UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE GEOFÍSICA

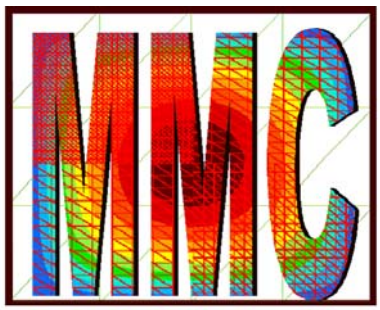
Y

GRUPO DE MODELACIÓN MATEMATICA Y COMPUTACIONAL

Método de Elementos Finitos

Antonio Carrillo Ledesma
Ismael Herrera Revilla
Robert Yates Smith

http://www.mmc.igeofcu.unam.mx/



INSTITUTO DE GEOFÍSICA
UNAM

2008

Índice

1. Sistemas Continuos y sus Modelos

Los fundamentos de la física macroscópica los proporciona la 'teoría de los medios continuos'. En este capítulo, con base en ella se introduce una formulación clara, general y sencilla de los modelos matemáticos de los sistemas continuos. Esta formulación es tan sencilla y tan general, que los modelos básicos de sistemas tan complicados y diversos como la atmósfera, los océanos, los yacimientos petroleros, o los geotérmicos, se derivan por medio de la aplicación repetida de una sola ecuación diferencial: 'la ecuación diferencial de balance'.

Dicha formulación también es muy clara, pues en el modelo general no hay ninguna ambigüedad; en particular, todas las variables y parámetros que intervienen en él, están definidos de manera unívoca. En realidad, este modelo general de los sistemas continuos constituye una realización extraordinaria de los paradigmas del pensamiento matemático. El descubrimiento del hecho de que los modelos matemáticos de los sistemas continuos, independientemente de su naturaleza y propiedades intrínsecas, pueden formularse por medio de balances, cuya idea básica no difiere mucho de los balances de la contabilidad financiera, fue el resultado de un largo proceso de perfeccionamiento en el que concurrieron una multitud de mentes brillantes.

1.1. Los Modelos

Un modelo de un sistema es un sustituto de cuyo comportamiento es posible derivar el correspondiente al sistema original. Los modelos matemáticos, en la actualidad, son los utilizados con mayor frecuencia y también los más versátiles. En las aplicaciones específicas están constituidos por programas de cómputo cuya aplicación y adaptación a cambios de las propiedades de los sistemas es relativamente fácil. También, sus bases y las metodologías que utilizan son de gran generalidad, por lo que es posible construirlos para situaciones y sistemas muy diversos.

Los modelos matemáticos son entes en los que se integran los conocimientos científicos y tecnológicos, con los que se construyen programas de cómputo que se implementan con medios computacionales. En la actualidad, la simulación numérica permite estudiar sistemas complejos y fenómenos naturales que sería muy costoso, peligroso o incluso imposible de estudiar por experimentación directa. En esta perspectiva la significación de los modelos matemáticos en ciencias e ingeniería es clara, porqué la modelación matemática constituye el método más efectivo de predecir el comportamiento de los diversos sistemas de interés. En nuestro país, ellos son usados ampliamente en la industria petrolera, en las ciencias y la ingeniería del agua y en muchas otras.

1.1.1. Física Microscópica y Física Macroscópica

La materia, cuando se le observa en el ámbito ultramicroscópico, está formada por moléculas y átomos. Estos a su vez, por partículas aún más pequeñas como los protones, neutrones y electrones. La predicción del comportamiento de

estas partículas es el objeto de estudio de la mecánica cuántica y la física nuclear. Sin embargo, cuando deseamos predecir el comportamiento de sistemas tan grandes como la atmósfera o un yacimiento petrolero, los cuales están formados por un número extraordinariamente grande de moléculas y átomos, su estudio resulta inaccesible con esos métodos y en cambio el enfoque macroscópico es apropiado.

Por eso en lo que sigue distinguiremos dos enfoques para el estudio de la materia y su movimiento. El primero -el de las moléculas, los átomos y las partículas elementales- es el enfoque microscópico y el segundo es el enfoque macroscópico. Al estudio de la materia con el enfoque macroscópico, se le llama física macroscópica y sus bases teóricas las proporciona la mecánica de los medios continuos.

Cuando se estudia la materia con este último enfoque, se considera que los cuerpos llenan el espacio que ocupan, es decir que no tienen huecos, que es la forma en que los vemos sin el auxilio de un microscopio. Por ejemplo, el agua llena todo el espacio del recipiente donde está contenida. Este enfoque macroscópico está presente en la física clásica. La ciencia ha avanzado y ahora sabemos que la materia está llena de huecos, que nuestros sentidos no perciben y que la energía también está cuantizada. A pesar de que estos dos enfoques para el análisis de los sistemas físicos, el microscópico y el macroscópico, parecen a primera vista conceptualmente contradictorios, ambos son compatibles, y complementarios, y es posible establecer la relación entre ellos utilizando a la mecánica estadística.

1.2. Cinemática de los Modelos de Sistemas Continuos

En la teoría de los sistemas continuos, los cuerpos llenan todo el espacio que ocupan. Y en cada punto del espacio físico hay una y solamente una partícula. Así, definimos como sistema continuo a un conjunto de partículas. Aún más, dicho conjunto es un subconjunto del espacio Euclidiano tridimensional. Un cuerpo es un subconjunto de partículas que en cualquier instante dado ocupa un dominio, en el sentido matemático, del espacio físico; es decir, del espacio Euclidiano tridimensional. Denotaremos por $B(t)$ a la región ocupada por el cuerpo $\mathcal{B}$, en el tiempo t, donde t puede ser cualquier número real.

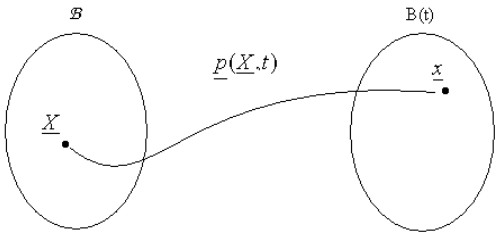


Figura 1: Representación del movimiento de partículas de un cuerpo $\mathcal{B}$, para un tiempo dado.

Frecuentemente, sin embargo, nuestro interés de estudio se limitará a un intervalo finito de tiempo. Dado un cuerpo $\mathcal{B}$, todo subdominio $\tilde{\mathcal{B}} \subset \mathcal{B}$, constituye a su vez otro cuerpo; en tal caso, se dice que $\tilde{\mathcal{B}} \subset \mathcal{B}$ es un subcuerpo de $\mathcal{B}$. De acuerdo con lo mencionado antes, una hipótesis básica de la teoría de los sistemas continuos es que en cualquier tiempo $t \in (-\infty, \infty)$ y en cada punto $x \in \mathcal{B}$ de la región ocupada por el cuerpo, hay una y sólo una partícula del cuerpo. Como en nuestra revisión se incluye no solamente la estática (es decir, los cuerpos en reposo), sino también la dinámica (es decir, los cuerpos en movimiento), un primer problema de la cinemática de los sistemas continuos consiste en establecer un procedimiento para identificar a las partículas cuando están en movimiento en el espacio físico.

Sea $\underline{X} \in \mathcal{B}$, una partícula y $p(\underline{X}, t)$ el vector de la posición que ocupa, en el espacio físico, dicha partícula en el instante t. Una forma, pero no la única, de identificar la partícula $\underline{X}$ es asociándole la posición que ocupa en un instante determinado. Tomaremos en particular el tiempo $t = 0$, en tal caso $\underline{p}(\underline{X}, 0) \equiv \underline{X}$.

A las coordenadas del vector $\underline{X} \equiv (x_1, x_2, x_3)$, se les llama las coordenadas materiales de la partícula. En este caso, las coordenadas materiales de una partícula son las coordenadas del punto del espacio físico que ocupaba la partícula en el tiempo inicial, $t = 0$. Desde luego, el tiempo inicial puede ser cualquier otro, si así se desea. Sea $\mathcal{B}$ el dominio ocupado por un cuerpo en el tiempo inicial, entonces $\underline{X} \in \mathcal{B}$ si y solamente si la partícula $\underline{X}$ es del cuerpo. Es decir, $\mathcal{B}$ caracteriza al cuerpo. Sin embargo, debido al movimiento, la región ocupada por el mismo cambia con el tiempo y será denotada por $\mathcal{B}(t)$.

Formalmente, para cualquier $t \in (-\infty, \infty)$, $\mathcal{B}(t)$ se define por

$$\mathcal{B}(t) \equiv \left\{ \underline{x} \in \mathbb{R}^3 \mid \exists \underline{X} \in \mathcal{B} \quad tal \quad que \quad \underline{x} = p(\underline{X}, t) \right\} \tag{1}$$

el vector posición $\underline{p}(\underline{X}, t)$ es función del vector tridimensional $\underline{X}$ y del tiempo. Si fijamos el tiempo t, $\underline{p}(\underline{X}, t)$ define una transformación del espacio Euclidiano $\mathbb{R}^3$ en si mismo y la Ec. (1) es equivalente a $\mathcal{B}(t) = \underline{p}(\mathcal{B}, t)$. Una notación utilizada para representar esta familia de funciones es $\underline{p}(\cdot, t)$. De acuerdo a la hipótesis de los sistemas continuos: En cualquier tiempo $t \in (-\infty, \infty)$ y en cada punto $\underline{x} \in \mathcal{B}$ de la región ocupada por el cuerpo hay una y sólo una partícula del cuerpo $\mathcal{B}$ para cada t fijo. Es decir, $\underline{p}(\cdot, t)$ es una función biunívoca, por lo que existe la función inversa $\underline{p}^{-1}(\cdot, t)$.

Si se fija la partícula $\underline{X}$ en la función $\underline{p}(\underline{X}, t)$ y se varía el tiempo t, se obtiene su trayectoria. Esto permite obtener la velocidad de cualquier partícula, la cual es un concepto central en la descripción del movimiento. Ella se define como la derivada con respecto al tiempo de la posición cuando la partícula se mantiene fija. Es decir, es la derivada parcial con respecto al tiempo de la función de posición $\underline{p}(\underline{X}, t)$. Por lo mismo, la velocidad como función de las coordenadas materiales de las partículas, está dada por

$$\underline{\mathrm{V}}(\underline{X}, t) \equiv \frac{\partial \underline{p}}{\partial t}(\underline{X}, t). \tag{2}$$

1.2.1. Propiedades Intensivas y sus Representaciones

En lo que sigue consideraremos funciones definidas para cada tiempo, en cada una de las partículas de un sistema continuo. A tales funciones se les llama 'propiedades intensivas'. Las propiedades intensivas pueden ser funciones escalares o funciones vectoriales. Por ejemplo, la velocidad, definida por la Ec. (2), es una función vectorial que depende de la partícula $\underline{X}$ y del tiempo t.

Una propiedad intensiva con valores vectoriales es equivalente a tres escalares, correspondientes a cada una de sus tres componentes. Hay dos formas de representar a las propiedades intensivas: la representación Euleriana y la representación Lagrangiana. Los nombres son en honor a los matemáticos Leonard Euler (1707-1783) y Joseph Louis Lagrange (1736-1813), respectivamente. Frecuentemente, el punto de vista Lagrangiano es utilizado en el estudio de los sólidos, mientras que el Euleriano se usa más en el estudio de los fluidos.

Considere una propiedad intensiva escalar, la cual en el tiempo t toma el valor $\phi(\underline{X}, t)$ en la partícula $\underline{X}$. Entonces, de esta manera se define una función $\phi : \mathcal{B} \to \mathbb{R}^1$, para cada $t \in (-\infty, \infty)$ a la que se denomina representación Lagrangiana de la propiedad intensiva considerada. Ahora, sea $\psi(\underline{x}, t)$ el valor que toma esa propiedad en la partícula que ocupa la posición $\underline{x}$, en el tiempo t. En este caso, para cada $t \in (-\infty, \infty)$ se define una función $\psi : \mathcal{B}(t) \to \mathbb{R}^1$ a la cual se denomina representación Euleriana de la función considerada. Estas dos representaciones de una misma propiedad están relacionadas por la siguiente identidad

$$\phi(\underline{X}, t) \equiv \psi(\underline{p}(\underline{X}, t), t). \tag{3}$$

Nótese que, aunque ambas representaciones satisfacen la Ec. (3), las funciones $\phi(\underline{X}, t)$ y $\psi(\underline{x}, t)$ no son idénticas. Sus argumentos $\underline{X}$ y $\underline{x}$ son vectores tridimensionales (es decir, puntos de $\mathbb{R}^3$); sin embargo, si tomamos $\underline{X} = \underline{x}$, en general

$$\phi(\underline{X}, t) \neq \psi(\underline{X}, t). \tag{4}$$

La expresión de la velocidad de una partícula dada por la Ec. (2), define a su representación Lagrangiana, por lo que utilizando la Ec. (3) es claro que

$$\frac{\partial p}{\partial t}(\underline{X}, t) = \underline{\mathrm{V}}(\underline{X}, t) \equiv \underline{\mathrm{v}}(\underline{p}(\underline{X}, t), t) \tag{5}$$

donde $\underline{\mathrm{v}}(\underline{x}, t)$ es la representación Euleriana de la velocidad. Por lo mismo

$$\underline{\mathrm{v}}(\underline{x}, t) \equiv \underline{\mathrm{V}}(\underline{p}^{-1}(\underline{x}, t), t). \tag{6}$$

Esta ecuación tiene la interpretación de que la velocidad en el punto $\underline{x}$ del espacio físico, es igual a la velocidad de la partícula que pasa por dicho punto en el instante t. La Ec. (6) es un caso particular de la relación

$$\psi(\underline{x}, t) \equiv \phi(\underline{p}^{-1}(\underline{x}, t), t)$$

de validez general, la cual es otra forma de expresar la relación de la Ec. (3) que existe entre las dos representaciones de una misma propiedad intensiva.

La derivada parcial con respecto al tiempo de la representación Lagrangiana $\phi(\underline{X},t)$ de una propiedad intensiva, de acuerdo a la definición de la derivada parcial de una función, es la tasa de cambio con respecto al tiempo que ocurre en una partícula fija. Es decir, si nos montamos en una partícula y medimos a la propiedad intensiva y luego los valores así obtenidos los derivamos con respecto al tiempo, el resultado final es $\frac{\partial\phi(\underline{X},t)}{\partial t}$. En cambio, si $\psi(\underline{x},t)$ es la representación Euleriana de esa misma propiedad, entonces $\frac{\partial\psi(\underline{x},t)}{\partial t}$ es simplemente la tasa de cambio con respecto al tiempo que ocurre en un punto fijo en el espacio. Tiene interés evaluar la tasa de cambio con respecto al tiempo que ocurre en una partícula fija, cuando se usa la representación Euleriana. Derivando con respecto al tiempo a la identidad de la Ec. (3) y la regla de la cadena, se obtiene

$$\frac{\partial\phi(\underline{X},t)}{\partial t}=\frac{\partial\psi}{\partial t}(\underline{p}(\underline{X},t),t)+\sum_{i=1}^{3}\frac{\partial\psi}{\partial x_i}(\underline{p}(\underline{X},t),t)\frac{\partial p_i}{\partial t}(\underline{X},t). \tag{7}$$

Se acostumbra definir el símbolo $\frac{D\psi}{Dt}$ por

$$\frac{D\psi}{Dt}=\frac{\partial\psi}{\partial t}+\sum_{i=1}^{3}v_i\frac{\partial\psi}{\partial x_i} \tag{8}$$

o, más brevemente,

$$\frac{D\psi}{Dt}=\frac{\partial\psi}{\partial t}+\underline{\mathrm{v}}\cdot\nabla\psi \tag{9}$$

utilizando esta notación, se puede escribir

$$\frac{\partial\phi(\underline{X},t)}{\partial t}=\frac{D\psi}{Dt}(\underline{p}(\underline{X},t)\equiv\left(\frac{\partial\psi}{\partial t}+\underline{\mathrm{v}}\cdot\nabla\psi\right)(\underline{p}(\underline{X},t),t). \tag{10}$$

Por ejemplo, la aceleración de una partícula se define como la derivada de la velocidad cuando se mantiene a la partícula fija. Aplicando la Ec. (9) se tiene

$$\frac{D\underline{\mathrm{v}}}{Dt}=\frac{\partial\underline{\mathrm{v}}}{\partial t}+\underline{\mathrm{v}}\cdot\nabla\underline{\mathrm{v}} \tag{11}$$

una expresión más transparente se obtiene aplicando la Ec. (9) a cada una de las componentes de la velocidad. Así, se obtiene

$$\frac{D\mathrm{v}_i}{Dt}=\frac{\partial\mathrm{v}_i}{\partial t}+\underline{\mathrm{v}}\cdot\nabla\mathrm{v}_i. \tag{12}$$

Desde luego, la aceleración, en representación Lagrangiana es simplemente

$$\frac{\partial}{\partial t}\underline{\mathrm{V}}(\underline{X},t)=\frac{\partial^2}{\partial t^2}\underline{p}(\underline{X},t). \tag{13}$$

1.2.2. Propiedades Extensivas

En la sección anterior se consideraron funciones definidas en las partículas de un cuerpo, más precisamente, funciones que hacen corresponder a cada partícula y cada tiempo un número real, o un vector del espacio Euclidiano tridimensional $\mathbb{R}^3$. En ésta, en cambio, empezaremos por considerar funciones que a cada cuerpo $\mathcal{B}$ de un sistema continuo, y a cada tiempo t le asocia un número real o un vector de $\mathbb{R}^3$. A una función de este tipo $\mathbb{E}(\mathcal{B}, t)$ se le llama 'propiedad extensiva' cuando esta dada por una integral

$$\mathbb{E}(\mathcal{B}, t) \equiv \int_{\mathcal{B}(t)} \psi(\underline{x}, t) d\underline{\mathrm{x}}. \tag{14}$$

Observe que, en tal caso, el integrando define una función $\psi(\underline{x}, t)$ y por lo mismo, una propiedad intensiva. En particular, la función $\psi(\underline{x}, t)$ es la representación Euleriana de esa propiedad intensiva. Además, la Ec. (14) establece una correspondencia biunívoca entre las propiedades extensivas y las intensivas, porqué dada la representación Eulereana $\psi(\underline{x}, t)$ de cualquier propiedad intensiva, su integral sobre el dominio ocupado por cualquier cuerpo, define una propiedad extensiva. Finalmente, la notación empleada en la Ec. (14) es muy explicita, pues ahí se ha escrito $\mathbb{E}(\mathcal{B}, t)$ para enfatizar que el valor de la propiedad extensiva corresponde al cuerpo $\mathcal{B}$. Sin embargo, en lo que sucesivo, se simplificara la notación omitiendo el símbolo $\mathcal{B}$ es decir, se escribirá $\mathbb{E}(t)$ en vez de $\mathbb{E}(\mathcal{B}, t)$.

Hay diferentes formas de definir a las propiedades intensivas. Como aquí lo hemos hecho, es por unidad de volumen. Sin embargo, es frecuente que se le defina por unidad de masa véase [16]. Es fácil ver que la propiedad intensiva por unidad de volumen es igual a la propiedad intensiva por unidad de masa multiplicada por la densidad de masa (es decir, masa por unidad de volumen), por lo que es fácil pasar de un concepto al otro, utilizando la densidad de masa.

Sin embargo, una ventaja de utilizar a las propiedades intensivas por unidad de volumen, en lugar de las propiedades intensivas por unidad de masa, es que la correspondencia entre las propiedades extensivas y las intensivas es más directa: dada una propiedad extensiva, la propiedad intensiva que le corresponde es la función que aparece como integrando, cuando aquélla se expresa como una integral de volumen. Además, del cálculo se sabe que

$$\psi(\underline{x}, t) \equiv \lim_{Vol \to 0} \frac{\mathbb{E}(t)}{Vol} = \lim_{Vol \to 0} \frac{\int_{\mathcal{B}(t)} \psi(\underline{\xi}, t) d\underline{\xi}}{Vol}. \tag{15}$$

La Ec. (15) proporciona un procedimiento efectivo para determinar las propiedades extensivas experimentalmente: se mide la propiedad extensiva en un volumen pequeño del sistema continuo de que se trate, se le divide entre le volumen y el cociente que se obtiene es una buena aproximación de la propiedad intensiva.

El uso que haremos del concepto de propiedad extensiva es, desde luego, lógicamente consistente. En particular, cualquier propiedad que satisface las condiciones de la definición de propiedad extensiva establecidas antes es, por

ese hecho, una propiedad extensiva. Sin embargo, no todas las propiedades extensivas que se pueden obtener de esta manera son de interés en la mecánica de los medios continuos. Una razón básica por la que ellas son importantes es porqué el modelo general de los sistemas continuos se formula en términos de ecuaciones de balance de propiedades extensivas, como se verá más adelante.

1.2.3. Balance de Propiedades Extensivas e Intensivas

Los modelos matemáticos de los sistemas continuos están constituidos por balances de propiedades extensivas. Por ejemplo, los modelos de transporte de solutos (los contaminantes transportados por corrientes superficiales o subterráneas, son un caso particular de estos procesos de transporte) se construyen haciendo el balance de la masa de soluto que hay en cualquier dominio del espacio físico. Aquí, el término balance se usa, esencialmente, en un sentido contable. En la contabilidad que se realiza para fines financieros o fiscales, la diferencia de las entradas menos las salidas nos da el aumento, o cambio, de capital. En forma similar, en la mecánica de los medios continuos se realiza, en cada cuerpo del sistema continuo, un balance de las propiedades extensivas en que se basa el modelo.

Ecuación de Balance Global Para realizar tales balances es necesario, en primer lugar, identificar las causas por las que las propiedades extensivas pueden cambiar. Tomemos como ejemplo de propiedad extensiva a las existencias de maíz que hay en el país. La primera pregunta es: ¿qué causas pueden motivar su variación, o cambio, de esas existencias?. Un análisis sencillo nos muestra que dicha variación puede ser debida a que se produzca o se consuma. También a que se importe o se exporte por los límites del país (fronteras o litorales). Y con esto se agotan las causas posibles; es decir, esta lista es exhaustiva. Producción y consumo son términos similares, pero sus efectos tienen signos opuestos, que fácilmente se engloban en uno solo de esos conceptos. De hecho, si convenimos en que la producción puede ser negativa, entonces el consumo es una producción negativa.

Una vez adoptada esta convención, ya no es necesario ocuparnos separadamente del consumo. En forma similar, la exportación es una importación negativa. Entonces, el incremento en las existencias $\Delta\mathbb{E}$ en un período Δt queda dado por la ecuación

$$\Delta\mathbb{E} = \mathbb{P} + \mathbb{I} \tag{16}$$

donde a la producción y a la importación, ambas con signo, se les ha representado por $\mathbb{P}$ y $\mathbb{I}$ respectivamente.

Similarmente, en la mecánica de los medios continuos, la lista exhaustiva de las causas por las que una propiedad extensiva de cualquier cuerpo puede cambiar, contiene solamente dos motivos:

i) Por producción en el interior del cuerpo; y

ii) Por importación (es decir, transporte) a través de la frontera.

Esto conduce a la siguiente ecuación de "balance global", de gran generalidad, para las propiedades extensivas

$$\frac{d\mathbb{E}}{dt}(t) = \int_{\mathcal{B}(t)} g(\underline{x},t)d\underline{x} + \int_{\partial\mathcal{B}(t)} q(\underline{x},t)d\underline{x} + \int_{\Sigma(t)} g_\Sigma(\underline{x},t)d\underline{x}. \tag{17}$$

Donde $g(\underline{x},t)$ es la generación en el interior del cuerpo, con signo, de la propiedad extensiva correspondiente, por unidad de volumen, por unidad de tiempo. Además, en la Ec. (17) se ha tomado en cuenta la posibilidad de que haya producción concentrada en la superficie $\Sigma(t)$, la cual está dada en esa ecuación por la última integral, donde $g_\Sigma(\underline{x},t)$ es la producción por unidad de área. Por otra parte $q(\underline{x},t)$ es lo que se importa o transporta hacia el interior del cuerpo a través de la frontera del cuerpo $\partial\mathcal{B}(t)$, en otras palabras, es el flujo de la propiedad extensiva a través de la frontera del cuerpo, por unidad de área, por unidad de tiempo. Puede demostrarse, con base en hipótesis válidas en condiciones muy generales, que para cada tiempo t existe un campo vectorial $\tau(\underline{x},t)$ tal que

$$q(\underline{x},t) \equiv \tau(\underline{x},t)\cdot\underline{n}(\underline{x},t) \tag{18}$$

donde $\underline{n}(\underline{x},t)$ es normal exterior a $\partial\mathcal{B}(t)$. En vista de esta relación, la Ec. (17) de balance se puede escribir como

$$\frac{d\mathbb{E}}{dt}(t) = \int_{\mathcal{B}(t)} g(\underline{x},t)d\underline{x} + \int_{\partial\mathcal{B}(t)} \tau(\underline{x},t)\cdot\underline{n}(\underline{x},t)d\underline{x} + \int_{\Sigma(t)} g_\Sigma(\underline{x},t)d\underline{x}. \tag{19}$$

La relación (19) se le conoce con el nombre de "ecuación general de balance global" y es la ecuación básica de los balances de los sistemas continuos. A la función $g(\underline{x},t)$ se le denomina el generación interna y al campo vectorial $\tau(\underline{x},t)$ el campo de flujo.

Condiciones de Balance Local Los modelos de los sistemas continuos están constituidos por las ecuaciones de balance correspondientes a una colección de propiedades extensivas. Así, a cada sistema continuo le corresponde una familia de propiedades extensivas, tal que, el modelo matemático del sistema está constituido por las condiciones de balance de cada una de las propiedades extensivas de dicha familia.

Sin embargo, las propiedades extensivas mismas no se utilizan directamente en la formulación del modelo, en su lugar se usan las propiedades intensivas asociadas a cada una de ellas. Esto es posible porqué las ecuaciones de balance global son equivalentes a las llamadas condiciones de balance local, las cuales se expresan en términos de las propiedades intensivas correspondientes. Las condiciones de balance local son de dos clases: 'las ecuaciones diferenciales de balance local' y 'las condiciones de salto'.

Las primeras son ecuaciones diferenciales parciales, que se deben satisfacer en cada punto del espacio ocupado por el sistema continuo, y las segundas son ecuaciones algebraicas que las discontinuidades deben satisfacer donde ocurren; es decir, en cada punto de Σ. Cabe mencionar que las ecuaciones diferenciales

de balance local son de uso mucho más amplio que las condiciones de salto, pues estas últimas solamente se aplican cuando y donde hay discontinuidades, mientras que las primeras en todo punto del espacio ocupado por el sistema continuo.

Una vez establecidas las ecuaciones diferenciales y de salto del balance local, e incorporada la información científica y tecnológica necesaria para completar el modelo (la cual por cierto se introduce a través de las llamadas 'ecuaciones constitutivas'), el problema matemático de desarrollar el modelo y derivar sus predicciones se transforma en uno correspondiente a la teoría de la ecuaciones diferenciales, generalmente parciales, y sus métodos numéricos.

Las Ecuaciones de Balance Local En lo que sigue se supone que las propiedades intensivas pueden tener discontinuidades, de salto exclusivamente, a través de la superficie $\Sigma(t)$. Se entiende por 'discontinuidad de salto', una en que el límite por ambos lados de $\Sigma(t)$ existe, pero son diferentes.

Se utilizará en lo que sigue los resultados matemáticos que se dan a continuación, ver [8].

Teorema 1 *Para cada $t > 0$, sea $\mathcal{B}(t) \subset \mathbb{R}^3$ el dominio ocupado por un cuerpo. Suponga que la 'propiedad intensiva' $\psi(\underline{x}, t)$ es de clase C^1, excepto a través de la superficie $\Sigma(t)$. Además, sean las funciones $\underline{v}(\underline{x}, t)$ y $\underline{v}_\Sigma(\underline{x}, t)$ esta última definida para $\underline{x} \in \Sigma(t)$ solamente, las velocidades de las partículas y la de $\Sigma(t)$, respectivamente. Entonces*

$$\frac{d}{dt}\int_{\mathcal{B}(t)} \psi d\underline{x} \equiv \int_{\mathcal{B}(t)} \left\{ \frac{\partial \psi}{\partial t} + \nabla \cdot (\underline{v}\psi) \right\} d\underline{x} + \int_\Sigma \left[(\underline{v} - \underline{v}_\Sigma)\, \psi \right] \cdot \underline{n} d\underline{x}. \tag{20}$$

Teorema 2 *Considere un sistema continuo, entonces, la 'ecuación de balance global' (19) se satisface para todo cuerpo del sistema continuo si y solamente si se cumplen las condiciones siguientes:*

i) La ecuación diferencial

$$\frac{\partial \psi}{\partial t} + \nabla \cdot (\underline{v}\psi) = \nabla \cdot \underline{\tau} + g \tag{21}$$

vale en todo punto $\underline{x} \in \mathbb{R}^3$, de la región ocupada por el sistema.

ii) La ecuación

$$\left[\psi\, (\underline{v} - \underline{v}_\Sigma) - \underline{\tau} \right] \cdot \underline{n} = g_\Sigma \tag{22}$$

vale en todo punto $\underline{x} \in \Sigma$.

A las ecuaciones (21) y (22), se les llama 'ecuación diferencial de balance local' y 'condición de salto', respectivamente.

Desde luego, el caso más general que se estudiará se refiere a situaciones dinámicas; es decir, aquéllas en que las propiedades intensivas cambian con el tiempo. Sin embargo, los estados estacionarios de los sistemas continuos son de sumo interés. Por estado estacionario se entiende uno en que las propiedades intensivas son independientes del tiempo. En los estados estacionarios, además,

las superficies de discontinuidad $\Sigma(t)$ se mantienen fijas (no se mueven). En este caso $\frac{\partial \psi}{\partial t} = 0$ y $\underline{v}_\Sigma = 0$. Por lo mismo, para los estados estacionarios, la ecuación de balance local y la condición de salto se reducen a

$$\nabla \cdot (\underline{v}\psi) = \nabla \cdot \underline{\tau} + g \tag{23}$$

que vale en todo punto $\underline{x} \in \mathbb{R}^3$ y

$$[\psi \underline{v} - \underline{\tau}] \cdot \underline{n} = g_\Sigma \tag{24}$$

que se satisface en todo punto de la discontinuidad $\Sigma(t)$ respectivamente.

1.3. Ejemplos de Modelos

Una de las aplicaciones más sencillas de las condiciones de balance local es para formular restricciones en el movimiento. Aquí ilustramos este tipo de aplicaciones formulando condiciones que se deben cumplir localmente cuando un fluido es incompresible. La afirmación de que un fluido es incompresible significa que todo cuerpo conserva el volumen de fluido en su movimiento. Entonces, se consideraran dos casos: el de un 'fluido libre' y el de un 'fluido en un medio poroso'. En el primer caso, el fluido llena completamente el espacio físico que ocupa el cuerpo, por lo que el volumen del fluido es igual al volumen del dominio que ocupa el cuerpo, así

$$V_f(t) = \int_{\mathcal{B}(t)} d\underline{x} \tag{25}$$

aquí, $V_f(t)$ es el volumen del fluido y $\mathcal{B}(t)$ es el dominio del espacio físico (es decir, de $\mathbb{R}^3$) ocupado por el cuerpo. Observe que una forma más explicita de esta ecuación es

$$V_f(t) = \int_{\mathcal{B}(t)} 1 d\underline{x} \tag{26}$$

porqué en la integral que aparece en la Ec. (25) el integrando es la función idénticamente 1. Comparando esta ecuación con la Ec. (14), vemos que el volumen del fluido es una propiedad extensiva y que la propiedad intensiva que le corresponde es $\psi = 1$.

Además, la hipótesis de incompresibilidad implica

$$\frac{dV_f}{dt}(t) = 0 \tag{27}$$

esta es el balance global de la Ec. (19), con $g = g_\Sigma = 0$ y $\tau = 0$, el cual a su vez es equivalente a las Ecs. (21) y (22). Tomando en cuenta además que $\psi = 1$, la Ec. (21) se reduce a

$$\nabla \cdot \underline{v} = 0. \tag{28}$$

Esta es la bien conocida condición de incompresibilidad para un fluido libre Además, aplicando la Ec. (22) donde haya discontinuidades, se obtiene $[\underline{v}] \cdot \underline{n} = 0$. Esto implica que si un fluido libre es incompresible, la velocidad de sus partículas es necesariamente continua.

El caso en que el fluido se encuentra en un 'medio poroso', es bastante diferente. Un medio poroso es un material sólido que tiene huecos distribuidos en toda su extensión, cuando los poros están llenos de un fluido, se dice que el medio poroso esta 'saturado'. Esta situación es la de mayor interés en la práctica y es también la más estudiada. En muchos de los casos que ocurren en las aplicaciones el fluido es agua o petróleo. A la fracción del volumen del sistema, constituido por la 'matriz sólida' y los huecos, se le llama 'porosidad' y se le representara por ϕ, así

$$\phi(x,t) = \lim_{V \to 0} \frac{\text{Volumen de huecos}}{\text{Volumen total}} \tag{29}$$

aquí hemos escrito $\phi(x,t)$ para enfatizar que la porosidad generalmente es función tanto de la posición como del tiempo. Las variaciones con la posición pueden ser debidas, por ejemplo, a heterogeneidad del medio y los cambios con el tiempo a su elasticidad; es decir, los cambios de presión del fluido originan esfuerzos en los poros que los dilatan o los encogen.

Cuando el medio esta saturado, el volumen del fluido V_f es igual al volumen de los huecos del dominio del espacio físico que ocupa, así

$$V_f(t) = \int_{\mathcal{B}(t)} \phi(x,t) d\underline{x}. \tag{30}$$

En vista de esta ecuación, la propiedad intensiva asociada al volumen de fluido es la porosidad $\phi(x,t)$ por lo que la condición de incomprensibilidad del fluido contenido en un medio poroso, esta dada por la ecuación diferencial

$$\frac{\partial \phi}{\partial t} + \nabla \cdot (\underline{v}\phi) = 0. \tag{31}$$

Que la divergencia de la velocidad sea igual a cero en la Ec. (28) como condición para que un fluido en su movimiento libre conserve su volumen, es ampliamente conocida. Sin embargo, este no es el caso de la Ec. (31), como condición para la conservación del volumen de los cuerpos de fluido contenidos en un medio poroso. Finalmente, debe observarse que cualquier fluido incompresible satisface la Ec. (28) cuando se mueve en el espacio libre y la Ec. (31) cuando se mueve en un medio poroso.

Cuando un fluido efectúa un movimiento en el que conserva su volumen, al movimiento se le llama 'isocorico'. Es oportuno mencionar que si bien cierto que cuando un fluido tiene la propiedad de ser incompresible, todos sus movimientos son isocoricos, lo inverso no es cierto: un fluido compresible en ocasiones puede efectuar movimientos isocoricos.

Por otra parte, cuando un fluido conserva su volumen en su movimiento satisface las condiciones de salto de Ec. (22), las cuales para este caso son

$$[\phi(\underline{v} - \underline{v}_\Sigma)] \cdot \underline{n} = 0. \tag{32}$$

En aplicaciones a geohidrología y a ingeniería petrolera, las discontinuidades de la porosidad están asociadas a cambios en los estratos geológicos y por esta

razón están fijas en el espacio; así, $\underline{v}_\Sigma = 0$ y la Ec. (32) se reduce a

$$[\phi \underline{v}] \cdot \underline{n} = 0 \tag{33}$$

o, de otra manera

$$\phi_+ v_{n_+} = \phi_- v_{n_-}. \tag{34}$$

Aquí, la componente normal de la velocidad es $v_n \equiv \underline{v} \cdot \underline{n}$ y los subíndices más y menos se utilizan para denotar los límites por los lado más y menos de Σ, respectivamente. Al producto de la porosidad por la velocidad se le conoce con el nombre de velocidad de Darcy $\underline{U}$, es decir

$$\underline{U} = \phi \underline{v} \tag{35}$$

utilizándola, las Ecs. (33) y (34) obtenemos

$$[\underline{U}] \cdot \underline{n} = 0 \quad \text{y} \quad \underline{U}_{n_+} = \underline{U}_{n_-} \tag{36}$$

es decir, 1.

La Ec. (34) es ampliamente utilizada en el estudio del agua subterránea (geohidrología). Ahí, es frecuente que la porosidad ϕ sea discontinua en la superficie de contacto entre dos estratos geológicos diferentes, pues generalmente los valores que toma esta propiedad dependen de cada estrato. En tal caso, $\phi_+ \neq \phi_-$ por lo que $v_{n_+} \neq v_{n_-}$ necesariamente.

Para más detalles de la forma y del desarrollo de algunos modelos usados en ciencias de la tierra, véase [8], [16], [24] y [14].

2. Ecuaciones Diferenciales Parciales

Cada una de las ecuaciones de balance da lugar a una ecuación diferencial parcial u ordinaria (en el caso en que el modelo depende de una sola variable independiente), la cual se complementa con las condiciones de salto, en el caso de los modelos discontinuos. Por lo mismo, los modelos de los sistemas continuos están constituidos por sistemas de ecuaciones diferenciales cuyo número es igual al número de propiedades intensivas que intervienen en la formulación del modelo básico.

Los sistemas de ecuaciones diferenciales se clasifican en elípticas, hiperbólicas y parabólicas. Es necesario aclarar que esta clasificación no es exhaustiva; es decir, existen sistemas de ecuaciones diferenciales que no pertenecen a ninguna de estas categorías. Sin embargo, casi todos los modelos de sistemas continuos, en particular los que han recibido mayor atención hasta ahora, si están incluidos en alguna de estas categorías.

2.1. Clasificación

Es importante clasificar a las ecuaciones diferenciales parciales y a los sistemas de tales ecuaciones, porqué muchas de sus propiedades son comunes a cada una de sus clases. Así, su clasificación es un instrumento para alcanzar el objetivo de unidad conceptual. La forma más general de abordar la clasificación de tales ecuaciones, es estudiando la clasificación de sistemas de ecuaciones. Sin embargo, aquí solamente abordaremos el caso de una ecuación diferencia de segundo orden, pero utilizando un método de análisis que es adecuado para extenderse a sistemas de ecuaciones.

La forma general de un operador diferencial cuasi-lineal de segundo orden definido en $\Omega \subset \mathbb{R}^2$ es

$$\mathcal{L}u \equiv a(x,y)\frac{\partial^2 u}{\partial x^2} + b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} = F(x,y,u,u_x,u_y) \tag{37}$$

para una función u de variables independientes x e y. Nos restringiremos al caso en que a, b y c son funciones sólo de x e y y no funciones de u.

Para la clasificación de las ecuaciones de segundo orden consideraremos una simplificación de la ecuación anterior en donde $F(x,y,u,u_x,u_y) = 0$ y los coeficientes a, b y c son funciones constantes, es decir

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = 0 \tag{38}$$

en la cual, examinaremos los diferentes tipos de solución que se pueden obtener para diferentes elecciones de a, b y c. Entonces iniciando con una solución de la forma

$$u(x,y) = f(mx + y) \tag{39}$$

para una función f de clase C^2 y para una constante m, que deben ser determinadas según los requerimientos de la Ec. (38). Usando un apostrofe para

denotar la derivada de f con respecto de su argumento, las requeridas derivadas parciales de segundo orden de la Ec. (38) son

$$\frac{\partial^2 u}{\partial x^2} = m^2 f'', \quad \frac{\partial^2 u}{\partial x \partial y} = m f'', \quad \frac{\partial^2 u}{\partial y^2} = f'' \tag{40}$$

sustituyendo la ecuación anterior en la Ec. (38) obtenemos

$$\left(am^2 + bm + c\right) f'' = 0 \tag{41}$$

de la cual podemos concluir que $f'' = 0$ ó $am^2 + bm + c = 0$ ó ambas. En el caso de que $f'' = 0$ obtenemos la solución $f = f_0 + mx + y$, la cual es una función lineal de x e y y es expresada en términos de dos constantes arbitrarias, f_0 y m. En el otro caso obtenemos

$$am^2 + bm + c = 0 \tag{42}$$

resolviendo esta ecuación cuadrática para m obtenemos las dos soluciones

$$m_1 = \frac{\left(-b + \sqrt[2]{b^2 - 4ac}\right)}{2a}, \quad m_2 = \frac{\left(-b - \sqrt[2]{b^2 - 4ac}\right)}{2a} \tag{43}$$

de donde es evidente la importancia de los coeficientes de la Ec. (38), ya que el signo del discriminante $\left(b^2 - 4ac\right)$ es crucial para determinar el número y tipo de soluciones de la Ec. (42). Así, tenemos tres casos a considerar:

Caso I. $\left(b^2 - 4ac\right) > 0$, **Ecuación Hiperbólica**.

La Ec. (42) tiene dos soluciones reales distintas, m_1 y m_2. Así cualquier función de cualquiera de los dos argumentos $m_1 x + y$ ó $m_2 x + y$ resuelven a la Ec. (38). Por lo tanto la solución general de la Ec. (38) es

$$u(x, y) = \mathcal{F}\left(m_1 x + y\right) + \mathcal{G}(m_2 x + y) \tag{44}$$

donde $\mathcal{F}$ y $\mathcal{G}$ son cualquier función de clase C^2. Un ejemplo de este tipo de ecuaciones es la ecuación de onda, cuya ecuación canónica es

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial t^2} = 0. \tag{45}$$

Caso II. $\left(b^2 - 4ac\right) = 0$, **Ecuación Parabólica**.

Asumiendo que $b \neq 0$ y $a \neq 0$ (lo cual implica que $c \neq 0$). Entonces se tiene una sola raíz degenerada de la Ec. (42) con el valor de $m_1 = \frac{-b}{2a}$ que resuelve a la Ec. (38). Por lo tanto la solución general de la Ec. (38) es

$$u(x, y) = \mathcal{F}\left(m_1 x + y\right) + y\mathcal{G}(m_1 x + y) \tag{46}$$

donde $\mathcal{F}$ y $\mathcal{G}$ son cualquier función de clase C^2. Si $b = 0$ y $a = 0$, entonces la solución general es

$$u(x, y) = \mathcal{F}\left(x\right) + y\mathcal{G}(x) \tag{47}$$

la cual es análoga si $b = 0$ y $c = 0$. Un ejemplo de este tipo de ecuaciones es la ecuación de difusión o calor, cuya ecuación canónica es

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = 0. \tag{48}$$

Caso III. $\left(b^2 - 4ac\right) < 0$, **Ecuación Elíptica**.

La Ec. (42) tiene dos soluciones complejas m_1 y m_2 las cuales satisfacen que m_2 es el conjugado complejo de m_1, es decir, $m_2 = m_1^*$. La solución general puede ser escrita en la forma

$$u(x, y) = \mathcal{F}\left(m_1 x + y\right) + \mathcal{G}(m_2 x + y) \tag{49}$$

donde $\mathcal{F}$ y $\mathcal{G}$ son cualquier función de clase C^2. Un ejemplo de este tipo de ecuaciones es la ecuación de Laplace, cuya ecuación canónica es

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{50}$$

Consideremos ahora el caso de un operador diferencial lineal de segundo orden definido en $\Omega \subset \mathbb{R}^n$ cuya forma general es

$$\mathcal{L}u = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} + cu \tag{51}$$

y consideremos también la ecuación homogénea asociada a este operador

$$\mathcal{L}u = 0 \tag{52}$$

además, sea $\underline{x} \in \Omega$ un punto del espacio Euclidiano y $V(x)$ una vecindad de ese punto. Sea una función u definida en $V(\underline{x})$ con la propiedad de que exista una variedad Σ de dimensión $n-1$ cerrada y orientada, tal que la función u satisface la Ec. (52) en $V(\underline{x})\backslash\Sigma$. Se supone además que existe un vector unitario $\underline{n}$ que apunta en la dirección positiva (único) está definido en Σ. Además, la función u y sus derivadas de primer orden son continuas a través de Σ, mientras que los límites de las segundas derivadas de u existen por ambos lados de Σ. Sea $\underline{x} \in \Sigma$ tal que

$$\left[\frac{\partial^2 u}{\partial x_i \partial x_j}(\underline{x})\right] \neq 0 \tag{53}$$

para alguna pareja $i, j = 1, ..., n$. Entonces decimos que la función u es una solución débil de esta ecuación en $\underline{x}$.

Teorema 3 *Una condición necesaria para que existan soluciones débiles de la ecuación homogénea (52) en un punto $\underline{x} \in \Sigma$ es que*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} n_i n_j = 0. \tag{54}$$

Así, si definimos a la matriz $\underline{\underline{A}} = (a_{ij})$ y observamos que

$$\underline{n} \cdot \underline{\underline{A}} \cdot \underline{n} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} n_i n_j \tag{55}$$

entonces podemos decir que:

I) Cuando todos los eigenvalores de la matriz $\underline{\underline{A}}$ son distintos de cero y además del mismo signo, entonces se dice que el operador es **Elíptico**.

II) Cuando todos los eigenvalores de la matriz $\underline{\underline{A}}$ son distintos de cero y además $n-1$ de ellos tienen el mismo signo, entonces se dice que el operador es **Hiperbólico**.

III) Cuando uno y sólo uno de los eigenvalores de la matriz $\underline{\underline{A}}$ es igual a cero, entonces se dice que el operador es **Parabólico**.

Para el caso en que $n = 2$, esta forma de clasificación coincide con la dada anteriormente.

2.2. Condiciones Iniciales y de Frontera

Dado un problema concreto de ecuaciones en derivadas parciales sobre un dominio Ω, si la solución existe, esta no es única ya que generalmente este tiene un número infinito de soluciones. Para que el problema tenga una y sólo una solución es necesario imponer condiciones auxiliares apropiadas y estas son las condiciones iniciales y condiciones de frontera.

En esta sección sólo se enuncian de manera general las condiciones iniciales y de frontera que son esenciales para definir un problema de ecuaciones diferenciales:

A) **Condiciones Iniciales**

Las condiciones iniciales expresan el valor de la función al tiempo inicial $t = 0$ (t puede ser fijada en cualquier valor)

$$u(\underline{x}, \underline{y}, 0) = \gamma(\underline{x}, \underline{y}). \tag{56}$$

B) **Condiciones de Frontera**

Las condiciones de frontera especifican los valores que la función $u(\underline{x}, \underline{y}, t)$ o $\nabla u(\underline{x}, \underline{y}, t)$ tomarán en la frontera $\partial\Omega$, siendo de tres tipos posibles:

1) **Condiciones tipo Dirichlet**

Especifica los valores que la función $u(\underline{x}, \underline{y}, t)$ toma en la frontera $\partial\Omega$

$$u(\underline{x}, \underline{y}, t) = \gamma(\underline{x}, \underline{y}). \tag{57}$$

2) **Condiciones tipo Neumann**

Aquí se conoce el valor de la derivada de la función $u(\underline{x}, \underline{y}, t)$ con respecto a la normal $\underline{n}$ a lo largo de la frontera $\partial\Omega$

$$\nabla u(\underline{x}, \underline{y}, t) \cdot \underline{n} = \gamma(\underline{x}, \underline{y}). \tag{58}$$

3) **Condiciones tipo Robin**

Está condición es una combinación de las dos anteriores

$$\alpha(\underline{x}, \underline{y})u(\underline{x}, \underline{y}, t) + \beta(\underline{x}, \underline{y})\nabla u(\underline{x}, \underline{y}, t) \cdot \underline{n} = g_{\partial}(\underline{x}, \underline{y}) \tag{59}$$

$\forall\ \underline{x},\ \underline{y} \in \partial\Omega$.

En un problema dado se debe prescribir las condiciones iniciales al problema y debe de existir alguno de los tipos de condiciones de frontera o combinación de ellas en $\partial\Omega$.

2.3. Modelos Completos

Los modelos de los sistemas continuos están constituidos por:

- Una colección de propiedades intensivas o lo que es lo mismo, extensivas.
- El conjunto de ecuaciones de balance local correspondientes (diferenciales y de salto).
- Suficientes relaciones que liguen a las propiedades intensivas entre sí y que definan a g, $\underline{\tau}$ y $\underline{v}$ en términos de estas, las cuales se conocen como leyes constitutivas.

Una vez que se han planteado las ecuaciones que gobiernan al problema, las condiciones iniciales, de frontera y mencionado los procesos que intervienen de manera directa en el fenómeno estudiado, necesitamos que nuestro modelo sea *completo.* Decimos que el modelo de un sistema es *completo* si define un problema *bien planteado.* Un problema de valores iniciales y condiciones de frontera es *bien planteado* si cumple que:

i) Existe una y sólo una solución y,

ii) La solución depende de manera continua de las condiciones iniciales y de frontera del problema.

Es decir, un modelo completo es aquél en el cual se incorporan condiciones iniciales y de frontera que definen conjuntamente con las ecuaciones diferenciales un problema bien planteado.

A las ecuaciones diferenciales definidas en $\Omega \subset \mathbb{R}^n$

$$\begin{aligned} \Delta u &= 0 \\ \frac{\partial^2 u}{\partial t^2} - \Delta u &= 0 \\ \frac{\partial u}{\partial t} - \Delta u &= 0 \end{aligned} \tag{60}$$

se les conoce con los nombres de ecuación de Laplace, ecuación de onda y ecuación del calor, respectivamente. Cuando se considera la primera de estas ecuaciones, se entiende que u es una función del vector $x \equiv (x_1, ..., x_n)$, mientras que cuando se considera cualquiera de las otras dos, u es una función del vector $x \equiv (x_1, ..., x_n, t)$. Así, en estos últimos casos el número de variables independientes es $n + 1$ y los conceptos relativos a la clasificación y las demás nociones discutidas con anterioridad deben aplicarse haciendo la sustitución $n \to n + 1$ e identificando $x_{n+1} = t$.

Ecuación de Laplace Para la ecuación de Laplace consideraremos condiciones del tipo Robin. En particular, condiciones de Dirichlet y condiciones de Neumann. Sin embargo, en este último caso, la solución no es única pues cualquier función constante satisface la ecuación de Laplace y también $\frac{\partial u}{\partial \underline{n}} = g_\partial$ con $g_\partial = 0$.

Ecuación de Onda Un problema general importante consiste en obtener la solución de la ecuación de onda, en el dominio del espacio-tiempo $\Omega \times [0, t]$, que satisface para cada $t \in (0, t]$ una condición de frontera de Robin en $\partial\Omega$ y las condiciones iniciales

$$u(\underline{x}, 0) = u_0(\underline{x}) \quad \text{y} \quad \frac{\partial u}{\partial t}(\underline{x}, 0) = v_0(\underline{x}), \quad \forall \underline{x} \in \Omega \tag{61}$$

aquí $u_0(\underline{x})$ y $v_0(\underline{x})$ son dos funciones prescritas. El hecho de que para la ecuación de onda se prescriban los valores iniciales, de la función y su derivada con respecto al tiempo, es reminiscente de que en la mecánica de partículas se necesitan las posiciones y las velocidades iniciales para determinar el movimiento de un sistema de partículas.

Ecuación de Calor También para la ecuación del calor un problema general importante consiste en obtener la solución de la ecuación de onda, en el dominio del espacio-tiempo $\Omega \times [0, t]$, que satisface para cada $t \in (0, t]$ una condición de frontera de Robin en y ciertas condiciones iniciales. Sin embargo, en este caso en ellas sólo se prescribe a la función

$$u(\underline{x}, 0) = u_0(\underline{x}), \quad \forall \underline{x} \in \Omega. \tag{62}$$

3. Análisis Funcional y Problemas Variacionales

En esta sección se detallan los conceptos básicos de análisis funcional y problemas variacionales con énfasis en problemas elípticos de orden par $2m$, para comenzar detallaremos lo que entendemos por un operador diferencial parcial elíptico de orden par $2m$ en n variables, para después definir a los espacios de Sobolev para poder tratar problemas variacionales con valor en la frontera.

En donde, restringiéndonos a problemas elípticos, contestaremos una cuestión central en la teoría de problemas elípticos con valores en la frontera, y está se relaciona con las condiciones bajo las cuales uno puede esperar que el problema tenga solución y esta es única, así como conocer la regularidad de la solución, para mayor referencia de estos resultados ver [13], [19] y [3].

3.1. Operador Lineal Elíptico

Definición 4 *Entenderemos por un dominio al conjunto $\Omega \subset \mathbb{R}^n$ que sea abierto y conexo.*

Para poder expresar de forma compacta derivadas parciales de orden m o menor, usaremos la definición siguiente.

Definición 5 *Sea $\mathbb{Z}_+^n$ el conjunto de todas las n-duplas de enteros no negativos, un miembro de $\mathbb{Z}_+^n$ se denota usualmente por α ó β (por ejemplo $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$. Denotaremos por $|\alpha|$ la suma $|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n$ y por $D^\alpha u$ la derivada parcial*

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} \tag{63}$$

así, si $|\alpha| = m$, entonces $D^\alpha u$ denota la m-ésima derivada parcial de u.

Sea $\mathcal{L}$ un operador diferencial parcial de orden par $2m$ en n variables y de la forma

$$\mathcal{L}u = \sum_{|\alpha|,|\beta| \leq m} (-1)^{|\alpha|} D^\alpha \left(a_{\alpha\beta}(\underline{x}) D^\beta u \right), \quad \underline{x} \in \Omega \subset \mathbb{R}^n \tag{64}$$

donde Ω es un dominio en $\mathbb{R}^n$. Los coeficientes $a_{\alpha\beta}$ son funciones suaves real valuadas de $\underline{x}$.

El operador $\mathcal{L}$ es asumido que aparece dentro de una ecuación diferencial parcial de la forma

$$\mathcal{L}u = f, \tag{65}$$

donde f pertenece al rango del operador $\mathcal{L}$.

La clasificación del operador $\mathcal{L}$ depende sólo de los coeficientes de la derivada más alta, esto es, de la derivada de orden $2m$, y a los términos involucrados en esa derivada son llamados la parte principal del operador $\mathcal{L}$ denotado por $\mathcal{L}_0$ y para el operador (64) es de la forma

$$\mathcal{L}_0 = \sum_{|\alpha|,|\beta| \leq m} a_{\alpha\beta} D^{\alpha+\beta} u. \tag{66}$$

Teorema 6 *Sea ξ un vector en $\mathbb{R}^n$, y sea $\xi^\alpha = \xi_n^{\alpha_1}...\xi_n^{\alpha_n}, \alpha \in \mathbb{Z}_n^+$. Entonces*
i) $\mathcal{L}$ es elíptico en $\underline{x_o} \in \Omega$ si

$$\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}\left(\underline{x_o}\right)\xi^{\alpha+\beta} \neq 0 \qquad \forall \xi \neq 0; \tag{67}$$

ii) $\mathcal{L}$ es elíptico si es elíptico en todos los puntos de Ω;
iii) $\mathcal{L}$ es fuertemente elíptico si existe un número $\mu > 0$ tal que

$$\left|\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}\left(\underline{x_o}\right)\xi^{\alpha+\beta}\right| \geq \mu\left|\xi\right|^{2m} \tag{68}$$

satisfaciéndose en todo punto $\underline{x_o} \in \Omega$, y para todo $\xi \in \mathbb{R}^n$. Aquí $|\xi| = \left(\xi_1^2 + ... + \xi_n^2\right)^{\frac{1}{2}}$.

Para el caso en el cual $\mathcal{L}$ es un operador de $2do$ orden ($m = 1$), la notación se simplifica, tomando la forma

$$\mathcal{L}u = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\underline{x})\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{n} a_j \frac{\partial u}{\partial x_j} + a_0 u = f \tag{69}$$

en Ω.

Para coeficientes adecuados a_{ij}, a_j y a_0 la condición para conocer si el operador es elíptico, es examinado por la condición

$$\sum_{i,j=1}^{n} a_{ij}(\underline{x_0})\xi_i\xi_j \neq 0 \qquad \forall \xi \neq 0 \tag{70}$$

y para conocer si el operador es fuertemente elíptico, es examinado por la condición

$$\sum_{i,j=1}^{n} a_{ij}(\underline{x_0})\xi_i\xi_j > \mu\left|\xi\right|^2. \tag{71}$$

3.2. Espacios de Sobolev

En esta subsección detallaremos algunos resultados de los espacios de Sobolev sobre el conjunto de números reales, en estos espacios son sobre los cuales trabajaremos tanto para plantear el problema elíptico como para encontrar la solución al problema. Primeramente definiremos lo que entendemos por un espacio L^2.

Definición 7 *Una función medible $u(\underline{x})$ definida sobre $\Omega \subset \mathbb{R}^n$ se dice que pertenece al espacio $L^2(\Omega)$ si*

$$\int_\Omega \left|u(\underline{x})\right|^2 d\underline{x} < \infty \tag{72}$$

es decir, es integrable.

La definición de los espacios medibles, espacios L^p, distribuciones y derivadas de distribuciones están dados en el apéndice, estos resultados son la base para poder definir a los espacios de Sobolev.

Definición 8 *El espacio de Sobolev de orden m, denotado por $H^m(\Omega)$, es definido*

$$H^m(\Omega) = \left\{ u \mid D^\alpha u \in L^2(\Omega) \quad \forall \alpha \text{ tal que } |\alpha| \leq m \right\}. \tag{73}$$

El producto escalar $\langle \cdot, \cdot \rangle$ de dos elementos u y $v \in H^m(\Omega)$ esta dado por

$$\langle u, v \rangle_{H^m} = \int_\Omega \sum_{|\alpha| \leq m} \left(D^\alpha u \right) \left(D^\alpha v \right) d\underline{x} \text{ para } u, v \in H^m(\Omega). \tag{74}$$

Nota: Es común que el espacio $L^2(\Omega)$ sea denotado por $H^0(\Omega)$.

Un espacio completo con producto interior es llamado un espacio de Hilbert, un espacio normado y completo es llamado espacio de Banach. Y como todo producto interior define una norma, entonces todo espacio de Hilbert es un espacio de Banach.

Definición 9 *La norma $\|\cdot\|_{H^m}$ inducida a partir del producto interior $\langle \cdot, \cdot \rangle_{H^m}$ queda definida por*

$$\|u\|_{H^m}^2 = \langle u, u \rangle_{H^m} = \int_\Omega \sum_{|\alpha| \leq m} \left(D^\alpha u \right)^2 d\underline{x}. \tag{75}$$

Ahora, con norma $\|\cdot\|_{H^m}$, el espacio $H^m(\Omega)$ es un espacio de Hilbert, esto queda plasmado en el siguiente resultado.

Teorema 10 *El espacio $H^m(\Omega)$ con la norma $\|\cdot\|_{H^m}$ es un espacio de Hilbert.*

Ya que algunas de las propiedades de los espacios de Sobolev sólo son validas cuando la frontera del dominio es suficientemente suave. Para describir al conjunto donde los espacios de Sobolev están definidos, es común pedirle algunas propiedades y así definimos lo siguiente.

Definición 11 *Una función f definida sobre un conjunto $\Gamma \subset \mathbb{R}^n$ es llamada Lipschitz continua si existe una constante $L > 0$ tal que*

$$|f(x) - f(y)| \leq L |x - y| \quad \forall x, y \in \Gamma. \tag{76}$$

Notemos que una función Lipschitz continua es uniformemente continua.

Sea $\Omega \subset \mathbb{R}^n$ $(n \geq 2)$ un dominio con frontera $\partial\Omega$, sea $x_0 \in \partial\Omega$ y construyamos la bola abierta con centro en x_0 y radio ε, i.e. $B(x_0, \varepsilon)$, entonces definiremos el sistema coordenado $(\xi_1, ..., \xi_n)$ tal que el segmento $\partial\Omega \cap B(x_0, \varepsilon)$ pueda expresarse como una función

$$\xi_n = f(\xi_1, ..., \xi_{n-1}) \tag{77}$$

entonces definimos.

Definición 12 *La frontera $\partial\Omega$ del dominio Ω es llamada de Lipschitz si f definida como en la Ec. (77) es una función Lipschitz continua.*

El siguiente teorema resume las propiedades más importantes de los espacios de Sobolev $H^m(\Omega)$.

Teorema 13 *Sea $H^m(\Omega)$ el espacio de Sobolev de orden m y sea $\Omega \subset \mathbb{R}^n$ un dominio acotado con frontera Lipschitz. Entonces*
i) $H^r(\Omega) \subset H^m(\Omega)$ si $r \geq m$
ii) $H^m(\Omega)$ es un espacio de Hilbert con respecto a la norma $\|\cdot\|_{H^m}$
iii) $H^m(\Omega)$ es la cerradura con respecto a la norma $\|\cdot\|_{H^m}$ del espacio $C^\infty(\overline{\Omega})$.

De la parte *iii*) del teorema anterior, se puede hacer una importante interpretación: Para toda $u \in H^m(\Omega)$ es siempre posible encontrar una función infinitamente diferenciable f, tal que este arbitrariamente cerca de u en el sentido que

$$\|u - f\|_{H^m} < \varepsilon \tag{78}$$

para algún $\varepsilon > 0$ dado.

Cuando $m = 0$, se deduce la propiedad $H^0(\Omega) = L^2(\Omega)$ a partir del teorema anterior.

Corolario 14 *El espacio $L^2(\Omega)$ es la cerradura, con respecto a la norma L^2, del espacio $C^\infty(\overline{\Omega})$.*

Otra propiedad, se tiene al considerar a cualquier miembro de $u \in H^m(\Omega)$, este puede ser identificado con una función en $C^m(\overline{\Omega})$, después de que posiblemente sean cambiados algunos valores sobre un conjunto de medida cero, esto queda plasmado en los dos siguientes resultados.

Teorema 15 *Sean X y Y dos espacios de Banach, con $X \subset Y$. Sea $i : X \to Y$ tal que $i(u) = u$. Si el espacio X tiene definida la norma $\|\cdot\|_X$ y el espacio Y tiene definida la norma $\|\cdot\|_Y$, decimos que X está inmersa continuamente en Y si*

$$\|i(u)\|_Y = \|u\|_Y \leq K \|u\|_X \tag{79}$$

para alguna constante $K > 0$.

Teorema 16 *(Inmersión de Sobolev)*
Sea $\Omega \subset \mathbb{R}^n$ un dominio acotado con frontera $\partial\Omega$ de Lipschitz. Si $(m - k) > n/2$, entonces toda función en $H^m(\Omega)$ pertenece a $C^k(\overline{\Omega})$, es decir, hay un miembro que pertenece a $C^k(\overline{\Omega})$. Además, la inmersión

$$H^m(\Omega) \subset C^k(\overline{\Omega}) \tag{80}$$

es continua.

3.2.1. Trazas de una Función en $H^m(\Omega)$.

Una parte fundamental en los problemas con valores en la frontera definidos sobre el dominio Ω, es definir de forma única los valores que tomará la función sobre la frontera $\partial\Omega$, en este apartado veremos bajo que condiciones es posible tener definidos de forma única los valores en la frontera $\partial\Omega$ tal que podamos definir un operador $tr(\cdot)$ continuo que actué en $\overline{\Omega}$ tal que $tr(u) = u_{|\partial\Omega}$.

El siguiente lema nos dice que el operador $tr(\cdot)$ es un operador lineal continuo de $C^1\left(\overline{\Omega}\right)$ a $C(\partial\Omega)$, con respecto a las normas $\|\cdot\|_{H^1(\Omega)}$ y $\|\cdot\|_{L^2(\partial\Omega)}$.

Lema 17 *Sea Ω un dominio con frontera $\partial\Omega$ de Lipschitz. La estimación*

$$\|tr(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)} \tag{81}$$

se satisface para toda función $u \in C^1\left(\overline{\Omega}\right)$, para alguna constante $C > 0$.

Ahora, para el caso $tr(\cdot) : H^1(\Omega) \to L^2(\partial\Omega)$, se tiene el siguiente teorema.

Teorema 18 *Sea Ω un dominio acotado en $\mathbb{R}^n$ con frontera $\partial\Omega$ de Lipschitz. Entonces:*

i) Existe un único operador lineal acotado $tr(\cdot) : H^1(\Omega) \to L^2(\partial\Omega)$, tal que

$$\|tr(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}, \tag{82}$$

con la propiedad que si $u \in C^1\left(\overline{\Omega}\right)$, entonces $tr(u) = u|_{\partial\Omega}$.

ii) El rango de $tr(\cdot)$ es denso en $L^2(\partial\Omega)$.

El argumento anterior puede ser generalizado para los espacios $H^m(\Omega)$, de hecho, cuando $m > 1$, entonces para toda $u \in H^m(\Omega)$ tenemos que

$$D^\alpha u \in H^1(\Omega) \qquad \text{para } |\alpha| \leq m-1, \tag{83}$$

por el teorema anterior, el valor de $D^\alpha u$ sobre la frontera está bien definido y pertenece a $L^2(\Omega)$, es decir

$$tr(D^\alpha u) \in L^2(\Omega), \qquad |\alpha| \leq m-1. \tag{84}$$

Además, si u es m-veces continuamente diferenciable, entonces $D^\alpha u$ es al menos continuamente diferenciable para $|\alpha| \leq m-1$ y

$$tr(D^\alpha u) = (D^\alpha u)|_{\partial\Omega}. \tag{85}$$

3.2.2. Espacios $H_0^m(\Omega)$.

Los espacio $H_0^m(\Omega)$ surgen comúnmente al trabajar con problemas con valor en la frontera y serán aquellos espacios que se nulifiquen en la frontera del dominio, es decir

Definición 19 *Definimos a los espacios* $H_0^m(\Omega)$ *como la cerradura, en la norma de Sobolev* $\|\cdot\|_{H^m}$, *del espacio* $C_0^m(\Omega)$ *de funciones con derivadas continuas del orden menor que* m, *todas las cuales tienen soporte compacto en* Ω, *es decir* $H_0^m(\Omega)$ *es formado al tomar la unión de* $C_0^m(\Omega)$ *y de todos los límites de sucesiones de Cauchy en* $C_0^m(\Omega)$ *que no pertenecen a* $C_0^m(\Omega)$.

Las propiedades básicas de estos espacios están contenidas en el siguiente resultado.

Teorema 20 *Sea* Ω *un dominio acotado en* $\mathbb{R}^n$ *con frontera* $\partial\Omega$ *suficientemente suave y sea* $H_0^m(\Omega)$ *la cerradura de* $C_0^\infty(\Omega)$ *en la norma* $\|\cdot\|_{H^m}$, *entonces*

a) $H_0^m(\Omega)$ *es la cerradura de* $C_0^\infty(\Omega)$ *en la norma* $\|\cdot\|_{H^m}$;

b) $H_0^m(\Omega) \subset H^m(\Omega)$;

c) Si $u \in H^m(\Omega)$ *pertenece a* $H_0^m(\Omega)$, *entonces*

$$D^\alpha u = 0, \text{ sobre } \partial\Omega, |\alpha| \leq m-1. \tag{86}$$

Teorema 21 *(Desigualdad de Poincaré-Friedrichs)*

Sea Ω *un dominio acotado en* $\mathbb{R}^n$. *Entonces existe una constante* $C > 0$ *tal que*

$$\int_\Omega |u|^2 \, dx \leq C \int_\Omega |\nabla u|^2 \, dx \tag{87}$$

para toda $u \in H_0^1(\Omega)$.

Introduciendo ahora una familia de semi-normas sobre $H^m(\Omega)$ (una semi-norma $|\cdot|$ satisface casi todos los axiomas de una norma excepto el de positivo definido), de la siguiente forma:

Definición 22 *La semi-norma* $|\cdot|_m$ *sobre* $H^m(\Omega)$, *se define como*

$$|u|_m^2 = \sum_{|\alpha|=m} \int_\Omega |D^\alpha u|^2 \, dx. \tag{88}$$

Esta es una semi-norma, ya que $|u|_m = 0$ *implica que* $D^\alpha u = 0$ *para* $|\alpha| = m$, *lo cual no implica que* $u = 0$.

La relevancia de esta semi-norma está al aplicar la desigualdad de Poincaré-Friedrichs ya que es posible demostrar que $|\cdot|_1$ es de hecho una norma sobre $H_0^1(\Omega)$.

Corolario 23 *La semi-norma* $|\cdot|_1$ *es una norma sobre* $H_0^1(\Omega)$, *equivalente a la norma estándar* $\|\cdot\|_{H^1}$.

Es posible extender el teorema anterior y su corolario a los espacios $H_0^m(\Omega)$ para cualquier $m \geq 1$, de la siguiente forma:

Teorema 24 *Sea Ω un dominio acotado en $\mathbb{R}^n$. Entonces existe una constante $C > 0$ tal que*

$$\|u\|_{L^2}^2 \leq C\,|u|_m^2 \tag{89}$$

para toda $u \in H_0^m(\Omega)$, además, $|\cdot|_m$ es una norma sobre $H_0^m(\Omega)$ equivalente a la norma estándar $\|\cdot\|_{H^m}$.

Definición 25 *Sea Ω un dominio acotado en $\mathbb{R}^n$. Definimos por $H^{-m}(\Omega)$ al espacio de todas las funcionales lineales acotadas sobre $H_0^m(\Omega)$, es decir, $H^{-m}(\Omega)$ será el espacio dual del espacio $H_0^m(\Omega)$.*

Teorema 26 *q será una distribución de $H^{-m}(\Omega)$ si y sólo si q puede ser expresada en la forma*

$$q = \sum_{|\alpha|<m} D^\alpha q_\alpha \tag{90}$$

donde q_α son funcionales en $L^2(\Omega)$.

3.3. Formulas de Green y Problemas Adjuntos

Una cuestión central en la teoría de problemas elípticos con valores en la frontera se relaciona con las condiciones bajo las cuales uno puede esperar una única solución a problemas de la forma

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \subset \mathbb{R}^n \\ &\left.\begin{aligned} B_o u &= g_0 \\ B_1 u &= g_1 \\ &\vdots \\ B_{m-1} u &= g_{m-1} \end{aligned}\right\} \quad \text{en } \partial\Omega \end{aligned} \tag{91}$$

donde $\mathcal{L}$ es un operador elíptico de orden $2m$, de forma

$$\mathcal{L}u = \sum_{|\alpha|\leq m} (-1)^{|\alpha|} D^\alpha \left(\sum_{|\beta|\leq m} a_{\alpha\beta}(\underline{x}) D^\beta u \right), \quad \underline{x} \in \Omega \subset \mathbb{R}^n \tag{92}$$

donde los coeficientes $a_{\alpha\beta}$ son funciones de $\underline{x}$ suaves y satisfacen las condiciones para que el operador sea elíptico, el conjunto $B_0, B_1, ..., B_{m-1}$ de operadores de frontera son de la forma

$$B_j u = \sum_{|\alpha|\leq q_j} b_\alpha^{(j)} D^\alpha u = g_j \tag{93}$$

y constituyen un conjunto de condiciones de frontera que cubren a $\mathcal{L}$. Los coeficientes $b_\alpha^{(j)}$ son asumidos como funciones suaves.

En el caso de problemas de segundo orden la Ec. (93) puede expresarse como una sola condición de frontera

$$Bu = \sum_{j=1}^{n} b_j \frac{\partial u}{\partial x_j} + cu = g \quad \text{en } \partial\Omega. \tag{94}$$

Antes de poder ver las condiciones bajo las cuales se garantice la existencia y unicidad es necesario introducir el concepto de formula de Green asociada con el operador $\mathcal{L}^*$, para ello definimos:

Definición 27 *Con el operador dado como en la Ec. (92), denotaremos por $\mathcal{L}^*$ al operador definido por*

$$\mathcal{L}^* u = \sum_{|\alpha| \leq m} (-1)^{|\alpha|} D^\alpha \left(\sum_{|\beta| \leq m} a_{\beta\alpha}(\underline{x}) D^\beta u \right) \tag{95}$$

y nos referiremos a $\mathcal{L}^*$ como el adjunto formal del operador $\mathcal{L}$.

La importancia del adjunto formal es que si aplicamos el teorema de Green (83) a la integral

$$\int_\Omega v \mathcal{L} u d\underline{x} \tag{96}$$

obtenemos

$$\int_\Omega v \mathcal{L} u d\underline{x} = \int_\Omega u \mathcal{L}^* v d\underline{x} + \int_{\partial\Omega} F(u,v) d\underline{s} \tag{97}$$

en la cual $F(u,v)$ representa términos de frontera que se nulifican al aplicar el teorema ya que la función $v \in H_0^1(\Omega)$. Si $\mathcal{L} = \mathcal{L}^*$; i.e. $a_{\alpha\beta} = a_{\beta\alpha}$ el operador es llamado de manera formal el auto-adjunto.

En el caso de problemas de segundo orden, dos sucesivas aplicaciones del teorema de Green (83) y obtenemos, para i y j fijos

$$\begin{aligned} -\int_\Omega v \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) d\underline{x} &= -\int_{\partial\Omega} v a_{ij} \frac{\partial u}{\partial x_j} n_i d\underline{s} + \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} d\underline{x} \\ &= -\int_{\partial\Omega} \left[v a_{ij} \frac{\partial u}{\partial x_j} n_i - u a_{ij} \frac{\partial v}{\partial x_i} n_j \right] d\underline{s} \\ &\quad -\int_\Omega u \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial v}{\partial x_i} \right) d\underline{x}. \end{aligned} \tag{98}$$

Pero sumando sobre i y j, obtenemos de la Ec. (97)

$$\mathcal{L}^* v = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(a_{ji}(\underline{x}) \frac{\partial v}{\partial x_j} \right) \tag{99}$$

y

$$F(u,v) = -\sum_{i,j=1}^{n} a_{ij} \left(v \frac{\partial u}{\partial x_j} n_i - u \frac{\partial v}{\partial x_i} n_j \right) \tag{100}$$

tal que $\mathcal{L}$ es formalmente el auto-ajunto si $a_{ji} = a_{ij}$.

Para hacer el tratamiento más simple, restringiremos nuestra atención al problema homogéneo, es decir, en el cual $g_0, g_1, ..., g_{m-1} = 0$ (esta no es una restricción real, ya que se puede demostrar que cualquier problema no-homogéneo

con condiciones de frontera puede convertirse en uno con condiciones de frontera homogéneo de una manera sistemática), asumiremos también que Ω es suave y la frontera $\partial\Omega$ de Ω es de clase C^{∞}.

Así, en lo que resta de la sección, daremos los pasos necesarios para poder conocer bajo que condiciones el problema elíptico con valores en la frontera del tipo

$$\begin{array}{rcl} \mathcal{L}u & = & f_{\Omega} \quad \text{en } \Omega \subset \mathbb{R}^n \\ & & \left.\begin{array}{c} B_o u = 0 \\ B_1 u = 0 \\ \vdots \\ B_{m-1} u = 0 \end{array}\right\} \quad \text{en } \partial\Omega \end{array} \tag{101}$$

donde el operador $\mathcal{L}$ y B_j estan dados como en (92) y (93), con $s \geq 2m$ tiene solución y esta es única. Para ello, necesitamos adoptar el lenguaje de la teoría de operadores lineales, algunos resultados clave de algebra lineal están detallados en el apéndice.

Primeramente denotemos $N(B_j)$ al espacio nulo del operador de frontera $B_j : H^s(\Omega) \to L^2(\Omega)$, entonces

$$N(B_j) = \{u \in H^s(\Omega) \mid B_j u = 0 \text{ en } \partial\Omega\} \tag{102}$$

para $j = 0, 1, 2, ..., m-1$.

Adicionalmente definimos al dominio del operador $\mathcal{L}$, como el espacio

$$\begin{array}{rcl} D(\mathcal{L}) & = & H^s(\Omega) \cap N(B_0) \cap ... \cap N(B_{m-1}) \\ & = & \{u \in H^s(\Omega) \mid B_j u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\}. \end{array} \tag{103}$$

Entonces el problema elíptico con valores en la frontera de la Ec. (101) con $s \geq 2m$, puede reescribirse como, dado $\mathcal{L} : D(\mathcal{L}) \to H^{s-2m}(\Omega)$, hallar u que satisfaga

$$\mathcal{L}u = f_{\Omega} \quad \text{en } \Omega. \tag{104}$$

Lo primero que hay que determinar es el conjunto de funciones f_{Ω} en $H^{s-2m}(\Omega)$ para las cuales la ecuación anterior se satisface, i.e. debemos identificar el rango $R(\mathcal{L})$ del operador $\mathcal{L}$. Pero como nos interesa conocer bajo que condiciones la solución u es única, entonces podemos definir el núcleo $N(\mathcal{L})$ del operador $\mathcal{L}$ como sigue

$$\begin{array}{rcl} N(\mathcal{L}) & = & \{u \in D(\mathcal{L}) \mid \mathcal{L}u = 0\} \\ & = & \{u \in H^s(\Omega) \mid \mathcal{L}u = 0 \quad \text{en } \Omega, B_j u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\}. \end{array} \tag{105}$$

Si el $N(\mathcal{L}) \neq \{0\}$, entonces no hay una única solución, ya que si u_0 es una solución, entonces $u_0 + w$ también es solución para cualquier $w \in N(\mathcal{L})$, ya que

$$\mathcal{L}(u_0 + w) = \mathcal{L}u_0 + \mathcal{L}w = \mathcal{L}u_0 = f_{\Omega}. \tag{106}$$

Así, los elementos del núcleo $N(\mathcal{L})$ de $\mathcal{L}$ deberán ser excluidos del dominio $D(\mathcal{L})$ del operador $\mathcal{L}$, para poder asegurar la unicidad de la solución u.

Si ahora, introducimos el complemento ortogonal $N(\mathcal{L})^{\perp}$ del núcleo $N(\mathcal{L})$ del operador $\mathcal{L}$ con respecto al producto interior L^2, definiéndolo como

$$N(\mathcal{L})^{\perp} = \{v \in D(\mathcal{L}) \mid (v, w) = 0 \ \forall w \in N(\mathcal{L})\}. \tag{107}$$

De esta forma tenemos que

$$D(\mathcal{L}) = N(\mathcal{L}) \oplus N(\mathcal{L})^{\perp} \tag{108}$$

i.e. para toda $u \in D(\mathcal{L})$, u se escribe como $u = v + w$ donde $v \in N(\mathcal{L})^{\perp}$ y $w \in N(\mathcal{L})$. Además $N(\mathcal{L}) \cap N(\mathcal{L})^{\perp} = \{0\}$.

De forma similar, podemos definir los espacios anteriores para el problema adjunto

$$\begin{array}{rcl} \mathcal{L}^* u & = & f_\Omega \quad \text{en } \Omega \subset \mathbb{R}^n \\ & & \left.\begin{array}{c} B_o^* u = 0 \\ B_1^* u = 0 \\ \vdots \\ B_{m-1}^* u = 0 \end{array}\right\} \quad \text{en } \partial\Omega \end{array} \tag{109}$$

y definimos

$$\begin{array}{rcl} D(\mathcal{L}^*) & = & H^s(\Omega) \cap N(B_0^*) \cap ... \cap N(B_{m-1}^*) \\ & = & \left\{u \in H^s(\Omega) \mid B_j^* u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\right\}. \end{array} \tag{110}$$

Entonces el problema elíptico con valores en la frontera de la Ec. (101) con $s \geq 2m$, puede reescribirse como, dado $\mathcal{L}^* : D(\mathcal{L}^*) \to H^{s-2m}(\Omega)$, hallar u que satisfaga

$$\mathcal{L}^* u = f_\Omega \quad \text{en } \Omega. \tag{111}$$

Definiendo para el operador $\mathcal{L}^*$

$$\begin{array}{rcl} N(\mathcal{L}^*) & = & \{u \in D(\mathcal{L}^*) \mid \mathcal{L}^* u = 0\} \\ & = & \left\{u \in H^s(\Omega) \mid \mathcal{L}^* u = 0 \quad \text{en } \Omega, B_j^* u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\right\}. \end{array} \tag{112}$$

y

$$N(\mathcal{L}^*)^{\perp} = \{v \in D(\mathcal{L}^*) \mid (v, w)_{L^2} = 0 \ \forall w \in N(\mathcal{L}^*)\}. \tag{113}$$

Así, con estas definiciones, es posible ver una cuestión fundamental, esta es, conocer bajo que condiciones el problema elíptico con valores en la frontera de la Ec. (101) con $s \geq 2m$ tiene solución y esta es única, esto queda resuelto en el siguiente teorema cuya demostración puede verse en [3] y [13].

Teorema 28 *Considerando el problema elíptico con valores en la frontera de la Ec. (101) con $s \geq 2m$ definido sobre un dominio Ω acotado con frontera $\partial\Omega$ suave. Entonces*

i) Existe al menos una solución si y sólo si $f \in N(\mathcal{L}^*)^{\perp}$, esto es, si

$$(f, v)_{L^2(\Omega)} = 0 \quad \forall v \in N(\mathcal{L}^*). \tag{114}$$

ii) Asumiendo que la solución u existe, esta es única si $u \in N(\mathcal{L})^{\perp}$, esto es, si

$$(u, w)_{L^2(\Omega)} = 0 \quad \forall w \in N(\mathcal{L}). \tag{115}$$

iii) Si existe una única solución, entonces existe una única constante $C > 0$, independiente de u, tal que

$$\|u\|_{H^s} \leq C \|f\|_{H^{s-2m}}. \tag{116}$$

Observación 29 *i) El teorema afirma que el operador $\mathcal{L}$ es un operador suprayectivo de $D(\mathcal{L})$ sobre el subespacio de funciones en H^{s-2m} que satisface (115). Además el operador $\mathcal{L}$ es inyectivo si el dominio es restringido al espacio de funciones que satisfagan a (114).*

ii) La parte (iii) del teorema puede interpretarse como un resultado de regularidad, en el sentido en que se muestra

$$u \in H^{s-2m}(\Omega) \text{ si } f \in H^s(\Omega). \tag{117}$$

Así, formalmente podemos definir el adjunto formal de la siguiente manera

Definición 30 *Sea $\mathcal{L}$ un Operador Diferencial, decimos que un operador $\mathcal{L}^*$ es su adjunto formal si satisface la siguiente condición*

$$w\mathcal{L}u - u\mathcal{L}^*w = \nabla \cdot \underline{\mathfrak{D}}(u, w) \tag{3.1}$$

tal que las funciones u y w pertenecen a un espacio lineal. Aquí $\underline{\mathfrak{D}}(u, w)$ es una funcional bilineal que representa términos de frontera.

Ejemplos de Operadores Adjuntos Formales A continuación se muestra mediante ejemplos el uso de la definición de operadores adjuntos formales y la parte correspondiente a términos de frontera.

A) **Operador de la derivada de orden cero**

La derivada de orden cero de una función u es tal que

$$\frac{d^n u}{dx^n} = u \tag{118}$$

es decir, $n = 0$, sea el operador

$$\mathcal{L}u = u \tag{119}$$

de la definición de operador adjunto tenemos que

$$w\mathcal{L}u = u\mathcal{L}^* w + \nabla \cdot \underline{\mathfrak{D}}\,(u, w) \tag{120}$$

entonces el término izquierdo es

$$w\mathcal{L}u = wu \tag{121}$$

de aquí

$$u\mathcal{L}^* w = uw \tag{122}$$

por lo tanto el operador adjunto formal es

$$\mathcal{L}^* w = w \tag{123}$$

nótese que el operador es auto-adjunto.

B) **Operador de la derivada de primer orden**

La derivada de primer orden en términos del operador es

$$\mathcal{L}u = c\frac{du}{dx} \tag{124}$$

de la definición de operador adjunto tenemos

$$w\mathcal{L}u = u\mathcal{L}w + \nabla \cdot \underline{\mathfrak{D}}\,(u, w) \tag{125}$$

desarrollando el lado izquierdo

$$\begin{aligned} w\mathcal{L}u &= wc\frac{du}{dx} \\ &= \frac{d\,(wcu)}{dx} - u\frac{d\,(cw)}{dx} \\ &= \frac{d\,(wcu)}{dx} - uc\frac{dw}{dx} \end{aligned} \tag{126}$$

por lo tanto, el operador adjunto formal es

$$\mathcal{L}^* w = -c\frac{dw}{dx} \tag{127}$$

y los términos de frontera son

$$\mathfrak{D}\,(u, w) = wcu \tag{128}$$

C) **Operador Elíptico**

El operador elíptico más sencillo es el Laplaciano

$$\mathcal{L}u \equiv -\triangle u = -\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) \tag{129}$$

de la ecuación del operador adjunto formal tenemos

$$\begin{aligned} w\mathcal{L}u &= -w\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) \\ &= -\frac{\partial}{\partial x_i}\left(w\frac{\partial u}{\partial x_i}\right)+\frac{\partial u}{\partial x_i}\frac{\partial w}{\partial x_i} \\ &= -\frac{\partial}{\partial x_i}\left(w\frac{\partial u}{\partial x_i}\right)+\frac{\partial}{\partial x_i}\left(u\frac{\partial w}{\partial x_i}\right)-u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \\ &= \frac{\partial}{\partial x_i}\left(u\frac{\partial w}{\partial x_i}-w\frac{\partial u}{\partial x_i}\right)-u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \end{aligned} \tag{130}$$

entonces, el operador adjunto formal es

$$\mathcal{L}^* w = -u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \tag{131}$$

es decir, el operador es autoadjunto. Notemos que la función bilineal $\underline{\mathfrak{D}}(u,w)$ es

$$\underline{\mathfrak{D}}\,(u,w) = u\frac{\partial w}{\partial x_i}-w\frac{\partial u}{\partial x_i}. \tag{132}$$

D) Consideremos el operador diferencial elíptico más general de segundo orden

$$\mathcal{L}u = -\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(\underline{b}u\right)+cu \tag{133}$$

de la definición de operador adjunto formal tenemos que

$$w\mathcal{L}u = u\mathcal{L}^* w+\nabla\cdot\underline{\mathfrak{D}}\,(u,w) \tag{134}$$

desarrollando el lado derecho de la ecuación anterior

$$\begin{aligned} w\mathcal{L}u &= w\left(-\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(\underline{b}u\right)+cu\right) \\ &= -w\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+w\nabla\cdot\left(\underline{b}u\right)+wcu \end{aligned} \tag{135}$$

aplicando la igualdad de divergencia a los dos primeros sumandos se tiene que la ecuación anterior es

$$\begin{aligned} w\mathcal{L}u &= -\nabla\cdot\left(w\underline{\underline{a}}\cdot\nabla u\right)+\underline{\underline{a}}\cdot\nabla u\cdot\nabla w+\nabla\cdot\left(w\underline{b}u\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \\ &= -\nabla\cdot\left(w\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(u\underline{\underline{a}}\nabla w\right)-u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right)+\nabla\cdot\left(w\underline{b}u\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \\ &= \nabla\cdot\left[\underline{\underline{a}}\left(u\nabla w-w\nabla u\right)\right]+\nabla\cdot\left(w\underline{b}u\right)-u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \end{aligned} \tag{136}$$

reordenando términos se tiene

$$w\mathcal{L}u = -u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right)-u\underline{b}\cdot\nabla w+ucw+\nabla\cdot\left[\underline{\underline{a}}\left(u\nabla w-w\nabla u\right)+\left(w\underline{b}u\right)\right] \tag{137}$$

por lo tanto, el operador adjunto formal es

$$\mathcal{L}^* w = -\nabla \cdot \left(\underline{\underline{a}} \cdot \nabla w\right) - \underline{b} \cdot \nabla w + cw \tag{138}$$

y el término correspondiente a valores en la frontera es

$$\underline{\mathfrak{D}}\left(u, w\right) = \underline{\underline{a}} \cdot \left(u \nabla w - w \nabla u\right) + \left(w \underline{b} u\right). \tag{139}$$

E) **La ecuación bi-armónica**

Consideremos el operador diferencial bi-armónico

$$\mathcal{L}u = \Delta^2 u \tag{140}$$

entonces se tiene que

$$w\mathcal{L}u = u\mathcal{L}^* w + \nabla \cdot \underline{\mathfrak{D}}\left(u, w\right) \tag{141}$$

desarrollemos el término del lado derecho

$$\begin{aligned} w\mathcal{L}u &= w\Delta^2 u \\ &= w\nabla \cdot \left(\nabla \Delta u\right) \end{aligned} \tag{142}$$

utilizando la igualdad de divergencia

$$\nabla \cdot \left(sV\right) = s\nabla \cdot V + V \cdot \nabla s \tag{143}$$

tal que s es función escalar y V vector, entonces sea $w = s$ y $\nabla \Delta u = V$, se tiene

$$\begin{aligned} & w\nabla \cdot \left(\nabla \Delta u\right) \\ = \;& \nabla \cdot \left(w \nabla \Delta u\right) - \nabla \Delta u \cdot \nabla w \end{aligned} \tag{144}$$

ahora sea $s = \Delta u$ y $V = \nabla w$, entonces tenemos

$$\begin{aligned} & \nabla \cdot \left(w \nabla \Delta u\right) - \nabla \Delta u \cdot \nabla w \\ = \;& \nabla \cdot \left(w \nabla \Delta u\right) + \Delta u \nabla \cdot \nabla w - \nabla \cdot \left(\Delta u \nabla w\right) \\ = \;& \Delta w \nabla \cdot \nabla u + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \end{aligned} \tag{145}$$

sea $s = \Delta w$ y $V = \nabla u$, entonces tenemos

$$\begin{aligned} & \Delta w \nabla \cdot \nabla u + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \\ = \;& \nabla \cdot \left(\Delta w \nabla u\right) - \nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \\ = \;& -\nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \end{aligned} \tag{146}$$

por último sea $s = u$ y $V = \nabla \left(\Delta w\right)$ y obtenemos

$$\begin{aligned} & -\nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \\ = \;& u \nabla \cdot \left(\nabla \left(\Delta w\right)\right) - \nabla \cdot \left(u \nabla \left(\Delta w\right)\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \end{aligned} \tag{147}$$

reordenando términos

$$w\mathcal{L}u = u\Delta^2 w + \nabla \cdot (w\nabla\Delta u + \Delta w\nabla u - \Delta u\nabla w - u\nabla\Delta w) \tag{148}$$

entonces se tiene que el operador adjunto formal es

$$\mathcal{L}^* w = \Delta^2 w \tag{149}$$

y los términos de frontera son

$$\underline{\mathfrak{D}}\,(u, w) = w\nabla\Delta u + \Delta w\nabla u - \Delta u\nabla w - u\nabla\Delta w. \tag{150}$$

3.4. Adjuntos Formales para Sistemas de Ecuaciones

En esta sección trabajaremos con funciones vectoriales; para ello necesitamos plantear la definición de operadores adjuntos formales para este tipo de funciones.

Definición 31 *Sea $\underline{\underline{\mathcal{L}}}$ un operador diferencial, decimos que un operador $\underline{\underline{\mathcal{L}}}^*$ es su adjunto formal si satisface la siguiente condición*

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} - \underline{u}\ \underline{\underline{\mathcal{L}}}^*\underline{w} = \nabla \cdot \underline{\mathfrak{D}}\,(\underline{u}, \underline{w}) \tag{151}$$

tal que las funciones $\underline{u}$ y $\underline{w}$ pertenecen a un espacio lineal. Aquí $\underline{\mathfrak{D}}(\underline{u}, \underline{w})$ representa términos de frontera.

Por lo tanto se puede trabajar con funciones vectoriales utilizando operadores matriciales.

A) **Operador diferencial vector-valuado con elasticidad estática**

$$\underline{\underline{\mathcal{L}}}\ \underline{u} = -\nabla \cdot \underline{\underline{\underline{C}}} : \nabla\underline{u} \tag{152}$$

de la definición de operador adjunto formal tenemos que

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} = \underline{u}\ \underline{\underline{\mathcal{L}}}^*\underline{w} + \nabla \cdot \underline{\mathfrak{D}}\,(\underline{u}, \underline{w}) \tag{153}$$

para hacer el desarrollo del término del lado derecho se utilizará notación indicial, es decir, este vector $\underline{w\underline{\underline{\mathcal{L}}}u}$ tiene los siguientes componentes

$$-w_i\left(\frac{\partial}{\partial x_j}\left(C_{ijpq}\frac{\partial u_p}{\partial x_q}\right)\right);\ i = 1, 2, 3 \tag{154}$$

utilizando la igualdad de divergencia tenemos

$$\begin{aligned} & -w_i\left(\frac{\partial}{\partial x_j}\left(C_{ijpq}\frac{\partial u_p}{\partial x_q}\right)\right) \qquad (155)\\ = \ & C_{ijpq}\frac{\partial u_p}{\partial x_q}\frac{\partial w_i}{\partial x_j} - \frac{\partial}{\partial x_j}\left(w_i C_{ijpq}\frac{\partial u_p}{\partial x_q}\right)\\ = \ & \frac{\partial}{\partial x_j}\left(u_i C_{ijpq}\frac{\partial w_i}{\partial x_j}\right) - u_i\frac{\partial}{\partial x_j}\left(C_{ijpq}\frac{\partial w_i}{\partial x_j}\right) - \frac{\partial}{\partial x_j}\left(w_i C_{ijpq}\frac{\partial u_p}{\partial x_q}\right) \end{aligned}$$

reordenado términos tenemos que la ecuación anterior es

$$\frac{\partial}{\partial x_j}\left(u_i C_{ijpq}\frac{\partial w_i}{\partial x_j} - w_i C_{ijpq}\frac{\partial u_p}{\partial x_q}\right) - u_i\frac{\partial}{\partial x_j}\left(C_{ijpq}\frac{\partial w_i}{\partial x_j}\right) \tag{156}$$

en notación simbólica tenemos que

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} = -\underline{u}\nabla\cdot\left(\underline{\underline{\underline{C}}} : \nabla\underline{w}\right) + \nabla\cdot\left(\underline{u}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{w} - \underline{w}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{u}\right) \tag{157}$$

por lo tanto el operador adjunto formal es

$$\underline{\underline{\underline{\mathcal{L}}}}^{*}\underline{w} = -\nabla\cdot\left(\underline{\underline{\underline{C}}} : \nabla\underline{w}\right) \tag{158}$$

y los términos de frontera son

$$\underline{\mathfrak{D}}\left(\underline{u},\underline{w}\right) = \underline{u}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{w} - \underline{w}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{u} \tag{159}$$

El operador de elasticidad es **auto-adjunto formal.**

B) **Métodos Mixtos a la Ecuación de Laplace**

Operador Laplaciano

$$\underline{\underline{\mathcal{L}}}\ \underline{u} = \Delta\underline{u} = \underline{f} \tag{160}$$

escrito en un sistema de ecuaciones se obtiene

$$\underline{\underline{\mathcal{L}}}\ \underline{u} = \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\begin{bmatrix} \underline{p} \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix} \tag{161}$$

consideraremos campos vectoriales de 4 dimensiones, estos son denotados por :

$$\underline{u} \equiv \left\{\underline{p}, u\right\}\ \text{ y }\ \underline{w} = \left\{\underline{q}, w\right\} \tag{162}$$

ahora el operador diferencial vector-valuado es el siguiente

$$\begin{aligned} \underline{\underline{\mathcal{L}}}\underline{u} &= \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\cdot\begin{bmatrix} \underline{p} \\ u \end{bmatrix} \\ &= \begin{bmatrix} \underline{p} - \nabla u \\ \nabla\cdot\underline{p} \end{bmatrix} \end{aligned} \tag{163}$$

entonces

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} = \begin{bmatrix} \underline{q} \\ w \end{bmatrix}\cdot\begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\cdot\begin{bmatrix} \underline{p} \\ u \end{bmatrix} \tag{164}$$

utilizando la definición de operador adjunto

$$\underline{w}\underline{\underline{\mathcal{L}}}\underline{u} = \underline{u}\underline{\underline{\mathcal{L}}}\underline{w} + \nabla\cdot\underline{\mathfrak{D}}\left(\underline{u},\underline{w}\right) \tag{165}$$

haciendo el desarrollo del término izquierdo se tiene que

$$\begin{aligned} \underline{w}\underline{\underline{\mathcal{L}}}\underline{u} &= \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \cdot \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \\ &= \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \cdot \begin{bmatrix} \underline{p} - \nabla u \\ \nabla \cdot \underline{p} \end{bmatrix} \\ &= \underline{qp} - \underline{q}\nabla \cdot u + w\nabla \cdot \underline{p} \end{aligned} \tag{166}$$

aquí se utiliza la igualdad de divergencia en los dos términos del lado derecho y obtenemos

$$\begin{aligned} & \underline{qp} - \underline{q}\nabla \cdot u + w\nabla \cdot \underline{p} \\ = \; & \underline{qp} + u\nabla \cdot \underline{q} - \nabla \cdot \left(\underline{q}u\right) - \underline{p} \cdot \nabla w + \nabla \cdot \left(w\underline{p}\right) \\ = \; & \underline{p}\left(\underline{q} - \nabla w\right) + u\nabla \cdot \underline{q} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \end{aligned} \tag{167}$$

si se agrupa los dos primeros términos en forma matricial, se tiene

$$\begin{aligned} & \underline{p}\left(\underline{q} - \nabla w\right) + u\nabla \cdot \underline{q} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \\ = \; & \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \begin{bmatrix} \underline{q} - \nabla w \\ w \end{bmatrix} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \\ = \; & \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \cdot \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{q} \\ w \end{bmatrix} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \end{aligned} \tag{168}$$

por lo tanto, el operador adjunto formal es

$$\begin{aligned} \underline{\underline{\mathcal{L}}}^{*}\underline{w} &= \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \\ &= \begin{bmatrix} \underline{q} - \nabla w \\ \nabla \cdot \underline{q} \end{bmatrix} \end{aligned} \tag{169}$$

y el término correspondiente a valores en la frontera es

$$\underline{\mathfrak{D}}\left(\underline{u}, \underline{w}\right) = w\underline{p} - u\underline{q}. \tag{170}$$

C) **Problema de Stokes**

El problema de Stokes es derivado de la ecuación de Navier-Stokes, la cual es utilizada en dinámica de fluidos viscosos. En este caso estamos suponiendo que el fluido es estacionario, la fuerza gravitacional es nula y el fluido incompresible**.** Entonces el sistema de ecuaciones a ser considerado es

$$\begin{aligned} -\Delta\underline{u} + \nabla p &= f \\ -\nabla \cdot \underline{u} &= 0 \end{aligned} \tag{171}$$

se considerará un campo vectorial de 4 dimensiones. Ellos serán denotados por

$$\underline{U} = \{\underline{u}, p\} \text{ y } \underline{W} = \{\underline{w}, q\} \tag{172}$$

ahora el operador diferencial vector-valuado es el siguiente

$$\underline{\underline{\mathcal{L}U}} = \begin{bmatrix} -\Delta & \nabla \\ -\nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{u} \\ p \end{bmatrix} \tag{173}$$

el desarrollo se hará en notación indicial, entonces tenemos que

$$\underline{W}\underline{\underline{\mathcal{L}U}} = \begin{cases} -\underline{w}\Delta\underline{u} + \underline{w}\nabla p \\ -q\nabla\cdot\underline{u} \end{cases} \tag{174}$$

usando notación indicial se obtiene

$$\begin{aligned} w_i\left(-\sum_j \frac{\partial^2 u_i}{\partial x_j^2} + \frac{\partial p}{\partial x_i}\right) &= -\sum_j w_i \frac{\partial^2 u_i}{\partial x_j^2} + w_i\frac{\partial p}{\partial x_i} \\ &= \sum_j \frac{\partial w_i}{\partial x_j}\frac{\partial u_i}{\partial x_j^2} - \sum_j \frac{\partial}{\partial x_j}\left(w_i\frac{\partial u_i}{\partial x_j}\right) - \\ &\quad p\frac{\partial w_i}{\partial x_i} + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{175}$$

desarrollando la primera suma como la derivada de dos funciones se tiene

$$\begin{aligned} &-\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2} + \sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j}\right) \\ &-\sum_j \frac{\partial}{\partial x_j}\left(w_i\frac{\partial u_i}{\partial x_j}\right) - p\frac{\partial w_i}{\partial x_i} + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{176}$$

reordenando términos tenemos

$$\begin{aligned} &-\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2} - p\frac{\partial w_i}{\partial x_i} + \\ &\sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j} - w_i\frac{\partial u_i}{\partial x_j}\right) + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{177}$$

Ahora consideremos la ecuación 2 en Ec. (174), tenemos

$$-q\nabla\cdot\underline{u} \tag{178}$$

en notación índicial se tiene

$$\begin{aligned} -q\sum_i \frac{\partial u_i}{\partial x_i} &= -\sum_i q\frac{\partial u_i}{\partial x_i} \\ &= \sum_i u_i\frac{\partial q}{\partial x_i} - \sum_i \frac{\partial}{\partial x_i}(q u_i) \end{aligned} \tag{179}$$

en la ecuación anterior se utilizó la igualdad de divergencia, entonces agrupando las ecuaciones Ec. (177) y Ec. (179) se tiene

$$\begin{aligned} & w_i \left(-\sum_j \frac{\partial^2 u_i}{\partial x_j^2} + \frac{\partial p}{\partial x_i} \right) - q \sum_i \frac{\partial u_i}{\partial x_i} \qquad (180) \\ = & -\sum_j u_i \frac{\partial^2 w_i}{\partial x_j^2} - p \frac{\partial w_i}{\partial x_i} + \\ & \sum_j \frac{\partial}{\partial x_j} \left(u_i \frac{\partial w_i}{\partial x_j} - w_i \frac{\partial u_i}{\partial x_j} \right) + \frac{\partial}{\partial x_i} (w_i p) + \\ & \sum_i u_i \frac{\partial q}{\partial x_i} - \sum_i \frac{\partial}{\partial x_i} (q u_i) \end{aligned}$$

ordenando los términos tenemos

$$\begin{aligned} & -\sum_j u_i \frac{\partial^2 w_i}{\partial x_j^2} + \sum_i u_i \frac{\partial q}{\partial x_i} - p \frac{\partial w_i}{\partial x_i} \qquad (181) \\ & + \sum_j \frac{\partial}{\partial x_j} \left(u_i \frac{\partial w_i}{\partial x_j} - w_i \frac{\partial u_i}{\partial x_j} \right) + \frac{\partial}{\partial x_i} (w_i p) - \sum_i \frac{\partial}{\partial x_i} (q u_i) \end{aligned}$$

escribiendo la ecuación anterior en notación simbólica, se obtiene

$$-\underline{u} \Delta \underline{w} + \underline{u} \nabla q - p \nabla \cdot \underline{w} + \nabla \cdot (\underline{u} \nabla \underline{w} - \underline{w} \nabla \underline{u} + \underline{w} p - \underline{u} q) \qquad (182)$$

por lo tanto, el operador adjunto formal es

$$\underline{\underline{\mathcal{L}}}^* \underline{W} = \begin{bmatrix} -\Delta & \nabla \\ -\nabla \cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{w} \\ q \end{bmatrix} \qquad (183)$$

y el término de valores de frontera es

$$\underline{\mathfrak{D}} (\underline{u}, \underline{w}) = \underline{u} \nabla \underline{w} - \underline{w} \nabla \underline{u} + \underline{w} p - \underline{u} q. \qquad (184)$$

3.5. Problemas Variacionales con Valor en la Frontera

Restringiéndonos ahora en problemas elípticos de orden 2 (problemas de orden mayor pueden ser tratados de forma similar), reescribiremos este en su forma variacional. La formulación variacional es más débil que la formulación convencional ya que esta demanda menor suavidad de la solución u, sin embargo cualquier problema variacional con valores en la frontera corresponde a un problema con valor en la frontera y viceversa.

Además, la formulación variacional facilita el tratamiento de los problemas al usar métodos numéricos de ecuaciones diferenciales parciales, en esta sección veremos algunos resultados clave como es la existencia y unicidad de la solución de este tipo de problemas, para mayores detalles, ver [3] y [13].

Si el operador $\mathcal{L}$ está definido por

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{185}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, el problema queda escrito como

$$\begin{aligned} -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega. \end{aligned} \tag{186}$$

Si multiplicamos a la ecuación $-\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu = f_\Omega$ por $v \in V = H_0^1(\Omega)$, obtenemos

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{187}$$

aplicando el teorema de Green (83) obtenemos la Ec. (98), que podemos re-escribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{188}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{189}$$

y la funcional lineal

$$l(v) = \langle f, v\rangle = \int_\Omega v f_\Omega d\underline{x} \tag{190}$$

podemos reescribir el problema dado por la Ec. (91) de orden 2, haciendo uso de la forma bilineal $a(\cdot,\cdot)$ y la funcional lineal $l(\cdot)$.

Entonces entenderemos en el presente contexto un problema variacional con valores de frontera (VBVP) por uno de la forma: hallar una función u que pertenezca a un espacio de Hilbert $V = H_0^1(\Omega)$ y que satisfaga la ecuación

$$a(u,v) = \langle f, v\rangle \tag{191}$$

para toda función $v \in V$ donde $a(\cdot,\cdot)$ es una forma bilineal y $l(\cdot)$ es una funcional lineal.

Definición 32 *Sea V un espacio de Hilbert y sea $\|\cdot\|_V$ la norma asociada a dicho espacio, decimos que una forma bilineal $a(\cdot,\cdot)$ es continua si existe una constante $M > 0$ tal que*

$$|a(u,v)| \leq M \|u\|_V \|v\|_V \quad \forall u, v \in V \tag{192}$$

y es V-elíptico si existe una constante $\alpha > 0$ tal que

$$a(v,v) \geq \alpha \|v\|_V^2 \quad \forall v \in V \tag{193}$$

donde $\|\cdot\|_V$ es la norma asociada al espacio V.

Esto significa que una forma $V-$ elíptico es una que siempre es no negativa y toma el valor de 0 sólo en el caso de que $v = 0$, i.e. es positiva definida.

Notemos que el problema (186) definido en $V = H_0^1(\Omega)$ reescrito como el problema (191) genera una forma bilineal V-elíptico cuyo producto interior sobre V es simétrico y positivo definido ya que

$$a(v,v) \geq \alpha \|v\|_V^2 > 0, \quad \forall v \in V, v \neq 0 \tag{194}$$

reescribiéndose el problema (191), en el cual debemos encontrar $u \in V$ tal que

$$a(u,v) = \langle f, v\rangle - a(u_0, v) \tag{195}$$

donde $u_0 = g$ en $\partial\Omega$, para toda $v \in V$.

Entonces, la cuestión fundamental, es conocer bajo que condiciones el problema anterior tiene solución y esta es única, el teorema de Lax-Milgram nos da las condiciones bajo las cuales el problema (186) reescrito como el problema (191) tiene solución y esta es única, esto queda plasmado en el siguiente resultado.

Teorema 33 *(Lax-Milgram)*
Sea V un espacio de Hilbert y sea $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ una forma bilineal continua V-elíptico sobre V. Además, sea $l(\cdot) : V \to \mathbb{R}$ una funcional lineal continua sobre V. Entonces
i) El VBVP de encontrar $u \in V$ que satisfaga

$$a(u,v) = \langle f, v\rangle, \forall v \in V \tag{196}$$

tiene una y sólo una solución;
ii) La solución depende continuamente de los datos, en el sentido de que

$$\|u\|_V \leq \frac{1}{\alpha}\|l\|_{V^*} \tag{197}$$

donde $\|\cdot\|_{V^*}$ es la norma en el espacio dual V^* de V y α es la constante de la definición de V-elíptico.

Más específicamente, considerando ahora V un subespacio cerrado de $H^m(\Omega)$ las condiciones para la existencia, unicidad y la dependencia continua de los datos queda de manifiesto en el siguiente resultado.

Teorema 34 *Sea V un subespacio cerrado de $H^m(\Omega)$, sea $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ una forma bilineal continua V-elíptico sobre V y sea $l(\cdot) : V \to \mathbb{R}$ una funcional lineal continua sobre V. Sea P un subespacio cerrado de V tal que*

$$a(u+p, v+\overline{p}) = a(u,v) \quad \forall u, v \in V \text{ y } p, \overline{p} \in P. \tag{198}$$

También denotando por Q el subespacio de V consistente de las funciones ortogonales a P en la norma L^2; tal que

$$Q = \left\{ v \in V \mid \int_\Omega upd\underline{x} = 0 \quad \forall p \in P \right\}, \tag{199}$$

y asumiendo que $a(\cdot,\cdot)$ *es* Q*-elíptico: existe una constante* $\alpha > 0$ *tal que*

$$a(q,q) \geq \alpha \|q\|_Q^2 \quad \textit{para } q \in Q, \tag{200}$$

la norma sobre Q *será la misma que sobre* V*. Entonces*

i) Existe una única solución al problema de encontrar $u \in Q$ *tal que*

$$a(u,v) = \langle l, v \rangle, \quad \forall v \in V \tag{201}$$

si y sólo si las condiciones de compatibilidad

$$\langle l, p \rangle = 0 \quad \textit{para } p \in P \tag{202}$$

se satisfacen.

ii) La solución u *satisface*

$$\|u\|_Q \leq \alpha^{-1} \|l\|_{Q^*} \tag{203}$$

(dependencia continua de los datos).

Otro aspecto importante es la regularidad de la solución, si la solución u al VBVP de orden $2m$ con $f \in H^{s-2m}(\Omega)$ donde $s \geq 2m$, entonces u pertenecerá a $H^s(\Omega)$ y esto queda de manifiesto en el siguiente resultado.

Teorema 35 *Sea* $\Omega \subset \mathbb{R}^n$ *un dominio suave y sea* $u \in V$ *la solución al VBVP*

$$a(u,v) = \langle f, v \rangle, \quad v \in V \tag{204}$$

donde $V \subset H^m(\Omega)$*. Si* $f \in H^{s-2m}(\Omega)$ *con* $s \geq 2m$*, entonces* $u \in H^s(\Omega)$ *y la estimación*

$$\|u\|_{H^s} \leq C \|f\|_{H^{s-2m}} \tag{205}$$

se satisface.

4. Solución de Grandes Sistemas de Ecuaciones

Es este trabajo se mostró como proceder para transformar un problema de ecuaciones diferenciales parciales con valores en la frontera en un sistema algebraico de ecuaciones y así poder hallar la solución resolviendo el sistema de ecuaciones lineales que se pueden expresar en la forma matricial siguiente

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{206}$$

donde la matriz $\underline{\underline{A}}$ es bandada (muchos elementos son nulos) y en problemas reales tiene grandes dimensiones.

Los métodos de resolución del sistema algebraico de ecuaciones $\underline{\underline{A}}\underline{u} = \underline{b}$ se clasifican en dos grandes grupos: los métodos directos y los métodos iterativos.

En los métodos directos la solución $\underline{u}$ se obtiene en un número fijo de pasos y sólo están sujetos a los errores de redondeo. En los métodos iterativos, se realizan iteraciones para aproximarse a la solución $\underline{u}$ aprovechando las características propias de la matriz $\underline{\underline{A}}$, tratando de usar un menor número de pasos que en un método directo.

Los métodos iterativos rara vez se usan para resolver sistemas lineales de dimensión pequeña (el concepto de dimensión pequeña es muy relativo), ya que el tiempo necesario para conseguir una exactitud satisfactoria rebasa el que requieren los métodos directos. Sin embargo, en el caso de sistemas grandes con un alto porcentaje de elementos cero, son eficientes tanto en el almacenamiento en la computadora como en el tiempo que se invierte en su solución. Por ésta razón al resolver éstos sistemas algebraicos de ecuaciones es preferible aplicar métodos iterativos tal como Gradiente Conjugado.

Cabe hacer mención de que la mayoría del tiempo de cómputo necesario para resolver el problema de ecuaciones diferenciales parciales (EDP), es consumido en la solución del sistema algebraico de ecuaciones asociado a la discretización, por ello es determinante elegir aquel método numérico que minimice el tiempo invertido en este proceso.

4.1. Métodos Directos

En estos métodos, la solución $\underline{u}$ se obtiene en un número fijo de pasos y sólo están sujetos a los errores de redondeo. Entre los métodos más importantes podemos encontrar: Eliminación Gausiana, descomposición LU, eliminación bandada y descomposición de Cholesky.

Los métodos antes mencionados, se colocaron en orden descendente en cuanto al consumo de recursos computacionales y ascendente en cuanto al aumento en su eficiencia.

Eliminación Gausiana Tal vez es el método más utilizado para encontrar la solución usando métodos directos. Este algoritmo sin embargo no es eficiente, ya que en general, un sistema de N ecuaciones requiere para su almacenaje en memoria de N^2 entradas para la matriz $\underline{\underline{A}}$, pero cerca de $N^3/3 + O(N^2)$

multiplicaciones y $N^3/3 + O(N^2)$ adiciones para encontrar la solución siendo muy costoso computacionalmente.

La eliminación Gausiana se basa en la aplicación de operaciones elementales a renglones o columnas de tal forma que es posible obtener matrices equivalentes.

Escribiendo el sistema de N ecuaciones lineales con N incógnitas como

$$\sum_{j=1}^{N} a_{ij}^{(0)} x_j = a_{i,n+1}^{(0)}, \quad i = 1, 2, ..., N \tag{207}$$

y si $a_{11}^{(0)} \neq 0$ y los pivotes $a_{ii}^{(i-1)}, i = 2, 3, ..., N$ de las demás filas, que se obtienen en el curso de los cálculos, son distintos de cero, entonces, el sistema lineal anterior se reduce a la forma triangular superior (eliminación hacia adelante)

$$x_i + \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j = a_{i,n+1}^{(i)}, \quad i = 1, 2, ..., N \tag{208}$$

donde

$$\begin{aligned} k &= 1, 2, ..., N; \{j = k+1, ..., N\{ \\ a_{kj}^{(k)} &= \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}}; \\ i &= k+1, ..., N+1\{ \\ a_{ij}^{(k)} &= a_{ij}^{(k-1)} - a_{kj}^{(k)} a_{ik}^{(k-1)} \}\}\} \end{aligned}$$

y las incógnitas se calculan por sustitución hacia atrás, usando las fórmulas

$$\begin{aligned} x_N &= a_{N,N+1}^{(N)}; \\ i &= N-1, N-2, ..., 1 \\ x_i &= a_{i,N+1}^{(i)} - \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j. \end{aligned} \tag{209}$$

En algunos casos nos interesa conocer $\underline{\underline{A}}^{-1}$, por ello si la eliminación se aplica a la matriz aumentada $\underline{\underline{A}} \mid \underline{\underline{I}}$ entonces la matriz $\underline{\underline{A}}$ de la matriz aumentada se convertirá en la matriz $\underline{\underline{I}}$ y la matriz $\underline{\underline{I}}$ de la matriz aumentada será $\underline{\underline{A}}^{-1}$. Así, el sistema $\underline{\underline{A}}\underline{u} = \underline{b}$ se transformará en $\underline{u} = \underline{\underline{A}}^{-1}\underline{b}$ obteniendo la solución de $\underline{u}$.

Descomposición LU Sea $\underline{\underline{U}}$ una matriz triangular superior obtenida de $\underline{\underline{A}}$ por eliminación bandada. Entonces $\underline{\underline{U}} = \underline{\underline{L}}^{-1}\underline{\underline{A}}$, donde $\underline{\underline{L}}$ es una matriz triangular inferior con unos en la diagonal. Las entradas de $\underline{\underline{L}}^{-1}$ pueden obtenerse de los coeficientes m_{ij} definidos en el método anterior y pueden ser almacenados estrictamente en las entradas de la diagonal inferior de $\underline{\underline{A}}$ ya que estas ya fueron eliminadas. Esto proporciona una factorización $\underline{\underline{LU}}$ de $\underline{\underline{A}}$ en la misma matriz $\underline{\underline{A}}$ ahorrando espacio de memoria.

El problema original $\underline{\underline{A}}\underline{u} = \underline{b}$ se escribe como $\underline{\underline{LU}}\underline{u} = \underline{b}$ y se reduce a la solución sucesiva de los sistemas lineales triangulares

$$\underline{\underline{L}}\underline{y} = \underline{b} \quad \text{y} \quad \underline{\underline{U}}\underline{u} = \underline{y}. \tag{210}$$

La descomposición $\underline{\underline{LU}}$ requiere también $N^3/3$ operaciones aritméticas para la matriz llena, pero sólo Nb^2 operaciones aritméticas para la matriz con un ancho de banda de b siendo esto más económico computacionalmente.

Nótese que para una matriz no singular $\underline{\underline{A}}$, la eliminación de Gausiana (sin redondear filas y columnas) es equivalente a la factorización LU.

Eliminación Bandada Cuando se usa la ordenación natural de los nodos, la matriz $\underline{\underline{A}}$ que se genera es bandada, por ello se puede ahorrar considerable espacio de almacenamiento en ella. Este algoritmo consiste en triangular a la matriz $\underline{\underline{A}}$ por eliminación hacia adelante operando sólo sobre las entradas dentro de la banda central no cero. Así el renglón j es multiplicado por $m_{ij} = a_{ij}/a_{jj}$ y el resultado es restado al renglón i para $i = j+1, j+2, ...$

El resultado es una matriz triangular superior $\underline{\underline{U}}$ que tiene ceros abajo de la diagonal en cada columna. Así, es posible resolver el sistema resultante al sustituir en forma inversa las incógnitas.

Descomposición de Cholesky Cuando la matriz es simétrica y definida positiva, se obtiene la descomposición $\underline{\underline{LU}}$ de la matriz $\underline{\underline{A}}$, así $\underline{\underline{A}} = \underline{\underline{LDU}} = \underline{\underline{LDL}}^T$ donde $\underline{\underline{D}} = diag(\underline{\underline{U}})$ es la diagonal con entradas positivas. La mayor ventaja de esta descomposición es que, en el caso en que es aplicable, el costo de cómputo es sustancialmente reducido, ya que requiere de $N^3/6$ multiplicaciones y $N^3/6$ adiciones.

4.2. Métodos Iterativos

En estos métodos se realizan iteraciones para aproximarse a la solución $\underline{u}$ aprovechando las características propias de la matriz $\underline{\underline{A}}$, tratando de usar un menor número de pasos que en un método directo, para más información de estos y otros métodos ver [16] y [25].

Un método iterativo en el cual se resuelve el sistema lineal

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{211}$$

comienza con una aproximación inicial $\underline{u}^0$ a la solución $\underline{u}$ y genera una sucesión de vectores $\left\{u^k\right\}_{k=1}^{\infty}$ que converge a $\underline{u}$. Los métodos iterativos traen consigo un proceso que convierte el sistema $\underline{\underline{A}}\underline{u} = \underline{b}$ en otro equivalente de la forma $\underline{u} = \underline{\underline{T}}\underline{u} + \underline{c}$ para alguna matriz fija $\underline{\underline{T}}$ y un vector $\underline{c}$. Luego de seleccionar el vector inicial $\underline{u}^0$ la sucesión de los vectores de la solución aproximada se genera calculando

$$\underline{u}^k = \underline{\underline{T}}\underline{u}^{k-1} + \underline{c} \quad \forall k = 1, 2, 3, ... \tag{212}$$

La convergencia a la solución la garantiza el siguiente teorema cuya solución puede verse en [26].

Teorema 36 *Si* $\left\|\underline{\underline{T}}\right\| < 1$*, entonces el sistema lineal* $\underline{u} = \underline{\underline{T}}\underline{u} + \underline{c}$ *tiene una solución única* $\underline{u}^*$ *y las iteraciones* $\underline{u}^k$ *definidas por la fórmula* $\underline{u}^k = \underline{\underline{T}}\underline{u}^{k-1} + \underline{c} \quad \forall k = 1, 2, 3, ...$ *convergen hacia la solución exacta* $\underline{u}^*$ *para cualquier aproximación lineal* $\underline{u}^0$.

Notemos que mientras menor sea la norma de la matriz $\underline{\underline{T}}$, más rápida es la convergencia, en el caso cuando $\left\|\underline{\underline{T}}\right\|$ es menor que uno, pero cercano a uno, la convergencia es muy lenta y el número de iteraciones necesario para disminuir el error depende significativamente del error inicial. En este caso, es deseable proponer al vector inicial $\underline{u}^0$ de forma tal que se mínimo el error inicial. Sin embargo, la elección de dicho vector no tiene importancia si la $\left\|\underline{\underline{T}}\right\|$ es pequeña ya que la convergencia es rápida.

Como es conocido, la velocidad de convergencia de los métodos iterativos dependen de las propiedades espectrales de la matriz de coeficientes del sistema de ecuaciones, cuando el operador diferencial $\mathcal{L}$ de la ecuación del problema a resolver es auto-adjunto se obtiene una matriz simétrica y positivo definida y el número de condicionamiento de la matriz $\underline{\underline{A}}$, es por definición

$$cond(\underline{\underline{A}}) = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} \geq 1 \tag{213}$$

donde $\lambda_{\text{máx}}$ y $\lambda_{\text{mín}}$ es el máximo y mínimo de los eigenvalores de la matriz $\underline{\underline{A}}$. Si el número de condicionamiento es cercano a 1 los métodos numéricos al solucionar el problema convergerá en pocas iteraciones, en caso contrario se requerirán muchas iteraciones. Frecuentemente al usar el método de elemento finito se tiene una velocidad de convergencia de $O\left(\frac{1}{h^2}\right)$ y en el caso de métodos de descomposición de dominio se tiene una velocidad de convergencia de $O\left(\frac{1}{h}\right)$ en el mejor de los casos, donde h es la máxima distancia de separación entre nodos continuos de la partición, es decir, que poseen una pobre velocidad de convergencia cuando $h \to 0$, para más detalles ver [2].

Entre los métodos más usados para el tipo de problemas tratados en el presente trabajo podemos encontrar: Jacobi, Gauss-Seidel, Richardson, relajación sucesiva, Gradiente Conjugado, Gradiente Conjugado precondicionado.

Los métodos antes mencionados se colocaron en orden descendente en cuanto al consumo de recursos computacionales y ascendente en cuanto al aumento en la eficiencia en su desempeño, describiéndose a continuación:

Jacobi Si todos los elementos de la diagonal principal de la matriz $\underline{\underline{A}}$ son diferentes de cero $a_{ii} \neq 0$ para $i = 1, 2, ...n$. Podemos dividir la $i-$ésima ecuación del sistema lineal (211) por a_{ii} para $i = 1, 2, ...n$, y después trasladamos todas las incógnitas, excepto x_i, a la derecha, se obtiene el sistema equivalente

$$\underline{u} = \underline{\underline{B}}\underline{u} + \underline{d} \tag{214}$$

donde

$$d_i = \frac{b_i}{a_{ii}} \quad \text{y} \quad B = \{b_{ij}\} = \begin{cases} -\frac{a_{ij}}{a_{ii}} & \text{si } j \neq i \\ 0 & \text{si } j = i \end{cases} .$$

Las iteraciones del método de Jacobi están definidas por la fórmula

$$x_i = \sum_{j=1}^{n} b_{ij} x_j^{(k-1)} + d_i \tag{215}$$

donde $x_i^{(0)}$ son arbitrarias $(i = 1, 2, ...n; k = 1, 2, ...)$.

También el método de Jacobi se puede expresar en términos de matrices. Supongamos por un momento que la matriz $\underline{\underline{A}}$ tiene la diagonal unitaria, esto es $diag(\underline{\underline{A}}) = \underline{\underline{I}}$. Si descomponemos $\underline{\underline{A}} = \underline{\underline{I}} - \underline{\underline{B}}$, entonces el sistema dado por la Ecs. (211) se puede reescribir como

$$\left(\underline{\underline{I}} - \underline{\underline{B}}\right) \underline{u} = \underline{b}. \tag{216}$$

Para la primera iteración asumimos que $\underline{k}$=$\underline{b}$; entonces la última ecuación se escribe como $\underline{u} = \underline{\underline{B}}\underline{u} + \underline{k}$. Tomando una aproximación inicial $\underline{u}^0$, podemos obtener una mejor aproximación remplazando $\underline{u}$ por la más resiente aproximación de $\underline{u}^m$. Esta es la idea que subyace en el método Jacobi. El proceso iterativo queda como

$$\underline{u}^{m+1} = \underline{\underline{B}}\underline{u}^m + \underline{k}. \tag{217}$$

La aplicación del método a la ecuación de la forma $\underline{\underline{A}}\underline{u} = \underline{b}$, con la matriz $\underline{\underline{A}}$ no cero en los elementos diagonales, se obtiene multiplicando la Ec. (211) por $D^{-1} = \left[diag(\underline{\underline{A}})\right]^{-1}$ obteniendo

$$\underline{\underline{B}} = \underline{\underline{I}} - \underline{\underline{D}}^{-1}\underline{\underline{A}}, \quad \underline{k} = \underline{\underline{D}}^{-1}\underline{b}. \tag{218}$$

Gauss-Seidel Este método es una modificación del método Jacobi, en el cual una vez obtenido algún valor de $\underline{u}^{m+1}$, este es usado para obtener el resto de los valores utilizando los valores más actualizados de $\underline{u}^{m+1}$. Así, la Ec. (217) puede ser escrita como

$$u_i^{m+1} = \sum_{j<i} b_{ij} u_j^{m+1} + \sum_{j>i} b_{ij} u_j^m + k_i. \tag{219}$$

Notemos que el método Gauss-Seidel requiere el mismo número de operaciones aritméticas por iteración que el método de Jacobi. Este método se escribe en forma matricial como

$$\underline{u}^{m+1} = \underline{\underline{E}}\underline{u}^{m+1} + \underline{\underline{F}}\underline{u}^m + \underline{k} \tag{220}$$

donde $\underline{\underline{E}}$ y $\underline{\underline{F}}$ son las matrices triangular superior e inferior respectivamente. Este método mejora la convergencia con respecto al método de Jacobi en un factor aproximado de 2.

Richardson Escribiendo el método de Jacobi como

$$\underline{u}^{m+1} - \underline{u}^{m} = \underline{b} - \underline{\underline{A}}\underline{u}^{m} \tag{221}$$

entonces el método Richardson se genera al incorporar la estrategia de sobrerrelajación de la forma siguiente

$$\underline{u}^{m+1} = \underline{u}^{m} + \omega \left(\underline{b} - \underline{\underline{A}}\underline{u}^{m} \right). \tag{222}$$

El método de Richardson se define como

$$\underline{u}^{m+1} = \left(\underline{\underline{I}} - \omega \underline{\underline{A}} \right) \underline{u}^{m} + \omega \underline{b} \tag{223}$$

en la práctica encontrar el valor de ω puede resultar muy costoso computacionalmente y las diversas estrategias para encontrar ω dependen de las características propias del problema, pero este método con un valor ω óptimo resulta mejor que el método de Gauss-Seidel.

Relajación Sucesiva Partiendo del método de Gauss-Seidel y sobrerrelajando este esquema, obtenemos

$$u_i^{m+1} = (1-\omega)\, u_i^{m} + \omega \left[\sum_{j=1}^{i-1} b_{ij} u_j^{m+1} + \sum_{j=i+1}^{N} b_{ij} u_j^{m} + k_i \right] \tag{224}$$

y cuando la matriz $\underline{\underline{A}}$ es simétrica con entradas en la diagonal positivas, éste método converge si y sólo si $\underline{\underline{A}}$ es definida positiva y $\omega \in (0, 2)$. En la práctica encontrar el valor de ω puede resultar muy costoso computacionalmente y las diversas estrategias para encontrar ω dependen de las características propias del problema.

4.3. Gradiente Conjugado

El método del Gradiente Conjugado ha recibido mucha atención en su uso al resolver ecuaciones diferenciales parciales y ha sido ampliamente utilizado en años recientes por la notoria eficiencia al reducir considerablemente en número de iteraciones necesarias para resolver el sistema algebraico de ecuaciones. Aunque los pioneros de este método fueron Hestenes y Stiefel (1952), el interés actual arranca a partir de que Reid (1971) lo planteara como un método iterativo, que es la forma en que se le usa con mayor frecuencia en la actualidad, esta versión está basada en el desarrollo hecho en [9].

La idea básica en que descansa el método del Gradiente Conjugado consiste en construir una base de vectores ortogonales y utilizarla para realizar la búsqueda de la solución en forma más eficiente. Tal forma de proceder generalmente no sería aconsejable porqué la construcción de una base ortogonal utilizando el procedimiento de Gramm-Schmidt requiere, al seleccionar cada nuevo elemento de la base, asegurar su ortogonalidad con respecto a cada uno de los vectores

construidos previamente. La gran ventaja del método de Gradiente Conjugado radica en que cuando se utiliza este procedimiento, basta con asegurar la ortogonalidad de un nuevo miembro con respecto al último que se ha construido, para que automáticamente esta condición se cumpla con respecto a todos los anteriores.

Definición 37 *Una matriz* $\underline{\underline{A}}$ *es llamada positiva definida si todos sus eigenvalores tienen parte real positiva o equivalentemente, si* $\underline{u}^T \underline{\underline{A}}\underline{u}$ *tiene parte real positiva para* $\underline{u} \in \mathbb{C}\backslash\{0\}$. *Notemos en este caso que*

$$\underline{u}^T \underline{\underline{A}}\underline{u} = \underline{u}^T \frac{\underline{\underline{A}} + \underline{\underline{A}}^T}{2} \underline{u} > 0, \ \ con\ \underline{u} \in \mathbb{R}^n \backslash \{0\}.$$

En el algoritmo de Gradiente Conjugado (CGM), se toma a la matriz $\underline{\underline{A}}$ como simétrica y positiva definida, y como datos de entrada del sistema

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{225}$$

el vector de búsqueda inicial $\underline{u}^0$ y se calcula $\underline{r}^0 = \underline{b} - \underline{\underline{A}}\underline{u}^0$, $\underline{p}^0 = \underline{r}^0$, quedando el método esquemáticamente como:

$$\begin{aligned} \beta^{k+1} &= \frac{\underline{\underline{A}}\underline{p}^k \cdot \underline{r}^k}{\underline{\underline{A}}\underline{p}^k \cdot \underline{p}^k} \\ \underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1}\underline{p}^k \\ \alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{\underline{A}}\underline{p}^{k+1} \cdot \underline{p}^{k+1}} \end{aligned} \tag{226}$$

$$\begin{aligned} \underline{u}^{k+1} &= \underline{u}^k + \alpha^{k+1}\underline{p}^{k+1} \\ \underline{r}^{k+1} &= \underline{r}^k - \alpha^{k+1}\underline{\underline{A}}\underline{p}^{k+1}. \end{aligned}$$

Si denotamos $\{\lambda_i, V_i\}_{i=1}^N$ como las eigensoluciones de $\underline{\underline{A}}$, i.e. $\underline{\underline{A}}V_i = \lambda_i V_i$, $i = 1, 2, ..., N$. Ya que la matriz $\underline{\underline{A}}$ es simétrica, los eigenvalores son reales y podemos ordenarlos por $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_N$. Definimos el número de condición por $Cond(\underline{\underline{A}}) = \lambda_N/\lambda_1$ y la norma de la energía asociada a $\underline{\underline{A}}$ por $\|\underline{u}\|_{\underline{\underline{A}}}^2 = \underline{u} \cdot \underline{\underline{A}}\underline{u}$ entonces

$$\left\|\underline{u} - \underline{u}^k\right\|_{\underline{\underline{A}}} \leq \left\|\underline{u} - \underline{u}^0\right\|_{\underline{\underline{A}}} \left[\frac{1 - \sqrt{Cond(\underline{\underline{A}})}}{1 + \sqrt{Cond(\underline{\underline{A}})}}\right]^{2k}. \tag{227}$$

El siguiente teorema nos da idea del espectro de convergencia del sistema $\underline{\underline{A}}\underline{u} = \underline{b}$ para el método de Gradiente Conjugado.

Teorema 38 *Sea* $\kappa = cond(\underline{\underline{A}}) = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} \geq 1$*, entonces el método de Gradiente Conjugado satisface la* $\underline{\underline{A}}-$*norma del error dado por*

$$\frac{\|e^n\|}{\|e^0\|} \leq \frac{2}{\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-n}\right]} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \tag{228}$$

donde $\underline{e}^m = \underline{u} - \underline{u}^m$ *del sistema* $\underline{\underline{A}}\underline{u} = \underline{b}$.

Notemos que para κ grande se tiene que

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \simeq 1 - \frac{2}{\sqrt{\kappa}} \tag{229}$$

tal que

$$\|\underline{e}^n\|_{\underline{\underline{A}}} \simeq \left\|\underline{e}^0\right\|_{\underline{\underline{A}}} \exp\left(-2\frac{n}{\sqrt{\kappa}}\right) \tag{230}$$

de lo anterior podemos esperar un espectro de convergencia del orden de $O(\sqrt{\kappa})$ iteraciones, para mayor referencia ver [26].

Definición 39 *Un método iterativo para la solución de un sistema lineal es llamado óptimo, si la razón de convergencia a la solución exacta es independiente del tamaño del sistema lineal.*

4.4. Precondicionadores

Una vía que permite mejorar la eficiencia de los métodos iterativos consiste en transformar al sistema de ecuaciones en otro equivalente, en el sentido de que posea la misma solución del sistema original pero que a su vez tenga mejores condiciones espectrales. Esta transformación se conoce como precondicionamiento y consiste en aplicar al sistema de ecuaciones una matriz conocida como precondicionador encargada de realizar el mejoramiento del número de condicionamiento.

Una amplia clase de precondicionadores han sido propuestos basados en las características algebraicas de la matriz del sistema de ecuaciones, mientras que por otro lado también existen precondicionadores desarrollados a partir de las características propias del problema que lo origina, un estudio más completo puede encontrarse en [2] y [18].

¿Qué es un Precondicionador? De una manera formal podemos decir que un precondicionador consiste en construir una matriz $\underline{\underline{C}}$, la cuál es una aproximación en algún sentido de la matriz $\underline{\underline{A}}$ del sistema $\underline{\underline{A}}\underline{u} = \underline{b}$, de manera tal que si multiplicamos ambos miembros del sistema de ecuaciones original por $\underline{\underline{C}}^{-1}$ obtenemos el siguiente sistema

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1}\underline{b} \tag{231}$$

donde el número de condicionamiento de la matriz del sistema transformado $\underline{\underline{C}}^{-1}\underline{\underline{A}}$ debe ser menor que el del sistema original, es decir

$$Cond(\underline{\underline{C}}^{-1}\underline{\underline{A}}) < Cond(\underline{\underline{A}}), \tag{232}$$

dicho de otra forma un precondicionador es una inversa aproximada de la matriz original

$$\underline{\underline{C}}^{-1} \simeq \underline{\underline{A}}^{-1} \tag{233}$$

que en el caso ideal $\underline{\underline{C}}^{-1} = \underline{\underline{A}}^{-1}$ el sistema convergería en una sola iteración, pero el coste computacional del cálculo de $\underline{\underline{A}}^{-1}$ equivaldría a resolver el sistema por un método directo. Se sugiere que $\underline{\underline{C}}$ sea una matriz lo más próxima a $\underline{\underline{A}}$ sin que su determinación suponga un coste computacional elevado.

Dependiendo de la forma de platear el producto de $\underline{\underline{C}}^{-1}$ por la matriz del sistema obtendremos distintas formas de precondicionamiento, estas son:

$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1}\underline{b}$	Precondicionamiento por la izquierda
$\underline{\underline{A}}\underline{\underline{C}}^{-1}\underline{\underline{C}}\underline{u} = \underline{b}$	Precondicionamiento por la derecha
$\underline{\underline{C}}_1^{-1}\underline{\underline{A}}\underline{\underline{C}}_2^{-1}\underline{\underline{C}}_2\underline{u} = \underline{\underline{C}}_1^{-1}\underline{b}$	Precondicionamiento por ambos lados si $\underline{\underline{C}}$ puede factorizarse como $\underline{\underline{C}} = \underline{\underline{C}}_1\underline{\underline{C}}_2$.

El uso de un precondicionador en un método iterativo provoca que se incurra en un costo de cómputo extra debido a que inicialmente se construye y luego se debe aplicar en cada iteración. Teniéndose que encontrar un balance entre el costo de construcción y aplicación del precondicionador versus la ganancia en velocidad en convergencia del método.

Ciertos precondicionadores necesitan poca o ninguna fase de construcción, mientras que otros pueden requerir de un trabajo substancial en esta etapa. Por otra parte la mayoría de los precondicionadores requieren en su aplicación un monto de trabajo proporcional al número de variables; esto implica que se multiplica el trabajo por iteración en un factor constante.

De manera resumida un buen precondicionador debe reunir las siguientes características:

i) Al aplicar un precondicionador $\underline{\underline{C}}$ al sistema original de ecuaciones $\underline{\underline{A}}\underline{u} = \underline{b}$, se debe reducir el número de iteraciones necesarias para que la solución aproximada tenga la convergencia a la solución exacta con una exactitud ε prefijada.

ii) La matriz $\underline{\underline{C}}$ debe ser fácil de calcular, es decir, el costo computacional de la construcción del precondicionador debe ser pequeño comparado con el costo total de resolver el sistema de ecuaciones $\underline{\underline{A}}\underline{u} = \underline{b}$.

iii) El sistema $\underline{\underline{C}}\underline{z} = \underline{r}$ debe ser fácil de resolver. Esto debe interpretarse de dos maneras:

a) El monto de operaciones por iteración debido a la aplicación del precondicionador $\underline{\underline{C}}$ debe ser pequeño o del mismo orden que las

que se requerirían sin precondicionamiento. Esto es importante si se trabaja en máquinas secuenciales.

b) El tiempo requerido por iteración debido a la aplicación del precondicionador debe ser pequeño.

En computadoras paralelas es importante que la aplicación del precondicionador sea paralelizable, lo cual eleva su eficiencia, pero debe de existir un balance entre la eficacia de un precondicionador en el sentido clásico y su eficiencia en paralelo ya que la mayoría de los precondicionadores tradicionales tienen un componente secuencial grande.

El método de Gradiente Conjugado por si mismo no permite el uso de precondicionadores, pero con una pequeña modificación en el producto interior usado en el método, da origen al método de Gradiente Conjugado precondicionado que a continuación detallaremos.

4.4.1. Gradiente Conjugado Precondicionado

Cuando la matriz $\underline{\underline{A}}$ es simétrica y definida positiva se puede escribir como

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{234}$$

y tomando la matriz $\underline{\underline{C}}^{-1}$ como un precondicionador de $\underline{\underline{A}}$ con la condición de que

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{C}}^{-1}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{235}$$

entonces la Ec. (225) se pude escribir como

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1}b \tag{236}$$

donde $\underline{\underline{C}}^{-1}\underline{\underline{A}}$ es también simétrica y definida positiva en el producto interior $\langle \underline{u}, \underline{v} \rangle = \underline{u} \cdot \underline{\underline{C}}\underline{v}$, porque

$$\begin{aligned} \left\langle \underline{u}, \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right\rangle &= \underline{u} \cdot \underline{\underline{C}} \left(\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right) \\ &= \underline{u} \cdot \underline{\underline{A}}\underline{v} \end{aligned} \tag{237}$$

que por hipótesis es simétrica y definida positiva en ese producto interior.

La elección del producto interior $\langle \cdot, \cdot \rangle$ quedará definido como

$$\langle \underline{u}, \underline{v} \rangle = \underline{u} \cdot \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \tag{238}$$

por ello las Ecs. (226[1]) y (226[3]), se convierten en

$$\alpha^{k+1} = \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \tag{239}$$

y

$$\beta^{k+1} = \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \tag{240}$$

generando el método de Gradiente Conjugado precondicionado con precondicionador $\underline{\underline{C}}^{-1}$. Es necesario hacer notar que los métodos Gradiente Conjugado y Gradiente Conjugado Precondicionado sólo difieren en la elección del producto interior.

Para el método de Gradiente Conjugado Precondicionado, los datos de entrada son un vector de búsqueda inicial $\underline{u}^0$ y el precondicionador $\underline{\underline{C}}^{-1}$. Calculándose $\underline{r}^0 = \underline{b} - \underline{\underline{A}}\underline{u}^0$, $\underline{p} = \underline{\underline{C}}^{-1}\underline{r}^0$, quedando el método esquemáticamente como:

$$\begin{aligned}
\beta^{k+1} &= \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \\
\underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1}\underline{p}^k \\
\alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \\
\underline{u}^{k+1} &= \underline{u}^k + \alpha^{k+1}\underline{p}^{k+1} \\
\underline{r}^{k+1} &= \underline{\underline{C}}^{-1}\underline{r}^k - \alpha^{k+1}\underline{\underline{A}}\underline{p}^{k+1}.
\end{aligned} \tag{241}$$

Algoritmo Computacional del Método Dado el sistema $\underline{\underline{A}}\underline{u} = \underline{b}$, con la matriz $\underline{\underline{A}}$ simétrica y definida positiva de dimensión $n \times n$. La entrada al método será una elección de $\underline{u}^0$ como condición inicial, $\varepsilon > 0$ como la tolerancia del método, N como el número máximo de iteraciones y la matriz de precondicionamiento $\underline{\underline{C}}^{-1}$ de dimensión $n \times n$, el algoritmo del método de Gradiente Conjugado Precondicionado queda como:

$\underline{r} = \underline{b} - \underline{\underline{A}}\underline{u}$
$\underline{w} = \underline{\underline{C}}^{-1}\underline{r}$
$\underline{v} = (\underline{\underline{C}}^{-1})^T\underline{w}$
$\alpha = \sum_{j=1}^{n} w_j^2$
$k = 1$
Mientras que $k \leq N$

Si $\|\underline{v}\|_\infty < \varepsilon$ Salir
$\underline{x} = \underline{\underline{A}}\underline{v}$
$t = \dfrac{\alpha}{\sum_{j=1}^{n} v_j x_j}$
$\underline{u} = \underline{u} + t\underline{v}$
$\underline{r} = \underline{r} - t\underline{x}$
$\underline{w} = \underline{\underline{C}}^{-1}\underline{r}$
$\beta = \sum_{j=1}^{n} w_j^2$

Si $\|\underline{r}\|_\infty < \varepsilon$ Salir

$s = \dfrac{\beta}{\alpha}$

$\underline{v} = \left(\underline{\underline{C}}^{-1}\right)^T \underline{w} + s\underline{v}$

$\alpha = \beta$

$k = k + 1$

La salida del método será la solución aproximada $\underline{u} = (u_1, ..., u_n)$ y el residual $\underline{r} = (r_1, ..., r_n)$.

En el caso del método sin precondicionamiento, $\underline{\underline{C}}^{-1}$ es la matriz identidad, que para propósitos de optimización sólo es necesario hacer la asignación de vectores correspondiente en lugar del producto de la matriz por el vector. En el caso de que la matriz $\underline{\underline{A}}$ no sea simétrica, el método de Gradiente Conjugado puede extenderse para soportarlas, para más información sobre pruebas de convergencia, resultados numéricos entre los distintos métodos de solución del sistema algebraico $\underline{\underline{A}}\underline{u} = \underline{b}$ generada por la discretización de un problema elíptico y como extender estos para matrices no simétricas ver [9] y [7].

Teorema 40 *Sean $\underline{\underline{A}}, \underline{\underline{B}}$ y $\underline{\underline{C}}$ tres matrices simétricas y positivas definidas entonces*

$$\kappa\left(\underline{\underline{C}}^{-1}\underline{\underline{A}}\right) \leq \kappa\left(\underline{\underline{C}}^{-1}\underline{\underline{B}}\right)\kappa\left(\underline{\underline{B}}^{-1}\underline{\underline{A}}\right).$$

Clasificación de los Precondicionadores En general se pueden clasificar en dos grandes grupos según su manera de construcción: los algebraicos o a posteriori y los a priori o directamente relacionados con el problema continuo que lo origina.

4.4.2. Precondicionador a Posteriori

Los precondicionadores algebraicos o a posteriori son los más generales, ya que sólo dependen de la estructura algebraica de la matriz $\underline{\underline{A}}$, esto quiere decir que no tienen en cuenta los detalles del proceso usado para construir el sistema de ecuaciones lineales $\underline{\underline{A}}\underline{u} = \underline{b}$. Entre estos podemos citar los métodos de precondicionamiento del tipo Jacobi, SSOR, factorización incompleta, inversa aproximada, diagonal óptimo y polinomial.

Precondicionador Jacobi El método precondicionador Jacobi es el precondicionador más simple que existe y consiste en tomar en calidad de precondicionador a los elementos de la diagonal de $\underline{\underline{A}}$

$$C_{ij} = \begin{cases} A_{ij} & si \quad i = j \\ 0 & si \quad i \neq j. \end{cases} \tag{242}$$

Debido a que las operaciones de división son usualmente más costosas en tiempo de cómputo, en la práctica se almacenan los recíprocos de la diagonal de $\underline{\underline{A}}$.

Ventajas: No necesita trabajo para su construcción y puede mejorar la convergencia.

Desventajas: En problemas con número de condicionamiento muy grande, no es notoria la mejoría en el número de iteraciones.

Precondicionador SSOR Si la matriz original es simétrica, se puede descomponer como en el método de sobrerrelajamiento sucesivo simétrico (SSOR) de la siguiente manera

$$\underline{\underline{A}} = \underline{\underline{D}} + \underline{\underline{L}} + \underline{\underline{L}}^T \tag{243}$$

donde $\underline{\underline{D}}$ es la matriz de la diagonal principal y $\underline{\underline{L}}$ es la matriz triangular inferior.

La matriz en el método SSOR se define como

$$\underline{\underline{C}}(\omega) = \frac{1}{2-w}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)\left(\frac{1}{\omega}\underline{\underline{D}}\right)^{-1}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)^T \tag{244}$$

en la práctica la información espectral necesaria para hallar el valor óptimo de ω es demasiado costoso para ser calculado.

Ventajas: No necesita trabajo para su construcción, puede mejorar la convergencia significativamente.

Desventajas: Su paralelización depende fuertemente del ordenamiento de las variables.

Precondicionador de Factorización Incompleta Existen una amplia clase de precondicionadores basados en factorizaciones incompletas. La idea consiste en que durante el proceso de factorización se ignoran ciertos elementos diferentes de cero correspondientes a posiciones de la matriz original que son nulos. La matriz precondicionadora se expresa como $\underline{\underline{C}} = \underline{\underline{LU}}$, donde $\underline{\underline{L}}$ es la matriz triangular inferior y $\underline{\underline{U}}$ la superior. La eficacia del método depende de cuán buena sea la aproximación de $\underline{\underline{C}}^{-1}$ con respecto a $\underline{\underline{A}}^{-1}$.

El tipo más común de factorización incompleta se basa en seleccionar un subconjunto S de las posiciones de los elementos de la matriz y durante el proceso de factorización considerar a cualquier posición fuera de éste igual a cero. Usualmente se toma como S al conjunto de todas las posiciones (i,j) para las que $A_{ij} \neq 0$. Este tipo de factorización es conocido como factorización incompleta LU de nivel cero, ILU(0).

El proceso de factorización incompleta puede ser descrito formalmente como sigue:

Para cada k, si $i, j > k$:

$$S_{ij} = \begin{cases} A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} & \text{Si} \quad (i,j) \in S \\ \\ A_{ij} & \text{Si} \quad (i,j) \notin S. \end{cases} \tag{245}$$

Una variante de la idea básica de las factorizaciones incompletas lo constituye la factorización incompleta modificada que consiste en que si el producto

$$A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} \neq 0 \tag{246}$$

y el llenado no está permitido en la posición (i, j), en lugar de simplemente descartarlo, esta cantidad se le substrae al elemento de la diagonal A_{ij}. Matemáticamente esto corresponde a forzar a la matriz precondicionadora a tener la misma suma por filas que la matriz original. Esta variante resulta de interés puesto que se ha probado que para ciertos casos la aplicación de la factorización incompleta modificada combinada con pequeñas perturbaciones hace que el número de condicionamiento espectral del sistema precondicionado sea de un orden inferior.

Ventaja: Puede mejorar el condicionamiento y la convergencia significativamente.

Desventaja: El proceso de factorización es costoso y difícil de paralelizar en general.

Precondicionador de Inversa Aproximada El uso del precondicionador de inversas aproximada se ha convertido en una buena alternativa para los precondicionadores implícitos debido a su naturaleza paralelizable. Aquí se construye una matriz inversa aproximada usando el producto escalar de Frobenius.

Sea $\mathcal{S} \subset C_n$, el subespacio de las matrices $\underline{\underline{C}}$ donde se busca una inversa aproximada explícita con un patrón de dispersión desconocido. La formulación del problema esta dada como: Encontrar $\underline{\underline{C}}_0 \in \mathcal{S}$ tal que

$$\underline{\underline{C}}_0 = \arg mín_{\underline{\underline{C}} \in \mathcal{S}} \left\| \underline{\underline{AC}} - \underline{\underline{I}} \right\|. \tag{247}$$

Además, esta matriz inicial $\underline{\underline{C}}_0$ puede ser una inversa aproximada de $\underline{\underline{A}}$ en un sentido estricto, es decir,

$$\left\| \underline{\underline{AC}}_0 - \underline{\underline{I}} \right\| = \varepsilon < 1. \tag{248}$$

Existen dos razones para esto, primero, la ecuación (248) permite asegurar que $\underline{\underline{C}}_0$ no es singular (lema de Banach), y segundo, esta será la base para construir un algoritmo explícito para mejorar $\underline{\underline{C}}_0$ y resolver la ecuación $\underline{\underline{A}}\underline{u} = \underline{b}$.

La construcción de $\underline{\underline{C}}_0$ se realiza en paralelo, independizando el cálculo de cada columna. El algoritmo permite comenzar desde cualquier entrada de la columna k, se acepta comúnmente el uso de la diagonal como primera aproximación. Sea r_k el residuo correspondiente a la columna k-ésima, es decir

$$r_k = \underline{\underline{AC}}_k - \underline{e}_k \tag{249}$$

y sea $\mathcal{I}_k$ el conjunto de índices de las entradas no nulas en r_k, es decir, $\mathcal{I}_k = \{i = \{1, 2, ..., n\} \mid r_{ik} \neq 0\}$. Si $\mathcal{L}_k = \{l = \{1, 2, ..., n\} \mid C_{lk} \neq 0\}$, entonces la nueva entrada se busca en el conjunto $\mathcal{J}_k = \{j \in \mathcal{L}_k^c \mid A_{ij} \neq 0, \forall i \in \mathcal{I}_k\}$. En realidad las únicas entradas consideradas en $\underline{C}_k$ son aquellas que afectan las entradas no

nulas de r_k. En lo que sigue, asumimos que $\mathcal{L}_k \cup \{j\} = \left\{i_1^k, i_2^k, ..., i_{p_k}^k\right\}$ es no vacío, siendo p_k el número actual de entradas no nulas de $\underline{C}_k$ y que $i_{p_k}^k = j$, para todo $j \in \mathcal{J}_k$. Para cada j, calculamos

$$\left\|\underline{\underline{A}}\underline{C}_k - \underline{e}_k\right\|_2^2 = 1 - \sum_{l=1}^{p_k} \frac{\left[\det\left(\underline{\underline{D}}_l^k\right)\right]^2}{\det\left(\underline{\underline{G}}_{l-2}^k\right)\det\left(\underline{\underline{G}}_l^k\right)} \tag{250}$$

donde, para todo k, $\det\left(\underline{\underline{G}}_0^k\right) = 1$ y $\underline{\underline{G}}_l^k$ es la matriz de Gram de las columnas $i_1^k, i_2^k, ..., i_{p_k}^k$ de la matriz $\underline{\underline{A}}$ con respecto al producto escalar Euclideano; $\underline{\underline{D}}_l^k$ es la matriz que resulta de remplazar la última fila de la matriz $\underline{\underline{G}}_l^k$ por $a_{ki_1^k}, a_{ki_2^k}, , ..., a_{ki_l^k}$, con $1 \leq l \leq p_k$. Se selecciona el índice j_k que minimiza el valor de $\left\|\underline{\underline{A}}\underline{C}_k - \underline{e}_k\right\|_2$.

Esta estrategia define el nuevo índice seleccionado j_k atendiendo solamente al conjunto $\mathcal{L}_k$, lo que nos lleva a un nuevo óptimo donde se actualizan todas las entradas correspondientes a los índices de $\mathcal{L}_k$. Esto mejora el criterio de (247) donde el nuevo índice se selecciona manteniendo las entradas correspondientes a los índices de $\mathcal{L}_k$. Así $\underline{C}_k$ se busca en el conjunto

$$\mathcal{S}_k = \left\{\underline{C}_k \in \mathbb{R}^n \mid C_{ik} = 0, \forall i \in \mathcal{L}_k \cup \{j_k\}\right\},$$

$$\underline{m}_k = \sum_{l=1}^{p_k} \frac{\det\left(\underline{\underline{D}}_l^k\right)}{\det\left(\underline{\underline{G}}_{l-2}^k\right)\det\left(\underline{\underline{G}}_l^k\right)} \tilde{\underline{m}}_l \tag{251}$$

donde $\tilde{\underline{C}}_l$ es el vector con entradas no nulas i_h^k $(1 \leq h \leq l)$. Cada una de ellas se obtiene evaluado el determinante correspondiente que resulta de remplazar la última fila del $\det\left(\underline{\underline{G}}_l^k\right)$ por e_h^t, con $1 \leq l \leq p_k$.

Evidentemente, los cálculos de $\left\|\underline{\underline{A}}\underline{C}_k - \underline{e}_k\right\|_2^2$ y de $\underline{C}_k$ pueden actualizarse añadiendo la contribución de la última entrada $j \in \mathcal{J}_k$ a la suma previa de 1 a $p_k - 1$. En la práctica, $\det\left(\underline{\underline{G}}_l^k\right)$ se calcula usando la descomposición de Cholesky puesto que $\underline{\underline{G}}_l^k$ es una matriz simétrica y definida positiva. Esto sólo involucra la factorización de la última fila y columna si aprovechamos la descomposición de $\underline{\underline{G}}_{l-1}^k$. Por otra parte, $\det\left(\underline{\underline{D}}_l^k\right) / \det\left(\underline{\underline{G}}_l^k\right)$ es el valor de la última incógnita del sistema $\underline{\underline{G}}_l^k \underline{d}_l = \left(a_{ki_1^k}, a_{ki_2^k}, , ..., a_{ki_l^k}\right)^T$ necesitándose solamente una sustitución por descenso. Finalmente, para obtener $\tilde{\underline{C}}_l$ debe resolverse el sistema $\underline{\underline{G}}_l^k \underline{v}_l = \underline{e}_l$, con $\tilde{\underline{C}}_{i_1^k l} = v_{hl}, (1 \leq h \leq l)$.

Ventaja: Puede mejorar el condicionamiento y la convergencia significativamente y es fácilmente paralelizable.

Desventaja: El proceso construcción es algo laborioso.

4.4.3. Precondicionador a Priori

Los precondicionadores a priori son más particulares y dependen para su construcción del conocimiento del proceso de discretización de la ecuación diferencial parcial, dicho de otro modo dependen más del proceso de construcción de la matriz $\underline{\underline{A}}$ que de la estructura de la misma.

Estos precondicionadores usualmente requieren de más trabajo que los del tipo algebraico discutidos anteriormente, sin embargo permiten el desarrollo de métodos de solución especializados más rápidos que los primeros.

Veremos algunos de los métodos más usados relacionados con la solución de ecuaciones diferenciales parciales en general y luego nos concentraremos en el caso de los métodos relacionados directamente con descomposición de dominio.

En estos casos el precondicionador $\underline{\underline{C}}$ no necesariamente toma la forma simple de una matriz, sino que debe ser visto como un operador en general. De aquí que $\underline{\underline{C}}$ podría representar al operador correspondiente a una versión simplificada del problema con valores en la frontera que deseamos resolver.

Por ejemplo se podría emplear en calidad de precondicionador al operador original del problema con coeficientes variables tomado con coeficientes constantes. En el caso del operador de Laplace se podría tomar como precondicionador a su discretización en diferencias finitas centrales.

Por lo general estos métodos alcanzan una mayor eficiencia y una convergencia óptima, es decir, para ese problema en particular el precondicionador encontrado será el mejor precondicionador existente, llegando a disminuir el número de iteraciones hasta en un orden de magnitud. Donde muchos de ellos pueden ser paralelizados de forma efectiva.

El Uso de la Parte Simétrica como Precondicionador La aplicación del método del Gradiente Conjugado en sistemas no auto-adjuntos requiere del almacenamiento de los vectores previamente calculados. Si se usa como precondicionador la parte simétrica

$$(\underline{\underline{A}} + \underline{\underline{A}}^T)/2 \tag{252}$$

de la matriz de coeficientes $\underline{\underline{A}}$, entonces no se requiere de éste almacenamiento extra en algunos casos, resolver el sistema de la parte simétrica de la matriz $\underline{\underline{A}}$ puede resultar más complicado que resolver el sistema completo.

El Uso de Métodos Directos Rápidos como Precondicionadores En muchas aplicaciones la matriz de coeficientes $\underline{\underline{A}}$ es simétrica y positivo definida, debido a que proviene de un operador diferencial auto-adjunto y acotado. Esto implica que se cumple la siguiente relación para cualquier matriz $\underline{\underline{B}}$ obtenida de una ecuación diferencial similar

$$c_1 \leq \frac{\underline{x}^T \underline{\underline{A}} \underline{x}}{\underline{x}^T \underline{\underline{B}} \underline{x}} \leq c_2 \quad \forall \underline{x} \tag{253}$$

donde c_1 y c_2 no dependen del tamaño de la matriz. La importancia de esta propiedad es que del uso de $\underline{\underline{B}}$ como precondicionador resulta un método iterativo cuyo número de iteraciones no depende del tamaño de la matriz.

La elección más común para construir el precondicionador $\underline{\underline{B}}$ es a partir de la ecuación diferencial parcial separable. El sistema resultante con la matriz $\underline{\underline{B}}$ puede ser resuelto usando uno de los métodos directos de solución rápida, como pueden ser por ejemplo los basados en la transformada rápida de Fourier.

Como una ilustración simple del presente caso obtenemos que cualquier operador elíptico puede ser precondicionado con el operador de Poisson.

Construcción de Precondicionadores para Problemas Elípticos Empleando DDM Existen una amplia gama de este tipo de precondicionadores, pero son específicos al método de descomposición de dominio usado, para el método de subestructuración, los más importantes se derivan de la matriz de rigidez y por el método de proyecciones, el primero se detalla en la sección (**??**) y el segundo, conjuntamente con otros precondicionadores pueden ser consultados en [12], [5], [4] y [2].

Definición 41 *Un método para la solución del sistema lineal generado por métodos de descomposición de dominio es llamado escalable, si la razón de convergencia no se deteriora cuando el número de subdominios crece.*

La gran ventaja de este tipo de precondicionadores es que pueden ser óptimos y escalables.

5. Métodos de Solución Aproximada para EDP

En el presente capítulo se prestará atención a varios aspectos necesarios para encontrar la solución aproximada de problemas variacionales con valor en la frontera (VBVP). Ya que en general encontrar la solución a problemas con geometría diversa es difícil y en algunos casos imposible usando métodos analíticos.

En este capítulo se considera el VBVP de la forma

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{254}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{255}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, como un caso particular del operador elíptico definido por la Ec. (91) de orden 2, con $\Omega \subset R^2$ un dominio poligonal, es decir, Ω es un conjunto abierto acotado y conexo tal que su frontera $\partial\Omega$ es la unión de un número finito de polígonos.

La sencillez del operador $\mathcal{L}$ nos permite facilitar la comprensión de muchas de las ideas básicas que se expondrán a continuación, pero tengamos en mente que esta es una ecuación que gobierna los modelos de muchos sistemas de la ciencia y la ingeniería, por ello es muy importante su solución.

Si multiplicamos a la ecuación $-\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu = f_\Omega$ por $v \in V = H_0^1(\Omega)$, obtenemos

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{256}$$

aplicando el teorema de Green (83) obtenemos la Ec. (98), que podemos reescribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{257}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{258}$$

y la funcional lineal

$$l(v) = \langle f, v\rangle = \int_\Omega v f_\Omega d\underline{x} \tag{259}$$

podemos reescribir el problema dado por la Ec. (254) de orden 2 en forma variacional, haciendo uso de la forma bilineal $a(\cdot,\cdot)$ y la funcional lineal $l(\cdot)$.

5.1. Método Galerkin

La idea básica detrás del método Galerkin es, considerando el VBVP, encontrar $u \in V = H_0^1(\Omega)$ que satisfaga

$$a(u,v) = \langle f, v\rangle \quad \forall v \in V \tag{260}$$

donde V es un subespacio de un espacio de Hilbert H (por conveniencia nos restringiremos a espacios definidos sobre los números reales).

El problema al tratar de resolver la Ec. (260) está en el hecho de que el espacio V es de dimensión infinita, por lo que resulta que en general no es posible encontrar el conjunto solución. En lugar de tener el problema en el espacio V, se supone que se tienen funciones linealmente independientes $\phi_1, \phi_2, ..., \phi_N$ en V y definimos el espacio V^h a partir del subespacio dimensionalmente finito de V generado por las funciones ϕ_i, es decir,

$$V^h = Generado\left\{\phi_i\right\}_{i=1}^{N}, \quad V^h \subset V. \tag{261}$$

El índice $h = 1/N$ es un parámetro que estará entre 0 y 1, cuya magnitud da alguna indicación de cuan cerca V^h esta de V, h se relaciona con la dimensión de V^h. Y como el número N de las funciones base se escoge de manera que sea grande y haga que h sea pequeño, en el límite, cuando $N \to \infty$, $h \to 0$.

Después de definir el espacio V^h, es posible trabajar con V^h en lugar de V y encontrar una función u_h que satisfaga

$$a\left(u_h, v_h\right) = \langle f, v_h \rangle \quad \forall v_h \in V^h. \tag{262}$$

Esta es la esencia del método Galerkin, notemos que u_h y v_h son sólo combinaciones lineales de las funciones base de V^h, tales que

$$u_h = \sum_{i=1}^{N} c_i \phi_i \quad \text{y} \quad v_h = \sum_{j=1}^{N} d_j \phi_j \tag{263}$$

donde v_h es arbitraria, como los coeficientes de d_j y sin perdida de generalidad podemos hacer $v_h = \phi_j$. Así, para encontrar la solución u_h sustituimos las Ecs. (263) en la Ec. (262) y usando el hecho que $a\left(\cdot,\cdot\right)$ es una forma bilineal y $l\left(\cdot\right)$ es una funcional lineal se obtiene la ecuación

$$\sum_{i=1}^{N} a\left(\phi_i, \phi_j\right) c_i = \left\langle f, \phi_j \right\rangle \tag{264}$$

o más concisamente, como

$$\sum_{i=1}^{N} K_{ij} c_i - F_j = 0 \quad j = 1, 2, ...N \tag{265}$$

en la cual

$$K_{ij} = a\left(\phi_i, \phi_j\right) \quad \text{y} \quad F_j = \left\langle f, \phi_j \right\rangle \tag{266}$$

notemos que tanto K_{ij} y F_j pueden ser evaluados, ya que ϕ_i, $a\left(\cdot,\cdot\right)$ y $l\left(\cdot\right)$ son conocidas.

Entonces el problema se reduce a resolver el sistema de ecuaciones lineales

$$\sum_{i=1}^{N} K_{ij} c_i - F_j, \quad j = 1, 2, ...N \tag{267}$$

o más compactamente

$$\underline{\underline{\mathbb{K}}}\underline{u} = \underline{F} \tag{268}$$

en la cual $\underline{\underline{\mathbb{K}}}$ y $\underline{F}$ son la matriz y el vector cuyas entradas son K_{ij} y F_j. Una vez que el sistema es resuelto, la solución aproximada u_h es encontrada.

Notemos que la forma bilineal $a(\cdot,\cdot)$ define un producto interior sobre V, si $a(\cdot,\cdot)$ es simetrica y $V-$elíptica, entonces las propiedades de linealidad y simetria son obvias, mientras que la propiedad de $V-$elíticidad de $a(\cdot,\cdot)$ es por

$$a(v,v) \geq \alpha \|v\|^2 > 0 \qquad \forall v \neq 0, \tag{269}$$

además, si $a(\cdot,\cdot)$ es continua, entonces la norma $\|v\|_a \equiv a(v,v)$ generada por este producto interior es equivalente a la norma estandar sobre V, tal que si V es completa con respecto a la norma estandar, esta también es completa con respecto a la norma $\|v\|_a$.

Por otro lado, si el conjunto de funciones base $\{\phi_i\}_{i=1}^N$ se eligen de tal forma que sean sean ortogonales entre si, entonces el sistema (265) se simplifica considerablemente, ya que

$$K_{ij} = a\left(\phi_i, \phi_j\right) = 0 \qquad \text{si } i \neq j \tag{270}$$

y

$$K_{ii}c_i = F_i \qquad \acute{o} \qquad c_i = F_i/K_{ii}. \tag{271}$$

Así, el problema (254) definido en $V^h = H_0^1(\Omega)$ reescrito como el problema (260) genera una forma bilineal V^h-elíptica cuyo producto interior sobre V^h es simétrico y positivo definido ya que

$$a(v_h, v_h) \geq \alpha \|v_h\|_{V^h}^2 > 0, \quad \forall v_h \in V^h, v_h \neq 0 \tag{272}$$

reescribiéndose el problema (262) como el problema aproximado en el cual debemos encontrar $u_h \in V^h \subset V$ tal que

$$a(u_h, v_h) = \langle f, v_h \rangle - a(u_0, v_h) \tag{273}$$

donde $u_0 = g = 0$ en $\partial\Omega$, para toda $v_h \in V^h$, es decir

$$\int_\Omega \left(\nabla v_h \cdot \underline{\underline{a}} \cdot \nabla u_h + c u_h v_h\right) dxdy = \int_\Omega f_\Omega v_h dxdy \tag{274}$$

para todo $v_h \in V^h$.

Entonces, el problema (254) al aplicarle el método Galerkin obtenemos (257), el cual podemos reescribirlo como (274). Aplicando el teorema de Lax-Milgram (33) a este caso particular, tenemos que este tiene solución única y esta depende continuamente de los datos.

Como un caso particular del teorema de Lax-Milgram (33) tenemoe el siguiente resultado

Teorema 42 *Sea V^h un subespacio de dimensión finita de un espacio de Hilbert V, sea $a(\cdot,\cdot) : V^h \times V^h \to \mathbb{R}$ una forma bilineal continua y V-elíptica, y $l(\cdot) : V^h \to \mathbb{R}$ una funcional lineal acotada. Entonces existe una única funcion $u_h \in V^h$ tal que satisface*

$$a(u_h, v_h) = \langle l, v_h \rangle \qquad \forall v_h \in V^h. \tag{275}$$

Además, si $l(\cdot)$ es de la forma

$$\langle l, v_h \rangle = \int_\Omega f_\Omega v_h d\underline{x} \tag{276}$$

con $f \in L^2(\Omega)$, entonces

$$\|u_h\|_V \leq \frac{1}{\alpha} \|f\|_{L^2}, \tag{277}$$

donde α es la constante en (269).

El siguiente resultado nos da una condición suficiente para que la aproximación u_h del método Galerkin converja a la solución u del problema dado por la Ec. (260), para más detalle véase [13] y [3].

Teorema 43 *Sea V un subespacio cerrado de un espacio de Hilbert, y sea la forma bilineal $a(\cdot,\cdot) : V^h \times V^h \to \mathbb{R}$ continua V-elíptica y sea $l(\cdot)$ una funcional lineal acotada. Entonces existe una constante C, independiente de h, tal que*

$$\|u - u_h\|_V \leq C \inf_{v_h \in V^h} \|u - v_h\|_V \tag{278}$$

donde u es solución de (260) y u_h es solución de (273), consecuentemente, una condición suficiente para que la aproximación u_h del método Galerkin converge a la solución u del problema dado por la Ec. (260) es que exista una familia $\{V^h\}$ de subespacios con la propiedad de que

$$\inf_{v_h \in V^h} \|u - v_h\|_V \to 0 \quad \textit{cuando} \quad h \to 0. \tag{279}$$

5.2. El Método de Residuos Pesados

Este método se basa en el método Galerkin, y se escogen subespacios U^h y V^h de tal manera que la dimensión $\dim U^h = \dim V^h = N$, eligiendo las bases como

$$\{\phi_i\}_{i=1}^N \text{ para } U^h \text{ y } \{\psi_j\}_{j=1}^N \text{ para } V^h \tag{280}$$

entonces

$$u_h = \sum_{i=1}^N c_i \phi_i \text{ y } v_h = \sum_{j=1}^N b_j \psi_j \tag{281}$$

donde los coeficientes b_j son arbitrarios ya que v_h es arbitraria.

Sustituyendo esta última expresión Ec. (281) en

$$(\mathcal{L}u_h - f, v_h) = 0 \tag{282}$$

se obtienen N ecuaciones simultaneas

$$\sum_{i=1}^{N} K_{ij} c_i = F_j \text{ con } j = 1, ..., N$$

en la cual en la cual $\underline{\underline{\mathbb{K}}}$ y $\underline{F}$ son la matriz y el vector cuyas entradas son

$$K_{ij} = \left(\mathcal{L}\phi_i, \psi_j\right) \text{ y } F_j = \left(f, \psi_j\right)$$

donde $(\cdot, \cdot)$ representa el producto interior asociado a L^2. A la expresión

$$\tau(u_h) \equiv \mathcal{L}u_h - f \tag{283}$$

se le llama el residuo; si u_h es la solución exacta, entonces por supuesto el residuo se nulifica.

5.3. Método de Elementos Finitos

El método Finite Elements Method (FEM) provee una manera sistemática y simple de generar las funciones base en un dominio con geometría Ω poligonal. Lo que hace al método de elemento finito especialmente atractivo sobre otros métodos, es el hecho de que las funciones base son polinomios definidos por pedazos (elementos Ω_i) que son no cero sólo en una pequeña parte de Ω, proporcionando a la vez una gran ventaja computacional al método ya que las matrices generadas resultan bandadas ahorrando memoria al implantarlas en una computadora.

Así, partiendo del problema aproximado (274), se elegirá una familia de espacios $V^h (h \in (0, 1))$ definido por el procedimiento de elementos finitos (descritos en las subsecciones siguientes en el caso de interpoladores lineales, para otros tipos de interpoladores, ver [16]), teniendo la propiedad de que V^h se aproxima a V cuando h se aproxima a cero en un sentido apropiado, esto es, por supuesto una propiedad indispensable para la convergencia del método Galerkin.

Mallado del dominio El Mallado o triangulación $\mathcal{T}_h$ del dominio Ω es el primer aspecto básico, y ciertamente el más característico, el dominio $\Omega \subset \mathbb{R}^2$ es subdividido en E subdominios o elementos Ω_e llamados elementos finitos, tal que

$$\overline{\Omega} = \bigcup_{e=1}^{E} \overline{\Omega}_e$$

donde:

- Cada $\Omega_e \in \mathcal{T}_h$ es un polígono (rectángulo o triángulo) con interior no vacío ($\mathring{\Omega}_e \neq \emptyset$) y conexo.
- Cada $\Omega_e \in \mathcal{T}_h$ tiene frontera $\partial\Omega_e$ Lipschitz continua.
- Para cada $\Omega_i, \Omega_j \in \mathcal{T}_h$ distintos, $\mathring{\Omega}_i \cap \mathring{\Omega}_j = \emptyset$.
- El diámetro $h_i = Diam(\Omega_e)$ de cada Ω_e satisface $Diam(\Omega_e) \leq h$ para cada $e = 1, 2, ..., E$.
- Cualquier cara de cualquier elemento $\Omega_i \in \mathcal{T}_h$ en la triangulación es también un subconjunto de la frontera $\partial\Omega$ del dominio Ω o una cara de cualquier otro elemento $\Omega_j \in \mathcal{T}_h$ de la triangulación, en este último caso Ω_i y Ω_j son llamados adyacentes.
- Los vértices de cada Ω_e son llamados nodos, teniendo N de ellos por cada elemento Ω_e.

Una vez que la triangulación $\mathcal{T}_h$ del dominio Ω es establecida, se procede a definir el espacio de elementos finitos $\mathbb{P}^h[k]$ a través del proceso descrito a continuación.

Funciones Base A continuación describiremos la manera de construir las funciones base usada por el método de elemento finito. En este procedimiento debemos tener en cuenta que las funciones base están definidas en un subespacio de $V = H^1(\Omega)$ para problemas de segundo orden que satisfacen las condiciones de frontera.

Las funciones base deberán satisfacer las siguientes propiedades:

- Las funciones base ϕ_i son acotadas y continuas, i.e $\phi_i \in C(\Omega_e)$.
- Existen ℓ funciones base por cada nodo del polígono Ω_e, y cada función ϕ_i es no cero solo en los elementos contiguos conectados por el nodo i.
- $\phi_i = 1$ en cada i nodo del polígono Ω_e y cero en los otros nodos.
- La restricción ϕ_i a Ω_e es un polinomio, i.e. $\phi_i \in \mathbb{P}_k[\Omega_e]$ para alguna $k \geq 1$ donde $\mathbb{P}_k[\Omega_e]$ es el espacio de polinomios de grado a lo más k sobre Ω_e.

Decimos que $\phi_i \in \mathbb{P}_k[\Omega_e]$ es una base de funciones y por su construcción es evidente que estas pertenecen a $H^1(\Omega)$. Al conjunto formado por todas las funciones base definidas para todo Ω_e de Ω será el espacio $\mathbb{P}^h[k]$ de funciones base, i.e.

$$\mathbb{P}^h[k] = \bigcup_{e=1}^{E} \mathbb{P}_k[\Omega_e]$$

estas formarán las funciones base globales.

Solución aproximada Para encontrar la solución aproximada elegimos el espacio $\mathbb{P}^h[k]$ de funciones base, como el espacio de funciones lineales ϕ_i definidas por pedazos de grado menor o igual a k (en nuestro caso $k = 1$), entonces el espacio a trabajar es

$$V^h = Generado\left\{\phi_i \in \mathbb{P}^h[k] \;\middle|\; \phi_i(x) = 0 \text{ en } \partial\Omega\right\}. \tag{284}$$

La solución aproximada de la Ec. (274) al problema dado por la Ec. (254) queda en términos de

$$\int_\Omega \left(\nabla\phi_i \cdot \underline{\underline{a}} \cdot \nabla\phi_j - c\phi_i\phi_j\right) dxdy = \int_\Omega f_\Omega \phi_j dxdy \tag{285}$$

si definimos el operador bilineal

$$K_{ij} \equiv a\left(\phi_i, \phi_j\right) = \int_\Omega \left(\nabla\phi_i \cdot a_{ij} \cdot \nabla\phi_j - c\phi_i\phi_j\right) dxdy \tag{286}$$

y la funcional lineal

$$F_j \equiv \left\langle f, \phi_j \right\rangle = \int_\Omega f_\Omega \phi_j dxdy \tag{287}$$

entonces la matriz $\underline{\underline{K}} \equiv [K_{ij}]$, los vectores $\underline{u} \equiv (u_1, ..., u_N)$ y $\underline{F} \equiv (F_1, ..., F_N)$ definen el sistema lineal (que es positivo definido)

$$\underline{\underline{K}}\underline{u} = \underline{F} \tag{288}$$

donde $\underline{u}$ será el vector solución a la Ec. (288) cuyos valores serán la solución al problema dado por la Ec. (274) que es la solución aproximada a la Ec. (254) en los nodos interiores de Ω.

Un Caso más General Sea el operador elíptico (caso simétrico) en el dominio Ω, y el operador definido por

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega\backslash\Sigma \\ u &= g \quad \text{en } \partial\Omega \\ [u]_\Sigma &= J_0 \\ [a_n \cdot \nabla u]_\Sigma &= J_1 \end{aligned} \tag{289}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{290}$$

conjuntamente con una partición $\prod = \{\Omega_1, ..., \Omega_E\}$ de Ω. Multiplicando por la función w obtenemos

$$w\mathcal{L}u = -w\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cwu = wf_\Omega \tag{291}$$

entonces si $w(x)$ es tal que $[w] = 0$ (es decir w es continua) y definimos

$$a(u,w)=\sum_{i=1}^{E}\int_{\Omega_i}\left(\nabla u\cdot\underline{\underline{a}}\cdot\nabla w+cwu\right)d\underline{x} \tag{292}$$

tal que $a(u,w)$ define un producto interior sobre

$$H^1(\Omega)=H^1(\Omega_1)\oplus H^1(\Omega_2)\oplus....\oplus H^1(\Omega_E).$$

Entonces, reescribimos la Ec. (291) como

$$\begin{aligned} a(u,w) &= \int_\Omega wfd\underline{x}+\sum_{i=1}^{E}\int_{\partial\Omega}wa_n\cdot\nabla ud\underline{s} \\ &= \int_\Omega wf_\Omega d\underline{x}+\int_{\partial\Omega}wa_n\cdot\nabla ud\underline{s}-\int_\Sigma w\left[a_n\cdot\nabla u\right]_\Sigma d\underline{s}. \end{aligned} \tag{293}$$

Sea $u_0(x)$ una función que satisface las condiciones de frontera y J_0 una función que satisface las condiciones de salto, tal que

i) $u_0(x)=g(x)$ en $\partial\Omega$

ii) $[u_0(x)]_\Sigma=J_0$

y sea $u(x)=u_0(x)+v(x)$. Entonces $u(x)$ satisface la Ec. (292) si y sólo si $v(x)$ satisface

$$a(u,w)=\int_\Omega wf_\Omega d\underline{x}-\langle u_0,w\rangle-\int_\Sigma J_1wd\underline{s} \tag{294}$$

para toda w tal que $w(x)=0$ en $\partial\Omega$. Sea $\{\phi_i\}$ una base de un subespacio de dimensión finita V^h definido como

$$V^h=\left\{\phi_i\mid\phi_i\in C^1(\Omega_i),\forall i,\phi_i=0\text{ en }\partial\Omega\text{ y }\phi_i\in C^0(\Omega)\right\}. \tag{295}$$

La solución por elementos finitos de (294) se obtiene al resolver el sistema lineal

$$\underline{\underline{K}}\underline{u}=\underline{F} \tag{296}$$

donde

$$K_{ij}=a\left(\phi_i,\phi_j\right) \tag{297}$$

y

$$F_j=\int_\Omega\phi_jf_\Omega d\underline{x}-a\left(u_0,\phi_j\right)-\int_\Sigma J_1\phi_jd\underline{s} \tag{298}$$

esta solución será la solución en los nodos interiores de Ω.En el presente capítulo se prestará atención a varios aspectos necesarios para encontrar la solución aproximada de problemas variacionales con valor en la frontera (VBVP). Ya que en general encontrar la solución a problemas con geometría diversa es difícil y en algunos casos imposible usando métodos analíticos.

6. Método de Elementos Finitos

En este capítulo se considera el VBVP de la forma

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{299}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{300}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, como un caso particular del operador elíptico definido por la Ec. (91) de orden 2, con $\Omega \subset R^2$ un dominio poligonal, es decir, Ω es un conjunto abierto acotado y conexo tal que su frontera $\partial\Omega$ es la unión de un número finito de polígonos.

La sencillez del operador $\mathcal{L}$ nos permite facilitar la comprensión de muchas de las ideas básicas que se expondrán a continuación, pero tengamos en mente que esta es una ecuación que gobierna los modelos de muchos sistemas de la ciencia y la ingeniería, por ello es muy importante su solución.

Si multiplicamos a la ecuación $-\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu = f_\Omega$ por $v \in V = H_0^1(\Omega)$, obtenemos

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{301}$$

aplicando el teorema de Green (83) obtenemos la Ec. (98), que podemos re-escribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{302}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{303}$$

y la funcional lineal

$$l(v) = \langle f, v\rangle = \int_\Omega v f_\Omega d\underline{x} \tag{304}$$

podemos reescribir el problema dado por la Ec. (299) de orden 2 en forma variacional, haciendo uso de la forma bilineal $a(\cdot,\cdot)$ y la funcional lineal $l(\cdot)$.

6.1. Triangulación

El Mallado o triangulación $\mathcal{T}_h$ del dominio Ω es el primer aspecto básico, y ciertamente el más característico, el dominio $\Omega \subset \mathbb{R}^2$ es subdividido en E subdominios o elementos Ω_e llamados elementos finitos, tal que

$$\overline{\Omega} = \bigcup_{e=1}^{E} \overline{\Omega}_e$$

donde:

• Cada $\Omega_e \in \mathcal{T}_h$ es un polígono (rectángulo o triángulo) con interior no vacío ($\mathring{\Omega}_e \neq \emptyset$) y conexo.

• Cada $\Omega_e \in \mathcal{T}_h$ tiene frontera $\partial\Omega_e$ Lipschitz continua.

• Para cada $\Omega_i, \Omega_j \in \mathcal{T}_h$ distintos, $\mathring{\Omega}_i \cap \mathring{\Omega}_j = \emptyset$.

• El diámetro $h_i = Diam(\Omega_e)$ de cada Ω_e satisface $Diam(\Omega_e) \leq h$ para cada $e = 1, 2, ..., E$.

• Los vértices de cada Ω_e son llamados nodos, teniendo N de ellos por cada elemento Ω_e.

Definición 44 *Una familia de triangulaciones $\mathcal{T}_h$ es llamada de forma-regular si existe una constante independiente de h, tal que*

$$h_K \leq C\rho_K, \ \textit{con } K \in \mathcal{T}_h,$$

donde ρ_K es el radio del circulo más grande contenido en K. El radio h_K/ρ_K es llamado esl aspect ratio de K.

Definición 45 *Una familia de triangulaciones $\mathcal{T}_h$ es llamada cuasi-uniforme si esta es de forma-regular y si existe una constante independiente de h, tal que*

$$h_K \leq Ch, \ \textit{con } K \in \mathcal{T}_h.$$

Una vez que la triangulación $\mathcal{T}_h$ del dominio Ω es establecida, se procede a definir el espacio de elementos finitos $\mathbb{P}^h[k]$ a través del proceso descrito a continuación.

6.2. Interpolación para el Método de Elementos Finitos

Funciones Base A continuación describiremos la manera de construir las funciones base usada por el método de elemento finito. En este procedimiento debemos tener en cuenta que las funciones base están definidas en un subespacio de $V = H^1(\Omega)$ para problemas de segundo orden que satisfacen las condiciones de frontera.

Las funciones base deberán satisfacer las siguientes propiedades:

i) Las funciones base ϕ_i son acotadas y continuas, i.e $\phi_i \in C(\Omega_e)$.

ii) Existen ℓ funciones base por cada nodo del polígono Ω_e, y cada función ϕ_i es no cero solo en los elementos contiguos conectados por el nodo i.

iii) $\phi_i = 1$ en cada i nodo del polígono Ω_e y cero en los otros nodos.

iv) La restricción ϕ_i a Ω_e es un polinomio, i.e. $\phi_i \in \mathbb{P}_k[\Omega_e]$ para alguna $k \geq 1$ donde $\mathbb{P}_k[\Omega_e]$ es el espacio de polinomios de grado a lo más k sobre Ω_e.

Decimos que $\phi_i \in \mathbb{P}_k[\Omega_e]$ es una base de funciones y por su construcción es evidente que estas pertenecen a $H^1(\Omega)$. Al conjunto formado por todas las funciones base definidas para todo Ω_e de Ω será el espacio $\mathbb{P}^h[k]$ de funciones base, i.e.

$$\mathbb{P}^h[k] = \bigcup_{e=1}^{E} \mathbb{P}_k[\Omega_e]$$

estas formarán las funciones base globales.

6.3. Método de Elemento Finito Usando Discretización de Rectángulos

Para resolver la Ec. (299), usando una discretización con rectángulos, primero dividimos el dominio $\Omega \subset \mathbb{R}^2$ en N_x nodos horizontales por N_y nodos verticales, teniendo $E = (N_x - 1)(N_y - 1)$ subdominios o elementos rectangulares Ω_e tales que $\overline{\Omega} = \cup_{e=1}^{E} \overline{\Omega_e}$ y $\overline{\Omega_i} \cap \overline{\Omega_j} \neq \varnothing$ si son adyacentes, con un total de $N = N_x N_y$ nodos.

Donde las funciones lineales definidas por pedazos en Ω_e en nuestro caso serán polinomios de orden uno en cada variable separadamente y cuya restricción de ϕ_i a Ω_e es $\phi_i^{(e)}$. Para simplificar los cálculos en esta etapa, supondremos que la matriz $\underline{\underline{a}} = a \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, entonces se tiene que la integral del lado izquierdo de la Ec. (285) queda escrita como

$$\int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy = \int_\Omega f_\Omega \phi_j dxdy \tag{305}$$

donde

$$\begin{aligned} K_{ij} &= \int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy \qquad (306) \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \nabla \phi_i^{(e)} \cdot \nabla \phi_j^{(e)} + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \end{aligned}$$

y el lado derecho como

$$\begin{aligned} F_j &= \int_\Omega f_\Omega \phi_j dxdy \qquad (307) \\ &= \sum_{e=1}^{E} \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy. \end{aligned}$$

Para cada Ω_e de Ω, la submatriz de integrales (matriz de carga local)

$$K_{ij} = \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{308}$$

tiene la estructura

$$\begin{bmatrix} K_{1,1}^{(e)} & K_{1,2}^{(e)} & K_{1,3}^{(e)} & K_{1,4}^{(e)} \\ K_{2,1}^{(e)} & K_{2,2}^{(e)} & K_{2,3}^{(e)} & K_{2,4}^{(e)} \\ K_{3,1}^{(e)} & K_{3,2}^{(e)} & K_{3,3}^{(e)} & K_{3,4}^{(e)} \\ K_{4,1}^{(e)} & K_{4,2}^{(e)} & K_{4,3}^{(e)} & K_{4,4}^{(e)} \end{bmatrix}$$

la cual deberá ser ensamblada en la matriz de carga global que corresponda a la numeración de nodos locales del elemento Ω_e con respecto a la numeración global de los elementos en Ω.

De manera parecida, para cada Ω_e de Ω se genera el vector de integrales (vector de carga local)

$$F_j = \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \tag{309}$$

con la estructura

$$\begin{bmatrix} F_1^{(e)} \\ F_2^{(e)} \\ F_3^{(e)} \\ F_4^{(e)} \end{bmatrix}$$

el cual también deberá ser ensamblado en el vector de carga global que corresponda a la numeración de nodos locales al elemento Ω_e con respecto a la numeración global de los elementos de Ω.

Montando los $K_{ij}^{(e)}$ en la matriz $\underline{\underline{\mathbb{K}}}$ y los $F_j^{(e)}$ en el vector $\underline{\mathbb{F}}$ según la numeración de nodos global, se genera el sistema $\underline{\underline{\mathbb{K}}} \underline{u_h} = \underline{\mathbb{F}}$ donde $\underline{u_h}$ será el vector cuyos valores serán la solución aproximada a la Ec. (299) en los nodos interiores de Ω. La matriz $\underline{\underline{\mathbb{K}}}$ generada de esta forma, tiene una propiedad muy importante, es bandada y el ancho de banda es de 9 elementos, esto es muy útil al momento de soportar la matriz en memoria.

Para implementar numéricamente en cada Ω_e las integrales

$$\int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{310}$$

y

$$\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy, \tag{311}$$

teniendo en mente el simplificar los cálculos computacionales, se considera un elemento de referencia $\hat{\Omega}$ en los ejes coordenados (ε, η) cuyos vértices están el $(-1,-1), (1,-1), (1,1)$ y $(-1,1)$ respectivamente, en el cual mediante una

función afín será proyectado cualquier elemento rectangular Ω_e cuyos vértices $(x_1^{(e)}, y_1^{(e)}), (x_2^{(e)}, y_2^{(e)}), (x_3^{(e)}, y_3^{(e)})$ y $(x_4^{(e)}, y_4^{(e)})$ están tomados en sentido contrario al movimiento de las manecillas del reloj como se muetra en la figura

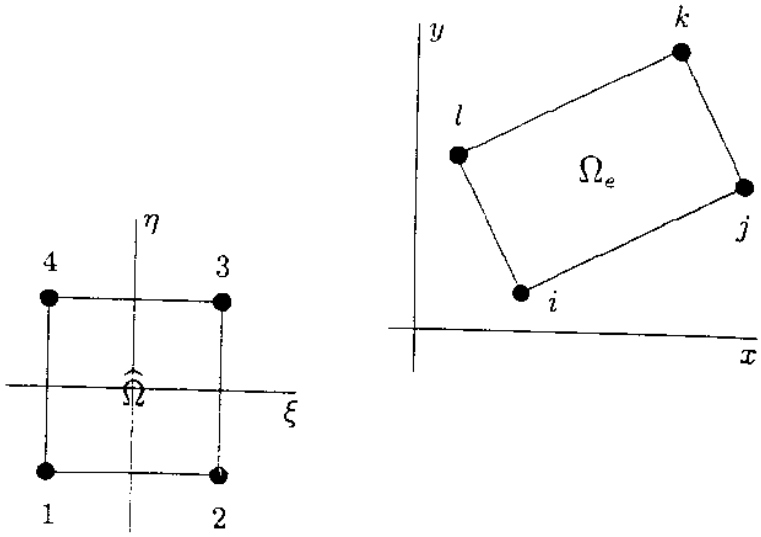


Figura 2:

mediante la transformación $f(x,y) = \underline{\underline{T}}(\varepsilon,\eta)+\underline{b}$, quedando dicha transformación como

$$\begin{aligned} x &= \frac{x_2^{(e)} - x_1^{(e)}}{2}\varepsilon + \frac{y_2^{(e)} - y_1^{(e)}}{2}\eta \\ y &= \frac{x_4^{(e)} - x_1^{(e)}}{2}\varepsilon + \frac{y_4^{(e)} - y_1^{(e)}}{2}\eta \end{aligned} \tag{312}$$

en la cual la matriz $\underline{\underline{T}}$ está dada por

$$\underline{\underline{T}} = \begin{pmatrix} \frac{x_2^{(e)} - x_1^{(e)}}{2} & \frac{y_2^{(e)} - y_1^{(e)}}{2} \\ \frac{x_4^{(e)} - x_1^{(e)}}{2} & \frac{y_4^{(e)} - y_1^{(e)}}{2} \end{pmatrix} \tag{313}$$

y el vector $\underline{b}= (b_1, b_2)$ es la posición del vector centroide del rectángulo Ω_e,

también se tiene que la transformación inversa es

$$\begin{aligned}
\varepsilon &= \frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2} \left[\frac{y - b_2}{\left(\frac{x_4^{(e)} - x_1^{(e)}}{2} \right) \left(\frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2}}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \right)} \right]}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \\
\eta &= \frac{y - b_2}{\left(\frac{x_4^{(e)} - x_1^{(e)}}{2} \right) \left(\frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2}}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \right) + \frac{y_4^{(e)} - y_1^{(e)}}{2}}.
\end{aligned} \tag{314}$$

Entonces las $\phi_i^{(e)}$ quedan definidas en términos de $\hat{\phi}_i$ como

$$\begin{aligned}
\hat{\phi}_1(\varepsilon, \eta) &= \frac{1}{4}(1 - \varepsilon)(1 - \eta) \\
\hat{\phi}_2(\varepsilon, \eta) &= \frac{1}{4}(1 + \varepsilon)(1 - \eta) \\
\hat{\phi}_3(\varepsilon, \eta) &= \frac{1}{4}(1 + \varepsilon)(1 + \eta) \\
\hat{\phi}_4(\varepsilon, \eta) &= \frac{1}{4}(1 - \varepsilon)(1 + \eta)
\end{aligned} \tag{315}$$

y las funciones $\phi_i^{(e)}$ son obtenidas por el conjunto $\phi_i^{(e)}(x, y) = \hat{\phi}_i(\varepsilon, \eta)$ con (x, y) y (ε,η) relacionadas por la Ec. (312), entonces se tendrian las siguientes integrales

$$\begin{aligned}
K_{ij}^{(e)} &= \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \\
&= \int_{\hat{\Omega}} \left(\left[a \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial x} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial x} \right) + \right. \right. \\
&\quad \left. \left. \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \right] + c \hat{\phi}_i \hat{\phi}_j \right) |J| \, d\varepsilon d\eta
\end{aligned} \tag{316}$$

donde el índice i y j varia de 1 a 4. En está última usamos la regla de la cadena y $dxdy = |J| \, d\varepsilon d\eta$ para el cambio de variable en las integrales, aquí $|J| = \det T$, donde T está dado como en la Ec. (313). Para resolver $\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy$ en cada Ω_e se genera las integrales

$$\begin{aligned}
F_j^{(e)} &= \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \\
&= \int_{\hat{\Omega}} f_\Omega \hat{\phi}_j \, |J| \, d\varepsilon d\eta
\end{aligned} \tag{317}$$

donde el índice i y j varia de 1 a 4.

Para realizar el cálculo numérico de las integrales en el rectángulo de referencia $\hat{\Omega} = [-1,1]\times[-1,1]$, debemos conocer $\frac{\partial \phi i}{\partial \varepsilon}$, $\frac{\partial \phi i}{\partial \eta}$, $\frac{\partial \varepsilon}{\partial x}$, $\frac{\partial \varepsilon}{\partial y}$, $\frac{\partial \eta}{\partial x}$ y $\frac{\partial \eta}{\partial y}$, entonces realizando las operaciones necesarias a la Ec. (315) obtenemos

$$\begin{array}{ll} \frac{\partial \phi_1}{\partial \varepsilon} = -\frac{1}{4}(1-\eta) & \frac{\partial \phi_1}{\partial \eta} = -\frac{1}{4}(1-\varepsilon) \\ \frac{\partial \phi_2}{\partial \varepsilon} = \frac{1}{4}(1-\eta) & \frac{\partial \phi_2}{\partial \eta} = -\frac{1}{4}(1+\varepsilon) \\ \frac{\partial \phi_3}{\partial \varepsilon} = \frac{1}{4}(1+\eta) & \frac{\partial \phi_3}{\partial \eta} = \frac{1}{4}(1+\varepsilon) \\ \frac{\partial \phi_4}{\partial \varepsilon} = -\frac{1}{4}(1+\eta) & \frac{\partial \phi_4}{\partial \eta} = \frac{1}{4}(1-\varepsilon) \end{array} \tag{318}$$

y también

$$\begin{array}{ll} \frac{\partial \varepsilon}{\partial x} = \left(\frac{y_4^{(e)} - y_1^{(e)}}{2\det T}\right) & \frac{\partial \varepsilon}{\partial y} = \left(\frac{x_4^{(e)} - x_1^{(e)}}{2\det T}\right) \\ \frac{\partial \eta}{\partial x} = \left(\frac{y_2^{(e)} - y_1^{(e)}}{2\det T}\right) & \frac{\partial \eta}{\partial y} = \left(\frac{x_2^{(e)} - x_1^{(e)}}{2\det T}\right) \end{array} \tag{319}$$

las cuales deberán de ser sustituidas en cada $\underline{\underline{K_{ij}^{(e)}}}$ y $\underline{F_j^{(e)}}$ para calcular las integrales en el elemento Ω_e. Estas integrales se harán en el programa usando cuadratura Gaussiana, permitiendo reducir el número de cálculos al mínimo pero manteniendo el balance entre precisión y número bajo de operaciones necesarias para realizar las integraciones.

Suponiendo que Ω fue dividido en E elementos, estos elementos generan N nodos en total, de los cuales N_d son nodos desconocidos y N_c son nodos conocidos con valor γ_j, entonces el algoritmo de ensamble de la matriz $\underline{\underline{K}}$ y el vector $\underline{F}$ se puede esquematizar como:

$K_{i,j} = (\phi_i, \phi_j) \quad \forall i = 1,2,...,E, j = 1,2,...,E$

$F_j = (f_\Omega, \phi_j) \quad \forall j = 1,2,...,E$

$\forall j = 1,2,...,N_d$:

$$b_j = b_j - \gamma_i K_{i,j} \quad \forall i = 1,2,...,E$$

Así, se construye una matriz global en la cual están representados los nodos conocidos y los desconocidos, tomando sólo los nodos desconocidos de la matriz $\underline{\underline{K}}$ formaremos una matriz $\underline{\underline{A}}$, haciendo lo mismo al vector $\underline{F}$ formamos el vector $\underline{b}$, entonces la solución al problema será la resolución del sistema de ecuaciones lineales $\underline{\underline{A}}\underline{x} = \underline{b}$, este sistema puede resolverse usando por ejemplo el método de Gradiente Conjugado. El vector $\underline{x}$ contendrá la solución buscada en los nodos desconocidos N_d.

6.4. Método de Elemento Finito Usando Discretización de Triángulos

Para resolver la Ec. (254), usando una discretización con triángulos, primero dividimos el dominio $\Omega \subset \mathbb{R}^2$ en N_x nodos horizontales por N_y nodos verticales, teniendo $E = 2(N_x - 1)(N_y - 1)$ subdominios o elementos triángulares Ω_e tales

que $\overline{\Omega} = \cup_{e=1}^{E} \overline{\Omega_e}$ y $\overline{\Omega_i} \cap \overline{\Omega_j} \neq \varnothing$ si son adyacentes, con un total de $N = N_x N_y$ nodos.

Donde las funciones lineales definidas por pedazos en Ω_e en nustro caso serán polinomios de orden uno en cada variable separadamente y cuya restricción de ϕ_i a Ω_e es $\phi_i^{(e)}$. Para simplificar los cálculos en esta etapa, supondremos que la matriz $\underline{\underline{a}} = a \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, entonces se tiene que la integral del lado izquierdo de la Ec. (285) queda escrita como

$$\int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy = \int_\Omega f_\Omega \phi_j dxdy \tag{320}$$

donde

$$\begin{aligned} K_{ij} &= \int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy \qquad (321) \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \nabla \phi_i^{(e)} \cdot \nabla \phi_j^{(e)} + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \end{aligned}$$

y el lado derecho como

$$\begin{aligned} \underline{F_j} &= \int_\Omega f_\Omega \phi_j dxdy \qquad (322) \\ &= \sum_{e=1}^{E} \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy. \end{aligned}$$

Para cada Ω_e de Ω la submatriz de integrales (matriz de carga local)

$$\underline{\underline{K_{ij}}} = \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{323}$$

tiene la estructura

$$\begin{bmatrix} k_{1,1}^{(e)} & k_{1,2}^{(e)} & k_{1,3}^{(e)} \\ k_{2,1}^{(e)} & k_{2,2}^{(e)} & k_{2,3}^{(e)} \\ k_{3,1}^{(e)} & k_{3,2}^{(e)} & k_{3,3}^{(e)} \end{bmatrix}$$

la cual deberá ser ensamblada en la matriz de carga global que corresponda a la numeración de nodos locales del elemento Ω_e con respecto a la numeración global de los elementos en Ω.

De manera parecida, para cada Ω_e de Ω se genera el vector de integrales (vector de carla gocal)

$$F_j = \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \tag{324}$$

con la estructura

$$\begin{bmatrix} F_1^{(e)} \\ F_2^{(e)} \\ F_3^{(e)} \end{bmatrix}$$

el cual también deberá ser ensamblado en el vector de carga global que corresponda a la numeriación de nodos locales al elemento Ω_e con respecto a la numeración global de los elementos de Ω.

Montando los $K_{ij}^{(e)}$ en la matriz $\underline{\underline{\mathbb{K}}}$ y los $F_j^{(e)}$ en el vector $\underline{\mathbb{F}}$ según la numeración de nodos global, se genera el sistema $\underline{\underline{\mathbb{K}}}\underline{u_h}=\underline{\mathbb{F}}$ donde $\underline{u_h}$ será el vector cuyos valores serán la solución aproximada a la Ec. (254) en los nodos interiores de Ω. La matriz $\underline{\underline{\mathbb{K}}}$ generada de esta forma, tiene una propiedad muy importante, es bandada y el ancho de banda es de 7 elementos, esto es muy útil al momento de soportar la matriz en memoria.

Para implementar numéricamente en cada Ω_e las integrales

$$\int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c\phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{325}$$

y

$$\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy$$

teniendo en mente el simplificar los cálculos computacionales se considera a un elemento de referencia $\hat{\Omega}$ en los ejes coordenados (ε, η) cuyos vertices estan en $(0,0), (1,0)$ y $(0,1)$ y en el cual mediante un mapeo afín será proyectado caulquier elemento triangular Ω_e cuyos vertices $(x_1^{(e)}, y_1^{(e)}), (x_2^{(e)}, y_2^{(e)}), (x_3^{(e)}, y_3^{(e)})$ están tomados en el sentido contrario al movimiento de las manecillas del reloj como se muetra en la figura mediante la transformación $f(\varepsilon, \eta) = \underline{\underline{T}}(\varepsilon, \eta) + \underline{b}$, quedando dicha transformación como

$$\begin{aligned} x &= x_1^{(e)}(1 - \varepsilon - \eta) + x_2^{(e)}\varepsilon + x_3^{(e)}\eta \\ y &= y_1^{(e)}(1 - \varepsilon - \eta) + y_2^{(e)}\varepsilon + y_3^{(e)}\eta \end{aligned} \tag{326}$$

y en la cual la matriz $\underline{\underline{T}}$ está dada por

$$\underline{\underline{T}} = \begin{pmatrix} x_2^{(e)} - x_1^{(e)} & x_3^{(e)} - x_1^{(e)} \\ y_2^{(e)} - y_1^{(e)} & y_3^{(e)} - y_1^{(e)} \end{pmatrix} \tag{327}$$

donde $\underline{b}$ es un vector constante

$$\underline{b} = \begin{pmatrix} x_1^{(e)} \\ y_1^{(e)} \end{pmatrix} \tag{328}$$

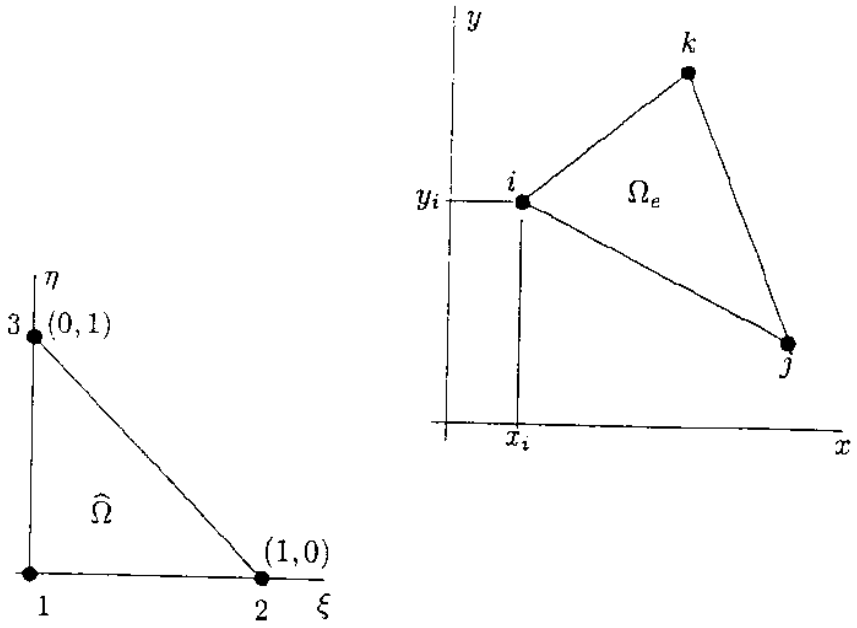


Figura 3:

también se tiene que la transformación inversa es

$$\begin{aligned} \varepsilon &= \frac{1}{2A_{\Omega_e}}\left[\left(y_3^{(e)}-y_1^{(e)}\right)\left(x-x_1^{(e)}\right)-\left(x_3^{(e)}-x_1^{(e)}\right)\left(y-y_1^{(e)}\right)\right] \quad (329)\\ \eta &= \frac{1}{2A_{\Omega_e}}\left[-\left(y_2^{(e)}-y_1^{(e)}\right)\left(x-x_1^{(e)}\right)-\left(x_2^{(e)}-x_1^{(e)}\right)\left(y-y_1^{(e)}\right)\right] \end{aligned}$$

donde

$$A_{\Omega_e} = \left|\det\begin{bmatrix} 1 & x_1^{(e)} & y_1^{(e)} \\ 1 & x_2^{(e)} & y_2^{(e)} \\ 1 & x_3^{(e)} & y_3^{(e)} \end{bmatrix}\right|. \quad (330)$$

Entoces las $\phi_i^{(e)}$ quedan definidas en términos de $\hat{\phi}_i$ como

$$\begin{aligned} \hat{\phi}_1(\varepsilon,\eta) &= 1-\varepsilon-\eta \quad (331)\\ \hat{\phi}_2(\varepsilon,\eta) &= \varepsilon \\ \hat{\phi}_3(\varepsilon,\eta) &= \eta \end{aligned}$$

entoces las funciones $\phi_i^{(e)}$ son obtenidas por el conjunto $\phi_i^{(e)}(x,y) = \hat{\phi}_i(\varepsilon,\eta)$ con (x,y) y (ε,η) relacionadas por la Ec. (326), entonces se tendrian las siguientes

integrales

$$
\begin{aligned}
k_{ij}^{(e)} &= \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \qquad (332) \\
&= \int_{\hat{\Omega}} \left(\left[a \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial x} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial x} \right) + \right. \right. \\
&\quad \left. \left. \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \right] + c \hat{\phi}_i \hat{\phi}_j \right) |J| \, d\varepsilon d\eta
\end{aligned}
$$

donde el índice i y j varia de 1 a 3. En está última usamos la regla de la cadena y $dxdy = |J| \, d\varepsilon d\eta$ para el cambio de variable en las integrales, aquí $|J| = \det T$, donde T está dado como en la Ec. (327). Para resolver $\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy$ en cada Ω_e se genera las integrales

$$
\begin{aligned}
F_j^{(e)} &= \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \qquad (333) \\
&= \int_{\hat{\Omega}} f_\Omega \hat{\phi}_j \, |J| \, d\varepsilon d\eta
\end{aligned}
$$

donde el indice i y j varia 1 a 3.

Para realizar el cálculo numérico de las integrales en el triángulo de referencia $\hat{\Omega}$, debemos conocer $\frac{\partial \phi i}{\partial \varepsilon}$, $\frac{\partial \phi i}{\partial \eta}$, $\frac{\partial \varepsilon}{\partial x}$, $\frac{\partial \varepsilon}{\partial y}$, $\frac{\partial \eta}{\partial x}$ y $\frac{\partial \eta}{\partial y}$, entonces realizando las operaciones necesarias a las Ec. (331) obtenemos

$$
\begin{array}{ll}
\frac{\partial \phi_1}{\partial \varepsilon} = -1 & \frac{\partial \phi_1}{\partial \eta} = -1 \\
\frac{\partial \phi_2}{\partial \varepsilon} = 1 & \frac{\partial \phi_2}{\partial \eta} = 0 \\
\frac{\partial \phi_3}{\partial \varepsilon} = 0 & \frac{\partial \phi_3}{\partial \eta} = 1
\end{array} \qquad (334)
$$

y también

$$
\begin{array}{ll}
\frac{\partial \varepsilon}{\partial x} = \frac{\left(y_3^{(e)} - y_1^{(e)} \right)}{2 A_{\Omega_e}} & \frac{\partial \varepsilon}{\partial y} = -\frac{\left(x_3^{(e)} - x_1^{(e)} \right)}{2 A_{\Omega_e}} \\
\frac{\partial \eta}{\partial x} = -\frac{\left(y_2^{(e)} - y_1^{(e)} \right)}{2 A_{\Omega_e}} & \frac{\partial \eta}{\partial y} = \frac{\left(x_2^{(e)} - x_1^{(e)} \right)}{2 A_{\Omega_e}}
\end{array} \qquad (335)
$$

las cuales deberán de ser sustituidas en cada $\underline{\underline{K_{ij}^{(e)}}}$ y $\underline{F_j^{(e)}}$ para calcular las integrales en el elemento Ω_e.

Suponiendo que Ω fue dividido en E elementos, estos elementos generan N nodos en total, de los cuales N_d son nodos desconocidos y N_c son nodos conocidos con valor γ_j, entonces el algoritmo de ensamble de la matriz $\underline{\underline{K}}$ y el vector $\underline{F}$ se puede esquematizar como:

$K_{i,j} = (\phi_i, \phi_j) \quad \forall i = 1, 2, ..., E, j = 1, 2, ..., E$

$F_j = (f_\Omega, \phi_j) \quad \forall j = 1, 2, ..., E$

$\forall j = 1, 2, ..., N_d$:

$$b_j = b_j - \gamma_i K_{i,j} \quad \forall i = 1, 2, ..., E$$

Así, se construye una matriz global en la cual están representados los nodos conocidos y los desconocidos, tomando sólo los nodos desconocidos de la matriz $\underline{\underline{K}}$ formaremos una matriz $\underline{\underline{A}}$, haciendo lo mismo al vector $\underline{F}$ formamos el vector $\underline{b}$, entonces la solución al problema será la resolución del sistema de ecuaciones lineales $\underline{\underline{A}}\underline{x} = \underline{b}$, este sistema puede resolverse usando por ejemplo el método de gradiente conjugado. El vector $\underline{x}$ contendrá la solución buscada en los nodos desconocidos N_d.

6.5. Implementación Computacional

A partir de los modelos matemáticos y los modelos numéricos en esta sección se describe el modelo computacional contenido en un programa de cómputo orientado a objetos en el lenguaje de programación C++ en su forma secuencial y en su forma paralela en C++ u-sando la interfaz de paso de mensajes (MPI) bajo el esquema maestro-esclavo.

Esto no sólo nos ayudará a demostrar que es factible la construcción del propio modelo computacional a partir del modelo matemático y numérico para la solución de problemas reales. Además, se mostrará los alcances y limitaciones en el consumo de los recursos computacionales, evaluando algunas de las variantes de los métodos numéricos con los que es posible implementar el modelo computacional y haremos el análisis de rendimiento sin llegar a ser exhaustivo esté.

También exploraremos los alcances y limitaciones de cada uno de los métodos implementados (FEM, DDM secuencial y paralelo) y como es posible optimizar los recursos computacionales con los que se cuente.

Primeramente hay que destacar que el paradigma de programación orientada a objetos es un método de implementación de programas, organizados como colecciones cooperativas de objetos. Cada objeto representa una instancia de alguna clase y cada clase es miembro de una jerarquía de clases unidas mediante relaciones de herencia, contención, agregación o uso.

Esto nos permite dividir en niveles la semántica de los sistemas complejos tratando así con las partes, que son más manejables que el todo, permitiendo su extensión y un mantenimiento más sencillo. Así, mediante la herencia, contención, agregación o usó nos permite generar clases especializadas que manejan eficientemente la complejidad del problema. La programación orientada a objetos organiza un programa entorno a sus datos (atributos) y a un conjunto de interfases bien definidas para manipular estos datos (métodos dentro de clases reusables) esto en oposición a los demás paradigmas de programación.

El paradigma de programación orientada a objetos sin embargo sacrifica algo de eficiencia computacional por requerir mayor manejo de recursos computacionales al momento de la ejecución. Pero en contraste, permite mayor flexibilidad al adaptar los códigos a nuevas especificaciones. Adicionalmente, disminuye notoriamente el tiempo invertido en el mantenimiento y búsqueda de errores dentro del código. Esto tiene especial interés cuando se piensa en la

cantidad de meses invertidos en la programación comparado con los segundos consumidos en la ejecución del mismo.

Para empezar con la implementación computacional, primeramente definiremos el problema a trabajar. Este, pese a su sencillez, no pierde generalidad permitiendo que el modelo mostrado sea usado en muchos sistemas de la ingeniería y la ciencia.

El Operador de Laplace y la Ecuación de Poisson Consideramos como modelo matemático el problema de valor en la frontera (BVP) asociado con el operador de Laplace en dos dimensiones, el cual en general es usualmente referido como la ecuación de Poisson, con condiciones de frontera Dirichlet, definido en Ω como:

$$\begin{aligned} -\nabla^2 u &= f_\Omega \text{ en } \Omega \\ u &= g_{\partial\Omega} \text{ en } \partial\Omega. \end{aligned} \tag{336}$$

Se toma está ecuación para facilitar la comprensión de las ideas básicas. Es un ejemplo muy sencillo, pero gobierna los modelos de muchos sistemas de la ingeniería y de la ciencia, entre ellos el flujo de agua subterránea a través de un acuífero isotrópico, homogéneo bajo condiciones de equilibrio y es muy usada en múltiples ramas de la física. Por ejemplo, gobierna la ecuación de la conducción de calor en un sólido bajo condiciones de equilibrio.

En particular consideramos el problema con Ω definido en:

$$\Omega = [-1, 1] \times [0, 1] \tag{337}$$

donde

$$f_\Omega = 2n^2\pi^2 \sin(n\pi x) * \sin(n\pi y) \quad \text{y} \quad g_{\partial\Omega} = 0 \tag{338}$$

cuya solución es

$$u(x, y) = \sin(n\pi x) * \sin(n\pi y). \tag{339}$$

Para las pruebas de rendimiento en las cuales se evalúa el desempeño de los programas realizados se usa $n = 10$, pero es posible hacerlo con $n \in \mathbb{N}$ grande. Por ejemplo para $n = 4$, la solución es $u(x, y) = \sin(4\pi x) * \sin(4\pi y)$, cuya gráfica se muestra a continuación:

Hay que hacer notar que al implementar la solución numérica por el método del elemento finito y el método de subestructuración secuencial en un procesador, un factor limitante para su operación es la cantidad de memoria disponible en la computadora, ya que el sistema algebraico de ecuaciones asociado a esté problema crece muy rápido (del orden de n^2), donde n es el número de nodos en la partición.

En todos los cálculos de los métodos numéricos usados para resolver el sistema lineal algebraico asociado se usó una tolerancia mínima de 1×10^{-10}. Ahora, veremos la implementación del método de elemento finito secuencial para después continuar con el método de descomposición de dominio tanto secuencial como paralelo y poder analizar en cada caso los requerimientos de cómputo, necesarios para correr eficientemente un problema en particular.

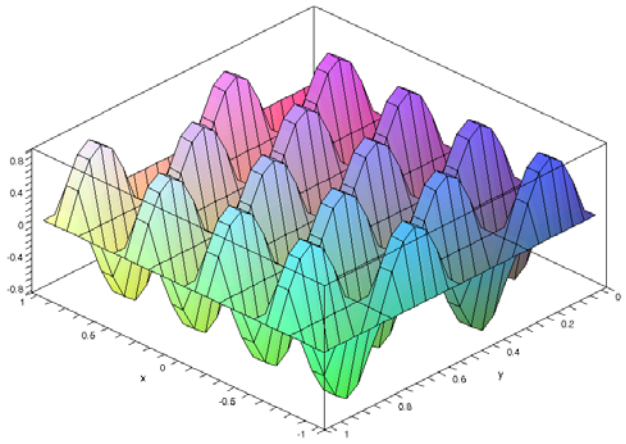


Figura 4: Solución a la ecuación de Poisson para n=4.

Método del Elemento Finito Secuencial A partir de la formulación del método de elemento finito visto en la sección (5.3), la implementación computacional que se desarrolló tiene la jerarquía de clases siguiente:

Donde las clases participantes en *FEM2D Rectángulos* son:

La clase *Interpolador Lineal* define los interpoladores lineales usados por el método de elemento finito.

La clase *Problema* define el problema a tratar, es decir, la ecuación diferencial parcial, valores de frontera y dominio.

La clase *Base FEM* ayuda a definir los nodos al usar la clase *Geometría* y mantiene las matrices generadas por el método y a partir de la clase *Resuelve Ax=B* se dispone de diversas formas de resolver el sistema lineal asociado al método.

La clase *FEM2D* controla lo necesario para poder hacer uso de la geometría en 2D y conocer los nodos interiores y de frontera, con ellos poder montar la matriz de rigidez y ensamblar la solución.

La clase *FEM2D Rectángulos* permite calcular la matriz de rigidez para generar el sistema algebraico de ecuaciones asociado al método.

Notemos que esta misma jerarquía permite trabajar problemas en una y dos dimensiones, en el caso de dos dimensiones podemos discretizar usando rectángulos o triángulos, así como usar varias opciones para resolver el sistema lineal algebraico asociado a la solución de EDP.

Como ya se menciono, el método de elemento finito es un algoritmo secuencial, por ello se implementa para que use un solo procesador y un factor limitante para su operación es la cantidad de memoria disponible en la computadora, por ejemplo:

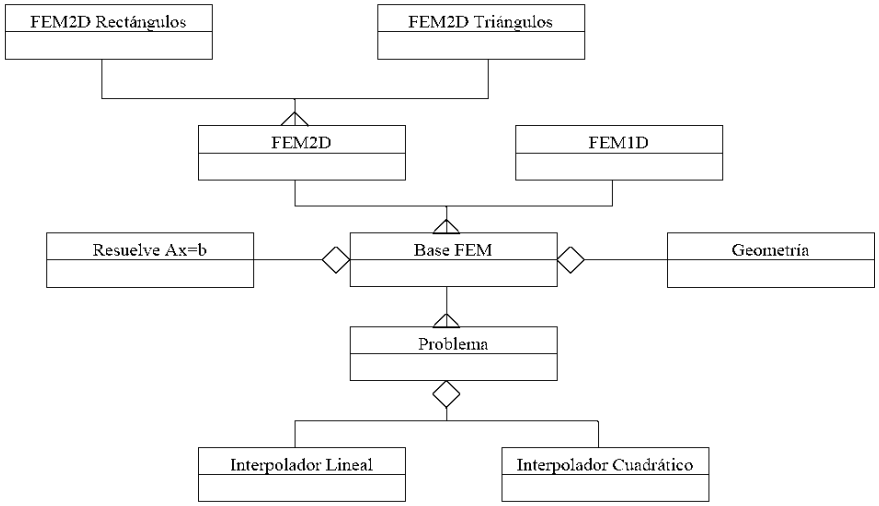


Figura 5: Jerarquía de clases para el método de elemento finito

Resolver la Ec. (**??**) con una partición rectangular de 513×513 nodos, genera 262144 elementos rectangulares con 263169 nodos en total, donde 261121 son desconocidos; así el sistema algebraico de ecuaciones asociado a este problema es de dimensión 261121×261121.

Usando el equipo secuencial, primeramente evaluaremos el desempeño del método de elemento finito con los distintos métodos para resolver el sistema algebraico de ecuaciones, encontrando los siguientes resultados:

Método Iterativo	Iteraciones	Tiempo Total
Jacobi	865037	115897 seg.
Gauss-Seidel	446932	63311 seg.
Gradiente Conjugado	761	6388 seg.

Como se observa el uso del método de Gradiente Conjugado es por mucho la mejor elección. En principio, podríamos quedarnos solamente con el método de Gradiente Conjugado sin hacer uso de precondicionadores por los buenos rendimientos encontrados hasta aquí, pero si se desea resolver un problema con un gran número de nodos, es conocido el aumento de eficiencia al hacer uso de precondicionadores.

Ahora, si tomamos ingenuamente el método de elemento finito conjuntamente con el método de Gradiente Conjugado con precondicionadores a posteriori (los más sencillos de construir) para resolver el sistema algebraico de ecuaciones, encontraremos los siguientes resultados:

Precondicionador	Iteraciones	Tiempo Total
Jacobi	760	6388 seg.
SSOR	758	6375 seg.
Factorización Incompleta	745	6373 seg.

Como es notorio el uso del método de Gradiente Conjugado precondicionado con precondicionadores a posteriori no ofrece una ventaja significativa que compense el esfuerzo computacional invertido al crear y usar un precondicionador en los cálculos por el mal condicionamiento del sistema algebraico. Existen también precondicionadores a priori para el método de elemento finito, pero no es costeable en rendimiento su implementación.

7. Apéndice A

En este apéndice se darán algunas definiciones que se usan a lo largo del presente trabajo, así como se detallan algunos resultados generales de algebra lineal y análisis funcional (en espacios reales) que se anuncian sin demostración pero se indica en cada caso la bibliografía correspondiente donde se encuentran estas y el desarrollo en detalle de cada resultado.

7.1. Nociones de Algebra Lineal

A continuación detallaremos algunos resultados de álgebra lineal, las demostraciones de los siguientes resultados puede ser consultada en [21].

Definición 46 *Sea V un espacio vectorial y sea $f(\cdot) : V \to \mathbb{R}$, f es llamada funcional lineal si satisface la condición*

$$f(\alpha v + \beta w) = \alpha f(v) + \beta f(w) \quad \forall v, w \in V \quad y \quad \alpha, \beta \in \mathbb{R}. \tag{340}$$

Definición 47 *Si V es un espacio vectorial, entonces el conjunto V^* de todas las funcionales lineales definidas sobre V es un espacio vectorial llamado espacio dual de V.*

Teorema 48 *Si $\{v_1, ..., v_n\}$ es una base para el espacio vectorial V, entonces existe una única base $\{v_1^*, ..., v_n^*\}$ del espacio vectorial dual V^* llamado la base dual de $\{v_1, ..., v_n\}$ con la propiedad de que $V_i^* = \delta_{ij}$. Por lo tanto V es isomorfo a V^*.*

Definición 49 *Sea $D \subset V$ un subconjunto del espacio vectorial V. El nulo de D es el conjunto $N(D)$ de todas las funcionales en V^* tal que se nulifican en todo el subconjunto D, es decir*

$$N(D) = \{f \in V^* \mid f(v) = 0 \quad \forall v \in D\}. \tag{341}$$

Teorema 50 *Sea V un espacio vectorial y V^* el espacio dual de V, entonces*

a) $N(D)$ es un subespacio de V^*

b) Si $M \subset V$ es un subespacio de dimensión m, V tiene dimensión n, entonces $N(M)$ tiene dimensión $n - m$ en V^*.

Corolario 51 *Si $V = L \oplus M$ (suma directa) entonces $V^* = N(L) \oplus N(M)$.*

Teorema 52 *Sean V y W espacios lineales, si $T(\cdot) : V \to W$ es lineal, entonces el adjunto T^* de T es un operador lineal $T^* : W^* \to V^*$ definido por*

$$T^*(w^*)(u) = w^*(Tu). \tag{342}$$

Teorema 53 *Si H es un espacio completo con producto interior, entonces $H^* = H$.*

Definición 54 *Si V es un espacio vectorial con producto interior y $T(\cdot): V \to V$ es una transformación lineal, entonces existe una transformación asociada a T llamada la transformación auto-adjunta T^* definida como*

$$\langle Tu, v\rangle = \langle u, T^* v\rangle . \tag{343}$$

Definición 55 *Sea V un espacio vectorial sobre los reales. Se dice que una función $\tau(\cdot,\cdot): V\times V \to \mathbb{R}$ es una forma bilineal sobre V, si para toda $x, y, z \in V$ y $\alpha, \beta \in \mathbb{R}$ se tiene*

$$\begin{aligned} \tau(\alpha x + \beta y, z) &= \alpha\tau(x, z) + \beta\tau(y, z) \\ \tau(x, \alpha y + \beta z) &= \alpha\tau(x, y) + \beta\tau(x, z) . \end{aligned} \tag{344}$$

Definición 56 *Si $\tau(\cdot,\cdot)$ es una forma bilineal sobre V, entonces la función $q_\tau(\cdot): V \to \mathbb{R}$ definida por*

$$q_\tau(x) = \tau(x, x) \quad \forall x \in V \tag{345}$$

se le llama la forma cuadrática asociada a τ.

Notemos que para una forma cuadrática $q_\tau(\cdot)$ se tiene que $q_\tau(\alpha x) = |\alpha|^2 q_\tau(x)$ $\forall x \in V$ y $\alpha \in \mathbb{R}$.

Definición 57 *Sea $V \subset \mathbb{R}^n$ un subespacio, $P \in \mathbb{R}^n \times \mathbb{R}^n$*

7.2. σ-Algebra y Espacios Medibles

A continuación detallaremos algunos resultados conjuntos de espacios σ-algebra, conjuntos de medida cero y funciones medibles, las demostraciones de los siguientes resultados puede ser consultada en [23] y [3].

Definición 58 *Una σ-algebra sobre un conjunto Ω es una familia ξ de subconjuntos de Ω que satisface*

- $\varnothing \in \xi$
- *Si* $\psi_n \in \xi$ *entonces* $\bigcup_{n=1}^{\infty} \psi_n \in \xi$
- *Si* $\psi \in \xi$ *entonces* $\psi^c \in \xi$.

Definición 59 *Si Ω es un espacio topológico, la familia de Borel es el conjunto σ-algebra más pequeño que contiene a los abiertos del conjunto Ω.*

Definición 60 *Una medida μ sobre Ω es una función no negativa real valuada cuyo dominio es una σ-algebra ξ sobre Ω que satisface*

- $\mu(\varnothing) = 0$ *y*

- *Si* $\{\psi_n\}$ *es una sucesión de conjuntos ajenos de* ξ *entonces*

$$\mu\left(\bigcup_{n=1}^{\infty}\psi_n\right)=\sum_{n=1}^{\infty}\mu\left(\psi_n\right). \tag{346}$$

Teorema 61 *Existe una función de medida* μ *sobre el conjunto de Borel de* $\mathbb{R}$ *llamada la medida de Lebesgue que satisface* $\mu\left([a,b]\right)=b-a$.

Definición 62 *Una función* $f:\Omega\to\mathbb{R}$ *es llamada medible si* $f^{-1}\left(U\right)$ *es un conjunto medible para todo abierto* U *de* Ω.

Definición 63 *Sea* $E\subset\Omega$ *un conjunto, se dice que el conjunto* E *tiene medida cero si* $\mu\left(E\right)=0$.

Teorema 64 *Si* α *es una medida sobre el espacio* X *y* β *es una medida sobre el espacio* Y, *podemos definir una medida* μ *sobre* $X\times Y$ *con la propiedad de que* $\mu\left(A\times B\right)=\alpha\left(A\right)\beta\left(B\right)$ *para todo conjunto medible* $A\in X$ *y* $B\in Y$.

Teorema 65 *(Fubini)*
Si $f\left(x,y\right)$ *es medible en* $X\times Y$ *entonces*

$$\int\limits_{X\times Y}f\left(x,y\right)d\mu=\int\limits_{X}\int\limits_{Y}f\left(x,y\right)d\beta d\alpha=\int\limits_{Y}\int\limits_{X}f\left(x,y\right)d\alpha d\beta \tag{347}$$

en el sentido de que cualquiera de las integrales existe y son iguales.

Teorema 66 *Una función* f *es integrable en el sentido de Riemann en* Ω *si y sólo si el conjunto de puntos donde* $f\left(\underline{x}\right)$ *es no continua tiene medida cero.*

Observación 67 *Sean* f *y* g *dos funciones definidas en* Ω, *decimos que* f *y* g *son iguales salvo en un conjunto de medida cero si* $f(x)\neq g(x)$ *sólo en un conjunto de medida cero.*

Definición 68 *Una propiedad* P *se dice que se satisface en casi todos lados, si existe un conjunto* E *con* $\mu\left(E\right)=0$ *tal que la propiedad se satisface en todo punto de* E^c.

7.3. Espacios L^p

Las definiciones y material adicional puede ser consultada en [13], [19] y [3].

Definición 69 *Una función medible* $f\left(\cdot\right)$ *(en el sentido de Lebesgue) es llamada integrable sobre un conjunto medible* $\Omega\subset\mathbb{R}^n$ *si*

$$\int_{\Omega}\left|f\right|d\underline{x}<\infty. \tag{348}$$

Definición 70 *Sea p un número real con $p \geq 1$. Una función $u(\cdot)$ definida sobre $\Omega \subset \mathbb{R}^n$ se dice que pertenece al espacio $L^p(\Omega)$ si*

$$\int_{\Omega} |u(\underline{x})|^p \, d\underline{x} \tag{349}$$

es integrable.

Al espacio $L^2(\Omega)$ se le llama cuadrado integrable.

Definición 71 *La norma $L^2(\Omega)$ se define como*

$$\|u\|_{L^2(\Omega)} = \left(\int_{\Omega} |u(\underline{x})|^2 \, d\underline{x} \right)^{\frac{1}{2}} < \infty \tag{350}$$

y el producto interior en la norma $L^2(\Omega)$ como

$$\langle u, v \rangle_{L^2(\Omega)} = \int_{\Omega} u(\underline{x}) v(\underline{x}) d\underline{x}. \tag{351}$$

Definición 72 *Si $p \to \infty$, entonces definimos al espacio $L^\infty(\Omega)$ como el espacio de todas las funciones medibles sobre $\Omega \subset \mathbb{R}^n$ que sean acotadas en casi todo Ω (excepto posiblemente sobre un conjunto de medida cero), es decir,*

$$L^\infty(\Omega) = \{u \mid |u(x)| \leq k\} \tag{352}$$

definida en casi todo Ω, para algún $k \in \mathbb{R}$.

7.4. Distribuciones

La teoría de distribuciones es la base para definir a los espacios de Sobolev, ya que permiten definir las derivadas parciales de funciones no continuas, pero esta es coincidente con las derivadas parciales clásica si las funciones son continuas, para mayor referencia de estos resultados ver [13], [19] y [3]

Definición 73 *Sea $\Omega \subset \mathbb{R}^n$ un dominio, al conjunto de todas las funciones continuas definidas en Ω se denotarán por $C^0(\Omega)$, o simplemente $C(\Omega)$.*

Definición 74 *Sea u una función definida sobre un dominio Ω la cual es no cero sólo en los puntos pertenecientes a un subconjunto propio $K \subset \Omega$. Sea $\overline{K}$ la clausura de K. Entonces $\overline{K}$ es llamado el soporte de u. Decimos que u tiene soporte compacto sobre Ω si su soporte $\overline{K}$ es compacto. Al conjunto de funciones continuas con soporte compacto se denota por $C_0(\Omega)$.*

Definición 75 *Sea $\mathbb{Z}_+^n$ el conjunto de todas las n-duplas de enteros no negativos, un miembro de $\mathbb{Z}_+^n$ se denota usualmente por α ó β (por ejemplo $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$. Denotaremos por $|\alpha|$ la suma $|\alpha| = \alpha_1 + \alpha_2 + ... + \alpha_n$ y por $D^\alpha u$ la derivada parcial*

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} \tag{353}$$

así, si $|\alpha| = m$, entonces $D^\alpha u$ denota la m-ésima derivada parcial de u.

Definición 76 *Sea $C^m(\Omega)$ el conjunto de todas las funciones $D^\alpha u$ tales que sean funciones continuas con $|\alpha| = m$. Y $C^\infty(\Omega)$ como el espacio de funciones en el cual todas las derivadas existen y sean continuas en Ω.*

Definición 77 *El espacio $\mathcal{D}(\Omega)$ será el subconjunto de funciones infinitamente diferenciales con soporte compacto, algunas veces se denota también como $C_0^\infty(\Omega)$.*

Definición 78 *Una distribución sobre un dominio $\Omega \subset \mathbb{R}^n$ es toda funcional lineal continua sobre $\mathcal{D}(\Omega)$.*

Definición 79 *El espacio de distribuciones es el espacio de todas las funcionales lineales continuas definidas en $\mathcal{D}(\Omega)$, denotado como $\mathcal{D}^*(\Omega)$, es decir el espacio dual de $\mathcal{D}(\Omega)$.*

Definición 80 *Un función $f(\cdot)$ es llamada localmente integrable, si para todo subconjunto compacto $K \subset \Omega$ se tiene*

$$\int_K |f(x)|\, dx < \infty. \tag{354}$$

Ejemplo de una distribución es cualquier función $f(\cdot)$ localmente integrable en Ω. La distribución F asociada a f se puede definir de manera natural como $F : \mathcal{D}(\Omega) \to \mathbb{R}$ como

$$\langle F, \phi \rangle = \int_\Omega f\phi dx \tag{355}$$

con $\phi \in \mathcal{D}(\Omega)$.

Si el soporte de ϕ es $K \subset \Omega$, entonces

$$|\langle F, \phi \rangle| = \left| \int_\Omega f\phi dx \right| = \left| \int_K f\phi dx \right| \leq \sup_{x \in K} |\phi| \int_\Omega |f(x)|\, dx \tag{356}$$

la integral es finita y $\langle F, \phi \rangle$ tiene sentido. Bajo estas circunstancias F es llamada una distribución generada por f.

Otro ejemplo de distribuciones es el generado por todas las funciones continuas acotadas, ya que estas son localmente integrables y por lo tanto generan una distribución.

Definición 81 *Si una distribución es generada por funciones localmente integrables es llamada una distribución regular. Si una distribución no es generada por una función localmente integrable, es llamada distribución singular (ejemplo de esta es la delta de Dirac).*

Es posible definir de manera natural en producto de una función y una distribución. Específicamente, si $\Omega \subset \mathbb{R}^n$, u pertenece a $C^\infty(\Omega)$, y si $f(\cdot)$ es una distribución sobre Ω, entonces entenderemos uf por la distribución que satisface

$$\langle (uf), \phi \rangle = \langle f, u\phi \rangle \tag{357}$$

para toda $\phi \in \mathcal{D}(\Omega)$. Notemos que la anterior ecuación es una generalización de la identidad

$$\int_{\Omega} \left[u\left(x\right) f\left(x\right)\right] \phi\left(x\right) dx = \int_{\Omega} f\left(x\right) \left[u\left(x\right) \phi\left(x\right)\right] dx \tag{358}$$

la cual se satisface si f es localmente integrable.

Derivadas de Distribuciones Funciones como la delta de Dirac y la Heaviside no tienen derivada en el sentido ordinario, sin embargo, si estas funciones son tratadas como distribuciones es posible extender el concepto de derivada de tal forma que abarque a dichas funciones, para ello recordemos que:

Teorema 82 *La versión clásica del teorema de Green es dada por la identidad*

$$\int_{\Omega} u \frac{\partial v}{\partial x_i} d\underline{x} = \int_{\partial\Omega} uvn_i d\underline{s} - \int_{\Omega} v \frac{\partial u}{\partial x_i} d\underline{x} \tag{359}$$

que se satisface para todas las funciones u, v en $C^1(\overline{\Omega})$, donde n_i es la i-ésima componente de la derivada normal del vector n en la frontera $\partial\Omega$ de un dominio Ω.

Una versión de la Ec. (359) en una dimensión se obtiene usando la formula de integración por partes, quedando como

$$\int_a^b uv'dx = \left[uv\right]|_a^b - \int_a^b vu'dx, \quad u, v \in C^1\left[a, b\right] \tag{360}$$

como un caso particular de la Ec. (359).

Este resultado es fácilmente generalizable a un resultado usando derivadas parciales de orden m de funciones $u, v \in C^m(\overline{\Omega})$ pero remplazamos u por $D^\alpha u$ en la Ec. (359) y con $|\alpha| = m$, entonces se puede mostrar que:

Teorema 83 *Otra versión del teorema de Green es dado por*

$$\int_{\Omega} \left(D^\alpha u\right) v d\underline{x} = (-1)^{|\alpha|} \int_{\Omega} u D^\alpha v d\underline{x} + \int_{\partial\Omega} h(u, v) d\underline{s} \tag{361}$$

donde $h(u, v)$ es una expresión que contiene la suma de productos de derivadas de u y v de orden menor que m.

Ahora remplazando v en la Ec. (361) por ϕ perteneciente a $\mathcal{D}(\Omega)$ y como $\phi = 0$ en la frontera $\partial\Omega$ tenemos

$$\int_{\Omega} \left(D^\alpha u\right) \phi d\underline{x} = (-1)^{|\alpha|} \int_{\Omega} u D^\alpha \phi d\underline{x} \tag{362}$$

ya que u es m-veces continuamente diferenciable, esta genera una distribución denotada por u, tal que

$$\langle u, \phi \rangle = \int_{\Omega} u\phi d\underline{x} \tag{363}$$

o, como $D^{\alpha}\phi$ también pertenece a $\mathcal{D}(\Omega)$, entonces

$$\langle u, D^{\alpha}\phi\rangle = \int_{\Omega} uD^{\alpha}\phi d\underline{x} \tag{364}$$

además, $D^{\alpha}u$ es continua, así que es posible generar una distribución regular denotada por $D^{\alpha}u$ satisfaciendo

$$\langle D^{\alpha}u, \phi\rangle = \int_{\Omega} \left(D^{\alpha}u\right)\phi d\underline{x} \tag{365}$$

entonces la Ec. (362) puede reescribirse como

$$\langle D^{\alpha}u, \phi\rangle = (-1)^{|\alpha|}\langle u, D^{\alpha}\phi\rangle, \quad \forall\phi \in D(\Omega). \tag{366}$$

Definición 84 *La derivada de cualquier distribución $f(\cdot)$ se define como: La $\alpha-$ésima derivada parcial distribucional o derivada generalizada de una distribución f es definida por una distribución denotada por $D^{\alpha}f$, que satisface*

$$\langle D^{\alpha}f, \phi\rangle = (-1)^{|\alpha|}\langle f, D^{\alpha}\phi\rangle, \quad \forall\phi \in \mathcal{D}(\Omega).$$

Nótese que si f pertenece a $C^{m}\left(\bar{\Omega}\right)$, entonces la derivada parcial distribucional coincide con la derivada parcial $\alpha-$ésima para $|\alpha| \leq m$.

Derivadas Débiles Supóngase que una función $u(\cdot)$ es localmente integrable que genere una distribución, también denotada por u, que satisface

$$\langle u, \phi\rangle = \int_{\Omega} u\phi dx \tag{367}$$

para toda $\phi \in \mathcal{D}(\Omega)$.

Además la distribución u posee derivada distribucional de todos los ordenes, en particular la derivada $D^{\alpha}u$ es definida por

$$\langle D^{\alpha}u, \phi\rangle = (-1)^{|\alpha|}\langle u, D^{\alpha}\phi\rangle, \quad \forall\phi \in D(\Omega). \tag{368}$$

por supuesto $D^{\alpha}u$ puede o no ser una distribución regular. Si es una distribución regular, entonces es generada por una función localmente integrable tal que

$$\langle D^{\alpha}u, \phi\rangle = \int_{\Omega} D^{\alpha}u(x)\,\phi(x)\,d\underline{x} \tag{369}$$

y se sigue que la función u y $D^{\alpha}u$ están relacionadas por

$$\int_{\Omega} D^{\alpha}u(x)\,\phi(x)\,d\underline{x} = (-1)^{|m|}\int_{\Omega} u(x)\,D^{\alpha}\phi(x)\,d\underline{x} \tag{370}$$

para $|\alpha| \leq m$.

Definición 85 *Llamamos a la función (o más precisamente, a la equivalencia de clases de funciones) $D^{\alpha}u$ obtenida en la Ec. (370), la $\alpha-$ésima derivada débil de la función u.*

Notemos que si u pertenece a $C^{m}\left(\bar{\Omega}\right)$, entonces la derivada $D^{\alpha}u$ coincide con la derivada clásica para $|\alpha| \leq m$.

8. Bibliografía

Referencias

[1] A. Quarteroni, A. Valli; *Domain Decomposition Methods for Partial Differential Equations*. Clarendon Press Oxford 1999.

[2] A. Toselli, O. Widlund; *Domain Decomposition Methods - Algorithms and Theory*. Springer, 2005.

[3] B. D. Reddy; *Introductory Functional Analysis - With Applications to Boundary Value Problems and Finite Elements*. Springer 1991.

[4] B. F. Smith, P. E. Bjørstad, W. D. Gropp; *Domain Decomposition, Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996.

[5] B. I. Wohlmuth; *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Springer, 2003.

[6] I. Foster; *Designing and Building Parallel Programs*. Addison-Wesley Inc., Argonne National Laboratory, and the NSF, 2004.

[7] G. Herrera; Análisis de Alternativas al Método de Gradiente Conjugado para Matrices no Simétricas. Tesis de Licenciatura, Facultad de Ciencias, UNAM, 1989.

[8] I. Herrera, M. Díaz; *Modelación Matemática de Sistemas Terrestres* (Notas de Curso en Preparación). Instituto de Geofísica, (UNAM).

[9] I. Herrera;*Un Análisis del Método de Gradiente Conjugado*. Comunicaciones Técnicas del Instituto de Geofísica, UNAM; Serie Investigación, No. 7, 1988.

[10] I. Herrera; *Método de Subestructuración* (Notas de Curso en Preparación). Instituto de Geofísica, (UNAM).

[11] I. Herrera y R. Yates; UnifieD Multipliers-Free Theory of Dual-Primal Domain Decomposition Methods, ...

[12] J. II. Bramble, J. E. Pasciak and A. II Schatz. *The Construction of Preconditioners for Elliptic Problems by Substructuring*. I. Math. Comput., 47, 103-134,1986.

[13] J. L. Lions & E. Magenes; *Non-Homogeneous Bonduary Value Problems and Applications Vol. I*, Springer-Verlag Berlin Heidelber New York 1972.

[14] K. Hutter & K. Jöhnk; *Continuum Methods of Physical Modeling*. Springer-Verlag Berlin Heidelber New York 2004.

[15] L. F. Pavarino, A. Toselli; *Recent Developments in Domain Decomposition Methods*. Springer, 2003.

[16] M.B. Allen III, I. Herrera & G. F. Pinder; *Numerical Modeling in Science And Engineering.* John Wiley & Sons, Inc . 1988.

[17] M. Diaz; *Desarrollo del Método de Colocación Trefftz-Herrera Aplicación a Problemas de Transporte en las Geociencias*. Tesis Doctoral, Instituto de Geofísica, UNAM, 2001.

[18] M. Diaz, I. Herrera; *Desarrollo de Precondicionadores para los Procedimientos de Descomposición de Dominio*. Unidad Teórica C, Posgrado de Ciencias de la Tierra, 22 pags, 1997.

[19] P.G. Ciarlet, J. L. Lions; *Handbook of Numerical Analysis, Vol. II*. North-Holland, 1991.

[20] R. L. Burden y J. D. Faires; *Análisis Numérico*. Math Learning, 7 ed. 2004.

[21] S. Friedberg, A. Insel, and L. Spence; *Linear Algebra*, 4th Edition, Prentice Hall, Inc. 2003.

[22] W. Gropp, E. Lusk, A. Skjellem, *Using MPI, Portable Parallel Programming Whit the Message Passing Interface*. Scientific and Engineering Computation Series, 2ed, 1999.

[23] W. Rudin; *Principles of Mathematical Analysis*. McGraw-Hill International Editions, 1976.

[24] X. O. Olivella, C. A. de Sacribar; *Mecánica de Medios Continuos para Ingenieros.* Ediciones UPC, 2000.

[25] Y. Saad; *Iterative Methods for Sparse Linear Systems*. SIAM, 2 ed. 2000.

[26] Y. Skiba; *Métodos y Esquemas Numéricos, un Análisis Computacional*. UNAM, 2005.