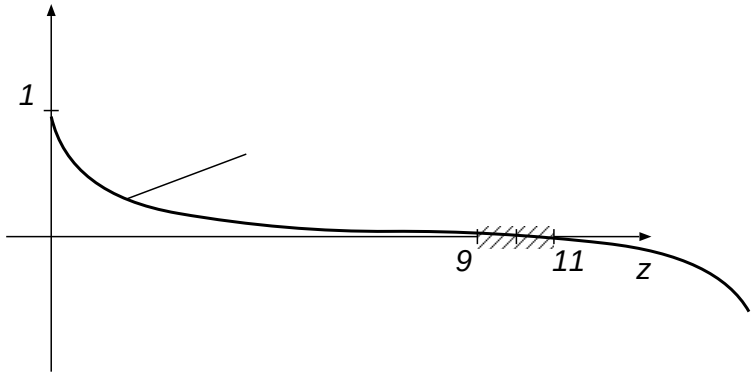


Figura 9.8: Polinomio residual apropiado para el caso (9.243)

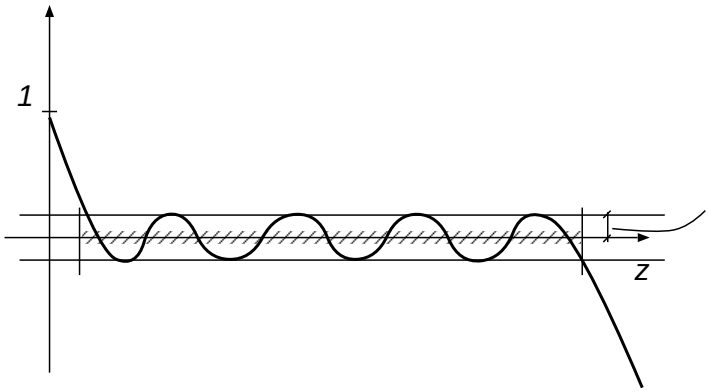


Figura 9.9: Polinomio residual basado en los polinomios de Tchebyshev

lo cual implica que la tasa de convergencia es $\gamma = 1/10$ y el número de iteraciones por orden de magnitud es

$$n = \frac{\log(10)}{\log(1/\gamma)} = 1 \tag{9.246}$$

Mientras tanto, para los residuos

$$\frac{\|Ax_k - b\|_2}{\|b\|_2} \leq \sqrt{1.22} \times 10^{-k} \tag{9.247}$$

$$= 1.10 \times 10^{-k} \tag{9.248}$$

$$\tag{9.249}$$

Para obtener la mejor estimación basada solamente en la información del autovalor máximo y mínimo debemos construir un polinomio de tal forma que tome los valores más bajos posibles en el intervalo $\lambda_1 \leq z \leq \lambda_n$ (ver figura 9.9). Estos polinomios pueden construirse en base a los *polinomios de Tchebyshev*.

Efectivamente, hagamos el cambio de variable

$$(z - \lambda_N)/(\lambda_1 - \lambda_N) = (1 - \cos\theta)/2 = (1 - x)/2 \tag{9.250}$$

de manera que $\theta = 0$ en $z = \lambda_N$ y $\theta = \pi$ en $z = \lambda_1$. Los polinomios de Tchebyshev $T_n(x)$ se definen como

$$T_n(x) = \cos n\theta \tag{9.251}$$

Por ejemplo

$$T_0(x) = 1 \tag{9.252}$$
$$T_1(x) = x \tag{9.253}$$
$$T_2(x) = 2x^2 - 1 \tag{9.254}$$

Por construcción vemos que $|T_n| \leq 1$ en $|x| \leq 1$, o sea $\lambda_n \leq z \leq \lambda_1$. Entonces tomamos $p_k(z) = T_n(x(z))/T_n(x(z=0))$. Como $x(z=0)$ cae fuera del intervalo $|x| \leq 1$ y T_n crece fuertemente (como x^n) fuera del mismo, es de esperar que $T_n(x(z=0)) \gg 1$ y entonces $|p_k(z)| \ll 1$ en $\lambda_n \leq z \leq \lambda_1$. Mediante estimaciones apropiadas para el valor de $T_n(x(z=0))$ puede demostrarse que

$$\|x_k - x^*\|_A \leq 2\,\|x_0 - x^*\|_A \left(\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1}\right)^k \tag{9.255}$$

Podemos ver que, si $\kappa \gg 1$, entonces podemos aproximar

$$\frac{\sqrt{\kappa_2} - 1}{\sqrt{\kappa_2} + 1} \approx 1 - \frac{2}{\sqrt{\kappa}} \tag{9.256}$$

y entonces

$$n \approx \frac{\log(10)}{2}\sqrt{\kappa} = 1.15\sqrt{\kappa} \tag{9.257}$$

que debe ser comparada con $n \approx 1.15\kappa$ para Richardson con $\omega = \omega_{\text{opt}}$. La ventaja es clara, teniendo en cuenta que (como veremos después) el costo (número de operaciones) de una iteración de gradientes conjugados es similar al de una iteración de Richardson.

Sin embargo, la convergencia puede ser mucho mejor que la dada por la estimación anterior, dependiendo de la distribución de los autovalores. Por ejemplo, asumamos que los autovalores están distribuidos en (ver figura 9.10)

$$1 \leq \lambda \leq 1.5 \text{ ó } 399 \leq \lambda \leq 400 \tag{9.258}$$

Entonces $\kappa = 400$ y la estimación anterior (basada sólo en el número de condición) da

$$n = 1.15\,\sqrt{400} = 23 \tag{9.259}$$

mientras que si tomamos el polinomio residual de la forma

$$\bar{p}_{3k} = \frac{(1.25 - z)^k\,(400 - z)^{2k}}{1.25^k\,400^{2k}} \tag{9.260}$$

Entonces debemos estimar

$$\max_{1 \leq z \leq 1.5} |p_{3k}(z)| = (0.25/1.25)^k = 0.2^k \tag{9.261}$$

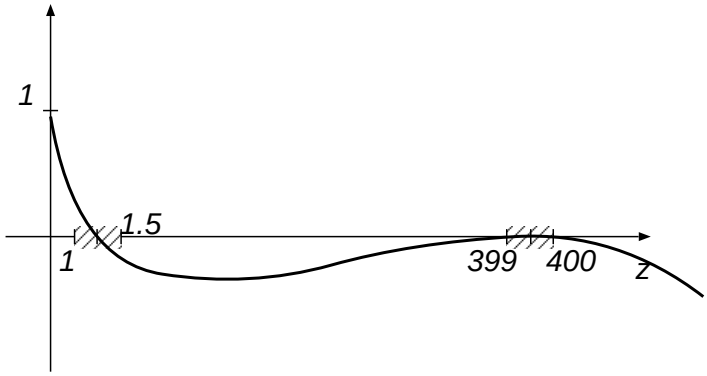


Figura 9.10: Polinomio residual apropiado para un espectro con dos clusters de autovalores

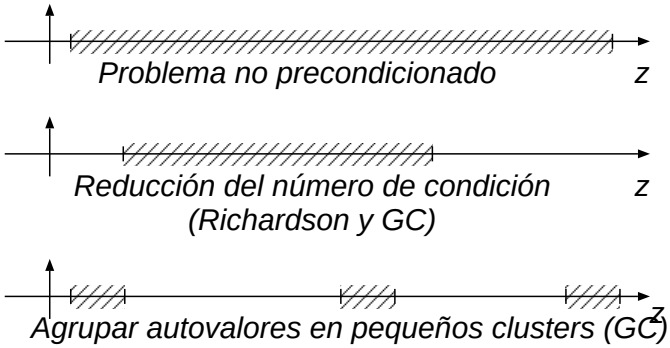


Figura 9.11: Posibles estrategias de precondicionamiento

y

$$\max_{399\leq z\leq 400} |p_{3k}(z)| \quad \leq (400/1.25)^k \, (1/400)^{2k} \tag{9.262}$$

$$= 1/(1.25 \times 400)^k \tag{9.263}$$

Por lo tanto,

$$\max_{z\in\sigma(A)} |\bar{p}_{3k}(z)| \leq 0.2^k \tag{9.264}$$

y $\gamma = 0.2^{1/3}$,

$$n = \frac{\log 10}{(1/3)\,\log(1/0.2)} = 4.29 \tag{9.265}$$

que representa una tasa de convergencia 5 veces más rapida que la predicha sólo en base al número de condición de A.

En el contexto de GC la convergencia se puede mejorar tratando de encontrar precondicionamientos que disminuyan el número de condición o bien que los autovalores están agrupados en pequeños *"clusters"* (ver figura 9.11).

9.2.4. Implementación de gradientes conjugados

La implementación de GC se basa en que conociendo x_k pueden ocurrir dos situaciones. O bien $x_{k+1} = x_k = x^*$, o bien

$$x_{k+1} = x_k + \alpha_{k+1}\, p_{k+1} \tag{9.266}$$

donde $p_{k+1} \neq 0$ es una *dirección de búsqueda* que puede obtenerse en forma sencilla y α_{k+1} es un escalar que se obtiene minimizando el funcional de GC sobre la recta,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\,\phi(x_k + \alpha p_{k+1}) = 0, \text{ en } \alpha = \alpha_{k+1} \tag{9.267}$$

Recordar que el funcional estaba definido como

$$\phi(x) = \frac{1}{2}\, x^T A x - x^T b \tag{9.268}$$

Entonces

$$\frac{\mathrm{d}\phi}{\mathrm{d}\alpha} = p_{k+1}^T\, \nabla\phi \tag{9.269}$$

$$= p_{k+1}^T\, [A(x_k + \alpha_{k+1}\, p_{k+1}) - b] = 0 \tag{9.270}$$

de donde

$$\alpha_{k+1} = \frac{p_{k+1}^T\, (b - A x_k)}{p_{k+1}^T\, A\, p_{k+1}} \tag{9.271}$$

Si $x_{k+1} = x_k$ entonces $\alpha = 0$, pero esto sólo ocurre si x_k es la solución.

Lema. $r_l \in \mathcal{K}_k$ para todo $l < k$

Demostración. Lo haremos por inducción en k. Para $k = 1$ $\mathcal{K}_k = \mathrm{span}\{r_0\}$ y obviamente se verifica que $r_0 \in \mathcal{K}_k$. Ahora asumiendo que es válido para k lo demostraremos para $k+1$. Para $l < k$, se cumple que $r_l \in \mathcal{K}_k \subset \mathcal{K}_{k+1}$, por lo tanto sólo resta demostrarlo para r_k. Pero

$$x_k = x_0 + \sum_{j=0}^{k-1} \alpha_j\, A^j\, r_0 \tag{9.272}$$

y entonces,

$$r_k = b - A x_k \tag{9.273}$$

$$= r_0 - \sum_{j=0}^{k-1} \alpha_j\, A^{j+1}\, r_0 \in \mathcal{K}_{k+1}\ \square. \tag{9.274}$$

Lema 2.4.1. Sea A spd. y $\{x_k\}$ las iteraciones de GC, entonces

$$r_k^T r_l = 0, \quad \text{para todo } 0 \leq l < k \tag{9.275}$$

Demostración. Como x_k minimiza ϕ en $x_0 + \mathcal{K}_k$, entonces para todo $\xi \in \mathcal{K}_k$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi(x_k + t\xi) = \nabla\phi(x_k + t\xi)^T\, \xi = 0, \quad \text{en } t = 0 \tag{9.276}$$

pero $\nabla\phi = -r$, entonces

$$-r_k^T\xi = 0, \quad \text{para todo } \xi \in \mathcal{K}_k \ \square. \tag{9.277}$$

Ahora bien, si $x_{k+1} = x_k$ entonces $r_k = r_{k+1}$ y

$$\|r_k\|_2^2 = r_k^T r_k = r_k^T r_{k+1} = 0 \tag{9.278}$$

por lo tanto $x_k = x^*$.De paso esto permite ver también que GC converge en N iteraciones (cosa que ya hemos demostrado basándonos en la propiedad de minimización.) Efectivamente, supongamos que $r_k \neq 0$ entonces

$$\dim \mathcal{K}_{k+1} = \dim \mathcal{K}_k + 1 \tag{9.279}$$

Pero para $k = N$ $\dim \mathcal{K}_k = N$ entonces $r_{k+1} = 0$.

Lema 2.4.2. Si $x_k \neq x^*$ entonces $x_{k+1} = x_k + \alpha_{k+1}\, p_{k+1}$ donde p_{k+1} está determinado (a menos de una constante multiplicatva) por

$$p_{k+1} \quad \in \mathcal{K}_{k+1} \tag{9.280}$$

$$p_{k+1}^T A\xi \quad = 0, \quad \text{para todo } \xi \in \mathcal{K}_k \tag{9.281}$$

Demostración. Como $\mathcal{K}_k \subset \mathcal{K}_{k+1}$

$$\nabla\phi(x_{k+1})^T\xi = 0, \quad \text{para todo } \xi \in \mathcal{K}_k \tag{9.282}$$

$$[Ax_k + \alpha_{k+1}Ap_{k+1} - b]^T\,\xi = 0 \tag{9.283}$$

pero $Ax_k - b = r_k \perp \xi$ para todo $\xi \in \mathcal{K}_k$ de manera que

$$p_{k+1}^T A\xi = 0 \tag{9.284}$$

y como $\dim \mathcal{K}_{k+1} = \dim \mathcal{K}_k + 1$, esto define únicamente a p_{k+1}.

Esta propiedad se llama que p_{k+1} es *"conjugado"* a $\mathcal{K}_k$. Ahora bien, tomando r_k y ortogonalizándolo con respecto a $\mathcal{K}_k$ podemos obtener p_{k+1}. Entonces,

$$p_{k+1} = r_k + w_k, \quad \text{con } w_k \in \mathcal{K}_k \tag{9.285}$$

Teorema 2.4.1. Sea A spd. y asumamos que $r_k \neq 0$. Sea $p_0 = 0$. Entonces $p_{k+1} = r_k + \beta_{k+1}p_k$ para algún escalar β_{k+1} y $k \geq 0$.

Demostración. ver libro.

Lema 2.4.3. Las siguientes fórmulas son también válidas para obtener los α_k y β_k.

$$\alpha_{k+1} = \frac{\|r_k\|_2^2}{p_{k+1}^T\, Ap_{k+1}}, \quad \beta_{k+1} = \frac{\|r_k\|_2^2}{\|r_{k-1}\|_2^2} \tag{9.286}$$

Demostración. Ver libro.

Algoritmo 2.4.1. $\texttt{cg}\,(x, b, A, \epsilon, k_{\max})$

1. $r = b - Ax$, $\rho_0 = \|r\|_2^2$, $k = 1$
2. Do while $\sqrt{\rho_{k-1}} > \epsilon \|b\|_2$ y $k < K_{\max}$
   a. If $k = 1$ then
      $p = r$
      else
      $\beta = \rho_{k-1}/\rho_{k-2}$
      $p = r + \beta p$
      endif
   b. $w = Ap$
   c. $\alpha = \rho_{k-1}/(p^T w)$
   d. $x = x + \alpha p$
   e. $r = r - \alpha w$
   f. $\rho_k = \|r_k\|_2$
   g. $k = k + 1$

Notar que no es necesario formar explícitamente la matriz A (ni siquiera en forma "sparse", es decir, la lista de $\{i, j, a_{ij}\}$), sino que solo es necesario definir una rutina que, dado x calcule Ax. Esto se llama una *operación matriz-vector*. Debido a esta propiedad de no necesitar tener la matriz almacenada, GC es llamado un método *mátrix free*.

Costo de GC. Las variables que se deben mantener almacenadas durante la iteración son x, w, p, r o sea $4N$ elementos. En cuanto al número de operaciones,

- 1 producto matriz vector (paso 2.b)
- 2 productos escalares (pasos 2.c y 2.f)
- 3 "axpys" (pasos 2.a, 2.d y 2.e).

Una operación "axpy" es aquella de la forma $y \leftarrow \alpha x + y$, es decir adicionar a un vector x un múltiplo escalar de otro vector y. (El nombre proviene de que la rutina de la librería BLAS que realiza esta operación y que a su vez es una descripción mnemotécnica de la operación que se realiza.). De todos, el que requiere más operaciones es generalmente el producto matriz vector, sobre todo en la versión *matrix free* y sobre todo cuanto más compleja es la física del problema. Por ejemplo, si x representa el vector de potenciales nodales guardados por columnas para una malla homogénea de paso h (como en la función `laplacian.m` usada en la Guía 1), entonces la implementación de Ax es

```
phi=reshape(phi,n-1,n-1);

% range of indices of internal nodes
II=(2:n);
JJ=II;

% Include the boundary nodes
Phi=zeros(n+1);
```

```
Phi((2:n),(2:n))=phi;

% Use the standard 5 point formula
lap_phi=((-Phi(II+1,JJ)...
         -Phi(II-1,JJ)...
         -Phi(II,JJ+1)...
         -Phi(II,JJ-1)...
         +4*Phi(II,JJ))/h^2);

lap_phi=lap_phi(:);
```

donde vemos que básicamente el número de operaciones necesario es $O(5N)$, ya que 5 es el número de puntos involucrados en el stencil de Poisson. Sin embargo, si la complejidad de la representación aumenta el cómputo se mantiene en general $O(N)$ pero con cada vez más operaciones por nodo. Por ejemplo, si la malla está refinada hacia los lados, es decir si el h no es constante, entonces hay que multiplicar cada fila por un h distinto. Si además permitimos que las líneas de la malla no sean paralelas a los ejes, entonces habrá que calcular los *coeficientes métricos* de la malla. Si consideramos el uso de mallas no estructuradas con el método de los elementos finitos el cálculo de las matrices de cada elemento aumentará todavía más el número de operaciones por nodo.

Costo de GC como método directo. Si bien el método de GC se usa raramente como método directo, por razones que veremos más adelante, vamos a estimar el número de operaciones necesarios y compararlo con eliminación de Gauss. Considerando un cuadrado de $N = n \times n$ nodos en 2D y un cubo de $N = n \times n \times n$ en 3D, tenemos que,

- Ancho de banda de la matriz: $m = n$ en 2D, $m = n^2$ en 3D.
- Número total de incógnitas: $N = n^2$ en 2D, $N = n^3$ en 3D.
- Número de op. Gauss: N^2 en 2D, $N^{2.33}$ en 3D
- Número de op. GC/directo: N^2 en 2D, N^2 en 3D

Hemos hecho uso de que el número de operaciones para eliminación de Gauss es Nm^2. Para GC hemos usado

$$\text{número de op.} = O(\text{número de iter.} \times \text{nro. de op. por iter.}) \tag{9.287}$$
$$= O(N^2) \tag{9.288}$$

Vemos que, incluso como método directo, GC llega a ser más competitivo en 3D. De todas formas, como ya hemos mencionado, por el problema de precisión finita (ver figura 9.7) en general para problemas grandes se llega a la precisión de la máquina antes de las N iteraciones.

En cuanto a la capacidad de almacenamiento, GC obviamente requiere mucho menos memoria ($O(N)$ para GC contra $O(N^{1.5})$ en 2D y $O(N^{1.66})$ en 3D para Gauss).

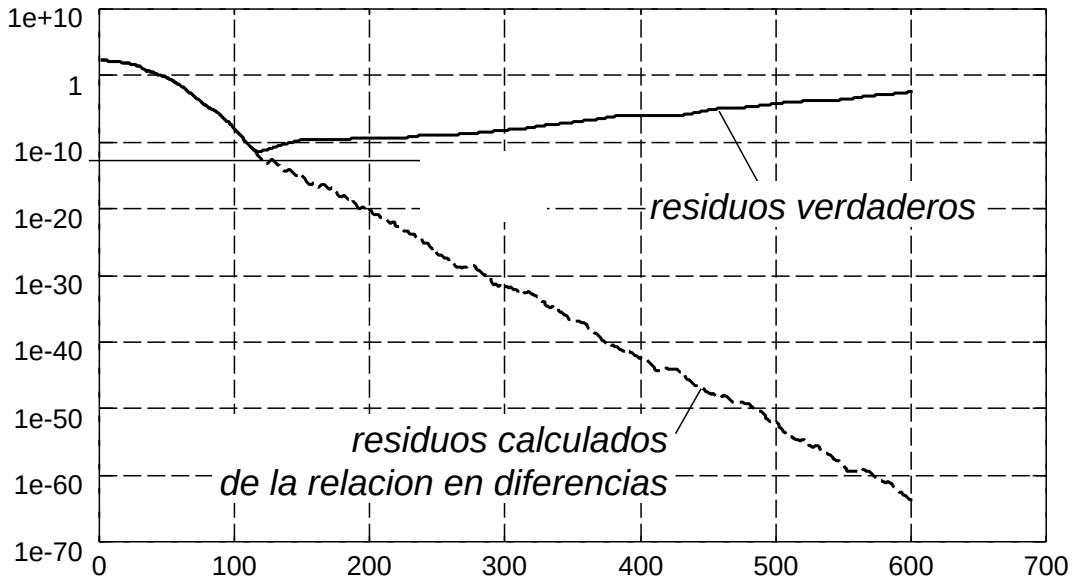


Figura 9.12: Gradientes Conjugados y los residuos verdaderos.

Comparación como método iterativo con Richardson. Tanto para Richardson como para GC el costo para bajar el error un order de magnitud es, básicamente

$$\begin{aligned}&\text{Nro. de opr. para bajar el residuo un orden de magnitud} = \\ &= n \times \text{nro. de oper. por iteración}\end{aligned} \tag{9.289}$$

Donde n es el número de iteraciones necesario para bajar el error un orden de magnitud. Asumiendo que el costo de las iteraciones es prácticamene el mismo para los dos métodos (lo cual es válido si hacemos una implementación *matrix free* con una formulación relativamente compleja, por ejemplo FEM), entonces los costos relativos están dados por las tasas de convergencia y, usando que en 3D $\kappa \sim n^2 = N^{2/3}$.

$$n(\text{Richardson con } \omega = \omega_{\text{opt}}) \quad \sim \kappa = N^{2/3} \tag{9.290}$$

$$n(\text{GC }) \quad \sim \sqrt{\kappa} = N^{1/3} \tag{9.291}$$

con lo cual la ganancia es evidente.

9.2.5. Los “verdaderos residuos”.

El algoritmo `cg(...)` (ver pág. 263) descripto más arriba tiene la particularidad de que los residuos no son calculados directamente usando (9.50) sino que son calculados en el paso (2e). Si bien las expresiones coinciden en una máquina de precisión infinita, debido a errores de redondeo los valores numéricos obtenidos con uno u otro método pueden diferir en una máquina de precisión finita. En la figura 9.12 vemos los residuos calculados en un experimento calculados de las dos formas. El comportamiento de los verdaderos residuos es similar al observado para Richardson en §9.1.4. Sin embargo, es todavía peor para GC.

El residuo no sólo no baja de el umbral $\|r\|_{\text{sat}}$ de saturación sino que empieza a crecer lentamente. Esto es muy peligroso desde el punta de vista práctico, ya que en el caso de Richardson el efecto de saturación sólo ocasiona un gasto de tiempo de cálculo innecesario, pero en el caso de GC puede ocasionar una pérdida de precisión. Podemos decir que Richardson es un método *estable* ante errores de redondeo, mientras que GC es *inestable*. Esta inestabilidad de GC se debe al hecho de que asume que los residuos van formando una base ortogonal y las direcciones de búsqueda van siendo conjugadas (ec. (9.281)), y estas propiedades se van perdiendo por errores de redondeo. Esta inestabilidad es compartida por todos los métodos que se basan en estas propiedades de conjugación, por ejemplo GMRES y los métodos que veremos después. Para peor los residuos calculados por la expresión recursiva (2e) tienden a seguir descendiendo, de manera que si el usuario no verifica el verdadero valor del residuo, puede creer que realmente el error ha bajado (después de 600 iteraciones) hasta 2×10^{-64}, mientras que el error verdadero está en el orden de 3×10^{-3}.

Cuando calculamos los residuos directamente a partir de (9.50) decimos que se trata de los *verdaderos residuos*. El porqué los residuos calculados según el algoritmo `cg(...)` (ver pág. 263) tienden a bajar, en vez de *rebotar* como lo hacen los verdeaderos residuos, puede entenderse más fácilmente en el algoritmo de Richardson. Efectivamente, puede demostrarse simplemente que los residuos también satisfacen una relación como la (9.101) a saber

$$r_{k+1} = (I - BA)\, r_k \tag{9.292}$$

De manera que también podríamos calcular los residuos utilizando recursivamente esta relación. Pero esta relación no involucra diferencias entre magnitudes grandes como en (9.50) y por lo tanto $\|r_k\|$ calculado por esta expresión sigue descendiendo, independientemente de la precisión de la máquina.

Una forma de corregir el el algoritmo `cg(...)` (ver pág. 263) es agregar una línea en cada iteración que calcule el verdadero residuo, *sólo a los efectos de chequear convergencia*, sin embargo esto involucra un producto matriz vector adicional por iteración, es decir prácticamente duplica el costo del método.

Precondicionamiento. Asumamos que tenemos una matriz M fácil de invertir y tal que $\kappa(MA) \ll \kappa(A)$ o tal que los autovalores están agrupados en clusters. Entonces podemos tratar de resolver

$$(MA)\, x = (Mb) \tag{9.293}$$

en vez del sistema original, ya que este está bien condicionado. Sin embargo, incluso si M es spd. el producto de dos matrices spd. no es necesariamente spd. Basta con ver el ejemplo,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \tag{9.294}$$

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{9.295}$$

A y M son spd. pero

$$MA = \begin{bmatrix} 2 & 2 \\ 1 & 4 \end{bmatrix} \tag{9.296}$$

no es simétrica. Una posibilidad es buscar $M = S^2$ y entonces redefinir

$$(SAS)\, y = (Sb) \tag{9.297}$$

y una vea obtenido y obtener $x = Sy$. Como SAS es equivalente a $S^2A = MA$ tienen la misma distribución de autovalores y SAS sí es spd. Pero falta resolver el problema de hallar S, lo cual puede ser más complicado

que hallar M y por otra parte el cálculo de SAS por un vector involucra el cálculo de dos productos matriz-vector adicionales, contra uno sólo si precondicionamos de la forma (9.293). La solución pasa por reproducir el algoritmo de Gradiente Conjugado pero reemplazando el producto escalar por $x^T Mx$ y finalmente resulta en el siguiente algoritmo de GC precondicionado,

Algoritmo 2.4.1. $\texttt{pcg}\,(x, b, A, M, \epsilon, k_{\max})$

1. $r = b - Ax$, $\tau_0 = r^T Mr$, $\rho_0 = \|r\|_2$, $k = 1$
2. Do while $\sqrt{\rho_{k-1}} > \epsilon \|b\|_2$ y $k < K_{\max}$
   a. If $k = 1$ then
      $z = Mr$
      else
      $\beta = \rho_{k-1}/\rho_{k-2}$
      $p = Mr + \beta p$
      endif
   b. $w = \alpha p$
   c. $\alpha = \rho_{k-1}/(p^T w)$
   d. $x = x + \alpha p$
   e. $r = r - \alpha w$
   f. $\rho_k = \|r_k\|_2^2$,$\tau_k = r_k^T Mr$
   g. $k = k + 1$

La modificación principal del algoritmo pasa por reemplazar p por Mp en las definiciones de los p y reemplazar los productos escalares $x^T y$ por la forma cuadrática $x^T My$. Además, ahora calculamos escalares τ_k para el cálculo de las direcciones p_k y escalares ρ_k para chequear la convergencia. (Mientras que en la versión no precondicionada, $\rho_k = \tau_k$).

El costo adicional más importante para el esquema precondicionado es un producto matriz-vector adicional con la matriz precondicionada. El costo del producto matriz-vector para el precondicionamiento puede variar mucho y depende de la complejidad del precondicionamiento. Por ejemplo, si tomamos M como la inversa de la parte diagonal de A (precondicionamiento Jacobi), entonces el costo es ínfimo, pero en el otro extremo podemos tomar $M = A^{-1}$. Entonces GC converge en una iteración pero el costo de calcular Mx es el costo de invertir A!! Tenemos entonces que

$$\text{Costo total} = n \times \text{nro de oper. por iteración} \tag{9.298}$$

Cuanto más complejo es el precondicionador el n tiende a disminuir pero el número de operaciones tiende a aumentar debido al costo del cálculo del precondicionamiento. Lo importante entonces, al evaluar la efectividad de un precondicionador es no sólo evaluar cómo aumenta la tasa de convergencia sino *también tener en cuenta el costo de evaluar* Mx*.*

Precondicionadores. Existen una variedad de precondicionadores propuestos para diferentes problemas. Algunos son muy específicos para ciertas aplicaciones otros son puramente algebraicos, es decir su aplicación es general para toda clase de problema (tal vez no su efectividad!). En el contexto de la resolución de sistemas lineales provenientes de la discretización de ecuaciones en derivadas parciales por diferencias finitas o volúmenes finitos podemos mencionar los siguientes

- Intrínsecos al problema
  - Resolvedores rápidos de Poisson
  - Multigrilla
  - Descomposición de dominios
  - ADI
- De tipo general
  - Jacobi
  - Factorización incompleta
  - Precondicionamiento polinomial

A continuación haremos una breve descripción de los mismos

Resolvedores rápidos de Poisson. Si consideramos la ecuación de Poisson 1D con condiciones de contorno periódicas la matriz resulta ser

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0 & \dots & -1 \\ -1 & 2 & -1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & 2 & -1 & \dots & 0 \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ -1 & 0 & 0 & \dots & 0 & -1 & 2 \end{bmatrix} \tag{9.299}$$

Sea x de la forma

$$(x^k)_i = \mathrm{e}^{\mathrm{i}2\pi ki/N} \tag{9.300}$$

donde i es la unidad imaginaria. Puede verse que $(x)_i$ es un autovector de A. Esto vale, no sólo para la matriz del laplaciano 1D, sino también para cualquier operador homogéneo (que no depende de x) discretizado sobre una malla homogénea. En este caso las filas de la matriz A son las mismas, sólo que se van corriendo una posición hacia la derecha a medida que bajamos una fila, es decir:

$$A_{ij} = \hat{A}_{i-j} \tag{9.301}$$

donde $\hat{A}_\mu$ es cíclico en p de período N, es decir $\hat{A}_{p+N} = \hat{A}_p$. Pero los x de la forma (9.300) son la base que induce la transformada de Fourier, de manera que si F es la matriz que tiene como columnas estos autovectores, vale que

$$Fz = \mathrm{fft}(z), \quad \forall z \tag{9.302}$$

donde $\mathrm{fft}(z)$ indica el operador de transformada de Fourier tal como está definido en Matlab. Como las columnas de F son autovectores de A, entonces $F^{-1}AF$ es diagonal y podemos tomar como precondicionamiento

$$M = F^{-1}\left[\mathrm{diag}(A)\right]^{-1}F \tag{9.303}$$

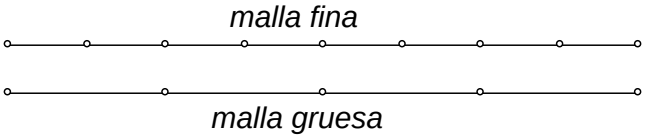


Figura 9.13: Mallas fina y gruesa para las técnicas de multigrilla.

por lo visto, para matrices periódicas, $M = A^{-1}$ y entonces GC convergerá en una iteración. La idea es que (9.303) puede ser un buen precondicionamiento incluso en condiciones más generales, por ejemplo cuando la matriz no es periódica o cuando la malla no es homogénea. En cuanto al costo del precondicionamiento, multiplicar por F y $^{-1}$ es equivalente a aplicar transformadas y antitransformadas de Fourier (aplicar las operaciones `fft( )` y `ifft( )` de Matlab. Estas requieren $O(N \log_2(N))$ operaciones y la inversión de la parte diagonal de A sólo requiere $O(N)$ operaciones. La idea puede extenderse fácilmente a 2D y 3D.

Multigrilla. Si hacemos una descomposición modal del error, puede verse que el error en las componentes correspondientes a los autovalores más pequeños converge en general mucho más lentamente que para los autovalores más altos. Esto es más evidente para el método de Richardson, donde los factores de amplificación $1 - \omega\lambda_i$ son muy cercanos a 1 para los autovalores más pequeños. A su vez, como ya hemos visto en los ejemplos, los autovalores altos están asociados a frecuencias altas (funciones que son muy oscilatorias espacialmente) mientras que los autovalores bajos están asociados a autofunciones suaves. Esto significa que después de un cierto número de iteraciones el error para las frecuencias altas se debe haber reducido mucho mientras que el grueso del error está en las componentes de baja frecuencia. Como las funciones suaves pueden ser restringidas a una malla más gruesa sin perder información, esto sugiere la idea de proyectar el problema a una malla más gruesa (digamos con $h' = 2h$ y por lo tanto con la mitad de nodos, ver figura 9.13) y realizar una serie de iteraciones sobre esta malla para obtener una corrección. Una vez obtenida ésta se interpola sobre la malla original y se agrega al vector de iteración. La idea es que la corrección va a ser tan buena como si hubiéramos iterado sobre la malla original pero a un costo igual a la $1/2^{n_d}$ ya que la malla gruesa tiene la mitad de nodos. Esta idea puede ser aplicar recursivamente, ya que al iterar en la malla gruesa va a volver a ocurrir que despues de una serie de iteraciones va a quedar una fuerte componente del error en las frecuencias bajas y estas las podemos corregir iterando sobre una malla $h'' = 2h' = 4h$ y así siguiendo. De esta manera, multigrilla se basa en resolver el problema en una serie de mallas con paso de malla $h, 2h, 4h, \ldots, 2^m h$, haciendo una serie de iteraciones en la malla fina seguida de iteraciones sobre la malla más gruesa y así siguiendo.

Descomposición de dominios. Ver §9.4.1

Precondicionamiento Jacobi. Consiste en simplemente tomar $M = (\operatorname{diag} A)^{-1}$. Como la matriz a invertir es diagonal el costo de invertirla es muy bajo ($O(N)$ opeaciones). Este precondicionamiento no aporta nada en situaciones donde los elementos de la diagonal son aproximadamente constantes (precondicionar con un múltiplo escalar de la identidad no puede nunca mejorar el número de condición). Esto ocurre por ejemplo para el Laplaciano (tanto en 1D como en más dimensiones) cuando el paso de la malla es constante. Cuando la malla no es homogénea (ver figura- 9.14) entonces el número de condición es

$$\kappa(A) = O\left(\left(\frac{L}{h_{\min}}\right)^2\right) \tag{9.304}$$

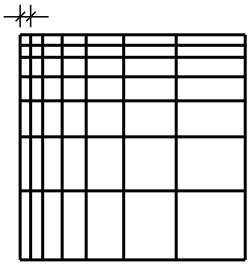


Figura 9.14: Malla refinada hacia una esquina.

donde $h_{\min}$ es el mínimo h sobre toda la malla y entonces puede ser bastante peor que $O(n^2)$ donde n es el número de elementos en una dirección característica. Los elementos diagonales de A son $\propto h_x^{-2} + h_y^{-2} = O(\min(h_x, h_y)^{-2})$ y puede verse que este precondicionamiento corrige el mal condicionamiento producido por el refinamiento de manera que

$$\kappa(\mathrm{diag}(A)^{-1}A) = O(n^2) \tag{9.305}$$

Factorización incompleta. Consideremos la factorización Cholesky A, $A = BB^T$. Entonces este precondicionamiento consiste en descartar elementos de B de manera de reducir su ancho de banda. En la implementación práctica el descarte se va haciendo a medida de que se va factorizando la matriz, basándose en el valor absoluto del elemento B_{ij} que se está evaluando y su distancia a la diagonal $|i - j|$. Por supuesto se trata de descartar aquellos elementos que tienen un menor valor absoluto y que se encuentran más alejados de la diagonal.

Precondicionamiento polinomial. Consiste en encontrar un polinomio $p(z)$ tal que $M = p(A) \approx A^{-1}$. El criterio para construir el polinomio en cuestión es similar al utilizado para construir los polinomios residuales que permiten obtener la estimación de convergencia (9.255) en base a los polinomios de Tchebyshev. El costo de un tal precondicionmiento es evaluar $Mx = p(A)x$ que involucra m productos matriz-vector, donde m es el orden del polinomio de Tchebyshev. Como es usual en la evaluación de polinomios, se utiliza la forma anidada, también llamada de Hörner, como en el siguiente script

```
y=0;
for k=0:m
  y=p(k)*x+A*y;
end
```

donde (usando la notación de Octave) $p(x) = p_1\, x^m + p_2\, x^{m-1} + \ldots + p_{m+1}$ y los coeficientes p_i están almacenados en un vector de longitud $m + 1$. En la versión *matrix free*, el producto Ax está definido por medio de una rutina. Sea `w=prodvec(x)` la rutina que retorna $w = Ax$ cuando se le pasa x entonces el algoritmo se escribe como

```
y=0;
for k=0:m
  y=p(k)*x+prodvec(y);
end
```

9.2.6. Métodos CGNR y CGNE

Si A es nosimétrica o no definida positiva, entonces podemos considerar resolver

$$(A^T A)\, x = (A^T b) \tag{9.306}$$

en el cual aparece la matriz $A^T A$ que es simétrica y definida positiva si A es no singular. Notar que en cada iteración de GC sobre este sistema estamos minimizando

$$\begin{aligned} \|x^* - x\|_{A^T A} \quad &= (x^* - x)^T A^T A\, (x^* - x) && (9.307)\\ &= (b - Ax)^T (b - Ax) = \|r\|_2^2 && (9.308) \end{aligned}$$

de ahí el nombre de *"Conjugate Gradient on the Normal Equations to minimize the Residual"* (CGNR). Otra posibilidad es hacer el cambio de variable $x = A^T y$ y resolver

$$(AA^T) y = b \tag{9.309}$$

para y. Una vez encontrado y, x se obtiene con una operación matriz vector adicional $x = A^T y$. La norma que se minimiza es en este caso

$$\begin{aligned} \|y^* - y\|_{AA^T} \quad &= (y^* - y)^T AA^T\, (y^* - y) && (9.310)\\ &= (A^T y^* - A^T y)^T\, (A^T y^* - A^T y) && (9.311)\\ &= (x^* - x)^T\, (x^* - x) = \|x^* - x\|_2^2 && (9.312) \end{aligned}$$

de ahí el nombre de *"Conjugate Gradient on the Normal Equations to minimize the Error"* (CGNE).

Observaciones

- En general ocurre que $\kappa(A^T A) \sim \kappa(A)^2$, lo cual augura muy bajas tasas de convergencia si el $\kappa(A)$ es grande.
- Se necesitan 2 productos matriz vector por iteración
- En la versión *matrix free* hay que programar no sólo una rutina que calcule Ax sino también una que calcule $A^T x$. Para problemas provenientes de discretizaciones por FEM o FDM de PDE's, esto puede resultar bastante complicado.

9.3. El método GMRES

9.3.1. La propiedad de minimización para GMRES y consecuencias

GMRES (por *"Generalized Minimum RESidual"* fue popuesto en 1986 por Y. Saad y m. Schulz como un método iterativo por subespacios de Krylov para sustemas no-simétricos. En contraposición con CGNR o CGNE *no* requiere el cálculo de productos de A^T con un vector, lo cual es una gran ventaja en muchos casos. pero es necesario guardar una base de $\mathcal{K}_k$ lo cual requiere un costo de almacenamiento adicional a medidad que la iteración progresa.

La iteración k-ésima ($k \geq 1$) es

$$x_k = \underset{x \in x_0 + \mathcal{K}_k}{\operatorname{argmin}} \|b - A\, x\|_2 \tag{9.313}$$

Nótese que esta propiedad de minimización es muy similar con la de Gradientes Conjugados, con la diferencia que en GC se aplica al funcional (9.181). Como aquí estamos contemplando la posibilidad que A no sea simétrica o definida positiva, entonces no podemos tomar este funcional. Si consideramos que minimizar $\|b - A\,x\|_2$ es equivalente a minimizar $\|b - A\,x\|_2^2$ el cual contiene en la forma cuadrática $A^T\,A$, entonces vemos que GMRES es en parecido a CGNR, pero con la diferencia de que el espacio de Krylov contiene $\mathrm{span}\{r_0, Ar_0, A^2r_0, ...\}$, no $\mathrm{span}\{r_0, A^TAr_0, (A^TA)^2r_0, ...\}$. Esta diferencia es muy importante ya que es la que hace que la tasa de convergencia de GMRES c

con la salvedad de que con GMRES no debemos calcular productos A^T-vector pero en contraparte debemos mantener una base del espacio de Krylov.

Como $x \in x_0 + \mathcal{K}_k$ puede ponerse de la forma

$$x = x_0 + \sum_{j=0}^{k-1} \gamma_j\,A^j\,r_0 \tag{9.314}$$

entonces

$$\begin{aligned} b - A\,x \quad &= b - A\,x_0 - \sum_{j=0}^{k-1} \gamma_j\,A^{j+1}\,r_0 && (9.315)\\ &= r_0 - \sum_{j=1}^{k} \gamma_{j-1}\,A^j\,r_0 && (9.316)\\ &= \bar{p}(A)\,r_0 && (9.317)\end{aligned}$$

donde $\bar{p} \in \mathcal{P}_k$ es un polinomio residual.

Teorema 3.1.1. Sea A no-singular y x_k la k-ésima iteración de GMRES. Entonces para todo $\bar{p} \in \mathcal{P}_k$

$$\|r_k\|_2 = \min_{p\in\mathcal{P}_k} \|p(A)\,r_0\|_2 \leq \|\bar{p}(A)\,r_0\|_2 \tag{9.318}$$

Corolario 3.1.1. Sea A no-singular, entonces

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \|\bar{p}_k(A)\|_2 \tag{9.319}$$

Teorema 3.1.2. Sea A no-singular. Entonces GMRES encuentra la solución dentro de las N iteraciones

Demostración. Usar $\bar{p}(z) = p(z)/p(0)$, donde $p(z) = \det(A - zI)$ es el polinomio característico. □

Para GC hemos usado el teorema espectral para encontrar estimaciones de la tasa de convergencia. Esto no puede asumirse en general si A no es simétrica. Sin embargo podemos restringir el análisis al caso de que A sea diagonalizable aunque los autovalores y autovectores pueden ser ahora comlejos.

Nota sobre la condición de diagonalizabilidad de matrices. Recordemos que

- A es diagonalizable si $A = V\Lambda V^{-1}$ con Λ diagonal. Cuando A es real y simétrica Λ es real y podemos elegir a V como real. Cuando A es no-simétrica tando Λ como V pueden ser complejos.
- En álgebra compleja muchos conceptos pueden extenderse fácilmente si reemplazamos la "tranpuesta" de A por la "transpuesta conjugada" de A que denotaremos como A^H.

- El producto escalar de dos vectores pertenecientes a C^N se define como $x^H y = \sum_{k=1}^{n} \overline{(x)_k}(y)_k$.
- V es *unitaria* si $V^H V = I$ (es la extensión del concepto de matriz ortogonal a complejos).
- A es *normal* si es diagonalizable y si la matriz de cambio de base correspondiente V es unitaria.
- Puede verse que si A conmuta con su transpuesta conjugada (es decir $A^H\, A = A\, A^H$) entonces A es normal.
- Es obvio que si A es hermítica (simétrica en el caso real) entonces A es normal

Teorema 3.1.3. para todo $\bar{p} - k \in \mathcal{P}_k$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \kappa_2(V) \max_{z \in \sigma(A)} |\bar{p}_k(Z)| \tag{9.320}$$

Demostración. Basta con ver que

$$\begin{aligned} \|\bar{p}_k(A)\|_2 &\leq \|V\|_2 \left\|V^{-1}\right\|_2 \|\bar{p}_k(\Lambda)\|_2 && (9.321)\\ &= \kappa_2(V) \max_{z \in \sigma(A)} |\bar{p}_k(Z)| \square. && (9.322) \end{aligned}$$

No está claro como puede estimarse el $\kappa_2(V)$, si existe. Si A es normal, entonces V es unitaria, preserva la norma y entonces $\|V\|_2 = \left\|V^{-1}\right\|_2 = \kappa_2(V) = 1$

$$\begin{aligned} \|V\|_2 &= \max_{x \neq 0} \frac{\|V\, x\|_2}{\|x\|_2} && (9.323)\\ &= \max_{x \neq 0} \frac{\sqrt{(V\, x)^H\, (V\, x)}}{\sqrt{x^H\, x}} && (9.324)\\ &= \max_{x \neq 0} \frac{\sqrt{x^H\, (V^H V)\, x}}{\sqrt{x^H\, x}} && (9.325)\\ &= 1 && (9.326) \end{aligned}$$

y similarmente para V^{-1}. Por otra parte, si A se aproxima a una matriz no diagonalizable, entonces $\kappa_A() \rightarrow \infty$. Esto puede verse con un ejemplo simple. La matriz

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{9.327}$$

es un *bloque de Jordan* y es claramente no diagonalizable. Perturbando ligeramente uno de los elementos diagonales de la forma

$$A = \begin{bmatrix} 0 & 1 \\ 0 & \epsilon \end{bmatrix} \tag{9.328}$$

con ϵ muy pequeño la matriz pasa a ser diagonalizable, ya que tiene dos autovalores distintos ($\lambda = 1, \epsilon$). Sin embargo podemos ver que la matriz de cambio de base V correspondiente tiene un número de condición que crece como $2/\epsilon$:

```
octave> a=[0 1;0 1e-5]
a =
  0.00000  1.00000
  0.00000  0.00001
octave> [v,d]=eig(a)
v =
  1.00000  1.00000
  0.00000  0.00001
d =
   0.0000e+00   0.0000e+00
   0.0000e+00   1.0000e-05
octave> cond(v)
ans =  2.0000e+05
octave>
```

Ahora consideremos que ocurre si utilizamos una rutina numérica de diagonalización (como el `eig( )` de Octave) sobre una matriz que no es diagonalizable. Lo deseable sería que la rutina numérica nos mostrara un mensaje de eror diciendo que la matriz no es diagonalizable. Sin embargo, debido a los errores de redondeo, la matriz aparece ser como diagonalizable desde el punto de vista numérico y entonces la forma efectiva de verificar *"cuán diagonalizable es una matriz"* es chequear $\kappa_2(V)$. Cuanto más grande es este número *"menos diagonalizable"* es la matriz.

```
octave>a=[0 1;0 0]
a =

  0  1
  0  0

octave> [v,d]=eig(a)
v =

   1.00000  -1.00000
   0.00000   0.00000

d =

  0  0
  0  0

octave> cond(v)
ans =  1.9958e+292
```

Esto es similar a cuando nos preguntamos si una matriz es inversible o no. Si calculamos el determinante de la matriz, por errores de redondeo este número siempre resulta ser diferente de cero. Además, la misma matriz multiplicada por una factor $a < 1$ tiene un determinante que es un factor a^N veces menor. Para

matrices grandes (digamos $N = 100$) un pequeño factor 0.1 puede representar un cambio en la magnitud del determinante por un factor 10^{-100}. Todo esto indica que el determinante no es un buen indicador de *"cuan singular es una matriz"* y un análisis más detallado muestra que el indicador correcto es el número de condición de la matriz: cuanto más alto es el número de condición *"más singular es la matriz"*.

Los siguientes Teoremas reproducen los resultados ya obtenidos para GC. Su demostración se basa nuevamente en la construcción de polinomios residuales apropiados que se anulan en los autovalores de la matriz.

Teorema 3.1.4. Sia A tienen autovalores distintos entonces GMRES converge en k iteraciones.

Teorema 3.1.5. Si r_0 es una combinación lineal de k autovectores de A entonces GMRES converge dentro de las k iteraciones.

9.3.2. Criterio de detención:

Pensando a GMRES como un método iterativo las estimaciones de la tasa de convergencia son similares a las de GC. Como en GC el criterio de detención es

$$\|r_k\|_2 \leq \texttt{tol}\ \|b\|_2$$

Una estimación bastante cruda de la tasa de convergencia se puede obtener asumiendo que existe un ω tal que $\|I - \omega A\|_2 = \rho < 1$. Tomando el polinomio de $\bar{p}_k(x) = (1 - \omega z)^k$ podemos obtener la estimación de convergencia

$$\|r_k\|_2 \leq \rho^k \|r_0\|_2 \tag{9.329}$$

Esta estimación de convergencia tiene algunos puntos en común con las estimaciones de convergencia que hemos obtenido para el método de Richardson. Notemos que bajoe estas mismas condiciones el método de Richardson predice la misma tasa de convergencia. Sin embargo la diferencia fundamental está en que con GMRES no es necesario conocer el valor de ω, de hecho, como (9.329) es válido para cualquier ω es válido para aquel que minimize $\|I - \omega A\|_2$, es decir que su convergencia es mejor que la de Richardson sobre-relajado con el mejor valor del parámetro de relajación que pudiéramos obtener, sin necesidad de conocer ningúna norma ni estimación de los autovalores de A. Por otra parte, GMRES tiene en su contra que debe guardar la base del espacio de Krylov. Pero si hacemos la estrategia del "restart" que veremos más adelante, según la cual se realizan un cierto número m (bajo) de iteraciones de GMRES y obtenido el x_m se vuelve a iterar GMRES de cero desde x_m, entonces este GMRES con restart tendría una tasa de convergencia mejor que la mejor que podríamos obtener con el mejor parámetro de relajación ω, a un costo similar por iteración y con un requerimiento de capacidad de almacenamiento no demasiado alto.

Como esta estimación no depende de una diagonalización de A podríamos esperar que nos de alguna estimación de la convergencia en el caso en que A no es diagonalizable. Deafortunadamente, puede verificarse que para un bloque de Jordan de la forma

$$A = \begin{bmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & 0 & \dots \\ \vdots & 0 & \lambda & 1 & \dots \\ 0 & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & \dots & 0 & \lambda \end{bmatrix} \tag{9.330}$$

vale que $\|I - \omega A\|_2 > 1$ para todo ω, es decir que no hay ω que haga convergente a Richardson y, a la vez, nos permita obtener una estimación de la tasa de convergencia para GMRES. Too esto haría pensar que si la matriz no es diagonalizable (o *"casi no diagonalizable"*) entonces GMRES no convergerá. Pero si la forma de Jordan de una Matriz incluye un pequeño bloque de Jordan de dimension k_J y el resto es diagonalizable, entonces basta con tomar polinomios residuales de la forma

$$\bar{p}_k(z) = \left[\frac{(z - \lambda_J)}{\lambda_J}\right]^{k_J} q_{k-k_J}(z) \tag{9.331}$$

para $k > k_J$, donde λ_J es el autovalor correspondiente al bloque de Jordan y q es un polinomio apropiado para estimar una buena convergencia sobre el espectro de autovalores restantes. Por ejemplo, si el resto de los autovalores es real y positivo, entonces podríamos usar los polinomios de Tchebyshev usados para estimar la tasa de convergencia de GC.

9.3.3. Precondicionamiento

La forma de implementar el precondicionamiento en GMRES difiere con GC en cuanto a que para GMRES no es necesario que el sistema precondicionado

$$(MA)\, x = (Mb) \tag{9.332}$$

sea ni simétrico ni definido positivo. Entonces basta con encontrar un precondicionamiento M tal que $\|I - MA\|_2 < 1$ y se resuelve directamente el sistema precondicionado. Debido a esto, la rutina de GMRES no incluye nada en especial en cuanto al precondicionamiento, sino que directamente se pasa Mb como miembro derecho, y la rutina que calcula el producto matriz vector retorna $(MA)\, x$ en vez de $A\, x$.

Por otra parte se pueden usar *precondicionadores por derecha* y *por izquierda*. El precondicionamiento por izquierda es como fue expuesto en el párrafo anterior mientras que el precondicionamiento po derecha consiste en encontrar un M tal que $\|I - AM\|_2 < 1$ y entonces hacer el cambio de variable $x = My$, resolver con GMRES el sistema

$$(AM)\, y = b \tag{9.333}$$

y finalmente, una vez encontrado y obtener x de $x = My$.

En cuanto a las ideas para desarrollar precondicionadores son similares a las utilizadas con GC.

9.3.4. Implementación básica de GMRES

Recordemos que x_k se obtiene del problema de minimización (9.313). Sea V_k una matriz que contiene los vectores que expanden $\mathcal{K}_k$ es decir

$$V_k = \left[r_0\ Ar_0\ \ldots\ A^{k-1}r_0\right] \tag{9.334}$$

Entonces $x - x_0$ es una combinación lineal de las columnas de V_k, es decir

$$x - x - 0 = V_k y, \quad y \in \mathbb{R}^k \tag{9.335}$$

Entonces y debe ser tal que minimize

$$\|b - A(x_0 + V_k y)\|_2 = \|r_0 - AV_k y\|_2 \tag{9.336}$$

Sea ahora

$$B_k = AV_k = \left[Ar_0 \; A^2 r_0 \; \ldots \; A^k r_0 \right] \tag{9.337}$$

entonces el cuadrado de la norma se escribe como

$$\begin{aligned} \|r_0 - B_k y\|_2 \quad &= (r_0 - B_k y)^T (r_0 - B_k y) && (9.338) \\ &= r_0^T r_0 - 2 r_0^T B y + y^T \left(B^T B\right) y && (9.339) \end{aligned}$$

que alcanza su mínimo cuando

$$-B_k^T \, r_0 + (B_k^T B_k) y = 0 \tag{9.340}$$

lo cual ocurre cuando

$$y = (B_k^T B_k)^{-1} \, B_k^T r_0 \tag{9.341}$$

Esto puede implementarse en Octave con el comando

```
y=(B'*B)\B*r0
```

pero aún más simplemente

```
y=B\r0
```

hace lo mismo ya que para matrices rectangulares el operador \ se interpreta como resolver el sistema de mínimos cuadrados asociado.

En cuanto a la implementación práctica, notemos que el vector

$$\begin{aligned} q_k \quad &= B_k^T r_0 && (9.342) \\ &= \begin{bmatrix} r_0 \, A \, r_0 \\ r_0 \, A^2 \, r_0 \\ \vdots \\ r_0 \, A^k \, r_0 \end{bmatrix} = \begin{bmatrix} q_{k-1} \\ r_0 \, A^k \, r_0 \end{bmatrix} && (9.343) \end{aligned}$$

de donde vemos que en cada iteración sólo necesitamos calcular el último elemento del vector, lo cual involucra $O(N)$ operaciones. Algo similar ocurre con el cálculo de la matriz $H_k = B_k^T B_k$ a invertir.

$$\begin{aligned} H_k \quad &= B_k^T \, B_k && (9.344) \\ &= \begin{bmatrix} B_{k-1}^T \\ (A^k r_0)^T \end{bmatrix} \left[B_{k-1} \; A^k r_0 \right] && (9.345) \\ &= \begin{bmatrix} H_{k-1} & B_{k-1}^T A^k r_0 \\ (B_{k-1}^T A^k r_0)^T & r_0^T (A^k)^T A^k r_0 \end{bmatrix} && (9.346) \end{aligned}$$

con lo cual sólo es necesario calcular la última columna y el elemento diagonal. La última fila es igual a la transpuesta de la última columna ya que H_k es simétrica y definida positiva por construcción. El cálculo de esta última columna requiere de $O(kN)$ operaciones. Finalmente la inversión del sistema cuya matriz es H_k requiere $O(k^3)$ operaciones. Mientras mantengamos $k \ll N$ (más estrictamente es $k^2 \ll N$) este costo es despreciable contra las k^N operaciones.

9.3.5. Implementación en una base ortogonal

En la práctica es muy común ir cosntruyendo una base ortogonal de $\mathcal{K}_k$ mediante el *proceso de ortogonalización de Gram-Schmidt*. El algoritmo es el siguiente

1. $r_0 = b - Ax_0$, $v_1 = r_0 / \|r_0\|_2$
2. Para $i = 1, \ldots, k-1$
   $w = Av_i - \sum_{j=1}^{i}((Av_i)^T v_j)\, v_j$
   $v_{i+1} = w_i / \|w_i\|_2$

Si en algún i sucede que $w = 0$ entonces decimos que el algoritmo *falla* o *colapsa* (*"breakdown"*). Asumiendo que los $r_0, Ar_0, \ldots, A^{k-1}r_0$ son linealmente independientes (o sea que: $\dim \mathcal{K}_k = k$), entonces podemos que

- El algoritmo no falla.
- Los $\{v_i\}_{i=1}^{k}$ así generados forman una base ortogonal de $\mathcal{K}_k$.
- $v_k = \alpha_k\, A^{k-1}r_0 + w_k$ con $w_k \in \mathcal{K}_{k-1}$ y $\alpha_k \neq 0$.

Esto se puede demostrar por inducción en i. Para $i = 1$ es obvio, ya que v_1 es igual a r_0 normalizado. Para demostrarlo para $i+1$, observemos que w es otogonal a todos los v_j para $j \leq i$

$$v_j^T\, A\, v_i - \sum_{l=1}^{i}((Av_i)^T v_l)\, v_l) \qquad = v_j^T\, A\, v_i - \sum_{l=1}^{i}(Av_i)^T v_l)\, \delta_{jl} \tag{9.347}$$

$$= v_j^T\, A\, v_i - (Av_i)^T v_j) \tag{9.348}$$

$$= 0 \tag{9.349}$$

Ahora bien Av_i pertenece a $\mathcal{K}_{i+1}$ y los v_l para $l = 1, \ldots, i$ pertenecen a $\mathcal{K}_i$. Si $w \neq 0$ entonces podemos normalizar w apra obtener v_{i+1} y $\{v_l\}_{l=1}^{i+1}$ es un conjunto de vectores ortonormales en $\mathcal{K}_{i+1}$. Como $\mathcal{K}_{i=1}$ es de dimensión a lo sumo $i+1$, v_{i+1} y $\{v_l\}_{l=1}^{i+1}$ es una base ortogonal. Sólo falta demostrar entonces que bajo las hipótesis asumida el algoritmo no falla.

Ahora bien, por inducción podemos asumir que

$$Av_i = \alpha_i\, A^i r_0 + Aw_i \tag{9.350}$$

Pero si el algoritmo falla, entonces quiere decir que Av_i es una combinación lineal de los $\{v_l\} l = 1^i$ y eso implicaría que $A^i r_0$ es una combinación lineal de los mismos e indicaría que los $\{A^{j-1}r_0\}_{j=1}^{i}$ son linealmente dependientes.

Colapso de GMRES (Breakdown)

Puede ocurrir que $w = 0$ para algún i, en cuyo caso (por lo visto en el párrafo anterior) indicaría que $\{A^{j-1}r_0\}_{j=1}^{i}$ son linealmente dependientes. Podemos ver que esto ocurre si $x^* - x_0 \in \mathcal{K}_k$.

Lemma 3.4.1. Si $w_{i+1} = 0$ entonces $x^* = A^{-1}b \in \mathcal{K}_i$

Demostración. Si $w = 0$ para algún i entonces eso implica

$$Av_i = \sum_{i=1}^{j} \alpha_j \, v_j \in \mathcal{K}_i \tag{9.351}$$

Por construcción $Av_j \in \mathcal{K}_i$ para $j < i$, con lo que resulta que $A\mathcal{K}_i \subset \mathcal{K}_i$. Pero como

$$V_i = [v_1 \; v_2 \; \ldots \; v)i] \tag{9.352}$$

es una base de $\mathcal{K}_i$ entonces

$$AV_i = V_i H \tag{9.353}$$

para una cierta matriz H de $i \times i$. La columna j-ésima de H son los coeficientes de la expansión de Av_j en término de los $\{v_l\}_{l=1}^{j+1}$. H es no singular ya que A no lo es. Efeectivamente si $z \neq= 0$, $z \in \mathbb{R}^i$ es tal que $Hz = 0$, entonces $V_i z$ es un vector no nulo y

$$A\,(V_i z) = V_i\, H\, z = 0 \tag{9.354}$$

Consideremos ahora el residuo en la i-ésima iteración

$$r_i = b - Ax_i = r_0 - A(x_i - x_0) = V_i y \tag{9.355}$$

con $y \in \mathbb{R}^i$ ya que $x_i - x_0 \in \mathcal{K}_i$. Además r_0 por construcción es proporcional a v_1 la primera columna de V_i lo cual puede escribirse como

$$r_0 = \beta V_i e_1 \tag{9.356}$$

con $e_1^T = [1 \; 0 \; \ldots \; 0]$. Como V_i es ortogonal, preserva la norma

$$\begin{aligned} \|r_i\|_2^2 &= \|V_i\,(\beta e_1 - Hy)\|_2^2 && (9.357)\\ &= (\beta e_1 - Hy)^T\, V_i^T\, V_i (\beta e_1 - Hy) && (9.358)\\ &= \|(\beta e_1 - Hy)\|_2^2 && (9.359) \end{aligned}$$

Pero basta con tomar $y = \beta H^{-1} e_1$ para el cual $\|r_i\|_2 = 0$ y GMRES ha convergido.

9.3.6. El algoritmo de Gram-Schmidt modificado

Uno de los principales problemas de GMRES es que por errores de redondeo se va perdiendo la ortogonalidad de los vectores v_i, por lo cual debe prestarse especial atención al proceso de ortogonalización. Una primera observación es que la siguiente versión modificada del proceso de ortogonalización de Gram-Schmidt resulta ser mucho más estable ante errores de redondeo

$v_{k+1} = Av_k$
for $j = 1, \ldots, k$
  $v_{k+1} = v_{k+1} - (v_{k+1}^T v_j) v_j$

9.3.7. Implementación eficiente

La matriz H que contiene los coeficientes que expanden Av_i en término de los v_l tiene una estructura particular que permite una implementación más eficiente del algoritmo. Consideremos la expresión

$$v_{i+1} = \|w_{i+1}\|_2^{-1} \; (Av_i - \sum_{j=1}^{i} \alpha_j \, v_j) \tag{9.360}$$

Esto quiere decir que Av_i es una combinación lineal de los $\{v_j\}_{j=1}^{i+1}$. Como la columna j-ésima de H_k son los coeficientes de la expansión de Av_j en término de los $\{v_l\}_{l=1}^{j+1}$

$$A\,V_k = V_K\,H_k \tag{9.361}$$

vemos los h_{lj} deben ser de la forma $h_{lj} = 0$ para $l > j + 1$. Es decir

$$H_k = \begin{bmatrix} h_{11} & h_{12} & h_{13} & \dots \\ h_{21} & h_{22} & h_{23} & \dots \\ 0 & h_{32} & h_{33} & \dots \\ 0 & 0 & h_{43} & \dots \\ \ddots & \ddots & \ddots & \ddots \end{bmatrix} \tag{9.362}$$

Este tipo de matriz se llama *Hessenberg superior* (*upper Hessenberg*).

El residuo se puede escribir como

$$r_k \quad = b - Ax_k = r_0 - A(x_k - x_0) \tag{9.363}$$
$$= V_{k+1}(\beta e_1 - H_k y^k) \tag{9.364}$$

y como V_k es ortogonal

$$\|r_k\|_2 = \left\|\beta e_1 - H_k y^k\right\|_2 \tag{9.365}$$

Para resolver este problema se recurre a una estrategia similar a la anterior. Es decir, en Octave `y=beta*H\e1` si usamos la solución por mínimos cuadrados interna de Octave, o `y=beta*(H'*H)\H*e1` si lo resolvemos explícitamente. Finalmente, existen algoritmos especiales que permiten factorizar la matriz con un algoritmo QR en forma más eficiente usando la estructura Hessenberg superior de H_k.

Independientemente de la implementación de la resolución del problema de cuadrados mínimos, el algoritmo de GMRES es

Algoritmo `gmresb` $(x, b, A, \epsilon, k_{\max}, \rho)$

1. $r = b - Ax$, $v_1 = r/\|r\|_2$, $\rho = \|r\|_2$, $\beta = \rho$, $k = 0$
2. While $\rho > \epsilon \|b\|_2$ y $k < k_{\max}$
   (a) $k = k + 1$
   (b) $v_{k+1} = Av_k$. Para $j = 1, \dots, k$
      (i) $h_{jk} = v_{k+1}^T v - j$
      (ii) $v_{k+1} = v_{k+1} - h_{jk}\, v_j$

(c) $h_{k+1,k} = \|v_{k+1}\|_2$
(d) $v_{k+1} = v_{k+1} / \|v_{k+1}\|_2$
(e) $e_1 = [1\,0\,0\,\ldots\,0]^T$
$y_k = \operatorname{argmin}_{y \in \mathbb{R}^k} \left\|\beta e_1 - H_k y^k\right\|_2$
(f) $\rho = \left\|\beta e_1 - H_k\, y^k\right\|_2$
3. $x_k = x_0 + V_k\, y^k$

9.3.8. Estrategias de reortogonalización

Se puede perder ortogonalidad por errores de redondeo. Una solución es hacer un segundo paso de ortogonalización sea entodas las iyeraciones, sea cuando algún indicador de pérdida de ortogonalidad se activa.

9.3.9. Restart

Como debemos almacenar una base para $\mathcal{K}_k$ esto requiere kN reales lo cual va creciendo con k y eventualmente excede la memoria de la máquina. Entonces la idea es iterar GMRES m iteraciones y empezar de nuevo a partir de la última iteración, es decir aplicar GMRES inicializando con $x_0 \leftarrow x_m$. El siguiente algoritmo `gmresm` refleja esto,

Algoritmo $\texttt{gmresm}(x, b, A, \epsilon, k_{\max}, m, \rho)$
1. $\texttt{gmres}(x, b, A, \epsilon, m, \rho)$
2. $k = m$
3. While $\rho > \epsilon \|b\|_2$ y $k < k_{\max}$
(a) $\texttt{gmres}(x, b, A, \epsilon, m, \rho)$
(b) $k = k + m$

9.3.10. Otros métodos para matrices no-simétricas

GMRES, CGNE y CGNR comparten las siguientes (buenas) características

- Son fáciles de implementar
- Se analizan con residuos polinomiales

Por otra parte los CGN* tienen la desventaja de que necesitan un producto $A^T x$ y su tasa de convergencia está regida por $\kappa(A^T A) \sim \kappa(A)^2$, mientras que GMRES sólo necesita calcular $A\,x$ y la convergencia está basada en $\kappa(A)$, pero es necesario guardar una base de $\mathcal{K}_k$, lo cual representa una capacidad de almacenamiento que va creciendo con las iteraciones. Como esto es virtualmente imposible para grandes sistemas desde el punto de vista práctico se utiliza el GMRESm, lo cual puede involucrar un serio deterioro de la convergencia.

El método ideal debería ser como CG:

- Sólo necesitar calcular $A\,x$ (no $A^T\,x$)
- Estar basado en una propiedad de minimización o conjugación

- Requerir baja capacidad de almacenamiento. (Sobre todo que no crezca con las iteraciones).
- Converger en N iteraciones.

En esta sección describiremos algunos métodos que tratan de aproximarse a estos requerimientos con menor o mayor éxito.

Bi-CG: (por *Biconjugate Gradient Squared*) No se basa en una propiedad de minimización sino de ortogonalidad

$$r_k^T w = 0, \quad \text{para todo } w \in \overline{\mathcal{K}_k} \tag{9.366}$$

donde $\overline{\mathcal{K}_k}$

$$\overline{\mathcal{K}_k} = \text{span}\{\hat{r}_0, A^T \hat{r}_0, \ldots, (A^T)^{k-1} \hat{r}_0\} \tag{9.367}$$

es el *espacio de Krylov conjugado*. $\hat{r}_0$ es un vector que debe ser provisto por el usuario, pero lo más usual es tomar $\hat{r}_0 = r_0$. El algoritmo genera secuencias de residuos y direcciones de búsqueda $\{r_k, p_k\}$ y sus correspondientes conjugados $\{\hat{r}_k, \hat{p}_k\}$ tales que hay biortogonalidad entre los residuos

$$\hat{r}_k^T r_l = 0, \quad k \neq l \tag{9.368}$$

y de bi-conjugación entre las direcciones de búsqueda

$$\hat{p}_k^T A p_l = 0, \quad k \neq l \tag{9.369}$$

Si A es simétrica y definida positiva y $\hat{r}_0 = r_0$, entonces Bi-CG es equivalente a GC (pero computa todo el doble, es decir que tiene el doble de costo por iteración). Bi-CG necesita un cálculo $A^T x$, pero la ventaja con respecto a los CGN* es que su tasa de convergencia está basada en $\kappa(A)$ no en $\kappa(A^T A)$. Es muy útil si A es aproximadamente spd, es decir si los autovalores de A están cerca de la recta real.

Podemos comparar Bi-CG con GMRES en cuanto a la tasa de convergencia si consideramos que resulta ser que

$$r_k = \bar{p}_k(A)\, r_0 \tag{9.370}$$

con $\bar{p}_k$ un polinomio residual, y entonces por la propiedad de minimización de GMRES

$$\left\| (r_k)^{\text{GMRES}} \right\|_2 \leq \left\| (r_k)^{\text{Bi-CG}} \right\|_2 \tag{9.371}$$

pero debe recordarse que Bi-CG comparado con GMRES no necesita un espacio de almacenamiento creciente. Además una iteración de Bi-CG requiere dos productos matriz vector, pero el costo de GMRES también crece con las iteraciones.

CGS (por *Conjugate Gradient Squared*). Este método trata de ser una extensión de Bi-CG pero tal que no necesita calcular $A^T x$. Puede verse que en la iteración k

$$r_k = \bar{p}_k(A)\, r_0, \quad \hat{r}_k = \bar{p}_k(A^T)\, \hat{r}_0 \tag{9.372}$$

entonces el factor $r_k^T \hat{r}_k$ (que juega el papel de ρ_k en GC, (ver el algoritmo `cg(...)` en pág. 263), puede reescribirse de la forma

$$r_k^T \hat{r}_k = (\bar{p}_k(A)\, r_0)^T (\bar{p}_k(A^T)\, \hat{r}_0) \tag{9.373}$$

$$= ([\bar{p}_k(A)]^2\, r_0)^T \hat{r}_0 \tag{9.374}$$

en el cual no es necesario calcular $A^T x$. Con manipulaciones similares se puede llegar a transformar todos las expresiones donde aparece A^T, pero el algoritmo resulta modificado, es decir las iteraciones de Bi-CG *no son las mismas* que para CGS.

Bi-CGstab (por Biconjugate Gradient Squared stabilized). Trata de mejorar CGS reemplazando

$$r_k = q(k)\,\bar{p}_(k)\,r_0 \tag{9.375}$$

donde

$$q_k(z) = \prod_{i=1}^{k}(1-\omega_i z) \tag{9.376}$$

donde ω_i se selecciona para minimizar

$$\|r_i\|_2 = \|q_i(A)\,\bar{p}_i(A)\,r_0\|_2 \tag{9.377}$$

como función de ω_i, esto es usualmente conocido como *line-searching*. Sea

$$\begin{aligned} r_i \quad &= (1-\omega_i A)\left[\prod_{i=1}^{k}(1-\omega_i z)\,\bar{p}_i(A)\,r_0\right] && (9.378)\\ &= (1-\omega_i A)\,w && (9.379)\\ &= w-\omega_i A\,w && (9.380)\\ & && (9.381)\end{aligned}$$

de manera que

$$\begin{aligned} \|r_i\|_2^2 \quad &= (w-\omega_i A\,w)^T\,(w-\omega_i A\,w) && (9.382)\\ &= \|w\|_2^2 - 2\omega_i\,w^T\,Aw + \omega_i^2 (Aw)^T\,Aw && (9.383)\end{aligned}$$

y el valor que minimiza es

$$\omega_i = \frac{w^T(Aw)}{\|Aw\|_2^2} \tag{9.384}$$

Bi-CGstab entonces no necesita del cálculo de $A^T x$ como CGS pero tiende a tener una tasa de convergencia mejor. El costo computacional involucrado por iteración es

- Almacenamiento para 7 vectores
- 4 productos escalares
- 2 productos matriz-vector (Ax)

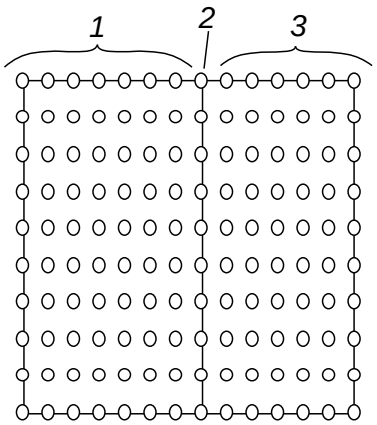


Figura 9.15: Descomposición de dominios

9.3.11. Guía Nro 3. GMRES

Consideremos la ec. de *advección-difusión*

$$k\phi'' - a\phi' = 0 \tag{9.385}$$
$$\phi(0) = 0 \tag{9.386}$$
$$\phi(1) = 1 \tag{9.387}$$

donde

- $\phi =$ temperatura del fluido
- $k =$ conductividad térmica del fluido
- $a =$ velocidad del fluido (puede ser positiva o negativa)

La solución exacta es

$$\phi(x) = \frac{\mathrm{e}^{2\mathrm{Pe}x} - 1}{\mathrm{e}^{2\mathrm{Pe}} - 1} \tag{9.388}$$

donde

$$\mathrm{Pe} = \frac{aL}{2k} \tag{9.389}$$

Aquí $L = 1$. En la figura 9.16 vemos Para $a \to 0$ el problema se acerca a la ec. de Laplace y la solución tiende a ser una recta que une los valores de contorno, mientras que para Pe grandes y positivos el fluido va en la dirección $+x$ y arrastra el valor de la condición *aguas arriba* una cierta longitud hasta que a una distancia pequeña δ sube rápidamente para tomar el valor dado por la condición en $x = 1$.

La discretización numérica pasa por reemplazar las derivadas por diferencias numéricas

$$\phi_j'' \approx (\phi_{j+1} - 2\phi_j + \phi_{j-1})/h^2 \tag{9.390}$$

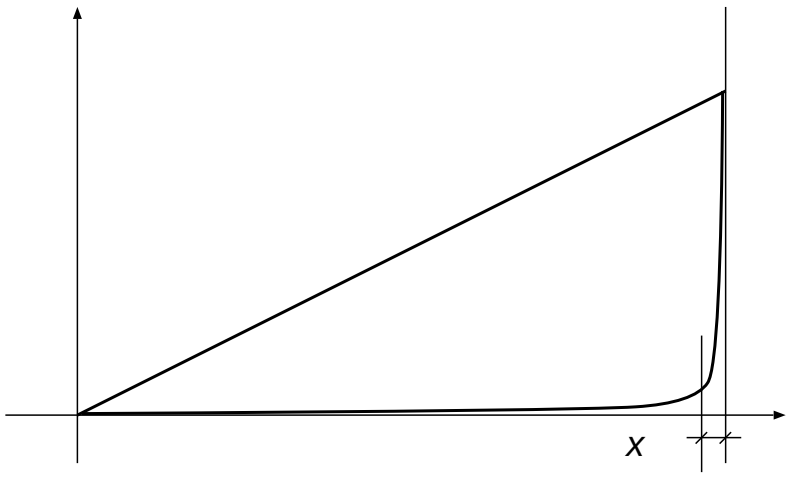


Figura 9.16: Solución para el problema de advección difusión, en los límites dominado por difusión y por advección

y

$$\phi'_j \approx= (\phi_{j+1} - \phi_{j-1})/(2h) \tag{9.391}$$

Lamentablemente, esta discretización falla cuando el Pe $\gg 1$, en el sentido de que produce fuertes oscilaciones numéricas. La solución usual es agregar una cierta cantidad de *difusión numérica*, es decir que se modifica la ecuación original a

$$(k + k_{\text{num}})\, \phi'' - a\phi' = 0 \tag{9.392}$$

donde

$$k_{\text{num}} = (ah/2) \left[\frac{1}{\text{Pe}_h} - \frac{1}{\tanh(\text{Pe}_h)} \right] \tag{9.393}$$

$$\text{Pe}_h = \frac{ah}{2k} \tag{9.394}$$

Realizar las siguientes tareas:

1. Calcular la matriz para $\text{Pe}_h = 0.01$, 1 y 100
2. Calcular los autovalores. Ver la distribución en el plano complejo.
3. Verificar que la matriz no es simétrica. Ver que la parte correspondiente a ϕ'' es simétrica mientras que la que corresponde a ϕ' es antisimétrica.
4. Ver a que tiende la matriz para $k \to 0$. Es diagonalizable?
5. Resolver por GMRES y CGNE.
6. Pensar como se podría hacer para calcular $A^T x$ sin construir A.

9.4. Descomposición de dominios.

Consideremos la solución de una ecuación en derivadas parciales en un dominio como muestra la figura 9.15. Descomponemos el problema en dos dominios de manera que podemos separar las incógnitas en x en tres grupos, aquellos que están estrictamente contenidos en el dominio de la izquierda x_1, los estricatamente contenidos en el dominio de la derecha x_3 y los que están sobre la interfase x_2. La ecuación se descompone en bloques de la siguiente forma

$$\begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & A_{23} \\ 0 & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{9.395}$$

La descomposición en bloques refleja el hecho de que como los grados de libertad contenidos en x_1 y x_3 no están compartidos por ningún elemento, los bloques correspondientes A_{13} y A_{31} son nulos. Podemos eliminar x_1 y x_3 de la primera y última línea de ecuaciones y obtener una ecuación para los grados de libertad de interfase x_2

$$(-A_{23}A_{33}^{-1}A_{32} + A_{22} - A_{21}A_{11}^{-1}A_{12})x_2 = b_2' \tag{9.396}$$

$$S\,x_2 = b_2' \tag{9.397}$$

Ahora consideremos resolver este sistema por GC. En cada iteración debemos calcular el producto de la matriz entre paréntesis por un vector x_2. Para los términos primero y tercero de la matriz es necesario factorizar y resolver un sistema lineal con las matrices A_{33} y A_{11}. Esto corresponde a resolver problemas independientes sobre cada uno de los dominios con condiciones Dirichlet sobre la interfase, por lo tanto estos problemas pueden ser resueltos en procesadores independientes lo cual implica un alto grado de paralelización del problema. Esto se puede extender a más procesadores, y en el límite podemos lograr que en cada procesador se resuelva un sistema lineal lo suficientemente pequeño como para ser resuelto en forma directa. La eficiencia del método puede ser mejorada aún más encontrando un buen precondicionamiento.

Si separamos la matriz S que aparece en el sistema (9.397) en la contribución por el dominio de la izquierda S_L y de la derecha S_R

$$S = S_R + S_L \tag{9.398}$$

$$S_L = (1/2)A_{22} - A_{21}A_{11}^{-1}A_{12} \tag{9.399}$$

$$S_R = -A_{23}A_{33}^{-1}A_{32} + (1/2)A_{22} \tag{9.400}$$

Entonces podemos poner como precondicionamiento

$$M = (1/4)\,(S_L^{-1} + S_R^{-1}) \tag{9.401}$$

Es M un buen precondicionamiento? O mejor dicho: Porqué habría M de parecerse a S^{-1}? Bien, si el operador es simétrico (el laplaciano lo es) y los dominios son iguales y la malla es simétrica, entonces puede verse que $S_L = S_R$ y entonces M y S^{-1} coinciden

$$M = S^{-1} = (1/2)S_L^{-1} \tag{9.402}$$

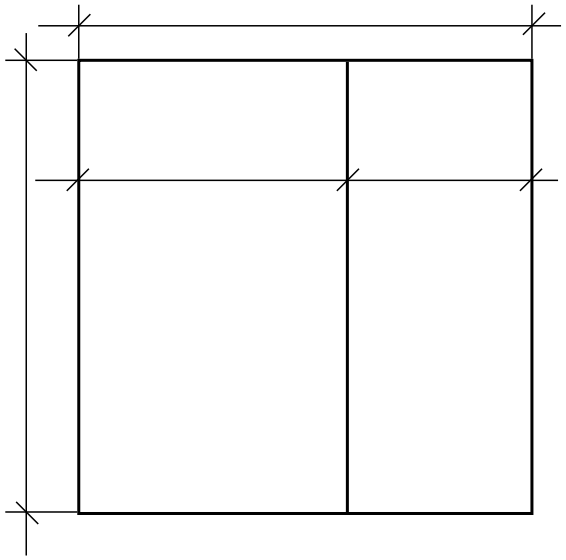


Figura 9.17: Descomposición de dominios. Problema del continuo.

Por otra parte resolver $x = My$ es equivalente a resolver

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & (1/2)\,A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_{2L} \end{bmatrix} = \begin{bmatrix} 0 \\ (1/2)\,y \end{bmatrix} \tag{9.403}$$

$$\begin{bmatrix} (1/2)\,A_{22} & A_{23} \\ A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x_{2R} \\ x_3 \end{bmatrix} = \begin{bmatrix} (1/2)\,y \\ 0 \end{bmatrix} \tag{9.404}$$

y después

$$x = (1/2)(x_{2L} + x_{2R}) \tag{9.405}$$

y el punto es que ambos sistemas (9.403) y (9.404) equivale a resolver problemas *independientes* con condiciones tipo Neumann sobre la interfase. De nuevo, el hecho de que ambos problemas sobre cada dominio sean independientes favorece la paralelización del algoritmo.

9.4.1. Condicionamiento del problema de interfase. Análisis de Fourier.

Obviamente la aplicabilidad de la descomposición de dominios descripta depende del número de iteraciones necesarias para resolver el problema de interfase y por lo tanto del número de condición de la matriz complemento de Schur S o de su precondicionada MS. Para hacer una estimación de estos números de condición nos basaremos en un análisis de Fourier del problema del continuo para la ecuación de Laplace. Las ecuaciones de gobierno son

$$\Delta\phi = -f \text{ en } \Omega \tag{9.406}$$

$$\phi = \bar{\phi} \text{ en } \Gamma_1 \tag{9.407}$$

$$\tag{9.408}$$

Consideremos la solución en dos dominios Ω_1, Ω_2. La restricción de ϕ a cada uno de los dominios debe satisfacer

$$\Delta\phi_1 = -f \text{ en } \Omega_1 \tag{9.409}$$
$$\phi_1 = \bar{\phi}_1 \text{ en } \Gamma_1 \tag{9.410}$$
$$\tag{9.411}$$

y

$$\Delta\phi_2 = -f \text{ en } \Omega_2 \tag{9.412}$$
$$\phi_2 = \bar{\phi}_2 \text{ en } \Gamma_2 \tag{9.413}$$
$$\tag{9.414}$$

y la continuidad de ϕ y su derivada normal a través de Γ_I

$$(\phi_1)_{\Gamma_I} = (\phi_2)_{\Gamma_I} \tag{9.415}$$
$$\left(\frac{\partial\phi_1}{\partial x}\right)_{\Gamma_I} = \left(\frac{\partial\phi_2}{\partial x}\right)_{\Gamma_I} \tag{9.416}$$

Ahora consideremos una descomposición $\phi = \psi + \tilde{\phi}$, de manera que $\psi = 0$ en Γ_I, es decir

$$\Delta\psi_1 = -f \text{ en } \Omega_1 \tag{9.417}$$
$$\psi_1 = \bar{\psi}_1 \text{ en } \Gamma_1 \tag{9.418}$$
$$\psi_1 = 0 \text{ en } \Gamma_I \tag{9.419}$$

y

$$\Delta\psi_2 = -f \text{ en } \Omega_2 \tag{9.420}$$
$$\psi_2 = \bar{\phi}_2 \text{ en } \Gamma_2 \tag{9.421}$$
$$\psi_2 = 0 \text{ en } \Gamma_I \tag{9.422}$$

y por lo tanto falta hallar $\tilde{\phi}$ definido por

$$\begin{aligned}\Delta\tilde{\phi}_1 &= 0 \text{ en } \Omega_1\\ \tilde{\phi}_1 &= 0 \text{ en } \Gamma_1\\ \tilde{\phi}_1 &= u \text{ en } \Gamma_I\end{aligned} \tag{9.423}$$

y

$$\begin{aligned}\Delta\tilde{\phi}_2 &= 0 \text{ en } \Omega_2\\ \tilde{\phi}_2 &= 0 \text{ en } \Gamma_2\\ \tilde{\phi}_2 &= u \text{ en } \Gamma_I\end{aligned} \tag{9.424}$$

donde u es una incógnita del problema y debe ser tal que

$$\left(\frac{\partial\tilde{\phi}_1}{\partial x}\right)_{\Gamma_I} - \left(\frac{\partial\tilde{\phi}_2}{\partial x}\right)_{\Gamma_I} = \tilde{b} \tag{9.425}$$

donde

$$\tilde{b} = -\left\{\left(\frac{\partial\psi_1}{\partial x}\right)_{\Gamma_I} - \left(\frac{\partial\psi_2}{\partial x}\right)_{\Gamma_I}\right\} \tag{9.426}$$

Definimos ahora el operador de *Stekhlov-Poincaré* $\mathcal{S}_1$ del dominio Ω_1 como

$$\mathcal{S}_1\{u\} = \frac{\partial\phi_1}{\partial n} = \frac{\partial\phi_1}{\partial x} \tag{9.427}$$

donde $\frac{\partial\phi_1}{\partial n}$ denota la derivada normal tomando la normal exterior al dominio en cuestión, que para el dominio Ω_1 coincide con la dirección de $+x$, y ϕ_1 es la solución de (9.423), para el correspondiente u. Análogamente se puede definir $\mathcal{S}_2$ por

$$\mathcal{S}_2\{u\} = \frac{\partial\phi_2}{\partial n} = -\frac{\partial\phi_2}{\partial x} \tag{9.428}$$

donde ϕ_2 está definido en función de u por (9.424) y el signo $-$ viene de que la normal exterior a Ω_2 sobre Γ_I va ahora según la dirección $-x$. Entonces

$$\mathcal{S}\{u\} = g \tag{9.429}$$

donde $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$.

Ec. (9.429) es la versión del continuo de (9.397). Calcularemos el número de condición de S a partir del cálculo de autovalores de $\mathcal{S}$.

Cada uno de los operadores de Stekhlov-Poincaré actúa sobre el espacio de funciones definidas sobre Γ_I. Podemos mostrar que las funciones de la forma

$$u_m = \sin(k_m y), \quad k_m = m\pi/L, \quad m = 1, 2, \ldots, \infty \tag{9.430}$$

son autofunciones de estos operadores. La solución $\tilde{\phi}_1$ que es solución de (9.423) correspondiente a $u = u_m$ es de la forma

$$\tilde{\phi}_1 = \hat{\phi}(x)\ \sin(k_m y), \tag{9.431}$$

Reemplazando en las ecuaciones que definen (9.423), la primera de las ecuaciones resulta ser

$$\hat{\phi}'' - k_m^2\hat{\phi} = 0 \tag{9.432}$$

de donde

$$\hat{\phi}(x) = a\, e^{k_m y} + b\, e^{-k_m y} \tag{9.433}$$

y de las condiciones de contorno en $x = 0$, L_1 resulta ser

$$\hat{\phi}(x) = \frac{\sinh k_m y}{\sinh k_m L_1} \tag{9.434}$$

la derivada normal en Γ_I es entonces

$$\frac{\partial \tilde{\phi}_1}{\partial n} = k_m \frac{\cosh k_m L_1}{\sinh k_m L_1} = \frac{k_m}{\tanh k_m L_1} \tag{9.435}$$

de manera que

$$\mathcal{S}_1\{u_m\} = \left(\frac{k_m}{\tanh k_m L_1} \right) u_m \tag{9.436}$$

Vemos entonces, que u_m es una autofunción de $\mathcal{S}_1$ con autovalor

$$\lambda_m^1 = \frac{k_m}{\tanh k_m L_1} \tag{9.437}$$

Análogamente, el operador $\mathcal{S}_2$ tiene autovalores

$$\lambda_m^2 = \frac{k_m}{\tanh k_m L_2} \tag{9.438}$$

El operador $\mathcal{S}$ tiene autovalores

$$\lambda_m = k_m \left(\frac{1}{\tanh k_m L_1} + \frac{1}{\tanh k_m L_2} \right) \tag{9.439}$$

Ahora bien, dijimos que estimaríamos el número de condición de la matriz S a partir de los autovalores de $\mathcal{S}$. En el continuo $\mathcal{S}$ tiene infinitos autovalores y los λ_m van a infinito para $m \to \infty$, de manera que el número de condición de $\mathcal{S}$ va a infinito. Obtendremos una estimación para el número de condición de S asumiendo que los autovalores de S se aproximan a los primeros N_I autovalores de $\mathcal{S}$, donde N_I es la dimensión de S. El autovalor más bajo es

$$\lambda_{\min} = \lambda_1 = \frac{\pi}{L} \left(\frac{1}{\tanh(L_1/L)} + \frac{1}{\tanh(L_2/L)} \right) \tag{9.440}$$

si $L_1 = L_2 = L/2$, entonces

$$\lambda_{\min} = \frac{\pi}{L} \frac{2}{\tanh(1/2)} = \frac{13.6}{L} \tag{9.441}$$

El autovalor máximo se obtiene para $m = N_I$ y asumiendo que $N_I \gg 1$ y $L_1/L, L_2/L$ no son demasiado pequeños, entonces $k_m L_1 = N_I \pi L_1/L \gg 1$ y $\tanh k_m L_1 \approx 1$ y similarmente $\tanh k_m L_2 \approx 1$ y por lo tanto

$$\lambda_{\max} = 2k_m = 2N_I \pi / L \tag{9.442}$$

y el número de condición es

$$\kappa(S) \approx \tanh(1/2)\, N_I = 0.46\, N_I \tag{9.443}$$

Notar que $\kappa(S)$ va como $1/h$ al refinar, mientras que recordemos que el número de condición de A (la matriz del sistema) va como $1/h^2$ (Guía de ejercicios Nro. 1).

Otro punto interesante a notar es que, para $m \to \infty$ los autovalores tienden a hacerse independientes de las longitudes de los dominios en la dirección normal a la interfase. Por ejemplo, para los autovalores λ_m^1 de $\mathcal{S}_1$, L_1 aparece en el argumento de la tangente hiperbólica y esta tiende a 1 para $m \to \infty$, independientemente del valor de L_1. Incluso puede verse que para $m \to \infty$ los autovalores se hacen independientes del

tipo de condición de contorno sobre el borde opuesto (es decir sobre $x = 0$). Esta observación se basa en el hecho de que el operador de Laplace es muy *local*. Si u es de la forma $\sin m\pi y/L$ sobre Γ_I, entonces la longitud de onda es $\beta = 2L/m$ la cual se va achicando a medida que $m \to \infty$. Las perturbaciones inducidas decaen como $e^{-n/\beta}$ donde n es una coordenada en la dirección normal a Γ_I y si β es muy pequeño la perturbación no llega al contorno opuesto.

Ahora consideremos los autovalores del problema precondicionado. Como todos las u_m de la forma (9.430) son autofunciones de los operadores $\mathcal{S}_1$, $\mathcal{S}_2$ y $\mathcal{S}$, entonces los autovalores de MS son aproximadamente los primeros N_I autovalores de $(1/4)(\mathcal{S}_1^{-1} + \mathcal{S}_2^{-1})(\mathcal{S}_1 + \mathcal{S}_2)$, es decir

$$\lambda_m^{\text{prec}} = (1/4)\left[(\lambda_m^1)^{-1} + (\lambda_m^2)^{-1}\right](\lambda_m^1 + \lambda_m^2) \tag{9.444}$$

Como mencionamos previamente (ver pág. 287), si el problema es simétrico (en el caso del continuo basta con $L_1 = L_2$, en el caso del discreto además la malla también tiene que ser simétrica alrededor de $x = 1/2$), entonces $M = S^{-1}$. En el caso del continuo ocurre lo mismo ya que si los dos dominios son iguales, entonces

$$\lambda_m^1 = \lambda_m^2 \tag{9.445}$$

y por lo tanto $\lambda_m^{\text{prec}} = 1$ para todo m. Si $L_1 \neq L_2$, entonces puede verse que para m grandes $\lambda_1 \approx \lambda_2$ ya que ambos se hacen independientes de $L_{1,2}$ y de la condición de contorno en el contorno opuesto y por lo tanto

$$\lambda_m^{\text{prec}} \to 1 \text{ para } m \to \infty \tag{9.446}$$

Pero entonces esto quiere decir que $\kappa(MS)$ se hace independiente de N_I para m suficientemente grandes. Este resultado es muy importante desde el punto de vista práctico, *el número de condición del problema precondicionado no se deteriora bajo refinamiento para el precondicionamiento Dirichlet-to-Neumann.*

9.5. Guía de Trabajos Prácticos

1. Generar aleatoriamente una matriz simétrica y positiva definida A y un miembro derecho b con
```
n=5; c=1;
a=rand(n);
a=(a+a');
a=expm(c*a);
b=rand(n,1);
```
Porqué es esta matriz spd?
2. Determinar para qué valores del parámetro de relajación ω el esquema de Richardson es convergente. Determinar la tasa de convergencia.

$$x_{k+1} = x_k + \omega\,(b - Ax_k) \tag{9.447}$$

3. Observar el comportamiento de una componente dada en función de las iteraciones (Graficar curvas de $(xk)_i$ en función de k. Que se observa para $\omega = \omega_{\text{opt}}$? Explicar.
4. **Ecuación de Poisson 1D:** Consideremos la ecuación de Poisson 1-dimensional en un dominio $0 < x < 1$ con condiciones Dirichlet en $x = 0, 1$:

$$\phi'' = -f \tag{9.448}$$

Consideramos una discretización por diferencias finitas de paso constante $h = 1/N$. Los nodos son $x_j = jh$, $j = 0, \ldots, N$ y las ecuaciones para los nodos interiores $1 < j < N - 1$ son

$$h^{-2}(-\phi_{j+1} + 2\phi_j - \phi_{j-1}) = f_j \tag{9.449}$$

ϕ_0 y ϕ_N son datos dados por las condiciones de contorno Dirichlet. El sistema puede ponerse de la forma $Ax = b$ con

$$A = h^{-2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ 0 & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \tag{9.450}$$

a) Verificar que A es simétrica y definida positiva.

b) Calcular los autovalores de A y el número de condición de A, para $N = 10, 30, 100$.

c) Verificar que el número de condición de A se comporta de la forma $\text{cond}(A) \sim N^2 = 1/h^2$, al refinar.

d) Graficar las autofunciones de A para los autovalores más bajos y para los más altos.

e) Verificar cuán buena es la aproximación de $\|A\|$ al radio espectral de A.

f) Efectuar experimentos numéricos con $\omega = \omega_{\text{opt}}, 0.7\omega_{\text{opt}}$, etc... como antes.

g) Evaluar las tasas de convergencia en forma experimental y teórica.

5. **Ecuación de Laplace 2-dimensional:** Lo mismo que en el ejercicio anterior pero en 2D en el dominio $0 < x, y < 1$ con condiciones dirichlet homogéneas en todo el contorno y $f = 1$. Una ventaja de los métodos iterativos es que no es necesario armar la matriz del sistema. En efecto, sólo necesitamos poder calcular el residuo r_k a partir del vector de estado x_k. Estudiar la implementación que se provee a través del script `poisson.m` que llama sucesivamente a la función `laplacian.m`. En `poisson.m` el vector de estado es de tamaño $(N-1)^2$ donde hemos puesto todos los valores de ϕ encolumnados. La función `laplacian(phi)` calcula el laplaciano del vector de iteración ϕ. El laplaciano es calculado convirtiendo primero el vector de entrada a una matriz cuadrada de $(N-1) \times (N-1)$ y despues se evalúa la aproximación estándar de 5 puntos al laplaciano

$$(\Delta\phi)_{ij} = h^{-2}(\phi_{i,j+1} + \phi_{i,j-1} + \phi_{i+1,j} + \phi_{i-1,j} - 4\phi_{ij}) \tag{9.451}$$

   a) Estimar el autovalor máximo con la norma 1 de A.

   b) Efectuar experimentos numéricos con varios valores de ω. Evaluar tasas de convergencia en forma experimental.

6. **Analogía entre el método de Richardson relajado y la solución seudo-temporal.** Consideremos el sistema ODE's

$$\frac{dx}{dt} = -Ax + b \tag{9.452}$$

   Entonces si A tiene autovalores con parte real positiva, la solución $x(t)$ tiende a la la solución de $Ax = b$ para $t \to \infty$. Entonces podemos generar métodos iterativos integrando este sistema en forma numérica. *Consigna:* Demostrar que aplicar el método de forward Euler a esta ecuación ($dx/dt \approx (x_{k+1} - x_k)/\Delta t$) equivale al método de Richardson, donde el paso de tiempo equivale al factor de relajación ω.
7. Si f(x) es una forma cuadrática $f(x) = \frac{1}{2}x^T Ax - b^T x + c$, donde A es una matriz en $\mathbb{R}^{n\times n}$, x y b son vectores en $\mathbb{R}^n$ y c es un escalar constate. Calcular $f'(x)$. Mostrar que si A es *spd*, $f(x)$ es minimizada por la solución de $Ax = b$.
8. Para el Método de Steepest Descent demostrar que si el error e_i en la iteración i es un autovalor de la matriz A cuyo autovalor es λ_e, se convergerá a la solución exacta en la próxima iteración, i.e., $i + 1$.
9. Dar una interpretación geométrica del ejercicio anterior y decir cuanto tiene que valer α (la long del paso en la dirección de búsqueda) para obtener convergencia inmediata.
10. **GC como método directo**.
    Resolver la ecuación de Poisson

$$\Delta\phi = -f, \text{ en } \Omega = \{x, y \,/\, 0 \le x, y \le 1\} \tag{9.453}$$

$$\phi = 0, \text{ en } \partial\Omega \tag{9.454}$$

    con una grilla de diferencias finitas de $(N+1) \times (N+1)$ puntos. Usar $N = 4, 6, 8$ y 10 y ver en cuantas iteraciones converge a precisión de máquina. Calcular los autovalores de A y ver cuantos distintos hay. Inicializar con `x0=rand(n,1)` y ver en cuantas iteraciones converge. Porqué?. Puede usar las rutinas de Octave provistas por la cátedra.
11. **GC como método iterativo**.
    Idem que el ej. anterior pero ahora para $N = 100$. (No tratar de construir la matriz! La matriz llena ocupa 800Mbytes y la banda 8Mbytes). Graficar la curva de convergencia y comparar el n experimental con

el teórico (basado en una estimación del número de condición de la matriz). Comparar la convergencia con el método de Richardson (con $\omega = \omega_{\text{ópt}}$), en números de iteraciones y en tiempo de CPU. Puede usar las rutinas de Octave provistas por la cátedra.

12. Resolver el punto anterior en el cluster en forma secuencial usando las rutinas de PETSc provistas, con y sin precondicionamiento de Jacobi (*point Jacobi*). Sacar conclusiones acerca de los resultados obtenidos en relación a los resultados de los puntos anteriores.
13. Resolver los puntos anteriores en el cluster usando 4 procesadores, con y sin precondicionamiento de Jacobi (*point Jacobi*). Sacar conclusiones acerca de los resultados obtenidos en relación a los resultados de los puntos anteriores.
14. Verificar la escalabilidad del Método CG (con prec. point Jacobi) para el problema de Poisson usando una grilla de $100 \cdot \sqrt{n_{\text{proc}}} \times 100 \cdot \sqrt{n_{\text{proc}}}$. Es decir el comportamiento de la cantidad de iteraciones de CG para bajar el residuo un cierto orden de magnitud (por ejemplo 5 órdenes) en función de la cantidad de procesadores cuando el problema crece en la misma proporción al número de procesadores n_{proc}. Sacar conclusiones acerca de la escalabilidad del algoritmo.

Capítulo 10

Flujo incompresible

10.1. Definición de flujo incompresible

Un flujo incompresible es áquel donde el fluido no se comprime, como es típicamente el caso de los líquidos, pero también puede pasar que bajo ciertas condiciones un fluido que es compresible (como los gases en general) no manifiesta efectos de compresibilidad para un patrón o régimen de flujo en particular. En ese caso se le asigna a la propiedad de flujo compresible o incompresible al patrón de flujo. Para los fluidos compresibles, puede demostrarse que los efectos compresibles van con el n'umero de Mach al cuadrado, es decir que la variación relativa de la densidad

$$\frac{\Delta\rho}{\rho} = O(M^2) \tag{10.1}$$

donde

$$\mathrm{M} = \frac{u}{c} \tag{10.2}$$

es el número de Mach, u es la velocidad del fluido y c es la velocidad del sonido. Podemos decir entonces que el flujo es compresible si el número de Mach es menor que un cierto valor, digamos 0.1. Por ejemplo, un auto a 100 Km/h en atmósfera estándar posee un Mach de approx. 0.1, con lo cual en esas condiciones podemos considerar que el flujo es incompresible.

Es de notar que si las variaciones de densidad son provocadas por otros efectos que no sean la presión mecánica como la dilatación térmica, expansión solutal (p.ej. salinidad), etc... entonces el patrón de flujo puede considerarse (con respecto a los efectos sobre los algoritmos numéricos) incompresible, aún si la densidad resulta no ser constante ni espacialmente ni en el tiempo. El término compresible/incompresible se aplica a las variaciones de densidad producidad exclusivamente por efecto de la presión.

Si bien en principio uno podría pensar que la incompresibilidad es una ventaja, ya que permite eliminar (en muchos casos) una variable (la densidad), desde el punto de vista numérico suele traer más problemas que soluciones.

10.2. Ecuaciones de Navier-Stokes incompresible

Las ecuaciones de Navier-Stokes incompresibles son

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \Delta \mathbf{u} \tag{10.3}$$

$$\nabla \cdot \mathbf{u} = 0 \tag{10.4}$$

La primera es la *"ecuación de momento"*, mientras que la segunda es la *"ecuación de continuidad"* o *"balance de masa"*. Es importante notar que en el límite de *"flujo reptante"* o *"flujo de Stokes"* (es decir, despreciando el término convectivo), las ecuaciones resultantes son exactamente iguales a las de elasticidad lineal incompresible isotrópica, si reemplazamos el vector de velocidad por el de desplazamiento y la viscosidad por el módulo de elasticidad.

Las siguientes observaciones nos permiten adelantar el problema ocasionado por la incompresibilidad:

- **La condición de incompresibilidad no tiene un término temporal:** Esto quiere decir que "la presión no tiene historia". El estado del fluido sólo stá dado por la velocidad. También podemos decir que la ecuación de continuidad aparece como una restricción, más que como una ecuación de evolución. La presión, pasa a ser el multiplicador de Lagrange asociado.
- **Las ecuaciones son no locales:** Esto es más fácil de ver en el caso de elasticidad. Se sabe bien que en el caso de elasticidad compresible el problema es elíptico, de manera que hay una cierta localidad de los efectos. Esto se pierde en el caso incompresible. Por ejemplo, consideremos un sólido incompresible que ocupa una región Ω (ver figura 10.1). Las condiciones son de desplazamiento nulo en toda la frontera, menos en una cierta parte Γ_1 donde se aplica un cierto desplazamiento uniforme, y otra cierta parte Γ_2 donde las condiciones son libres, es decir tracción nula. En el caso compresible, el operador es elíptico, local, y la influencia del desplazamiento impuesto sobre el dominio Γ_1 en el dominio Γ_2 dependerá de la distancia entre ambas regiones, sus tamaños relativos, etc... Si el tamaño de ambas regiones es similar y muy pequeños con respecto a la distancia que los separa, entonces los desplazamientos en Γ_2 serán despreciables. Por el contrario, en el caso incompresible, el cambio de volumen total en Γ_2 debe ser igual al impuesto en Γ_1, por lo tanto los desplazamientos en Γ_2 serán del mismo orden que aquellos impuestos en Γ_1 (asumiendo que ambas regiones de la frontera tienen dimensiones similares).
- **Cambia el cáracter matemático de las ecuaciones:** También en el caso elástico, estacionario las ecuaciones dejan de ser elípticas al pasar al caso incompresible. Esto se debe a que la ecuación de continuidad *"no tiene término en derivadas segundas"*.
- **La ecuación de la energía se desacopla de la de momento y continuidad:** El campo de temperaturas se puede obtener a posteriori a partir de el campo de velocidades obtenido.

10.3. Formulación vorticidad-función de corriente

La vorticidad se define como

$$\boldsymbol{\Omega} = \nabla \times \mathbf{u} \tag{10.5}$$

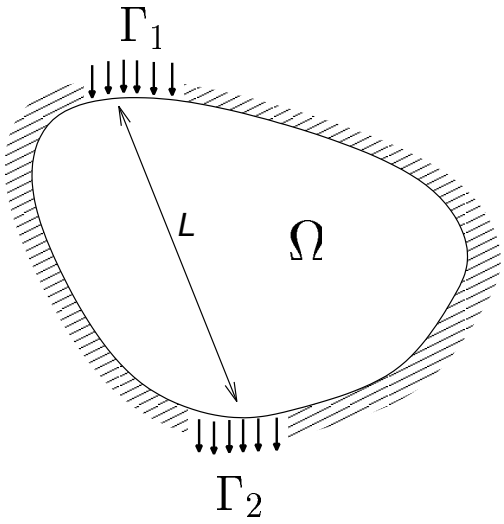


Figura 10.1: No localidad en elasticidad incompresible.

el cual, para un flujo bidimensional se reduce a

$$\Omega = \Omega_z = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \tag{10.6}$$

En 2D se puede encontrar una función de corriente ψ tal que

$$u = \frac{\partial \psi}{\partial y} \tag{10.7}$$

$$v = -\frac{\partial \psi}{\partial x} \tag{10.8}$$

Tomando rotor de (10.3) se llega, despues de un cierto trabajo algebraico, a

$$\frac{\partial \mathbf{\Omega}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{\Omega} - (\mathbf{\Omega} \cdot \nabla)\mathbf{u} = \nu \Delta \mathbf{\Omega} \tag{10.9}$$

pero (sólo en 2D!) el tercer término es nulo, ya que $\nabla \mathbf{u}$ debe estar en el plano y $\mathbf{\Omega}$ está fuera del plano, de manera que la ecuación se reduce a una ecuación de advección difusión para la vorticidad

$$\frac{\partial \Omega}{\partial t} + (\mathbf{u} \cdot \nabla)\Omega = \nu \Delta \Omega \tag{10.10}$$

Por otra parte, recombinando (10.6) con (10.7) se llega a una ecuación de Poisson para la función de corriente:

$$\Delta \psi = -\Omega \tag{10.11}$$

La *"formulación vorticidad/función de corriente"* consiste en resolver (10.10) y (10.11) en forma acoplada.

Las ventajas y desventajas de la formulación, con respecto a la formulación en variables primitivas (10.3-10.4) son

- La extensión a 3D de la formulación vorticidad/función de corriente es muy compleja.
- La formulación vorticidad/función de corriente tiene un grado de libertad menos por nodo.
- Las condiciones de contorno para la presión son desconocidas para la formulación en variables primitivas.
- Las condiciones de contorno para la vorticidad son desconocidas para la formulación vorticidad/función de corriente .
- La formulación vorticidad/función de corriente requiere de cierto cuidado en cuanto a la discretización.

10.4. Discretización en variables primitivas

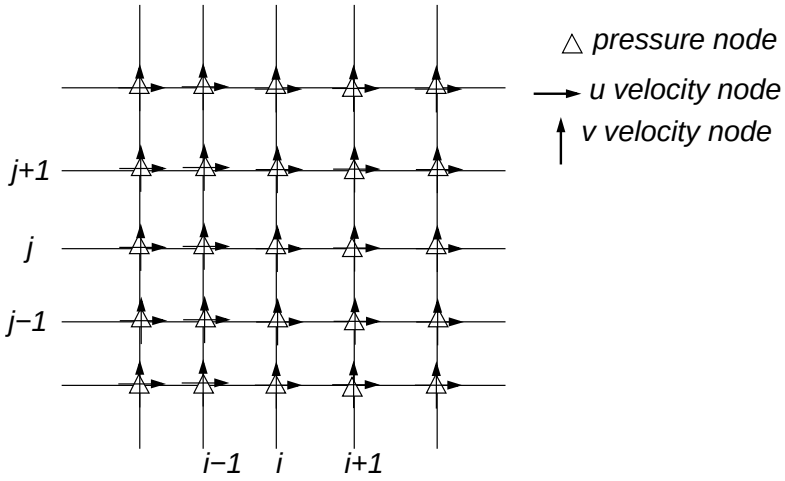


Figura 10.2: Grilla estándar (no staggered) usada para flujo incompresible en variables primitivas.

Si despreciamos el término convectivo (problema de Stokes) y consideramos el caso estacionario en una malla de paso homogéneo h, la siguiente discretización (espacial) de segundo orden parece ser un buen punto de partida (ver figura 10.2) :

$$\begin{aligned} \nu(\Delta_h u)_{ij} - \frac{p_{i+1,j} - p_{i-1,j}}{2\rho h} &= 0 \\ \nu(\Delta_h v)_{ij} - \frac{p_{i,j+1} - p_{i,j-1}}{2\rho h} &= 0 \\ \frac{u_{i+1,j} - u_{i-1,j}}{2h} + \frac{v_{i,j+1} - v_{i,j-1}}{2h} &= 0 \end{aligned} \tag{10.12}$$

donde Δ_h reresenta el operador de Laplace discreto estándar de 5 puntos

$$(\Delta_h u)_{ij} = \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{ij}}{h^2} \tag{10.13}$$

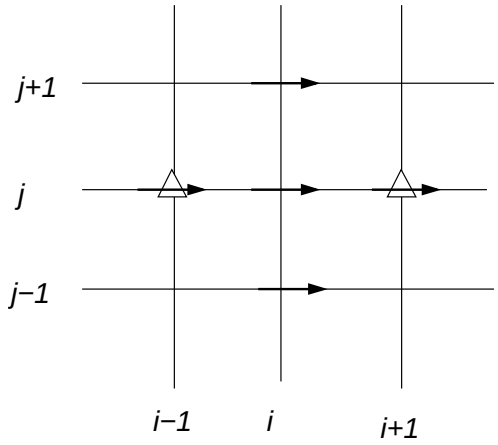


Figura 10.3: Grilla no staggered. Stencil para la ecuación de momento según x

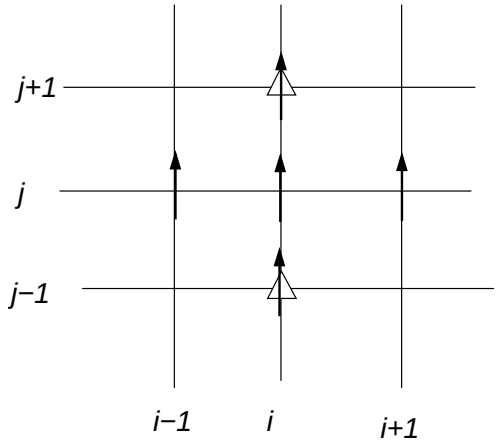


Figura 10.4: Grilla no staggered. Stencil para la ecuación de momento según y

En las figuras 10.3, 10.4 y 10.5 pueden verse los stenciles usados para cada una de las ecuaciones. Pero resulta ser que las presiones en los nodos impares se desacopla de los pares dando lugar a modos *"checkerboard"* en la presión. Las formas des resolver esto es

- Resolver una ecuación alternativa para la presión llamada PPE (Poisson Pressure Equation).
- Usar mallas "staggered" (en español "desparramadas" (???))
- Usar métodos de compresibilidad artificial.

Discutiremos a continuación el uso de mallas staggered.

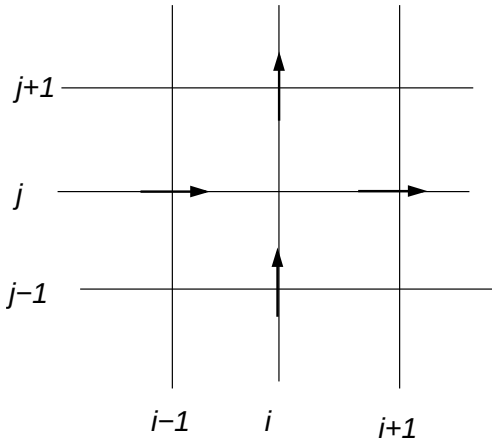


Figura 10.5: Grilla no staggered. Stencil para la ecuación de continuidad

10.5. Uso de mallas staggered

Si consideramos la ecuación de momento según x, entonces vemos que lo ideal sería tener una malla para los nodos de velocidad x desplazada en $h/2$ con respecto a la malla de los nodos de presión, en ese caso podríamos tener una ecuación de la forma (ver figura 10.7)

$$\nu(\Delta_h u)_{i+1/2,j} - \frac{p_{i+1,j} - p_{i,j}}{\rho h} = 0 \tag{10.14}$$

Similarmente, para la ecuación de momento según y tenemos (ver figura 10.8)

$$\nu(\Delta_h v)_{i,j+1/2} - \frac{p_{i,j+1} - p_{i,j}}{\rho h} = 0 \tag{10.15}$$

Esto evita el desacoplamiento de las presiones entre nodos pares e impares. Entonces tenemos 3 redes "staggered" a saber (ver figura 10.6)

- Los nodos de presión: $p_{ij} \approx p(ih, jh)$
- Los nodos de velocidad x: $u_{i+1/2,j} \approx u((i+1/2)h, jh)$
- Los nodos de velocidad y: $v_{i,j+1/2} \approx v(ih, (j+1/2)h)$

con i, j enteros. Por otra parte, la ecuación de continuidad también queda en un stencil más compacto, si lo imponemos sobre los nodos de presión (ver figura 10.9)

$$\frac{u_{i+1/2,j} - u_{i-1/2,j}}{h} + \frac{v_{i,j+1/2} - v_{i,j-1/2}}{h} = 0 \tag{10.16}$$

Por otra parte, las condiciones de contorno también se simplifican algo, en cuanto a las condiciones sobre la presión, ya que utilizando sólo contornos que coinciden con lineas semienteras (i, j=entero+$1/2$).

El método de mallas staggered es probablemente el más robusto y prolijo para tratar flujo incompresible por diferencias finitas.

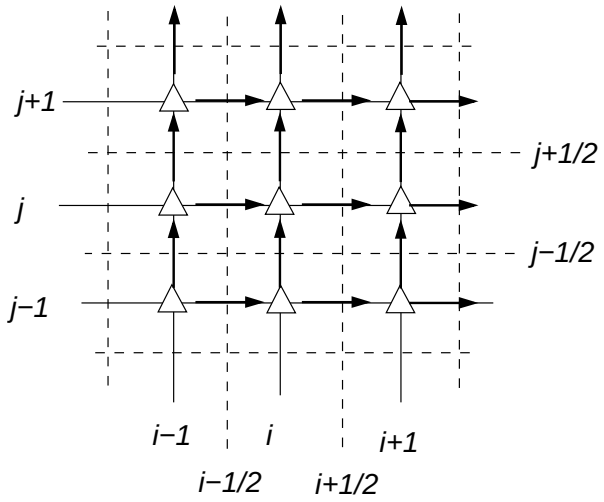


Figura 10.6: Grilla staggered usada para flujo incompresible en variables primitivas.

10.6. Discretización por elementos finitos

Considerando el caso estacionario, flujo reptante, un término forzante $\mathbf{f}$ y condiciones de contorno Dirichlet, las ecuaciones de gobierno son

$$\nu\Delta\mathbf{u} - \nabla p = \mathbf{f} \quad \text{en } \Omega \tag{10.17}$$
$$\nabla \cdot \mathbf{u} = 0 \quad \text{en } \Omega \tag{10.18}$$
$$\mathbf{u} = \bar{\mathbf{u}}, \quad \text{en } \Gamma \tag{10.19}$$

y espacios de interpolación

$$X_h = \text{span}\{N_{p\mu}, \mu = 1 \dots N\} \tag{10.20}$$
$$V_h = \text{span}\{N_{u\mu}, \mu = 1 \dots N\} \tag{10.21}$$

La formulación débil Galerkin se obtiene pesando la ecuación de momento por una función de interpolación de velocidad y pesando la ecuación de continuidad con las funciones de interpolación de presión.

$$\int_\Omega \phi\,(\nabla \cdot \mathbf{u})\,\mathrm{d}\Omega = 0, \;\; \forall\phi \in X_h \tag{10.22}$$
$$\int_\Omega (\nabla\cdot\phi)p\,\mathrm{d}\Omega + \int_\Omega \nu(\nabla\mathbf{v} : \nabla\mathbf{u})\,\mathrm{d}\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,\mathrm{d}\Omega + \int_\Gamma \mathbf{v}\cdot\mathbf{t}\cdot\hat{\mathbf{n}}\,\mathrm{d}\Gamma, \;\; \forall\mathbf{v} \in V_h \tag{10.23}$$

Notar que, como no aparecen derivadas de p ni ϕ entonces es posible utilizar aproximaciones discontinuas para p.

El sistema al que se llega es;

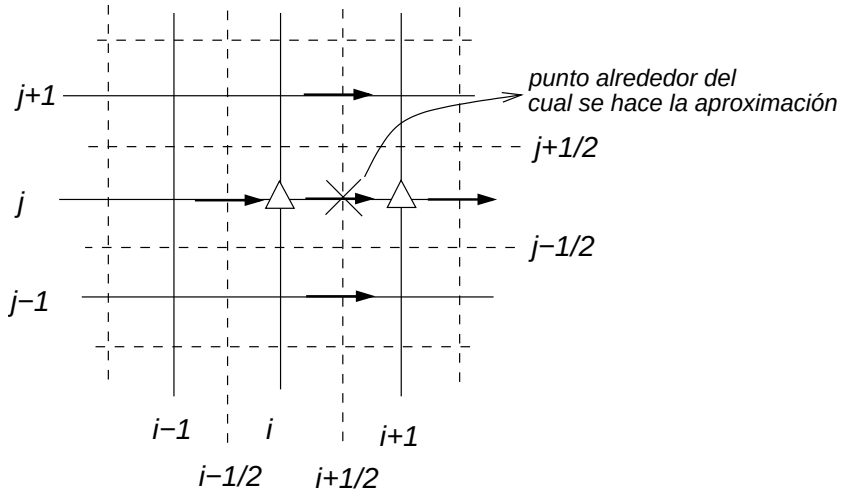


Figura 10.7: Grilla staggered. Stencil para la ecuación de momento según x

$$\begin{bmatrix} \mathbf{0} & -\mathbf{Q}^T \\ -\mathbf{Q} & \nu\mathbf{K} \end{bmatrix} \begin{bmatrix} \mathbf{P} \\ \mathbf{U} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{F} \end{bmatrix} \tag{10.24}$$

o

$$\mathbf{AX} = \mathbf{B} \tag{10.25}$$

donde

$$p_h = \sum_\mu p_\mu N_{p\mu} \tag{10.26}$$

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{bmatrix} \tag{10.27}$$

$$u_h = \sum_\mu u_\mu N_{u\mu} \tag{10.28}$$

$$\mathbf{P} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \tag{10.29}$$

$$Q_{\mu k\nu} = \int_\Omega N_{u\mu,k}\, N_{p\nu}\, \mathrm{d}\Omega \tag{10.30}$$

$$K_{i\mu j\nu} = \int_\Omega N_{u\mu,k}\, \delta_{ij}\, N_{u\nu,k}\, \mathrm{d}\Omega \tag{10.31}$$

Nótese que la matriz $\mathbf{K}$ es simétrica y definida positiva, mientras que la matriz total $\mathbf{A}$ sólo es simétrica y de hecho no puede ser definida positiva ya que tiene elementos diagonales (en el bloque $\mathbf{0}$) nulos.

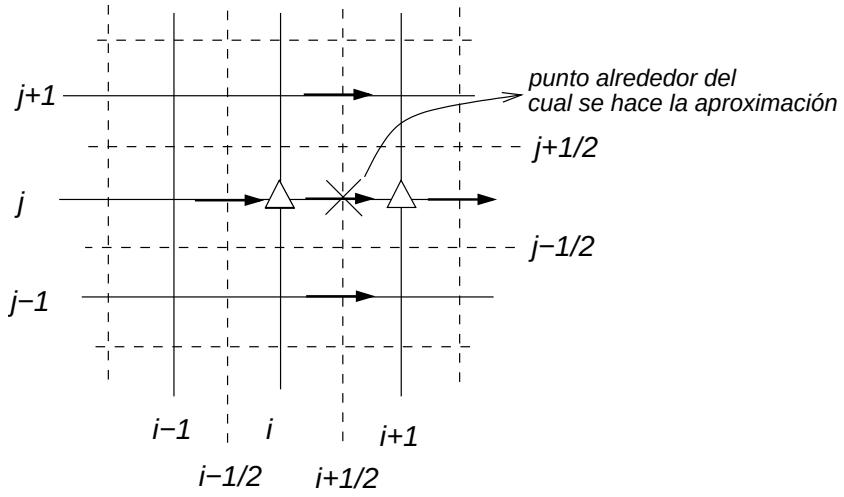


Figura 10.8: Grilla staggered. Stencil para la ecuación de momento según y

Como K es no-singular podemos eliminar $\mathbf{U}$ de la ecuación de momento e insertarla en la ecuación de continuidad obteniendo una ecuación para $\mathbf{P}$ de la forma

$$\mathbf{HP} = (\mathbf{Q}^T\mathbf{K}^{-1}\mathbf{Q})\,\mathbf{P} = -\mathbf{Q}^T\mathbf{K}^{-1}\mathbf{F} \tag{10.32}$$

La matriz $\mathbf{H}$ es simétrica y semidefinida positiva. Para que el problema este bien planteado debemos al menos exigir que la matriz sea no-singular. Podemos ver que esto ocurre si y sólo si $\mathbf{Q}$ tiene rango de filas (el número de filas linealmente independiente) N_p (el número de grados de libertad de presión). Efectivamente, si $\mathbf{Q}$ tiene rango menor que N_p entonces existe algun vector $\mathbf{P}$ tal que $\mathbf{QP} = \mathbf{0}$ y entonces $\mathbf{HP} = 0$. Por otra parte, si $\mathbf{Q}$ tiene rango igual a N_p entonces para todo $\mathbf{P} \neq \mathbf{0}$ vale que $\mathbf{u} = \mathbf{QP} \neq 0$ y entonces

$$\mathbf{P}^T\,(\mathbf{Q}^T\mathbf{K}^{-1}\mathbf{Q})\,\mathbf{P} = \mathbf{u}^T\,\mathbf{K}^{-1}\,\mathbf{u} > 0 \tag{10.33}$$

con lo cual $\mathbf{H}$ resulta ser definida positiva y por lo tanto no-singular.

10.7. El test de la parcela

Ahora bien $\mathbf{Q}$ es de dimensión $N_u \times N_p$, de manera que, para que $\mathbf{Q}$ sea no singular debemos pedir que al menos $N_u \geq N_p$. Si bien esto parece un requerimiento bastante simple, en realidad sirve para descartar toda una serie de familias de interpolación y da lugar al famoso *"test de la parcela"* (*"patch test"*).

Consideremos por ejemplo la interpolación más simple que se nos pueda ocurrir es $P1/P0$ para triángulos, es decir velocidades lineales continuas y presiones constantes por elemento (ver figura 10.10). (La convención aquí es poner primero el espacio de interpolación para velocidades y después el que se usa para presiones. En general, a menos que se mencione lo contrario el espacio para velocidades se asume continuo y el de presiones discontinuo. Pn denota el espacio de funciones que es polinomial de grado n por elemento, mientras que Qn denota el espacio de funciones bilineales (trilineales en 3D) de grado n.) En una malla estructurada de cuadrángulos, donde dividimos cada cuadrángulo en dos triángulos, tenemos (para

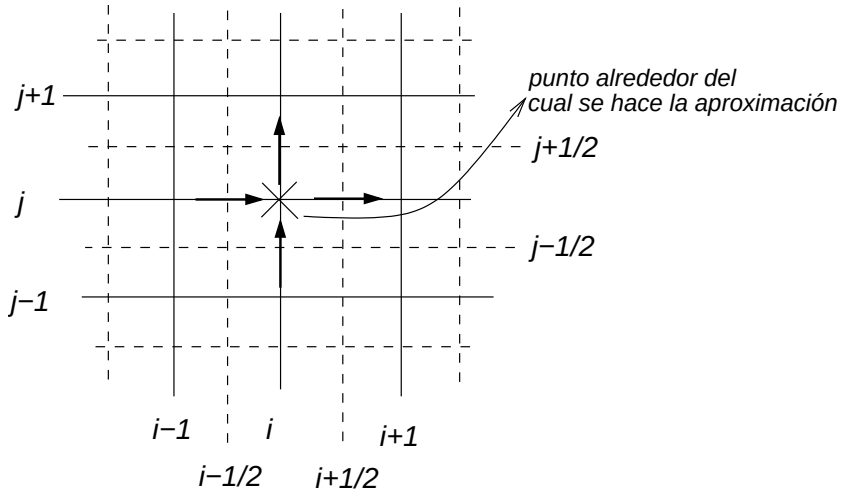


Figura 10.9: Grilla staggered. Stencil para la ecuación de continuidad.

una malla suficientmente grande) N_p=2 grados de libertad de presión por cada cuadrángulo y un nodo de velocidad (es decir $N_u = 2$), por lo tanto no se satisface el test de la parcela y la aproximación es inestable. Si tomamos parcelas más pequeñas la situación es peor, ya que el N_u es mayor o igual al N_u asintótico pero imponiendo las condiciones de contorno *"más inestables posibles"*, es decir todo el contorno de la parcela con velocidades impuestas el N_u resulta ser

$$\begin{aligned} N_u = (N_u\text{por elemento}) \times (\text{número de elementos}) \\ + \ (\text{número de nodos adicionales de contorno}) \\ - (\text{número de condiciones de contorno}) \\ \leq (N_u\text{por elemento}) \times (\text{número de elementos}) = (N_u \text{ asintótico}) \end{aligned} \tag{10.34}$$

Entonces, si bien el test de la parcela asintótico permite descartar una serie de familias de interpolación, el test aplicado sobre parcelas más pequeño resulta ser más restrictivo. Por ejemplo para la interpolación $Q1/P0$ el análisis asintótico da N_u por celda $= 2$, N_p por celda $= 1$ lo cual en principio está bien, pero cuando vamos a una parcela de $2 \times 2 = 4$ elementos cuadrangulares tenemos $N_u = 2$ (sólo el nodo de velocidad del medio está libre), $N_p = 3$ (uno de los nodos de presión siempre está restingido) lo cual está mal.

Sin embargo, puede verse que un macroelemento triangular formado por 3 elementos $Q1/P0$ es estable.

La relación asintótica más apropiada parece ser $N_u = 2N_p$.

10.8. La condición de Brezzi-Babuska

Si bien el test de la parcela es muy útil para descartar posibles familias de interpolación, no es suficiente para asegurar la convrgencia. Ríos de tinta han corrido en cuanto a cual es la condición para asegurar convergencia en problemas de este tipo y la respuesta es la conocida *"condición de Brezzi-Babûska"* tambien conocida como condicion *"inf-sup"*.

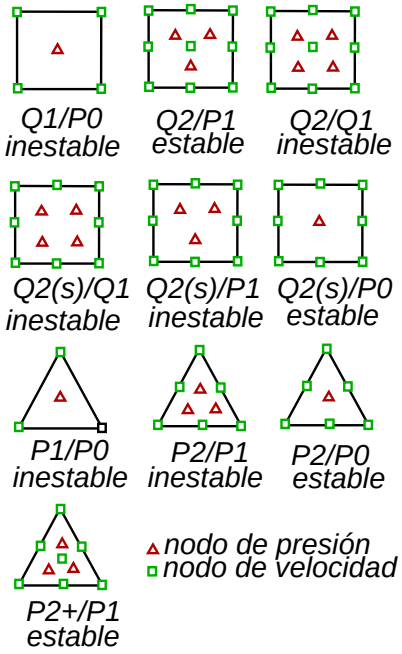


Figura 10.10: Combinaciones de espacios de interpolaci 'on de elementos finitos. Velocidades continuas, presiones discontinuas.

$$\inf_{q_h \in X_h - 0} \sup_{\mathbf{v}_h \in \mathbf{v}_h - 0} \frac{\int_\Omega q_h \nabla \cdot \mathbf{v}_h \, \mathrm{d}\Omega}{\left(\int_\Omega |\nabla \mathbf{v}_h|^2 \, \mathrm{d}\Omega\right)^{1/2} \left(\int_\Omega |q_h^2| \, \mathrm{d}\Omega\right)^{1/2}} = \overline{BB} \geq C \neq C(h) \tag{10.35}$$

Trabajando un poco esta ecuación puede llegarse a la conclusión de que

$$\overline{BB} = \text{mínimo autovalor de} \{\mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T\mathbf{M}_p^{-1}\} \tag{10.36}$$

donde M_p es la *"matriz de masa"* de las funciones de presión es decir

$$M_{p\mu\nu} = \int_\Omega N_{p\mu} N_{p\nu} \, \mathrm{d}\Omega \tag{10.37}$$

Como M_p es una "matriz de masa", tiene un número de condición muy bajo y no afecta mucho la expresión anterior, de manera que podemos tomar también

$$\overline{BB} = \text{mínimo autovalor de} \{\mathbf{Q}\mathbf{K}^{-1}\mathbf{Q}^T\} \tag{10.38}$$

Notemos que la matriz en cuestión es la misma que se obtiene si condensamos el sistema total en el vector de presiones, ec. (10.33). De manera que el criterio de BB parece relacionar la convergencia del espacio de interpolación con el comportamiento de los autovalores para $h \to 0$.

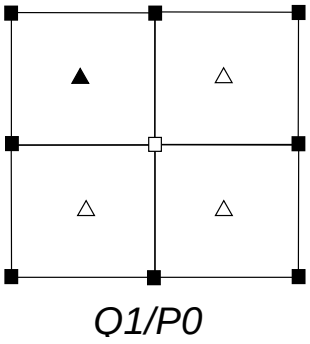


Figura 10.11: Parcela de 2×2 elementos cuadrangulares. Nodos pintados de negro indican grados de libertad restringidos.

10.9. Métodos FEM estabilizados

Podemos modificar el sistema de ecuaciones a resolver si agregamos a la ecuación de momento un término de estabilización que proviene de tomar el gradiente de la ecuación de continuidad y otro proporcional a la divergencia de la ecuación de momento a la ecuación de continuidad. El sistema modificado es

$$\nu \Delta \mathbf{u} - \nabla p - \beta \nabla (\nabla \cdot \mathbf{u}) = \mathbf{f} \tag{10.39}$$

$$\nabla \cdot \mathbf{u} - \alpha \Delta p = -\alpha (\nabla \cdot \mathbf{f}) \tag{10.40}$$

El sistema de ecuaciones es

$$\begin{bmatrix} -\alpha \mathbf{L} & -\mathbf{Q}^T \\ -\mathbf{Q} & \nu \mathbf{K} + \beta \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{P} \\ \mathbf{U} \end{bmatrix} = \begin{bmatrix} -\alpha \mathbf{S} \\ \mathbf{F} \end{bmatrix} \tag{10.41}$$

donde

$$H_{\mu\nu} = \int N_{p\mu,k} N_{p\nu,k} \, \mathrm{d}\Omega \tag{10.42}$$

$$G_{i\mu,j\nu} = \int N_{u\mu,i} N_{u\nu,j} \, \mathrm{d}\Omega \tag{10.43}$$

son aproximaciones a los operadores de Laplace (vectorial) y *"grad-div"*.

Notar que con el agregado de estos términos de estabilización ahora el bloque que antes era nulo ya no lo es. Por otra parte el sistema es ahora elíptico, con lo cual deberían imponerse también condiciones de contorno sobre la presión. (Lo cual no es trivial). En la práctica *"se deja la presión libre"* lo cual es equivalente a imponer derivada normal de la presión nula ($\frac{\partial p}{\partial n} = 0$) lo cual **no es cierto** pero se absorbe en una capa límite de espesor h.

Índice alfabético