Unidad Teórica B: Estudio de los métodos Galerkin Discontinuo, Discontinuo Enriquecido, FETI y Trefftz-Herrera para el tratamiento de ecuaciones diferenciales parciales elípticas.

Antonio Carrillo Ledesma

9 de diciembre de 2008

Índice

1. Análisis Funcional y Problemas Variacionales

En este capítulo 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 [12], [18], [3], [**?**] y [**?**].

1.1. Operador Lineal Elíptico

Definición 1 *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 2 *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*

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} \tag{1}$$

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

$$\mathcal{L}u = \sum_{|\alpha|,|\beta| \leq m} (-1)^{|\alpha|} D^\alpha \left(a_{\alpha\beta}(\underline{x}) D^\beta u \right), \quad \underline{x} \in \Omega \subset \mathbb{R}^n \tag{2}$$

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 con la forma

$$\mathcal{L}u = f, \tag{3}$$

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 (2) es de la forma

$$\mathcal{L}_0 = \sum_{|\alpha|,|\beta| \leq m} a_{\alpha\beta} D^{\alpha+\beta} u. \tag{4}$$

Teorema 3 *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

$$\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}\left(\underline{x_o}\right)\xi^{\alpha+\beta} \neq 0 \qquad \forall \xi \neq 0; \tag{5}$$

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

$$\left|\sum_{|\alpha|,|\beta|=m} a_{\alpha\beta}\left(\underline{x_o}\right)\xi^{\alpha+\beta}\right| \geq \mu\left|\xi\right|^{2m} \tag{6}$$

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

$$\mathcal{L}u = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\underline{x})\frac{\partial u}{\partial x_j}\right) + \sum_{j=1}^{n} a_j \frac{\partial u}{\partial x_j} + a_0 u = f \tag{7}$$

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

$$\sum_{i,j=1}^{n} a_{ij}(\underline{x_0})\xi_i\xi_j \neq 0 \qquad \forall \xi \neq 0 \tag{8}$$

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

$$\sum_{i,j=1}^{n} a_{ij}(\underline{x_0})\xi_i\xi_j > \mu\left|\xi\right|^2. \tag{9}$$

1.2. 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 4 *Una función medible $u(\underline{x})$ definida sobre $\Omega \subset \mathbb{R}^n$ se dice que pertenece al espacio $L^2(\Omega)$ si*

$$\int_\Omega \left|u(\underline{x})\right|^2 d\underline{x} < \infty \tag{10}$$

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

$$H^m(\Omega) = \left\{ u \mid D^\alpha u \in L^2(\Omega) \quad \forall \alpha \text{ tal que } |\alpha| \leq m \right\}. \tag{11}$$

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

$$\langle u, v \rangle_{H^m} = \int_\Omega \sum_{|\alpha| \leq m} \left(D^\alpha u\right) \left(D^\alpha v\right) d\underline{x} \text{ para } u, v \in H^m\left(\Omega\right). \tag{12}$$

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

$$\|u\|_{H^m}^2 = \langle u, u \rangle_{H^m} = \int_\Omega \sum_{|\alpha| \leq m} \left(D^\alpha u\right)^2 d\underline{x}. \tag{13}$$

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

Teorema 7 *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 8 *Una función f definida sobre un conjunto $\Gamma \subset \mathbb{R}^n$ es llamada Lipschitz continua si existe una constante $L > 0$ tal que*

$$|f(x) - f(y)| \leq L\,|x - y| \quad \forall x, y \in \Gamma. \tag{14}$$

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

$$\xi_n = f(\xi_1, ..., \xi_{n-1}) \tag{15}$$

entonces definimos.

Definición 9 *La frontera $\partial\Omega$ del dominio Ω es llamada de Lipschitz si f definida como en la Ec. (15) es una función Lipschitz continua.*

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

Teorema 10 *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

$$\|u - f\|_{H^m} < \varepsilon \tag{16}$$

para algún $\varepsilon > 0$ dado.

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

Corolario 11 *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 12 *Sean X y Y dos espacios de Banach, con $X \subset Y$. Sea $f : X \to Y$ tal que $f(u) = u$. 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*

$$\|f(u)\|_Y = \|u\|_Y \leq K \|u\|_X \tag{17}$$

para alguna constante $K > 0$.

Teorema 13 *(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

$$H^m(\Omega) \subset C^k(\overline{\Omega}) \tag{18}$$

es continua.

1.2.1. 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 14 *Sea Ω un dominio con frontera $\partial\Omega$ de Lipschitz. La estimación*

$$\|tr(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)} \tag{19}$$

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 15 *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

$$\|tr(u)\|_{L^2(\partial\Omega)} \leq C\|u\|_{H^1(\Omega)}, \tag{20}$$

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

$$D^\alpha u \in H^1(\Omega) \qquad \text{para } |\alpha| \leq m-1, \tag{21}$$

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

$$tr(D^\alpha u) \in L^2(\Omega), \qquad |\alpha| \leq m-1. \tag{22}$$

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

$$tr(D^\alpha u) = (D^\alpha u)|_{\partial\Omega}. \tag{23}$$

1.2.2. 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 16 *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 17 *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*

$$D^\alpha u = 0, \text{ sobre } \partial\Omega, |\alpha| \leq m-1. \tag{24}$$

Teorema 18 *(Desigualdad de Poincaré-Friedrichs)*

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

$$\int_\Omega |u|^2 \, dx \leq C \int_\Omega |\nabla u|^2 \, dx \tag{25}$$

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

$$|u|_m^2 = \sum_{|\alpha|=m} \int_\Omega |D^\alpha u|^2 \, dx. \tag{26}$$

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 20 *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 21 *Sea Ω un dominio acotado en $\mathbb{R}^n$. Entonces existe una constante $C > 0$ tal que*

$$\|u\|_{L^2}^2 \leq C\,|u|_m^2 \tag{27}$$

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 22 *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 23 *q será una distribución de $H^{-m}(\Omega)$ si y sólo si q puede ser expresada en la forma*

$$q = \sum_{|\alpha|<m} D^\alpha q_\alpha \tag{28}$$

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

Algunos Comentarios y Precisiones Sea Ω un dominio tal que la frontera $\partial\Omega$ es suficientemente suave (considerando sólo frontera Lipschitz continua), entonces existe un operador $\gamma_0 : H^1(\Omega) \to L^2(\Omega)$ lineal y continuo, tal que $\gamma_0 v = tr(v)$ sobre $\partial\Omega$ para toda v suave (por ejemplo $v \in C^1\left(\overline{\Omega}\right)$), un análisis más profundo muestra que tomando las trazas de todas las funciones de $H^1(\Omega)$ uno no obtiene el espacio completo de $L^2(\Omega)$, sólo obtiene un subespacio de este. Tenemos entonces

$$H^1(\partial\Omega) \subset \gamma_0\left(H^1(\Omega)\right) \subset L^2(\partial\Omega) \equiv H^0(\partial\Omega) \tag{29}$$

donde cada inclusión es estricta.

Finalmente reconocemos que el espacio $\gamma_0\left(H^1(\Omega)\right)$ pertenece a la familia de espacios $H^s(\partial\Omega)$ y corresponde exactamente a el valor de $s = 1/2$. De tal forma que

$$H^{1/2}(\partial\Omega) = \gamma_0\left(H^1(\Omega)\right) \tag{30}$$

con

$$\|g\|_{H^{1/2}(\partial\Omega)} = \inf_{v \in H^1(\Omega) \text{ y } \gamma_0 v = g} \|v\|_{H^1(\Omega)} \tag{31}$$

de forma similar, se ve que las trazas de las funciones en $H^2(\Omega)$ pertenecen a el espacio $H^s(\partial\Omega)$ para $s = 3/2$, por lo tanto tenemos que

$$H^{3/2}(\partial\Omega) = \gamma_0\left(H^2(\Omega)\right) \tag{32}$$

$$\|g\|_{H^{3/2}(\partial\Omega)} = \inf_{v \in H^2(\Omega) \text{ y } \gamma_0 v = g} \|v\|_{H^2(\Omega)}. \tag{33}$$

Esto puede generalizarse a las trazas de derivadas de orden alto. Por ejemplo, si la frontera $\partial\Omega$ es suficientemente suave, podemos definir $\frac{\partial v}{\partial n}_{|\partial\Omega} \in H^{1/2}(\partial\Omega)$ para $v \in H^2(\Omega)$.

Por otro lado, para ejemplificar algunos casos de H_0^m notemos que

$$\begin{aligned} H_0^1(\Omega) &= \left\{v \mid v \in H^1(\Omega), v_{|\partial\Omega} = 0\right\} \\ H_0^2(\Omega) &= \left\{v \mid v \in H^2(\Omega), v_{|\partial\Omega} = 0 \text{ y } \frac{\partial v}{\partial n}_{|\partial\Omega} = 0\right\}. \end{aligned} \tag{34}$$

Además, en algunas ocasiones necesitamos considerar funciones que se nulifican en alguna parte de la frontera, supongamos que $\partial\Omega = D \cup N$, donde D es frontera tipo Dirichlet y N es frontera tipo Neumann y $D \cap N = \emptyset$, entonces podemos definir

$$H_{0,D}^1(\Omega) = \left\{v \mid v \in H^1(\Omega), v_{|_D} = 0\right\} \tag{35}$$

y donde tenemos que $H_0^1(\Omega) \subset H_{0,D}^1(\Omega) \subset H^1(\Omega)$.

1.2.3. Espacios $H(div, \Omega)$

Definición 24 *Sea Ω un dominio acotado en $\mathbb{R}^n$. Definimos a $\left(L^2(\Omega)\right)^n$ al espacio*

$$\left(L^2(\Omega)\right)^n = \left\{grad\ H^1(\Omega)\right\} \oplus \left\{rot\ H_0^1(\Omega)\right\}. \tag{36}$$

Definición 25 *Sea Ω un dominio acotado en $\mathbb{R}^n$. Definimos por $H(div, \Omega)$ al espacio*

$$H(div, \Omega) = \left\{\underline{q} \mid \underline{q} \in \left(L^2(\Omega)\right)^n,\ div\ \underline{q} \in L^2(\Omega)\right\}. \tag{37}$$

Definición 26 *La norma $\|\cdot\|^2_{H(div,\Omega)}$ de $H(div, \Omega)$, se define como*

$$\left\|\underline{q}\right\|^2_{H(div,\Omega)} = \left\|\underline{q}\right\|^2_{0,\Omega} + \left\|div\ \underline{q}\right\|^2_{0,\Omega}. \tag{38}$$

Cuando $H(div, \Omega)$ es equipada con la norma $\|\cdot\|^2_{H(div,\Omega)}$ el correspondiente producto interior se convierte en un espacio de Hilbert.

Notemos que, si Ω es un dominio acotado en $\mathbb{R}^n$, con frontera suave $\partial\Omega$, si $\underline{n}$ es un vector normal a $\partial\Omega$ y sea $\underline{q} \in H(div, \Omega)$, entonces los vectores de $H(div, \Omega)$ admiten una norma de la traza sobre $\partial\Omega$. Esta norma de la traza $\underline{q} \cdot \underline{n}$ pertenece a $H^{-1/2}(\partial\Omega)$ y esto se sigue de la formula de integración por partes

$$\int_\Omega \underline{q} \cdot \text{grad}\ v dx + \int_\Omega \text{div}\ \underline{q} v dx = \left\langle v, \underline{q} \cdot \underline{n} \right\rangle_{H^{1/2}(\partial\Omega) \times H^{-1/2}(\partial\Omega)} \tag{39}$$

para toda $\underline{q} \in H(div, \Omega)$ y cualquier $v \in H^1(\Omega)$. Pudiendo escribir formalmente $\int_{\partial\Omega} v \underline{q} \cdot \underline{n} ds$ en lugar del producto dual $\left\langle v, \underline{q} \cdot \underline{n} \right\rangle$.

Lema 27 *Sea $\underline{q} \in H(div, \Omega)$, podemos definir $\underline{q} \cdot \underline{n}_{|\partial\Omega} \in H^{-1/2}(\partial\Omega)$ y por la formula de Green*

$$\int_\Omega div\ \underline{q} v dx + \int_\Omega \underline{q} \cdot grad\ v dx = \left\langle v, \underline{q} \cdot \underline{n} \right\rangle \tag{40}$$

para toda $v \in H^1(\Omega)$.

Lema 28 *La traza del operador* $\underline{q} \in H(div,\Omega) \rightarrow \underline{q}\cdot\underline{n}_{|_{\partial\Omega}} \in H^{-1/2}(\partial\Omega)$ *es suprayectivo.*

Sea Ω un dominio con frontera suave $\partial\Omega$, además supongamos que es frontera tipo Neumann $N=\partial\Omega$, entonces podemos definir

$$H_{0,N}(div,\Omega)=\left\{\underline{q}\mid \underline{q}\in H(div,\Omega),\left\langle v,\underline{q}\cdot\underline{n}\right\rangle=0,\forall v\in H^1_{0,D}(\Omega)\right\}. \tag{41}$$

este espacio contiene funciones del espacio $H(div,\Omega)$ cuyas trazas normales se nulifican en la frontera N.

Definición 29 *Un subespacio importante de* $H(div,\Omega)$ *es* $N^0(div,\Omega)$, *que se define como*

$$N^0(div,\Omega)=\left\{\underline{q}\mid \underline{q}\in H(div,\Omega),\ div\ \underline{q}=0\right\}. \tag{42}$$

Lema 30 *El operador de traza normal* $\underline{q}\rightarrow\underline{q}\cdot\underline{n}_{|_{\partial\Omega}}$ *es una forma suprayectiva* $N^0(div,\Omega)$ *sobre* $\left\{\mu\mid\mu\in H^{-1/2}(\partial\Omega),\langle\mu,1\rangle=0\right\}$.

1.3. 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

$$\begin{array}{rcll}\mathcal{L}u & = & f_\Omega & \text{en } \Omega\subset\mathbb{R}^n \\ & & \left.\begin{array}{c} B_o u=g_0 \\ B_1 u=g_1 \\ \vdots \\ B_{m-1}u=g_{m-1}\end{array}\right\} & \text{en } \partial\Omega\end{array} \tag{43}$$

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

$$\mathcal{L}u=\sum_{|\alpha|\leq m}(-1)^{|\alpha|}D^\alpha\left(\sum_{|\beta|\leq m}a_{\alpha\beta}(\underline{x})D^\beta u\right),\quad \underline{x}\in\Omega\subset\mathbb{R}^n \tag{44}$$

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

$$B_j u=\sum_{|\alpha|\leq q_j}b_\alpha^{(j)}D^\alpha u=g_j \tag{45}$$

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. (45) puede expresarse como una sola condición de frontera

$$Bu = \sum_{j=1}^{n} b_j \frac{\partial u}{\partial x_j} + cu = g \quad \text{en } \partial\Omega. \tag{46}$$

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 31 *Con el operador dado como en la Ec. (44), denotaremos por $\mathcal{L}^*$ al operador definido por*

$$\mathcal{L}^* u = \sum_{|\alpha| \leq m} (-1)^{|\alpha|} D^\alpha \left(\sum_{|\beta| \leq m} a_{\beta\alpha}(\underline{x}) D^\beta u \right) \tag{47}$$

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 (115) a la integral

$$\int_\Omega v \mathcal{L} u d\underline{x} \tag{48}$$

obtenemos

$$\int_\Omega v \mathcal{L} u d\underline{x} = \int_\Omega u \mathcal{L}^* v d\underline{x} + \int_{\partial\Omega} F(u,v) d\underline{s} \tag{49}$$

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 (115) y obtenemos, para i y j fijos

$$\begin{aligned} -\int_\Omega v \frac{\partial}{\partial x_i} \left(a_{ij} \frac{\partial u}{\partial x_j} \right) d\underline{x} &= -\int_{\partial\Omega} v a_{ij} \frac{\partial u}{\partial x_j} n_i d\underline{s} + \int_\Omega a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} d\underline{x} \\ &= -\int_{\partial\Omega} \left[v a_{ij} \frac{\partial u}{\partial x_j} n_i - u a_{ij} \frac{\partial v}{\partial x_i} n_j \right] d\underline{s} \\ &\quad -\int_\Omega u \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial v}{\partial x_i} \right) d\underline{x}. \end{aligned} \tag{50}$$

Pero sumando sobre i y j, obtenemos de la Ec. (49)

$$\mathcal{L}^* v = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(a_{ji}(\underline{x}) \frac{\partial v}{\partial x_j} \right) \tag{51}$$

y

$$F(u,v) = -\sum_{i,j=1}^{n} a_{ij} \left(v \frac{\partial u}{\partial x_j} n_i - u \frac{\partial v}{\partial x_i} n_j \right) \tag{52}$$

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

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \subset \mathbb{R}^n \\ &\left.\begin{array}{c} B_o u = 0 \\ B_1 u = 0 \\ \vdots \\ B_{m-1} u = 0 \end{array}\right\} \quad \text{en } \partial\Omega \end{aligned} \tag{53}$$

donde el operador $\mathcal{L}$ y B_j estan dados como en (44) y (45), 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

$$N(B_j) = \{u \in H^s(\Omega) \mid B_j u = 0 \text{ en } \partial\Omega\} \tag{54}$$

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

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

$$\begin{aligned} D(\mathcal{L}) &= H^s(\Omega) \cap N(B_0) \cap ... \cap N(B_{m-1}) \\ &= \{u \in H^s(\Omega) \mid B_j u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\}. \end{aligned} \tag{55}$$

Entonces el problema elíptico con valores en la frontera de la Ec. (53) con $s \geq 2m$, puede reescribirse como, dado $\mathcal{L} : D(\mathcal{L}) \to H^{s-2m}(\Omega)$, hallar u que satisfaga

$$\mathcal{L}u = f_\Omega \quad \text{en } \Omega. \tag{56}$$

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

$$\begin{aligned} N(\mathcal{L}) &= \{u \in D(\mathcal{L}) \mid \mathcal{L}u = 0\} \\ &= \{u \in H^s(\Omega) \mid \mathcal{L}u = 0 \quad \text{en } \Omega, B_j u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\}. \end{aligned} \tag{57}$$

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

$$\mathcal{L}(u_0 + w) = \mathcal{L}u_0 + \mathcal{L}w = \mathcal{L}u_0 = f_\Omega. \tag{58}$$

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

$$N(\mathcal{L})^\perp = \{v \in D(\mathcal{L}) \mid (v, w) = 0 \ \forall w \in N(\mathcal{L})\}. \tag{59}$$

De esta forma tenemos que

$$D(\mathcal{L}) = N(\mathcal{L}) \oplus N(\mathcal{L})^\perp \tag{60}$$

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

$$\begin{array}{rcl} \mathcal{L}^* u & = & f_\Omega \quad \text{en } \Omega \subset \mathbb{R}^n \\ & & \left.\begin{array}{c} B_o^* u = 0 \\ B_1^* u = 0 \\ \vdots \\ B_{m-1}^* u = 0 \end{array}\right\} \quad \text{en } \partial\Omega \end{array} \tag{61}$$

y definimos

$$\begin{array}{rcl} D(\mathcal{L}^*) & = & H^s(\Omega) \cap N(B_0^*) \cap ... \cap N(B_{m-1}^*) \\ & = & \left\{u \in H^s(\Omega) \mid B_j^* u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\right\}. \end{array} \tag{62}$$

Entonces el problema elíptico con valores en la frontera de la Ec. (53) con $s \geq 2m$, puede reescribirse como, dado $\mathcal{L}^* : D(\mathcal{L}^*) \to H^{s-2m}(\Omega)$, hallar u que satisfaga

$$\mathcal{L}^* u = f_\Omega \quad \text{en } \Omega. \tag{63}$$

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

$$\begin{array}{rcl} N(\mathcal{L}^*) & = & \{u \in D(\mathcal{L}^*) \mid \mathcal{L}^* u = 0\} \\ & = & \left\{u \in H^s(\Omega) \mid \mathcal{L}^* u = 0 \quad \text{en } \Omega, B_j^* u = 0 \text{ en } \partial\Omega, j = 0, 1, .., m-1\right\}. \end{array} \tag{64}$$

y

$$N(\mathcal{L}^*)^\perp = \{v \in D(\mathcal{L}^*) \mid (v, w)_{L^2} = 0 \ \forall w \in N(\mathcal{L}^*)\}. \tag{65}$$

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. (53) con $s \geq 2m$ tiene solución y esta es única, esto queda resuelto en el siguiente teorema cuya demostración puede verse en [3] y [12].

Teorema 32 *Considerando el problema elíptico con valores en la frontera de la Ec. (53) 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

$$(f, v)_{L^2(\Omega)} = 0 \quad \forall v \in N(\mathcal{L}^*). \tag{66}$$

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

$$(u, w)_{L^2(\Omega)} = 0 \quad \forall w \in N(\mathcal{L}). \tag{67}$$

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

$$\|u\|_{H^s} \leq C \|f\|_{H^{s-2m}}. \tag{68}$$

Observación 33 *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 (67). Además el operador $\mathcal{L}$ es inyectivo si el dominio es restringido al espacio de funciones que satisfagan a (66).*

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

$$u \in H^{s-2m}(\Omega) \;\; si \; f \in H^s(\Omega). \tag{69}$$

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

Definición 34 *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

$$\frac{d^n u}{dx^n} = u \tag{70}$$

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

$$\mathcal{L}u = u \tag{71}$$

de la definición de operador adjunto tenemos que

$$w\mathcal{L}u = u\mathcal{L}^* w + \nabla \cdot \underline{\mathfrak{D}}\,(u, w) \tag{72}$$

entonces el término izquierdo es

$$w\mathcal{L}u = wu \tag{73}$$

de aquí

$$u\mathcal{L}^* w = uw \tag{74}$$

por lo tanto el operador adjunto formal es

$$\mathcal{L}^* w = w \tag{75}$$

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

$$\mathcal{L}u = c\frac{du}{dx} \tag{76}$$

de la definición de operador adjunto tenemos

$$w\mathcal{L}u = u\mathcal{L}w + \nabla \cdot \underline{\mathfrak{D}}\,(u, w) \tag{77}$$

desarrollando el lado izquierdo

$$\begin{aligned} w\mathcal{L}u &= wc\frac{du}{dx} \\ &= \frac{d\,(wcu)}{dx} - u\frac{d\,(cw)}{dx} \\ &= \frac{d\,(wcu)}{dx} - uc\frac{dw}{dx} \end{aligned} \tag{78}$$

por lo tanto, el operador adjunto formal es

$$\mathcal{L}^* w = -c\frac{dw}{dx} \tag{79}$$

y los términos de frontera son

$$\mathfrak{D}\,(u, w) = wcu \tag{80}$$

C) **Operador Elíptico**

El operador elíptico más sencillo es el Laplaciano

$$\mathcal{L}u \equiv -\triangle u = -\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) \tag{81}$$

de la ecuación del operador adjunto formal tenemos

$$\begin{aligned} w\mathcal{L}u &= -w\frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) \\ &= -\frac{\partial}{\partial x_i}\left(w\frac{\partial u}{\partial x_i}\right)+\frac{\partial u}{\partial x_i}\frac{\partial w}{\partial x_i} \\ &= -\frac{\partial}{\partial x_i}\left(w\frac{\partial u}{\partial x_i}\right)+\frac{\partial}{\partial x_i}\left(u\frac{\partial w}{\partial x_i}\right)-u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \\ &= \frac{\partial}{\partial x_i}\left(u\frac{\partial w}{\partial x_i}-w\frac{\partial u}{\partial x_i}\right)-u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \end{aligned} \tag{82}$$

entonces, el operador adjunto formal es

$$\mathcal{L}^*w = -u\frac{\partial}{\partial x_i}\left(\frac{\partial w}{\partial x_i}\right) \tag{83}$$

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

$$\underline{\mathfrak{D}}\,(u,w) = u\frac{\partial w}{\partial x_i}-w\frac{\partial u}{\partial x_i}. \tag{84}$$

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

$$\mathcal{L}u = -\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(\underline{b}u\right)+cu \tag{85}$$

de la definición de operador adjunto formal tenemos que

$$w\mathcal{L}u = u\mathcal{L}^*w+\nabla\cdot\underline{\mathfrak{D}}\,(u,w) \tag{86}$$

desarrollando el lado derecho de la ecuación anterior

$$\begin{aligned} w\mathcal{L}u &= w\left(-\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(\underline{b}u\right)+cu\right) \\ &= -w\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla u\right)+w\nabla\cdot\left(\underline{b}u\right)+wcu \end{aligned} \tag{87}$$

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

$$\begin{aligned} w\mathcal{L}u &= -\nabla\cdot\left(w\underline{\underline{a}}\cdot\nabla u\right)+\underline{\underline{a}}\cdot\nabla u\cdot\nabla w+\nabla\cdot\left(w\underline{b}u\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \\ &= -\nabla\cdot\left(w\underline{\underline{a}}\cdot\nabla u\right)+\nabla\cdot\left(u\underline{\underline{a}}\nabla w\right)-u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right)+\nabla\cdot\left(w\underline{b}u\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \\ &= \nabla\cdot\left[\underline{\underline{a}}\left(u\nabla w-w\nabla u\right)\right]+\nabla\cdot\left(w\underline{b}u\right)-u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right) \\ &\quad -\underline{b}u\cdot\nabla w+wcu \end{aligned} \tag{88}$$

reordenando términos se tiene

$$\begin{aligned} w\mathcal{L}u &= -u\nabla\cdot\left(\underline{\underline{a}}\cdot\nabla w\right)-u\underline{b}\cdot\nabla w+ucw+ \\ &\quad \nabla\cdot\left[\underline{\underline{a}}\left(u\nabla w-w\nabla u\right)+\left(w\underline{b}u\right)\right] \end{aligned} \tag{89}$$

por lo tanto, el operador adjunto formal es

$$\mathcal{L}^* w = -\nabla \cdot \left(\underline{\underline{a}} \cdot \nabla w\right) - \underline{b} \cdot \nabla w + cw \tag{90}$$

y el término correspondiente a valores en la frontera es

$$\underline{\mathfrak{D}}\left(u, w\right) = \underline{\underline{a}} \cdot \left(u \nabla w - w \nabla u\right) + \left(w \underline{b} u\right). \tag{91}$$

E) **La ecuación bi-armónica**

Consideremos el operador diferencial bi-armónico

$$\mathcal{L}u = \Delta^2 u \tag{92}$$

entonces se tiene que

$$w\mathcal{L}u = u\mathcal{L}^* w + \nabla \cdot \underline{\mathfrak{D}}\left(u, w\right) \tag{93}$$

desarrollemos el término del lado derecho

$$\begin{aligned} w\mathcal{L}u &= w\Delta^2 u \\ &= w\nabla \cdot \left(\nabla \Delta u\right) \end{aligned} \tag{94}$$

utilizando la igualdad de divergencia

$$\nabla \cdot \left(sV\right) = s\nabla \cdot V + V \cdot \nabla s \tag{95}$$

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

$$\begin{aligned} & w\nabla \cdot \left(\nabla \Delta u\right) \\ = \;& \nabla \cdot \left(w \nabla \Delta u\right) - \nabla \Delta u \cdot \nabla w \end{aligned} \tag{96}$$

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

$$\begin{aligned} & \nabla \cdot \left(w \nabla \Delta u\right) - \nabla \Delta u \cdot \nabla w \\ = \;& \nabla \cdot \left(w \nabla \Delta u\right) + \Delta u \nabla \cdot \nabla w - \nabla \cdot \left(\Delta u \nabla w\right) \\ = \;& \Delta w \nabla \cdot \nabla u + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \end{aligned} \tag{97}$$

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

$$\begin{aligned} & \Delta w \nabla \cdot \nabla u + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \\ = \;& \nabla \cdot \left(\Delta w \nabla u\right) - \nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u - \Delta u \nabla w\right) \\ = \;& -\nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \end{aligned} \tag{98}$$

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

$$\begin{aligned} & -\nabla u \cdot \nabla \left(\Delta w\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \\ = \;& u \nabla \cdot \left(\nabla \left(\Delta w\right)\right) - \nabla \cdot \left(u \nabla \left(\Delta w\right)\right) + \nabla \cdot \left(w \nabla \Delta u + \Delta w \nabla u - \Delta u \nabla w\right) \end{aligned} \tag{99}$$

reordenando términos

$$w\mathcal{L}u = u\Delta^2 w + \nabla \cdot (w\nabla\Delta u + \Delta w\nabla u - \Delta u\nabla w - u\nabla\Delta w) \tag{100}$$

entonces se tiene que el operador adjunto formal es

$$\mathcal{L}^* w = \Delta^2 w \tag{101}$$

y los términos de frontera son

$$\underline{\mathfrak{D}}\,(u, w) = w\nabla\Delta u + \Delta w\nabla u - \Delta u\nabla w - u\nabla\Delta w. \tag{102}$$

1.4. 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 35 *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*

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} - \underline{u}\ \underline{\underline{\mathcal{L}}}^*\underline{w} = \nabla \cdot \underline{\mathfrak{D}}\,(\underline{u}, \underline{w}) \tag{103}$$

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**

Sea

$$\underline{\underline{\mathcal{L}}}\ \underline{u} = -\nabla \cdot \underline{\underline{\underline{C}}} : \nabla\underline{u} \tag{104}$$

de la definición de operador adjunto formal tenemos que

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} = \underline{u}\ \underline{\underline{\mathcal{L}}}^*\underline{w} + \nabla \cdot \underline{\mathfrak{D}}\,(\underline{u}, \underline{w}) \tag{105}$$

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

$$-w_i \left(\frac{\partial}{\partial x_j} \left(C_{ijpq} \frac{\partial u_p}{\partial x_q} \right) \right);\ i = 1, 2, 3 \tag{106}$$

utilizando la igualdad de divergencia tenemos

$$\begin{aligned} & -w_i \left(\frac{\partial}{\partial x_j} \left(C_{ijpq} \frac{\partial u_p}{\partial x_q} \right) \right) && (107) \\ = \ & C_{ijpq} \frac{\partial u_p}{\partial x_q} \frac{\partial w_i}{\partial x_j} - \frac{\partial}{\partial x_j} \left(w_i C_{ijpq} \frac{\partial u_p}{\partial x_q} \right) \\ = \ & \frac{\partial}{\partial x_j} \left(u_i C_{ijpq} \frac{\partial w_i}{\partial x_j} \right) - u_i \frac{\partial}{\partial x_j} \left(C_{ijpq} \frac{\partial w_i}{\partial x_j} \right) - \frac{\partial}{\partial x_j} \left(w_i C_{ijpq} \frac{\partial u_p}{\partial x_q} \right) \end{aligned}$$

reordenado términos tenemos que la ecuación anterior es

$$\frac{\partial}{\partial x_j}\left(u_i C_{ijpq}\frac{\partial w_i}{\partial x_j} - w_i C_{ijpq}\frac{\partial u_p}{\partial x_q}\right) - u_i\frac{\partial}{\partial x_j}\left(C_{ijpq}\frac{\partial w_i}{\partial x_j}\right) \tag{108}$$

en notación simbólica tenemos que

$$\underline{w}\ \underline{\underline{\mathcal{L}}}\ \underline{u} = -\underline{u}\nabla\cdot\left(\underline{\underline{\underline{C}}} : \nabla\underline{w}\right) + \nabla\cdot\left(\underline{u}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{w} - \underline{w}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{u}\right) \tag{109}$$

por lo tanto el operador adjunto formal es

$$\underline{\underline{\underline{\mathcal{L}}}}^{*}\,\underline{w} = -\nabla\cdot\left(\underline{\underline{\underline{C}}} : \nabla\underline{w}\right) \tag{110}$$

y los términos de frontera son

$$\underline{\mathfrak{D}}\left(\underline{u},\underline{w}\right) = \underline{u}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{w} - \underline{w}\cdot\underline{\underline{\underline{C}}} : \nabla\underline{u} \tag{111}$$

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

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

Operador Laplaciano

$$\underline{\underline{\underline{\mathcal{L}}}}\ \underline{u} = \Delta\underline{u} = \underline{f} \tag{112}$$

escrito en un sistema de ecuaciones se obtiene

$$\underline{\underline{\underline{\mathcal{L}}}}\ \underline{u} = \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\begin{bmatrix} \underline{p} \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix} \tag{113}$$

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

$$\underline{u} \equiv \left\{\underline{p}, u\right\}\ \text{ y }\ \underline{w} = \left\{\underline{q}, w\right\} \tag{114}$$

ahora el operador diferencial vector-valuado es el siguiente

$$\begin{aligned} \underline{\underline{\underline{\mathcal{L}}}}\underline{u} &= \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\cdot\begin{bmatrix} \underline{p} \\ u \end{bmatrix} \\ &= \begin{bmatrix} \underline{p} - \nabla u \\ \nabla\cdot\underline{p} \end{bmatrix} \end{aligned} \tag{115}$$

entonces

$$\underline{w}\ \underline{\underline{\underline{\mathcal{L}}}}\ \underline{u} = \begin{bmatrix} \underline{q} \\ w \end{bmatrix}\cdot\begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix}\cdot\begin{bmatrix} \underline{p} \\ u \end{bmatrix} \tag{116}$$

utilizando la definición de operador adjunto

$$\underline{w}\underline{\underline{\mathcal{L}}}\underline{u} = \underline{u}\underline{\underline{\mathcal{L}}}\underline{w} + \nabla\cdot\underline{\mathfrak{D}}\left(\underline{u},\underline{w}\right) \tag{117}$$

haciendo el desarrollo del término izquierdo se tiene que

$$\begin{aligned} \underline{w}\underline{\underline{\mathcal{L}}}\underline{u} &= \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \cdot \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \\ &= \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \cdot \begin{bmatrix} \underline{p} - \nabla u \\ \nabla \cdot \underline{p} \end{bmatrix} \\ &= \underline{qp} - \underline{q}\nabla \cdot u + w\nabla \cdot \underline{p} \end{aligned} \tag{118}$$

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

$$\begin{aligned} & \underline{qp} - \underline{q}\nabla \cdot u + w\nabla \cdot \underline{p} \\ = \; & \underline{qp} + u\nabla \cdot \underline{q} - \nabla \cdot \left(\underline{q}u\right) - \underline{p} \cdot \nabla w + \nabla \cdot \left(w\underline{p}\right) \\ = \; & \underline{p}\left(\underline{q} - \nabla w\right) + u\nabla \cdot \underline{q} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \end{aligned} \tag{119}$$

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

$$\begin{aligned} & \underline{p}\left(\underline{q} - \nabla w\right) + u\nabla \cdot \underline{q} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \\ = \; & \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \begin{bmatrix} \underline{q} - \nabla w \\ w \end{bmatrix} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \\ = \; & \begin{bmatrix} \underline{p} \\ u \end{bmatrix} \cdot \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{q} \\ w \end{bmatrix} + \nabla \cdot \left(w\underline{p} - u\underline{q}\right) \end{aligned} \tag{120}$$

por lo tanto, el operador adjunto formal es

$$\begin{aligned} \underline{\underline{\mathcal{L}}}^{*}\underline{w} &= \begin{bmatrix} 1 & -\nabla \\ \nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{q} \\ w \end{bmatrix} \\ &= \begin{bmatrix} \underline{q} - \nabla w \\ \nabla \cdot \underline{q} \end{bmatrix} \end{aligned} \tag{121}$$

y el término correspondiente a valores en la frontera es

$$\underline{\mathfrak{D}}\left(\underline{u}, \underline{w}\right) = w\underline{p} - u\underline{q}. \tag{122}$$

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

$$\begin{aligned} -\Delta\underline{u} + \nabla p &= f \\ -\nabla \cdot \underline{u} &= 0 \end{aligned} \tag{123}$$

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

$$\underline{U} = \{\underline{u}, p\} \text{ y } \underline{W} = \{\underline{w}, q\} \tag{124}$$

ahora el operador diferencial vector-valuado es el siguiente

$$\underline{\underline{\mathcal{L}}}\underline{U} = \begin{bmatrix} -\Delta & \nabla \\ -\nabla\cdot & 0 \end{bmatrix} \cdot \begin{bmatrix} \underline{u} \\ p \end{bmatrix} \tag{125}$$

el desarrollo se hará en notación indicial, entonces tenemos que

$$\underline{W}\underline{\underline{\mathcal{L}}}\underline{U} = \begin{Bmatrix} -\underline{w}\Delta\underline{u} + \underline{w}\nabla p \\ -q\nabla\cdot\underline{u} \end{Bmatrix} \tag{126}$$

usando notación indicial se obtiene

$$\begin{aligned} w_i\left(-\sum_j \frac{\partial^2 u_i}{\partial x_j^2} + \frac{\partial p}{\partial x_i}\right) &= -\sum_j w_i \frac{\partial^2 u_i}{\partial x_j^2} + w_i\frac{\partial p}{\partial x_i} \\ &= \sum_j \frac{\partial w_i}{\partial x_j}\frac{\partial u_i}{\partial x_j^2} - \sum_j \frac{\partial}{\partial x_j}\left(w_i\frac{\partial u_i}{\partial x_j}\right) - \\ &\quad p\frac{\partial w_i}{\partial x_i} + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{127}$$

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

$$\begin{aligned} &-\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2} + \sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j}\right) \\ &-\sum_j \frac{\partial}{\partial x_j}\left(w_i\frac{\partial u_i}{\partial x_j}\right) - p\frac{\partial w_i}{\partial x_i} + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{128}$$

reordenando términos tenemos

$$\begin{aligned} &-\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2} - p\frac{\partial w_i}{\partial x_i} + \\ &\sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j} - w_i\frac{\partial u_i}{\partial x_j}\right) + \frac{\partial}{\partial x_i}(w_i p) \end{aligned} \tag{129}$$

Ahora consideremos la ecuación 2 en Ec. (126), tenemos

$$-q\nabla\cdot\underline{u} \tag{130}$$

en notación índicial se tiene

$$\begin{aligned} -q\sum_i \frac{\partial u_i}{\partial x_i} &= -\sum_i q\frac{\partial u_i}{\partial x_i} \\ &= \sum_i u_i\frac{\partial q}{\partial x_i} - \sum_i \frac{\partial}{\partial x_i}(q u_i) \end{aligned} \tag{131}$$

en la ecuación anterior se utilizó la igualdad de divergencia, entonces agrupando las ecuaciones Ec. (129) y Ec. (131) se tiene

$$\begin{aligned} & w_i\left(-\sum_j \frac{\partial^2 u_i}{\partial x_j^2}+\frac{\partial p}{\partial x_i}\right)-q\sum_i \frac{\partial u_i}{\partial x_i} \qquad (132)\\ = & -\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2}-p\frac{\partial w_i}{\partial x_i}+\\ & \sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j}-w_i\frac{\partial u_i}{\partial x_j}\right)+\frac{\partial}{\partial x_i}\left(w_i p\right)+\\ & \sum_i u_i\frac{\partial q}{\partial x_i}-\sum_i \frac{\partial}{\partial x_i}\left(q u_i\right) \end{aligned}$$

ordenando los términos tenemos

$$\begin{aligned} & -\sum_j u_i\frac{\partial^2 w_i}{\partial x_j^2}+\sum_i u_i\frac{\partial q}{\partial x_i}-p\frac{\partial w_i}{\partial x_i} \qquad (133)\\ & +\sum_j \frac{\partial}{\partial x_j}\left(u_i\frac{\partial w_i}{\partial x_j}-w_i\frac{\partial u_i}{\partial x_j}\right)+\frac{\partial}{\partial x_i}\left(w_i p\right)-\sum_i \frac{\partial}{\partial x_i}\left(q u_i\right) \end{aligned}$$

escribiendo la ecuación anterior en notación simbólica, se obtiene

$$-\underline{u}\Delta\underline{w}+\underline{u}\nabla q-p\nabla\cdot\underline{w}+\nabla\cdot(\underline{u}\nabla\underline{w}-\underline{w}\nabla\underline{u}+\underline{w}p-\underline{u}q) \qquad (134)$$

por lo tanto, el operador adjunto formal es

$$\underline{\underline{\mathcal{L}}}^{*}\underline{W}=\begin{bmatrix} -\Delta & \nabla \\ -\nabla\cdot & 0 \end{bmatrix}\cdot\begin{bmatrix} \underline{w} \\ q \end{bmatrix} \qquad (135)$$

y el término de valores de frontera es

$$\underline{\mathfrak{D}}\left(\underline{u},\underline{w}\right)=\underline{u}\nabla\underline{w}-\underline{w}\nabla\underline{u}+\underline{w}p-\underline{u}q. \qquad (136)$$

1.5. 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 [12].

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

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{137}$$

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

$$\begin{aligned} -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega. \end{aligned} \tag{138}$$

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

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{139}$$

aplicando el teorema de Green (115) obtenemos la Ec. (50), que podemos re-escribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{140}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{141}$$

y la funcional lineal

$$l(v) = \langle f, v \rangle = \int_\Omega v f_\Omega d\underline{x} \tag{142}$$

podemos reescribir el problema dado por la Ec. (43) 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

$$a(u,v) = \langle f, v \rangle \tag{143}$$

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

Definición 36 *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*

$$|a(u,v)| \leq M \|u\|_V \|v\|_V \quad \forall u, v \in V \tag{144}$$

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

$$a(v,v) \geq \alpha \|v\|_V^2 \quad \forall v \in V \tag{145}$$

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 (138) definido en $V = H_0^1(\Omega)$ reescrito como el problema (143) genera una forma bilineal V-elíptico cuyo producto interior sobre V es simétrico y positivo definido ya que

$$a(v,v) \geq \alpha \|v\|_V^2 > 0, \quad \forall v \in V, v \neq 0 \tag{146}$$

reescribiéndose el problema (143), en el cual debemos encontrar $u \in V$ tal que

$$a(u,v) = \langle f, v\rangle - a(u_0, v) \tag{147}$$

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 (138) reescrito como el problema (143) tiene solución y esta es única, esto queda plasmado en el siguiente resultado.

Teorema 37 *(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

$$a(u,v) = \langle f, v\rangle, \forall v \in V \tag{148}$$

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

$$\|u\|_V \leq \frac{1}{\alpha}\|l\|_{V^*} \tag{149}$$

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 38 *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*

$$a(u+p, v+\overline{p}) = a(u,v) \quad \forall u, v \in V \text{ y } p, \overline{p} \in P. \tag{150}$$

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

$$Q = \left\{ v \in V \mid \int_\Omega upd\underline{x} = 0 \quad \forall p \in P \right\}, \tag{151}$$

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

$$a(q,q) \geq \alpha \|q\|_Q^2 \quad \textit{para } q \in Q, \tag{152}$$

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*

$$a(u,v) = \langle l, v\rangle, \quad \forall v \in V \tag{153}$$

si y sólo si las condiciones de compatibilidad

$$\langle l, p\rangle = 0 \quad \textit{para } p \in P \tag{154}$$

se satisfacen.

ii) La solución u *satisface*

$$\|u\|_Q \leq \alpha^{-1} \|l\|_{Q^*} \tag{155}$$

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

$$a(u,v) = \langle f, v\rangle, \quad v \in V \tag{156}$$

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*

$$\|u\|_{H^s} \leq C \|f\|_{H^{s-2m}} \tag{157}$$

se satisface.

2. Métodos de Solución Aproximada para EDP

Ya que en general encontrar la solución a problemas con geometría diversa es difícil y en algunos casos imposible usando métodos analíticos. 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).

En este capítulo se considera el VBVP de la forma

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{158}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{159}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, como un caso particular del operador elíptico definido por la Ec. (43) 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

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{160}$$

aplicando el teorema de Green (115) obtenemos la Ec. (50), que podemos reescribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{161}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{162}$$

y la funcional lineal

$$l(v) = \langle f, v\rangle = \int_\Omega v f_\Omega d\underline{x} \tag{163}$$

podemos reescribir el problema dado por la Ec. (158) de orden 2 en forma variacional, haciendo uso de la forma bilineal $a(\cdot,\cdot)$ y la funcional lineal $l(\cdot)$.

2.1. Método Galerkin

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

$$a(u,v) = \langle f, v\rangle \quad \forall v \in V \tag{164}$$

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. (164) 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,

$$V^h = Generado\left\{\phi_i\right\}_{i=1}^{N}, \quad V^h \subset V. \tag{165}$$

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

$$a\left(u_h, v_h\right) = \langle f, v_h \rangle \quad \forall v_h \in V^h. \tag{166}$$

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

$$u_h = \sum_{i=1}^{N} c_i \phi_i \quad \text{y} \quad v_h = \sum_{j=1}^{N} d_j \phi_j \tag{167}$$

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. (167) en la Ec. (166) 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

$$\sum_{i=1}^{N} a\left(\phi_i, \phi_j\right) c_i = \left\langle f, \phi_j \right\rangle \tag{168}$$

o más concisamente, como

$$\sum_{i=1}^{N} K_{ij} c_i - F_j = 0 \quad j = 1, 2, ...N \tag{169}$$

en la cual

$$K_{ij} = a\left(\phi_i, \phi_j\right) \quad \text{y} \quad F_j = \left\langle f, \phi_j \right\rangle \tag{170}$$

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

$$\sum_{i=1}^{N} K_{ij} c_i - F_j, \quad j = 1, 2, ...N \tag{171}$$

o más compactamente

$$\underline{\underline{\mathbb{K}}}\underline{u} = \underline{F} \tag{172}$$

en la cual $\underline{\underline{\mathbb{K}}}$ y $\underline{F}$ son la matriz y el vector cuyas entradas son K_{ij} y F_j respectivamente. 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

$$a(v,v) \geq \alpha \|v\|^2 > 0 \qquad \forall v \neq 0, \tag{173}$$

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 (169) se simplifica considerablemente, ya que

$$K_{ij} = a\left(\phi_i, \phi_j\right) = 0 \qquad \text{si } i \neq j \tag{174}$$

y

$$K_{ii}c_i = F_i \qquad \acute{o} \qquad c_i = F_i/K_{ii}. \tag{175}$$

Así, el problema (158) definido en $V^h = H_0^1(\Omega)$ reescrito como el problema (164) genera una forma bilineal V^h-elíptica cuyo producto interior sobre V^h es simétrico y positivo definido ya que

$$a(v_h, v_h) \geq \alpha \|v_h\|_{V^h}^2 > 0, \quad \forall v_h \in V^h, v_h \neq 0 \tag{176}$$

reescribiéndose el problema (166) como el problema aproximado en el cual debemos encontrar $u_h \in V^h \subset V$ tal que

$$a(u_h, v_h) = \langle f, v_h \rangle - a(u_0, v_h) \tag{177}$$

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

$$\int_\Omega \left(\nabla v_h \cdot \underline{\underline{a}} \cdot \nabla u_h + c u_h v_h\right) dxdy = \int_\Omega f_\Omega v_h dxdy \tag{178}$$

para todo $v_h \in V^h$.

Entonces, el problema (158) al aplicarle el método Galerkin obtenemos (161), el cual podemos reescribirlo como (178). Aplicando el teorema de Lax-Milgram (37) 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 (37) tenemoe el siguiente resultado

Teorema 40 *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*

$$a(u_h, v_h) = \langle l, v_h \rangle \qquad \forall v_h \in V^h. \tag{179}$$

Además, si $l(\cdot)$ es de la forma

$$\langle l, v_h \rangle = \int_\Omega f_\Omega v_h d\underline{x} \tag{180}$$

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

$$\|u_h\|_V \leq \frac{1}{\alpha} \|f\|_{L^2}, \tag{181}$$

donde α es la constante en (173).

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. (164), para más detalle véase [12] y [3].

Teorema 41 *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*

$$\|u - u_h\|_V \leq C \inf_{v_h \in V^h} \|u - v_h\|_V \tag{182}$$

donde u es solución de (164) y u_h es solución de (177), consecuentemente, una condición suficiente para que la aproximación u_h del método Galerkin converge a la solución u del problema dado por la Ec. (164) es que exista una familia $\{V^h\}$ de subespacios con la propiedad de que

$$\inf_{v_h \in V^h} \|u - v_h\|_V \to 0 \quad \textit{cuando} \quad h \to 0. \tag{183}$$

2.1.1. 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

$$\{\phi_i\}_{i=1}^N \text{ para } U^h \text{ y } \{\psi_j\}_{j=1}^N \text{ para } V^h \tag{184}$$

entonces

$$u_h = \sum_{i=1}^N c_i \phi_i \text{ y } v_h = \sum_{j=1}^N b_j \psi_j \tag{185}$$

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

Sustituyendo esta última expresión Ec. (185) en

$$(\mathcal{L}u_h - f, v_h) = 0 \tag{186}$$

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

$$\tau(u_h) \equiv \mathcal{L}u_h - f \tag{187}$$

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

2.1.2. Método de Elemento Finito

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 (178), 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 [15]), 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

$$V^h = Generado\left\{\phi_i \in \mathbb{P}^h[k] \;\middle|\; \phi_i(x) = 0 \text{ en } \partial\Omega\right\}. \tag{188}$$

La solución aproximada de la Ec. (178) al problema dado por la Ec. (158) queda en términos de

$$\int_\Omega \left(\nabla\phi_i \cdot \underline{\underline{a}} \cdot \nabla\phi_j - c\phi_i\phi_j\right) dxdy = \int_\Omega f_\Omega \phi_j dxdy \tag{189}$$

si definimos el operador bilineal

$$K_{ij} \equiv a\left(\phi_i, \phi_j\right) = \int_\Omega \left(\nabla\phi_i \cdot a_{ij} \cdot \nabla\phi_j - c\phi_i\phi_j\right) dxdy \tag{190}$$

y la funcional lineal

$$F_j \equiv \left\langle f, \phi_j \right\rangle = \int_\Omega f_\Omega \phi_j dxdy \tag{191}$$

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)

$$\underline{\underline{K}}\,\underline{u} = \underline{F} \tag{192}$$

donde $\underline{u}$ será el vector solución a la Ec. (192) cuyos valores serán la solución al problema dado por la Ec. (178) que es la solución aproximada a la Ec. (158) 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

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega\backslash\Sigma \\ u &= g \quad \text{en } \partial\Omega \\ [u]_\Sigma &= J_0 \\ [a_n \cdot \nabla u]_\Sigma &= J_1 \end{aligned} \tag{193}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{194}$$

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

$$w\mathcal{L}u = -w\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cwu = wf_\Omega \tag{195}$$

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

$$a(u,w) = \sum_{i=1}^{E} \int_{\Omega_i} \left(\nabla u \cdot \underline{\underline{a}} \cdot \nabla w + cwu \right) d\underline{x} \tag{196}$$

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. (195) como

$$\begin{aligned} a(u,w) &= \int_{\Omega} wf d\underline{x} + \sum_{i=1}^{E} \int_{\partial\Omega} wa_n \cdot \nabla u d\underline{s} \\ &= \int_{\Omega} wf_{\Omega} d\underline{x} + \int_{\partial\Omega} wa_n \cdot \nabla u d\underline{s} - \int_{\Sigma} w \left[a_n \cdot \nabla u \right]_{\Sigma} d\underline{s}. \end{aligned} \tag{197}$$

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. (196) si y sólo si $v(x)$ satisface

$$a(u,w) = \int_{\Omega} wf_{\Omega} d\underline{x} - \langle u_0, w \rangle - \int_{\Sigma} J_1 w d\underline{s} \tag{198}$$

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

$$V^h = \left\{ \phi_i \mid \phi_i \in C^1(\Omega_i), \forall i, \phi_i = 0 \text{ en } \partial\Omega \text{ y } \phi_i \in C^0(\Omega) \right\}. \tag{199}$$

La solución por elementos finitos de (198) se obtiene al resolver el sistema lineal

$$\underline{\underline{K}}\underline{u} = \underline{F} \tag{200}$$

donde

$$K_{ij} = a\left(\phi_i, \phi_j\right) \tag{201}$$

y

$$F_j = \int_{\Omega} \phi_j f_{\Omega} d\underline{x} - a\left(u_0, \phi_j\right) - \int_{\Sigma} J_1 \phi_j d\underline{s} \tag{202}$$

esta solución será la solución en los nodos interiores de Ω.

2.2. Método de Penalización Interior

En la década de los 70s se desarrollo el método de Interior Penalty (IP) de forma independiente del método Galerkin para ecuaciones elípticas y parabólicas resultando en dos métodos independientes en los que se usan elementos finitos discontinuos, y un número grande de variantes se han introducido y estudiado. Las penalizaciones fueron primeramente introducidas en el método de Elementos finitos como una forma de imponer condiciones de frontera tipo Dirichlet débiles más que incorporarlas a las condiciones de frontera dentro del espacio de elementos finitos [**?**].

Sea Ω un dominio y sea el operador elíptico

$$\begin{aligned} -\Delta u &= f \quad \text{en } \Omega \\ u &= 0 \quad \text{en } \partial\Omega \end{aligned} \tag{203}$$

claramente

$$\int_\Omega \nabla u \cdot \nabla v d\underline{x} - \int_{\partial\Omega} \frac{\partial u}{\partial n} v d\underline{s} = \int_\Omega f v d\underline{x} \tag{204}$$

para toda función de prueba v suficientemente suave. Puesto que u se nulifica en la frontera, tenemos también que $B(u,v) = \int f v d\underline{x}$ donde

$$B(u,v) = \int_\Omega \nabla u \cdot \nabla v d\underline{x} - \int_{\partial\Omega} \frac{\partial u}{\partial n} v d\underline{s} - \int_{\partial\Omega} \frac{\partial v}{\partial n} u d\underline{s} + \int_{\partial\Omega} \eta u v d\underline{s} \tag{205}$$

para cualquier función de peso η. El método entonces determina una solución aproximada u_h en un subespacio de elementos finitos de $H^1(\Omega)$ tal que $B(u_h, v_h) = \int f v_h d\underline{x}$ para todo v_h en el mismo espacio. Notemos que el segundo término de la forma bilineal B surge para asegurar que el método es consistente. El tercer término fue adicionado para que el problema discreto sea simétrico (y de manera que el método es verdaderamente variacional -la solución discreta minimiza $B(u,u)/2 - \int fu$ sobre el espacio de elementos finitos). Finalmente el último término es el término de penalización, el cual es necesario para garantizar la estabilidad.

Se muestra que si η es tomado como C/h donde h es el tamaño del elemento y C es una constante suficientemente grande, entonces la solución discreta converge a la solución exacta con orden óptimo en H^1 y L^2.

Un método de penalidad diferente para imponer condiciones de frontera tipo Dirichlet no incluye cualquiera de los términos segundo o tercero en la Ec.(205), y usa como peso de penalización $h^{-\sigma}$ para alguna $\sigma \geq 0$. A causa de la omisión del término de consistencia, el método y su análisis incluyen un error de consistencia.

Otra interesante posibilidad es la de incluir todos los términos en la Ec.(205) pero cambiando el signo del tercer término en B. La forma bilineal ya no será simétrica, pero tiene una propiedad coercitiva favorable, a saber, $B(u,u) \geq \int |\nabla u|^2$, no importando cual $\eta \geq 0$ se escoja.

3. Método Galerkin Discontinuo

En el presente capítulo se dará un esquema para el entendimiento, comparación y análisis de varios métodos de Galerkin Discontinuo que han sido propuestos para el tratamiento de problemas elípticos. Esta clase incluyen los llamados métodos de penalización interior [**?**].

Mallado del dominio El Mallado o triangulación $\mathcal{T}_h$ del dominio Ω, sin perdida de generalidad, consideremos el dominio $\Omega \subset \mathbb{R}^2$, el cual es subdividido en E subdominios o elementos Ω_e llamados elementos, tal que

$$\overline{\Omega} = \bigcup_{e=1}^{E} \overline{\Omega}_e \tag{206}$$

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.

3.1. Generalización del Método Galerkin Discontinuo

Sea Ω un dominio y se el operador elíptico

$$\begin{aligned} -\Delta u &= f \quad \text{en } \Omega \\ u &= 0 \quad \text{en } \partial\Omega \end{aligned} \tag{207}$$

donde el dominio Ω se asume como un dominio poligonal y f es una función dada en $L^2(\Omega)$. Para obtener la formulación débil sobre la cual la discretización se basa, reescribimos el anterior problema como sigue

$$\sigma = \nabla u, \quad -\nabla \cdot \sigma = f \quad \text{en } \Omega \tag{208}$$

$$u = 0 \quad \text{en } \partial\Omega. \tag{209}$$

Sea K la clausura de un subconjunto abierto de Ω con frontera por pedazos suave. Si multiplicamos la anterior ecuación por funciones de prueba e integramos formalmente sobre K, tenemos

$$\int_K \sigma \cdot \tau dx = -\int_K u\nabla \cdot \tau dx + \int_{\partial K} un_k \cdot \tau ds \tag{210}$$

$$\int_K \sigma \cdot \nabla v dx = \int_K fv dx + \int_{\partial K} \sigma \cdot n_K v ds \tag{211}$$

donde n_K es el vector normal unitario exterior a ∂K. Esta es la formulación débil que buscábamos. Con esto en mente podemos ahora definir el método generalizado Galerkin Discontinuo.

Denotamos por $\mathcal{T}_h$ una triangulación del dominio Ω en polígonos K, y por $P(K)$ un espacio de dimensión finita de funciones suaves, típicamente polinomios, definidos sobre el polígono K. Este espacio es usado para aproximar la variable u. Además, denotamos por $\Sigma(K)$ otro espacio de dimensión finita de funciones suaves que serán usadas para aproximar la variable auxiliar σ.

Sean

$$V_h = \left\{ v \in L^2(\Omega) \mid v_{|_K} \in P(k), \qquad \forall K \in \mathcal{T}_h \right\} \tag{212}$$

$$\Sigma_h = \left\{ \tau \in \left(L^2(\Omega)\right)^2 \mid \tau_{|_K} \in \Sigma(k), \qquad \forall K \in \mathcal{T}_h \right\} \tag{213}$$

y consideramos la siguiente formulación débil:

Encontrar $u_h \in V_h$ y $\sigma_h \in \Sigma_h$ tal que para toda $K \in \mathcal{T}_h$ tenemos

$$\int_K \sigma_h \cdot \tau dx = -\int_K u_h \nabla \cdot \tau dx + \sum_{e \subset \partial K} \int_e h_u^{e,K} n_k \cdot \tau ds, \forall \tau \in \Sigma(K) \tag{214}$$

$$\int_K \sigma_h \cdot \nabla v dx = \int_K fv dx + \sum_{e \subset \partial K} \int_e h_\sigma^{e,K} \cdot n_k v ds, \forall v \in P(K) \tag{215}$$

donde las sumas son tomadas sobre los bordes del polígono K, y el flujo numérico $h_\sigma^{e,K}$ y $h_u^{e,K}$ son aproximaciones a $\sigma_{|_e} = \nabla u_{|_e}$ y a $u_{|_e}$ respectivamente sobre las caras de la triangulación.

Por ejemplo, para elementos triangulares, podemos tomar $P(K)$ como el conjunto de polinomios de grado $p \geq 1$ y $\Sigma(K)$ como el conjunto de todos los polinomios del campo vectorial de grado $p-1$ o p. La elección de la forma de construir los flujos es crucial, algunas propiedades básicas que deben de compartir todas las elecciones de flujo se dan a continuación:

1. Localidad.- Sea $K = K_1$ un elemento de la triangulación, y sea e uno de sus bordes. Asumimos primero que e es un borde interior de nuestra triangulación, tal que existe un segundo elemento K_2 que comparte el borde e con K_1. Entonces asumimos que $h_\sigma^{e,K}$ y $h_u^{e,K}$ dependen de las restricciones $u_{h|_{K_i}}$ y $\sigma_{h|_{K_i}}$ de u_h y σ_h a K_i, $i = 1, 2$. Más precisamente, de modo local tenemos

$$h_\sigma^{e,K} = h_\sigma^{e,K}\left(u_{h|_{K_1}}, \sigma_{h|_{K_1}}, u_{h|_{K_2}}, \sigma_{h|_{K_2}}\right) \tag{216}$$

en los ejemplos, estas funcionales dependen de la forma particular de $h_\sigma^{e,K}$ y $h_u^{e,K}$, ya que dependen sólo de las trazas de $u_{h|_{K_i}}$, $\nabla u_{h|_{K_i}}$ y $\sigma_{h|_{K_i}}$ sobre el borde e. Ya que $u_h, \nabla u_h$ y σ_h son en general discontinuas a través de e, la traza de $u_{h|_{K_1}}$ sobre e será diferente que la traza de $u_{h|_{K_2}}$ sobre e, y de forma similar ∇u_h y σ_h las cuales tendrán dos diferentes trazas sobre e. Así que, $h_\sigma^{e,K}$ y $h_u^{e,K}$ dependen linealmente de seis cantidades

$$\left(u_{h|_{K_1}}\right)_{|_e}, \left(\nabla u_{h|_{K_1}}\right)_{|_e}, \left(\sigma_{h|_{K_1}}\right)_{|_e}, \tag{217}$$

$$\left(u_{h|_{K_2}}\right)_{|_e}, \left(\nabla u_{h|_{K_2}}\right)_{|_e}, \left(\sigma_{h|_{K_2}}\right)_{|_e}. \tag{218}$$

En nuestro particular caso de un problema homogéneo con condiciones de frontera tipo Dirichlet, los flujos sobre los bordes de la frontera tienen la misma dependencia funcional sobre esas seis trazas, siempre que se interpreten las trazas próximas a K_2 como sigue:

$$\left(u_{h|_{K_2}}\right)_{|_e} = 0, \tag{219}$$

$$\left(\nabla u_{h|_{K_2}}\right)_{|_e} = \left(\nabla u_{h|_{K_1}}\right)_{|_e} \text{ y } \left(\sigma_{h|_{K_2}}\right)_{|_e} = \left(\sigma_{h|_{K_1}}\right)_{|_e}. \tag{220}$$

Finalmente, es importante notar que en todos los métodos se analizará $h_u^{e,K}$ la cual no depende de $\sigma_{h|_{K_i}}$(ni sobre $\nabla u_{h|Ki}$, la cual es menos importante).

2. Consistencia.- En todos los métodos se considera como consistente en el sentido que, en una forma funcional descrito como

$$h_\sigma^{e,K}\left(u_{|_{K_1}}, \nabla u_{|_{K_1}}, u_{|_{K_2}}, \nabla u_{|_{K_2}}\right) = \nabla u_{|e} \tag{221}$$

$$h_u^{e,K}\left(u_{|_{K_1}}, \nabla u_{|_{K_1}}, u_{|_{K_2}}, \nabla u_{|_{K_2}}\right) = u_{|e} \tag{222}$$

donde u es una función suave que satisface las condiciones de frontera.

3. Conservación.- Todos los métodos satisfacen

$$h_\sigma^{e,K_1} = h_\sigma^{e,K_2} \tag{223}$$

donde e es un borde que comparten los elementos K_1 y K_2, y de tal forma que podemos escribir de forma simplificada h_σ^e. De tal forma que la propiedad de conservación la podemos escribir como: Si S es la unión de alguna colección de elementos, entonces, tomando a v como idénticamente la unidad en la Ec.(215) y sumando sobre K contenida en S tenemos

$$\int_S f dx + \sum_{e \subset \partial S} \int_e h_\sigma^e \cdot n ds = 0. \tag{224}$$

Cerramos esta sección con varios comentarios adicionales concernientes a las propiedades antes mencionadas.

- Como se vio, si $h_u^{e,K}$ no depende de σ_h, entonces la variable auxiliar σ_h puede ser eliminada localmente en términos de u_h y ∇u_h, usando la Ec.(214). Cuando se usan triángulos, se usa la base ortonormal de Dubiner convirtiendo esta eliminación en trivial.

- En todos los métodos se considera h_σ^e depende de cualquiera de las dos, de las trazas de ∇u_h o de σ_h, pero no de ambas. Aquéllas categorías, para las cuales la matriz de carga tiende a ser matriz dispersa incluye a los métodos de Penalización Interior y Baumann y Oden.

- La mayoría de los métodos satisfacen adicionalmente a la propiedad de conservación dada por la Ec.(223), la propiedad análoga $h_u^{e,K_1} = h_u^{e,K_2}$ en cuyo caso escribimos h_u^e. Y nos referiremos a esta como métodos completamente conservativos, estos métodos generan después de eliminar σ_h una matriz de carga simétrica excepto para los métodos de Baumann y Oden y en este caso son conocidos como métodos de penalidad pura. Todos los métodos que se consideran aquí son completamente conservativos.

- Notemos que, en vista de la Ec.(215) sólo la componente normal $h_\sigma^{e,K} \cdot n_k$ de $h_\sigma^{e,K}$ participa en el método, la componente tangencial es irrelevante. En la práctica, la componente normal depende sólo de las trazas normales.

3.2. Flujos Numéricos Independientes de ∇u_h

Sea e un borde que comparten los elementos K_1 y K_2. Definiendo también los vectores normales n_1 y n_2 sobre e apuntando hacia el exterior de K_1 y K_2 respectivamente. Si v es una función sobre $K_1 \cup K_2$ pero con posibles discontinuidades a través de e, sea v_i que denota $\left(v_{|K_i}\right)_{|_e}$, $i = 1, 2$.

Para una función escalar v definimos

$$\dot{v} = \frac{1}{2}\left(v_1 + v_2\right), \qquad [[v]] = v_1 n_1 + v_2 n_2 \tag{225}$$

si τ es una función vector valuada, tenemos

$$\dot{\tau} = \frac{1}{2}\left(\tau_1 + \tau_2\right), \qquad [[\tau]] = \tau_1 . n_1 + \tau_2 \cdot n_2 \tag{226}$$

notemos que el salto $[[v]]$ de una función escalar v es un vector paralelo a n y que $[[\tau]]$ es el salto de la componente normal de la función vectorial τ, siendo esta una cantidad escalar. La ventaja de esta definición es que no depende del asignamiento de un orden a los elementos K_i.

Aquí consideraremos que el método es determinado por la siguiente elección del flujo numérico

$$\begin{aligned} h_\sigma^{e,K} &= \dot{\sigma}_h - \alpha^e [[u_h]] + \beta^e [[\sigma_h]] \\ h_u^{e,K} &= \dot{u}_h + \gamma^e \cdot [[u_h]] \end{aligned} \tag{227}$$

donde β^e y γ^e son funciones vector valuadas sobre e. A menudo ellas son constantes y en muchos métodos ellas se toman como nulas. El término $\alpha^e\left([[u_h]]\right)$ puede ser tomada simplemente como

$$\alpha^e\left([[u_h]]\right) = \eta^e\left[[u_h]\right] \tag{228}$$

para alguna constante o función η^e. Otra posibilidad es definir el operador $r_e : L^1(e) \to \Sigma_h$ definida como

$$\int_\Omega r_e(q) \cdot \tau dx = -\int_e q \cdot \dot{\tau} ds \tag{229}$$

para todo $\tau \in \Sigma_h$ y $q \in L^1(e)$, y el conjunto

$$\alpha^e\left([[u_h]]\right) = \eta^e \widehat{r^e\,[[u_h]]}. \tag{230}$$

Primero reescribimos el método insertando el flujo de la Ec.(227) dentro de la ecuación de Galerkin dadas por las Ecs.(214) y (215) y tomando sobre $K \in \mathcal{T}_h$. Denotando por $\mathcal{E}_h$ el conjunto de todas los bordes, obteniendo

$$\begin{aligned} \int_\Omega \sigma_h \cdot \tau dx &= \sum_K \int_K \nabla u_h \cdot \tau dx + \\ &\quad \sum_{e\in\mathcal{E}_h} \int_e \left(\gamma^e \cdot [[u_h]]\,[[\tau]] - [[u_h]] \cdot \dot{\tau}\right) ds \end{aligned} \tag{231}$$

$$\begin{aligned} \sum_K \int_K \sigma_h \cdot \nabla v dx &= \int_\Omega f v dx + \\ &\quad \sum_{e\in\mathcal{E}_h} \int_e \left(\dot{\sigma}_h - \alpha^e\left([[u_h]]\right) + \beta^e\left[[u_h]\right]\right) \cdot [[v]]\, ds \end{aligned} \tag{232}$$

para toda $\tau \in \Sigma_h, v \in V_h$. Si tomamos que todos los α^e, β^e y γ^e se nulifican, recuperamos el método de Galerkin Discontinuo original. Este método puede ser inestable al menos para mallas uniformes, sin embargo la estabilidad se logra si α^e es un operador positivo. Definiendo α^e por la Ec.(230) con $\eta^e > 0$ (pueden ser β^e y γ^e zero) obteniendo las variantes de Bassi y Rebay, definiendo α^e por la Ec.(228), $\eta^e > 0$ obtenemos el método LDG.

Continuando, podemos eliminar σ_h para reescribir el método en términos de u_h solamente (esto es lo usualmente preferido en la implementación). Para hacer esto, definimos dos operadores R y L. El operador $R : V_h \to \Sigma_h$ dado por $R(v) = \sum_{e\in\mathcal{E}_h} r_e\left([[u_h]]\right)$, o equivalentemente,

$$\int_\Omega R(\varphi) \cdot \tau dx = -\sum_{e\in\mathcal{E}_h} \int_e [[\varphi]] \cdot \dot{\tau} ds \tag{233}$$

para toda $\tau \in \Sigma_h$ y el operador $L : L^1(\cup \mathcal{E}_h) \to \Sigma_h$ que es definido por

$$\int_\Omega L(\varphi) \cdot \tau dx = \sum_{e \in \mathcal{E}_h} \int_e \varphi \cdot [[\tau]] \, ds \tag{234}$$

para toda $\tau \in \Sigma_h$.

Denotamos por P_Σ la L^2-proyección sobre Σ_h, entonces podemos reescribir la Ec.(231) como

$$\sigma_h = P_\Sigma(\nabla u_h) + R(u_h) + L(\gamma \cdot [[u_h]]) \tag{235}$$

y la Ec.(232) como

$$\begin{aligned} \sum_K \int_K \sigma_h \cdot \nabla v dx &= \int_\Omega f v dx + \int_\Omega \sigma_h \cdot (-R(v) + L(\beta \cdot [[v]])) \\ &\quad - \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]] \, ds \end{aligned} \tag{236}$$

donde β y γ son funciones sobre $\cup \mathcal{E}_h$ las cuales están dadas por β^e y γ^e respectivamente, sobre cada borde e. Finalmente insertando la Ec.(235) en la Ec.(236), obtenemos

$$\begin{aligned} &\sum_K \int_K (P_\Sigma(\nabla u_h) + R(u_h) + L(\gamma \cdot [[u_h]])) \cdot (\nabla v + R(v) + L(\beta \cdot [[v]])) \, dx \\ &\quad + \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]] \, ds = \int_\Omega f v dx. \end{aligned} \tag{237}$$

Notemos que la segunda suma del lado izquierdo de la Ec.(237) es simétrica con respecto a u_h y v ya que

$$\sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]] \, ds = \tag{238}$$

$$= \begin{cases} \sum_{e \in \mathcal{E}_h} \int_e \eta^e [[u_h]] \cdot [[v]] \, ds, \text{ si } \alpha^e \text{ es definido por Ec.(228)} \\ \sum_{e \in \mathcal{E}_h} \int_e \eta^e r_e ([[u_h]]) \cdot r_e [[v]] \, ds, \text{ si } \alpha^e \text{ es definido por Ec.(230).} \end{cases}$$

de este modo es claro que la matriz de carga simétrica es obtenida si elegimos $\beta^e = -\gamma^e$ para toda e. Esta elección es usada por el método LDG.

En la practica la inclusión de $\nabla P(K) \subset \Sigma(K)$ generalmente es suficiente. En el caso de la proyección P_Σ no es requerida en la Ec.(237).

Finalmente, notemos que el soporte de v es contenido en un solo elemento K, entonces el soporte de $R(v)$ el cual generalmente contiene a todos los elementos que contienen el borde de K. Consecuentemente el producto $R(u_h) \cdot R(v)$ en la Ec.(237) la cual generalmente tiene un gran impacto negativo en la dispersión de la matriz de carga. Este problema es mucho menos severo cuando el flujo numérico es independiente de σ_h.

3.3. Flujos Numéricos Independientes de σ_h

Primeramente consideremos, en lugar de la Ec.(227) el siguiente flujo numérico

$$\begin{aligned} h_{\sigma}^{e,K} &= \dot{\widehat{\nabla u_h}} - \alpha^e [[u_h]] + \beta^e [[\nabla u_h]] \\ h_{u}^{e,K} &= \dot{u}_h + \gamma^e [[u_h]] \end{aligned} \tag{239}$$

donde β^e y γ^e son funciones vector valuadas sobre e. Procediendo a la eliminación de la variable σ_h como en la sección anterior. Pero usando las definiciones de R y L en las Ecs.(233) y (234) respectivamente, obtenemos

$$\begin{aligned} &\sum_K \int_K (P_\Sigma(\nabla u_h) + R(u_h) + L(\gamma \cdot [[u_h]])) \cdot \nabla v \\ &+\nabla u_h \cdot (R(v) - L(\beta \cdot [[v]]))\, dx \\ &+ \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]]\, ds = \int_\Omega f v dx. \end{aligned} \tag{240}$$

escogiendo $\beta = \gamma = 0$ y α escogido como la Ec.(228), recuperamos el método de Penalización Interior, mientras que para $\beta = \gamma = 0$ y α escogido como la Ec.(230) procedemos a recuperar el método original de Galerkin Discontinuo de Bassi y Rebay sobre algunas suposiciones generales, y para elementos triangulares, el esquema de estabilidad y convergencia optima se da siempre y cuando $\eta^e > 3$, donde este número representa, en esencia el numero de bordes por elemento.

Notemos que el número de entradas no cero de la matriz de carga es reducida a un mínimo, esto es debido a que el término $R(u_h) \cdot R(v)$ que aparece en la Ec.(237) ya no esta presente en la Ec.(240).

Considerando ahora otra familia de flujos numéricos, consideremos

$$\begin{aligned} h_{\sigma}^{e} &= \zeta \dot{\widehat{\nabla u_h}} - \alpha^e ([[u_h]]) \\ h_{u}^{e,K} &= \dot{u}_h + \delta [[u_h]] \cdot n_K \end{aligned} \tag{241}$$

donde ζ y δ son parámetros reales. Diferentes opciones de estos parámetros son seleccionadas en los diferentes métodos de Galerkin Discontinuo. Notemos que para $\delta \neq 0$ corresponde a métodos en los cuales no es totalmente conservativo y para $\zeta \neq 1$ la consistencias es violada.

Usando la Ec.(241) en las Ecs(214) y (215) y procedemos a eliminar σ_h como antes, obtenemos

$$\begin{aligned} &\sum_K \int_K (P_\Sigma(\nabla u_h) \cdot \nabla v + (1 - 2\delta) R(u_h) \cdot \nabla v + \zeta \nabla u_h \cdot R(v))\, dx \\ &+ \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]]\, ds = \int_\Omega f v dx \end{aligned} \tag{242}$$

para $\delta = 1, \zeta = 1, \alpha^e = 0$ y $\nabla P(K) \subset \Sigma(K)$ (tal que se $\nabla P(K)$, P_Σ se reduce al operador inclusión y puede ser suprimido), esto es exactamente el método Galerkin Discontinuo de Baumann y Oden. Para ver esto, la anterior ecuación puede ser reescrita. Para iniciar notemos que

$$\int_\Omega \nabla u \cdot R(v) dx = -\sum_{e \in \mathcal{E}_h} \int_e [[v]] \, \dot{\widehat{\nabla u}} ds = -\sum_K \int_{\partial K} ([v]) \frac{\partial u}{\partial n_K} ds \tag{243}$$

donde seleccionamos en cada elemento K, para cada $e \in \partial K$

$$([v]) = \frac{1}{2} \left(v^{int} - v^{ext} \right)_e \tag{244}$$

con obvio entendimiento de los símbolos. Con esta notación y cuando $\nabla P(K) \subset \Sigma(K)$, la Ec.(242) puede ser rescrita como

$$\sum_K \int_K \nabla u_h \cdot \nabla v dx + \int_{\partial K} \left((2\delta - 1) ([u_h]) \frac{\partial v}{\partial n} - \zeta ([u_h]) \frac{\partial u_h}{\partial n} \right) ds \tag{245}$$

$$+ \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]] \, ds = \int_\Omega f v dx$$

el cual es el método Galerkin Discontinuo de Baumann y Oden cuando $\delta = \zeta = 1$ y $\alpha^e = 0$. Este método requiere algunas suposiciones adicionales, como es el hecho de que los polinomios deberán de ser de grado mayo o igual a dos. La situación cobra importancia cuando α^e es tomada como en la Ec.(228) o Ec.(230) con $\eta^e > 0$.

Por otro lado, tomando $\delta = 1/2$ y $\zeta = 0$ en la Ec.(241), la Ec.(242) quedaría como

$$\sum_K \int_K P_\Sigma (\nabla u_h) \cdot \nabla v dx + \sum_{e \in \mathcal{E}_h} \int_e \alpha^e ([[u_h]]) \cdot [[v]] \, ds = \int_\Omega f v dx. \tag{246}$$

esto, cuando $\nabla P(K) \subset \Sigma(K)$, puede ser visto como una extensión del método de Babuška-Zlámal de Penalización Interior.

3.4. Distintos tipos de Métodos Galerkin Discontinuo

En esta unificación de diversos métodos de Galerkin Discontinuo, en esta sección se resumen las diferentes opciones de flujo que se necesitan para obtener las diversas variantes del método. Para todas las variantes del método $P(K)$ es el espacio de polinomios estándar y $\Sigma(K)$ es tomado tal que contiene $\nabla P(K)$.

Podemos ver que esta división en clases subdivide de forma natural aquellos métodos completamente conservativos y los parcialmente conservativos, por otro lado, divide aquellos cuyo flujo es independiente de σ_h y aquellos que no lo son. Podemos decir que los métodos completamente conservativos generan problemas simétricos cuando los parámetros de su flujo numérico están adecuadamente definidos, y los métodos parcialmente conservativos generan métodos no simétricos.

También notemos que cuyos métodos en los cuales el flujo numérico es independiente de σ_h produce matrices de carga con un marcado número de estradas distintas de cero.

Método	$h_\sigma^{e,K}$	$h_u^{e,K}$
Bassi-Rebay 1	$\dot{\sigma}_h$	$\dot{u}_h$
Brezze et al. 1	$\dot{\sigma}_h - \eta^e \dot{\widehat{r_e\left[[u_h]\right]}}$	$\dot{u}_h$
LDG	$\dot{\sigma}_h - \eta^e \left[[u_h]\right] + \beta^e \left[[\sigma_h]\right]$	$\dot{u}_h + \gamma^e \left[[u_h]\right]$
IP	$\dot{u}_h - \eta^e \left[[u_h]\right]$	$\dot{u}_h$
Bassi-Rebay 2	$\dot{u}_h - \eta^e \dot{\widehat{r_e\left[[u_h]\right]}}$	$\dot{u}_h$
Baumman-Oden	$\dot{u}_h$	$\dot{u}_h - \left[[u_h]\right] \cdot n_K$
Babuška-Zlámal	$-\eta^e \left[[u_h]\right]$	$u_{h\vert_K}$
Brezze et al. 2	$-\eta^e \dot{\widehat{r_e\left[[u_h]\right]}}$	$u_{h\vert_K}$

4. Método Discontinuo Enriquecido

El método estándar de elementos finitos que se basa en polinomios continuos definidos por pedazos mediante la aproximación Galerkin, esta es optima para el operador de Laplace, en el sentido de que este minimiza el error en la norma de energía o en la semi-norma H^1, esta propiedad asegura buen desempeño en el cálculo sobre mallado no muy fino. Sin embargo, un buen desempeño sobre cualquier mallado no esta garantizado para el método de elementos finitos, principalmente en presencia de gradientes grandes y oscilaciones rápidas.

Numerosos métodos se han desarrollado para salvar estas deficiencias, la gran mayoría de ellos se basan en modificaciones del método Galerkin, motivada por el método FETI para descomposición de dominio no conforme con el uso de multiplicadores de Lagrange, el método de Discontinuo Enriquecido [**?**] propone una discretización con elementos finitos estándar con un campo polinomial en el cual cada elemento es enriquecido por un espacio libre de soluciones que gobiernan el problema homogéneo con coeficientes constantes. Este enriquecimiento es fácil de obtener y es virtualmente independiente de la geometría y el orden del polinomio usado en la discretización. De este modo, características de las ecuaciones diferenciales son incluidas en la aproximación.

El concepto de métodos de elementos finitos con multiplicadores de Lagrange para hacer cumplir las restricciones de frontera son bien conocidos y estos han sido exitosamente aplicados a el análisis estructural de sistemas modelados por diferentes tipos de elementos.

Sea $\Omega \subset \mathbb{R}^n$ un dominio con frontera suave $\partial\Omega$, por simplicidad consideraremos el siguiente problema con condiciones de frontera Dirichlet: Encontrar $u : \overline{\Omega} \to \mathbb{R}$ tal que

$$\mathcal{L}u = f \text{ en } \Omega \tag{247}$$

$$u = g \text{ sobre } \partial\Omega \tag{248}$$

donde $f : \Omega \to \mathbb{R}$ y $g : \partial\Omega \to \mathbb{R}$ son funciones dadas, el operador $\mathcal{L}$ es considerado como de segundo orden.

Particionando el dominio Ω en E subdominios $\{\Omega_1, ..., \Omega_E\}$ sin traslape con fronteras $\partial\Omega_i$, con $i = 1, ..., E$ tal que

$$\overline{\Omega} = \bigcup_{i=1}^{E} \overline{\Omega} i \tag{249}$$

donde

$$\bigcap_{i=1}^{E} \Omega_i = \emptyset \tag{250}$$

y denotamos a la unión de los elementos interiores por

$$\widetilde{\Omega} = \bigcup_{i=1}^{E} \Omega_i \tag{251}$$

similarmente, la unión de los elementos de la frontera es denotado por

$$\widetilde{\partial\Omega} = \bigcup_{i=1}^{E} \partial\Omega_i \tag{252}$$

y a los elementos en la interfase o los elementos de la frontera interior es

$$\Gamma = \widetilde{\partial\Omega} \backslash \partial\Omega. \tag{253}$$

Un ejemplo de un dominio Ω y su descomposición en subdominios Ω_i y cada Ω_i a su vez descompuesto en Ω_e subdominios se muestra en la figura:

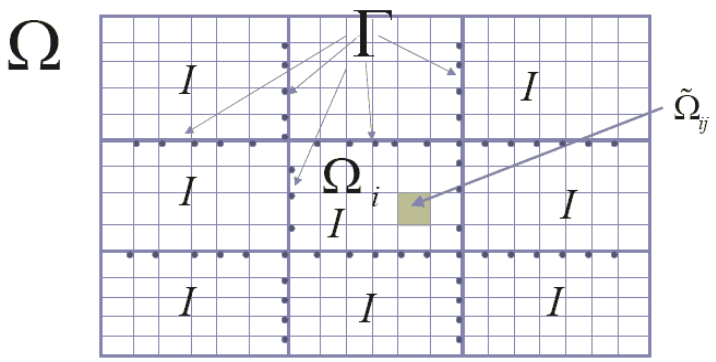

Figura 1: Dominio Ω descompuesto en subdominios Ω_i, con $i = 1, 2, ..., 9$.

Sea $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$ donde $\partial\Omega_i$ y $\partial\Omega_j$ son las fronteras de dos subregiones adyacentes, entonces definimos como la traza a la restricción de v^i a Γ_{ij}. Pero como Γ_{ij}, para dos subregiones vecinas hay dos trazas definidas una que corresponde a v^i y otra a v^j, entonces se requiere introducir la siguiente notación para poderlas distinguir entre si:

$$v_+ \equiv Tr(v^i) \tag{254}$$

cuando Ω_i cae del lado positivo de Γ_{ij} y

$$v_- \equiv Tr(v^j) \tag{255}$$

en caso contrario. Aquí $Tr(v)$ designa al operador traza de la función v. En general $v_+ \neq v_-$ ya que se trabaja con espacios de funciones definidas por tramos.

Observación 42 *Notemos que al considerar una función w en Ω, su definición en Γ es innecesaria, ya que la medida de Lebesgue de Γ es cero. Si la traza de w_α es definida en casi todos lados salvo un conjunto de medida cero sobre $\partial\Omega_\alpha$*

para $\alpha = 1, ..., E$, entonces tal traza es también definida en Γ. En particular, si la traza de w_α esta definida sobre $\partial\Omega_\alpha$ para cada $\alpha = 1, ..., E$, entonces ellas definen dos funciones definidas en casi todos lados salvo un conjunto de medida cero sobre Γ, denotadas por (w_+, w_-) correspondientes a los lados de trazas positivas y negativas de Γ respectivamente.

Definición 43 *El salto de v sobre Γ de funciones definidas por pedazos como*

$$[[w]] \equiv w_+ - w_- \tag{256}$$

y el promedio como

$$\dot{w} \equiv \frac{1}{2}(w_+ + w_-) \tag{257}$$

respectivamente.

4.1. Formulación Variacional Hibrida con Continuidad Débil

La formula variacional del problema con condiciones de frontera dado por la Ecs.(247) y (248) esta dado en términos de el conjunto de soluciones de prueba

$$\mathcal{V} = L^2(\Omega) \cap H^1\left(\widetilde{\partial\Omega}\right) \tag{258}$$

estas funciones posiblemente sean discontinuas a través de los elementos de frontera, similarmente, las funciones no necesariamente satisfacen las condiciones de frontera Dirichlet.

La continuidad entre elementos y las condiciones de frontera Dirichlet son ambas impuestas débilmente por los multiplicadores de Lagrange en $H^{-1/2}\left(\widetilde{\partial\Omega}\right)$, sea

$$H(div;\Omega) = \left\{\mathbf{p} \mid \mathbf{p} \in \left(L^2(\Omega)\right)^n, div\mathbf{p} \in L^2(\Omega)\right\} \tag{259}$$

y tomando $\mathbf{p} \in W = H(div;\Omega)$ entonces las trazas normales de $\mathbf{p}$ sobre $\partial\Omega_i$ son tomados como los multiplicadores de Lagrange. Estas trazas normales $\mathbf{p}\cdot\mathbf{n}$ están bien definidas por pertenecer a $H^{-1/2}\left(\widetilde{\partial\Omega}\right)$ y satisfacen

$$\langle \mathbf{p}\cdot\mathbf{n}, v\rangle_{\widetilde{\partial\Omega}} = (\nabla v, \mathbf{p})_{\Omega_i} + (v, div\mathbf{p})_{\Omega_i} \tag{260}$$

aquí $\langle\cdot,\cdot\rangle$ es la dualidad entre los pares $H^{-1/2}(\partial\Omega)$ y $H^{1/2}(\partial\Omega)$ y los subíndices denotan dominios de integración distintos de $\partial\Omega$, y $(\cdot,\cdot)$ es el producto interior en $L^2(\Omega)$ y los subíndices denotan los dominios de integración distintos de Ω. El vector normal unitario que apunta hacia afuera de la frontera es denotado por $\mathbf{n}$. No son requeridos grados de libertad adicionales en la aproximación por multiplicadores del Lagrange como la traza normal de $\mathbf{p}$, comparada con las funciones escalares definidas sobre los elementos de la frontera.

Ahora, buscaremos el punto estacionario $u \in \mathcal{V}$ y $\mathbf{p} \in \mathcal{W}$ de los multiplicadores de Lagrange

$$\Pi(u,\mathbf{p}) = \frac{1}{2}a(u,u) - \langle \mathbf{p}\cdot\mathbf{n}, v\rangle_{\widetilde{\partial\Omega}} - L(u) - L_b(\mathbf{p}) \tag{261}$$

admitiendo para las discontinuidades, el operador bilineal $a(\cdot,\cdot)$ definido sobre los elementos interiores $\widetilde{\Omega}$, satisfaciendo

$$a(v,u) = (v, \mathcal{L}u)_{\widetilde{\Omega}} + \langle \mathcal{L}_b u, v\rangle_{\widetilde{\partial\Omega}} \tag{262}$$

aquí, $\mathcal{L}_b$ es el operador de frontera correspondiente a $\mathcal{L}$. Los términos representando los datos son

$$\begin{aligned} L(v) &= (v, f) \\ L_b(\mathbf{q}) &= -\langle \mathbf{q}\cdot\mathbf{n}, g\rangle \end{aligned} \tag{263}$$

para funciones f y g suficientemente suaves.

Múltiples métodos de estabilización incluyendo saltos supone el operador de frontera a través de los elementos de la interfase. Tales términos son derivados directamente de las ecuaciones gobernantes del método variacional multiescala. La presente formulación impone continuidad de el campo en si mismo.

4.2. Formulación Débil

El punto estacionario de la funcional dada por le Ec.(261) es obtenida por el establecimiento en la primera variación a cero. En forma particionada, esto conduce a

$$a(v,u) - \langle \mathbf{p}\cdot\mathbf{n}, v\rangle_{\widetilde{\partial\Omega}} = L(v) \tag{264}$$

$$-\langle \mathbf{q}\cdot\mathbf{n}, v\rangle_{\widetilde{\partial\Omega}} = L_b(\mathbf{q}) \tag{265}$$

aquí, $v \in \mathcal{V}$ y $\mathbf{q} \in \mathcal{W}$ son variaciones arbitrarias de u y $\mathbf{p}$, respectivamente.

La clave de las condiciones de estabilidad para la formulación mixta y hibrida son descritos por el Teorema de Brezzi. Ellas se necesitan verificar para el problema dimensional finito. Estas condiciones restringe la selección de la interpolación de elementos finitos que puede usarse para un aplicación particular. La discretización de las Ecs.(264) y (265) conducen a una diagonal por bloque típicamente de ceros.

La correspondiente ecuación Euler-Lagrange típica es

$$\mathcal{L}u = f, \text{ en } \widetilde{\Omega} \tag{266}$$

$$[[u]] = 0, \text{ sobre } \Gamma \tag{267}$$

$$u = g, \text{ sobre } \partial\Omega \tag{268}$$

$$\mathbf{p}\cdot\mathbf{n} = \mathcal{L}_b u, \text{ sobre } \widetilde{\partial\Omega} \tag{269}$$

esta última ecuación facilita una interpretación de los multiplicadores de Lagrange. Por ejemplo, si $\mathcal{L}_b$ es la derivada normal, entonces $\mathbf{p} = \nabla u$ en $\widetilde{\Omega}$.

4.3. Aproximación Galerkin

Buscamos aproximar la solución $u^h \in \mathcal{V}^h \subset \mathcal{V}$ de la forma

$$u^h = u^P + u^Q \tag{270}$$

aquí, $u^P \in \mathcal{V}^P \subset H^1(\Omega)$ son las funciones polinomiales continuas estándar definidas por pedazos de elementos finitos en la escala gruesa y $u^Q \in \mathcal{V}^Q$ es el campo enriquecido. Distinto es en la escala fina, las cuales tienen un rol similar, u^Q puede ser discontinua a través de los elementos de frontera. Esto nos permite burlar los inconvenientes al intentar aproximar la escala fina global de las funciones de Green del método variacional multiescala, y la perdida de los efectos globales esperados por la restricción de los residuales libres que se nulifican en la traza de los elementos de frontera. Un adicional beneficio potencial es que mejora la aproximación de soluciones discontinuas.

Esto proporciona gran flexibilidad en la selección de $\mathcal{V}^Q$ en este esquema. Asumimos que la aproximación de soluciones particulares para $\mathcal{V}^P$ es satisfactoria. El enriquecimiento debe por lo consiguiente contener soluciones de la ecuación parcial homogénea que no son representadas por el subyacente campo polinomial, en este método, el campo enriquecido puede enteramente capturar la solución homogénea, más que simplemente mejorar el campo polinomial.

Débil imposición de continuidad permite el uso de espacios libres de soluciones como base para el enriquecimiento. Consecuentemente la potencial dificultad de problemas con valor de frontera a nivel de elemento no pueda ser resuelto, ni analítica ni numéricamente. El relativamente simple espacio libre de soluciones es aplicable a prácticamente cualquier geometría y orden polinomial que se le aplique al elemento.

Resumiendo, se emplea la suma directa de las relaciones $\mathcal{V}^h = \mathcal{V}^P \oplus \mathcal{V}^Q$, donde $u^Q \in \mathcal{V}^Q$ y $\mathcal{V}^Q$ es generada por las soluciones de

$$\mathcal{L}u^Q = 0 \text{ en } \mathbb{R}^n \tag{271}$$

que no están contempladas en las bases polinomiales. Entonces estas funciones son empleadas a un nivel de elementos, típicamente se emplean soluciones del caso con coeficientes constantes, el cual es fácil de obtener.

El tratamiento de funciones de peso es consistente con la Ec.(270), a saber $v^h = v^P + v^Q$ y $\mathbf{q} \in \mathcal{W}^h$. Por el método Galerkin, se busca $u^h \in \mathcal{V}^h$ y $\mathbf{p} \in \mathcal{W}^h$ tal que para todo $\left\{v^h, \mathbf{q}^h\right\} \in \mathcal{V}^h \times \mathcal{W}^h$ satisfagan

$$a\left(v^h, u^h\right) - \left\langle \mathbf{p}^h \cdot \mathbf{n}, v^h \right\rangle_{\widetilde{\partial\Omega}} = L\left(v^h\right) \tag{272}$$

$$-\left\langle \mathbf{q}^h \cdot \mathbf{n}, v^h \right\rangle_{\widetilde{\partial\Omega}} = L_b\left(\mathbf{q}^h\right) \tag{273}$$

estas ecuaciones pueden ser descompuestas como sigue

$$a\left(v^P, u^P\right) + a\left(v^P, u^Q\right) - \left\langle \mathbf{p}^h \cdot \mathbf{n}, v^P \right\rangle_{\widetilde{\partial\Omega}} = L\left(v^P\right) \tag{274}$$

$$a\left(v^Q, u^P\right) + a\left(v^Q, u^Q\right) - \left\langle \mathbf{p}^h \cdot \mathbf{n}, v^Q \right\rangle_{\widetilde{\partial\Omega}} = L\left(v^Q\right) \tag{275}$$

$$-\left\langle \mathbf{q}^h \cdot \mathbf{n}, u^P \right\rangle_{\widetilde{\partial\Omega}} - \left\langle \mathbf{q}^h \cdot \mathbf{n}, v^E \right\rangle_{\widetilde{\partial\Omega}} = L_b\left(\mathbf{q}^h\right) \tag{276}$$

Debido a la naturaleza discontinua de $\mathcal{V}^Q$, la Ec.(275) puede ser usada para eliminar u^Q por condensación estática en el nivel de elementos. Este procedimiento proporciona una aproximación local (y por tanto económica) de efecto global a escala fina sobre las escalas gruesas. La escala fina es derivada por los elementos interiores residuales $L\left(v^Q\right) - a\left(v^Q, u^P\right)$, y el inter elemento y las discontinuidades de frontera $\left\langle \mathbf{p}^h \cdot \mathbf{n}, v^Q \right\rangle_{\widetilde{\partial\Omega}}$.

4.4. Condensación Estática

Más que meramente un recurso conceptual, la eliminación local de u^Q, conduce a la formulación $u^P - \mathbf{p}$, es propuesto como un procedimiento práctico que simplifica y condiciona la formulación, a fin de reducir el costo computacional. De esta forma, el costo de resolver la matriz del problema que resulta del método es virtualmente independiente de la dimensión de u^Q.

El campo de enriquecimiento generalmente contiene varios grados de libertad en cada elemento. Consecuentemente, la condensación estática es presentada en esta sección en términos de ecuaciones discretas. Para advección-difusión el enriquecimiento puede contener un solo grado de libertad.

Consideremos una partición del sistema global de las ecuaciones discretas

$$\begin{bmatrix} \mathbf{K}^{PP} & \mathbf{K}^{PQ} & \mathbf{K}^{PC} \\ \mathbf{K}^{QP} & \mathbf{K}^{QQ} & \mathbf{K}^{QC} \\ \mathbf{K}^{CP} & \mathbf{K}^{CQ} & 0 \end{bmatrix} \begin{Bmatrix} \mathbf{u}^p \\ \mathbf{u}^Q \\ \mathbf{p} \end{Bmatrix} = \begin{Bmatrix} \mathbf{F}^P \\ \mathbf{F}^Q \\ \mathbf{F}^C \end{Bmatrix} \tag{277}$$

aquí, $\mathbf{u}^p, \mathbf{u}^Q$ y $\mathbf{p}$ son vectores conteniendo los grados de libertad de u^P, u^Q y p^h respectivamente. Las matrices de la Ec.(277) provienen de los términos de las ecuaciones de la formulación Galerkin de acuerdo a las siguientes relaciones

$$\begin{gathered} a(v^P, u^P) \to \mathbf{K}^{PP} \\ a(v^P, u^Q) \to \mathbf{K}^{PQ} \\ -\left\langle \mathbf{p}^h \cdot \mathbf{n}, v^P \right\rangle \to \mathbf{K}^{PC} \\ a(v^Q, u^P) \to \mathbf{K}^{QP} \\ a(v^Q, u^Q) \to \mathbf{K}^{QQ} \\ -\left\langle \mathbf{p}^h \cdot \mathbf{n}, v^Q \right\rangle_{\widetilde{\partial\Omega}} \to \mathbf{K}^{QC} \\ -\left\langle \mathbf{q}^h \cdot \mathbf{n}, v^P \right\rangle \to \mathbf{K}^{CP} \\ -\left\langle \mathbf{q}^h \cdot \mathbf{n}, v^Q \right\rangle_{\widetilde{\partial\Omega}} \to \mathbf{K}^{CQ} \\ L\left(v^P\right) \to \mathbf{F}^P \\ L\left(v^Q\right) \to \mathbf{F}^Q \\ L_b\left(\mathbf{q}^h\right) \to \mathbf{F}^C \end{gathered} \tag{278}$$

debido a la continuidad de u^P, los arreglos $\mathbf{K}^{PC}$ y $\mathbf{K}^{CP}$ son vacíos excepto a lo largo de la frontera del dominio $\partial\Omega$.

El sistema global es obtenido del ensamble de los elementos de los arreglos, el montaje de los nodos polinomiales de los grados de libertad es convencional,

los coeficientes del enriquecimiento son la generalización de los grados de libertad internos a cada elemento. La restricción de los grados de libertad están definidos en los elementos de frontera: Vértices, esquinas y caras en una, dos y tres dimensiones respectivamente. El arreglo de elementos es

$$\mathbf{k}^i = \begin{bmatrix} \mathbf{k}^{PP} & \mathbf{k}^{PQ} & \mathbf{k}^{PC} \\ \mathbf{k}^{QP} & \mathbf{k}^{QQ} & \mathbf{k}^{QC} \\ \mathbf{k}^{CP} & \mathbf{k}^{CQ} & 0 \end{bmatrix} \tag{279}$$

con la obvia correspondencia entre la matriz global y de elementos. Notemos que para resultados óptimos del siguiente procedimiento a nivel de elementos, los términos provienen de $\left\langle \mathbf{p}^h \cdot \mathbf{n}, v^P \right\rangle_{\partial\Omega_i}$ y $\left\langle \mathbf{q}^h \cdot \mathbf{n}, u^P \right\rangle_{\partial\Omega_i}$ deben de esta contenidas en $\mathbf{k}^{PC}$ y $\mathbf{k}^{CP}$ respectivamente, aunque el ensamble cancele a casi todas las entradas a lo largo de la frontera del dominio.

Los grados de libertad del enriquecimiento son eliminados a nivel de elementos para obtener

$$\widetilde{\mathbf{k}}^i = \begin{bmatrix} \widetilde{\mathbf{k}}^{PP} & \widetilde{\mathbf{k}}^{PC} \\ \widetilde{\mathbf{k}}^{CP} & \widetilde{\mathbf{k}}^{CC} \end{bmatrix} \tag{280}$$

donde

$$\begin{aligned} \widetilde{\mathbf{k}}^{PP} &= \mathbf{k}^{PP} - \mathbf{k}^{PQ} \left(\mathbf{k}^{QQ}\right)^{-1} \mathbf{k}^{QP} \\ \widetilde{\mathbf{k}}^{PC} &= \mathbf{k}^{PC} - \mathbf{k}^{PQ} \left(\mathbf{k}^{QQ}\right)^{-1} \mathbf{k}^{QC} \\ \widetilde{\mathbf{k}}^{CP} &= \mathbf{k}^{CP} - \mathbf{k}^{CQ} \left(\mathbf{k}^{QQ}\right)^{-1} \mathbf{k}^{QP} \\ \widetilde{\mathbf{k}}^{CC} &= -\mathbf{k}^{CQ} \left(\mathbf{k}^{QQ}\right)^{-1} \mathbf{k}^{QC} \end{aligned} \tag{281}$$

la condensación estática elimina la diagonal por bloques de ceros de la matriz sin condensación.

El sistema global que resulta, reduce $u^P - \mathbf{p}$ la formulación la cual es obtenida como un ensamble del arreglo de elementos. Este sistema en particular es adecuado para soluciones iterativas. La solución para el campo eliminado es obtenida como un post-procesamiento con cada elemento.

4.5. Aproximación de los Multiplicadores de Lagrange

Una amplia revisión de técnicas de aproximación de los multiplicadores de Lagrange se han empleado en este método, nos concentraremos en el escalamiento de las funciones base de los multiplicadores de Lagrange usando el factor de escala s con la dimensión de $\mathcal{L}_b$. Para $\mathcal{L}$ dada, s se escoge tal que los coeficientes de las entradas de las matrices correspondientes a $\mathbf{p}$ y a u sean del mismo orden de magnitud, con la idea de mejorar el condicionamiento de la matriz.

Consideremos a los multiplicadores de Lagrange que sean constantes a lo largo de las caras de un triángulo. Esto se consigue en el presente esquema como las trazas normales de

$$\mathbf{p}^h(x,y) = s \left\{ \begin{array}{c} c_1 + c_3 x \\ c_2 + c_3 y \end{array} \right\}, \text{ con } (x,y) \in \Omega_i \tag{282}$$

en las caras del triángulo, originalmente denotada por RT_0. En este caso $div\mathbf{p}^h = const$ en Ω_i.

En un triángulo con los multiplicadores de Lagrange que varían linealmente a lo largo de las caras, denotado por BDM, se consigue al considerar nodos en los tres vértices, con la interpolación lineal estándar de los valores en los nodos de $\mathbf{p}^h$. Los seis grados de libertad de estos elementos puede ser remplazada por los seis componentes normales de $\mathbf{p}^h$ sobre $\partial\Omega_i$ (dos por cara). Sin embargo, la representación nodal es particular de la adecuada estructura convencional de datos de los elementos finitos.

La aproximación para cuadriláteros se define en términos de las coordenadas naturales en el cuadrado de referencia que está alineado con los ejes cartesianos. En el caso de que la aproximación se especifique en términos de la componente normal de las caras de los elementos, el mapeo del dominio físico es llevado a cabo por un cambio de variables conocida como la transformación de Piola tal que la componente normal es preservada. En el caso, cuando los valores nodales son usados en conjunción con la integración de acuerdo a la integración del lado derecho de la Ec.(260) muy probablemente se deberá de usar funciones isoparamétricas.

Considerando multiplicadores de Lagrange que son constantes a lo largo de las caras del cuadrado de referencia. Es obtenida en el presente esquema como las trazas normales de

$$\mathbf{p}^h(x,y) = s\left\{\begin{array}{c} c_1 + c_2 x \\ c_3 + c_4 y \end{array}\right\} \tag{283}$$

en las caras del cuadrado original denotado BDFM y la cual coincide con RT_0 para rectángulos. En este caso $div\mathbf{p}^h = const$ en Ω_i como para RT_0. La traza normal es constante, como es requerido y la aproximación puede ser especificada por las cuatro componentes normales de $\mathbf{p}^h$ sobre la frontera.

4.6. Condiciones de Frontera Neumann y Robin

Hasta aquí se considero sólo el caso de condiciones de frontera Dirichlet. Sin embargo la formulación preserva la estructura de las matrices a nivel de elementos de las Ecs.(279) y (280) en presencia de condiciones de frontera tipo Neumann y Robin.

Considerando una partición de la frontera del dominio $\partial\Omega = \partial\Omega_D \cup \partial\Omega_R$ donde $\partial\Omega_D \cap \partial\Omega_R = \emptyset$. Se asume que las condiciones sobre la frontera tipo Dirichlet sobre toda la frontera del dominio dada por la Ec.(248) es remplazada por

$$u = g \text{ sobre } \partial\Omega_D \tag{284}$$

$$\mathcal{L}_b u + \alpha u = \beta \text{ sobre } \partial\Omega_R \tag{285}$$

aquí, $g : \partial\Omega_D \to \mathbb{R}$, $\alpha : \partial\Omega_R \to \mathbb{R}$ y $\beta : \partial\Omega_R \to \mathbb{R}$ son funciones dadas. La Ec.(285) representa las condiciones de frontera tipo Robin, también las condiciones tipo Neumann en el caso espacial de $\alpha = 0$.

Extendiendo $\mathbf{p}$ a $\partial\Omega_R$ como sigue

$$\mathcal{W} = \{\mathbf{p} \mid \mathbf{p} \in \mathbf{H}(div, \Omega), \mathbf{p} \cdot \mathbf{n} = -\beta \text{ sobre } \partial\Omega_R\} \tag{286}$$

y la funcional dada por la Ec.(261) es modificada por

$$\Pi(u, \mathbf{p}) = \frac{1}{2} a(u, u) + \frac{1}{2} \langle \alpha u, u \rangle_{\partial\Omega_R} - \langle \mathbf{p} \cdot \mathbf{n}, v \rangle_{\widetilde{\partial\Omega}} - L(u) - L_b(\mathbf{p}). \tag{287}$$

Esto conduce a la modificación de la forma débil

$$a(v, u) + \langle \alpha u, v \rangle_{\partial\Omega_R} - \langle \mathbf{p} \cdot \mathbf{n}, v \rangle_{\widetilde{\partial\Omega}} = L(v) \tag{288}$$

$$- \langle \mathbf{q} \cdot \mathbf{n}, u \rangle_{\widetilde{\partial\Omega}} = -L_b(\mathbf{q}) \tag{289}$$

donde,

$$\mathbf{q} \in \mathcal{W}_0 = \{\mathbf{q} \mid \mathbf{q} \in \mathbf{H}(div, \Omega), \mathbf{q} \cdot \mathbf{n} = 0 \text{ sobre } \partial\Omega_R\}. \tag{290}$$

La Ec.(269) de Euler-Lagrange es ahora remplazada por

$$\mathbf{p} \cdot \mathbf{n} = \mathcal{L}_b u \text{ sobre } \Gamma \cup \partial\Omega_D \tag{291}$$

$$\mathbf{p} \cdot \mathbf{n} = \mathcal{L}_b u + \alpha u \text{ sobre } \partial\Omega_R \tag{292}$$

para condiciones de frontera tipo Neumann ($\alpha = 0$) la definición de $\mathbf{p}$ no es cambiada.

La discretización de las formulaciones anteriores están contenidas en las matrices a nivel de elementos de las Ecs.(279) y (280). El proceso de ensamble ahora cuenta con la imposición de los valores $\mathbf{p} \cdot \mathbf{n}$ sobre $\partial\Omega_R$ como una condición esencial de frontera, de la misma manera que las condiciones de frontera tipo Dirichlet son impuestas en el método de elementos finitos convencional. En otras palabras, los grados de libertad asociados al nivel de elementos con $\mathbf{p} \cdot \mathbf{n}$ sobre $\partial\Omega_R$ no son ensamblados dentro de los coeficientes de la matriz global. En cambio, para datos inhomogéneos ($\beta \neq 0$), se usan los términos del lado derecho. El proceso de discretización es el dado por las siguientes relaciones

$$\begin{aligned}
a(v^P, u^P) + \left\langle \alpha u^P, v^P \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{K}^{PP} \\
a(v^P, u^Q) + \left\langle \alpha u^Q, v^P \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{K}^{PQ} \\
- \left\langle \mathbf{p}^h \cdot \mathbf{n}, v^P \right\rangle_{\partial\Omega_D} &\rightarrow \mathbf{K}^{PC} \\
a(v^Q, u^P) + \left\langle \alpha u^P, v^Q \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{K}^{QP} \\
a(v^Q, u^Q) + \left\langle \alpha u^Q, v^Q \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{K}^{QQ} \\
- \left\langle \mathbf{p}^h \cdot \mathbf{n}, v^Q \right\rangle_{\Gamma \cup \partial\Omega_D} &\rightarrow \mathbf{K}^{QC} \\
- \left\langle \mathbf{q}^h \cdot \mathbf{n}, v^P \right\rangle_{\partial\Omega_D} &\rightarrow \mathbf{K}^{CP} \\
- \left\langle \mathbf{q}^h \cdot \mathbf{n}, v^Q \right\rangle_{\Gamma \cup \partial\Omega_D} &\rightarrow \mathbf{K}^{CQ} \\
L\left(v^P\right) + \left\langle \beta, v^P \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{F}^P \\
L\left(v^Q\right) + \left\langle \beta, v^Q \right\rangle_{\partial\Omega_R} &\rightarrow \mathbf{F}^Q \\
L_b\left(\mathbf{q}^h\right) &\rightarrow \mathbf{F}^C
\end{aligned} \tag{293}$$

siendo estas las modificaciones para condiciones de frontera tipo Neumann y Robin.

5. Método FETI

Algunas de las formulaciones más conocidas de los métodos de descomposición de domino están basadas en un análisis de las condiciones de transmisión entre las interfaces del subdominio, los cuales usan al operador de Steklov-Poincaré. Nos interesa el estudio sistemático de algunas de las mejores formulaciones de métodos de descomposición de domino que se basen en la aplicación del operador de Steklov-Poincaré. Estos métodos están basados en una aproximación especial de las formulas de Green aplicable a funciones discontinuas.

El concepto unificador básico de la teoría, consiste en interpretar los métodos de descomposición de dominio como procedimientos para obtener información acerca de la solución en la frontera interior la cual separa el subdominio de cada uno de los otros, suficiente para definir problemas bien planteados en cada uno de los subdominios - referidos como problemas locales-. De esta manera, la solución puede ser reconstruida al resolver cada uno de los problemas locales exclusivamente.

La familia de algoritmos FETI (Finite Element Tearing and Interconnecting) y Neumann-Neumann [5] y [2] son de los métodos mejor conocidos y más probados para la resolución de ecuaciones diferenciales parciales elípticas. Ellos son métodos iterativos de subestructuración Sec.(10.4) y comparten muchos componentes algorítmicos, tales como soluciones locales para ambos problemas con condiciones de frontera Neumann y Dirichlet sobre las subregiones en donde el problema fue particionado.

5.1. Conceptos Básicos

En este capítulo se considerarán problemas con valor en la frontera (VBVP) de la forma

$$\begin{aligned} -\nabla u &= f \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{294}$$

entonces el problema dado por la Ec. (294) se reescribe como: hallar $\underline{u} \in H_0^1(\Omega)$ tal que $a(\underline{u}, \underline{v}) = l(\underline{v})$.

Consideremos el problema dado por la Ec. (294) en el dominio Ω, el cual es en general subdividido en E subdominios Ω_i, $i = 1, ..., E$ sin traslape, es decir

$$\Omega_i \cap \Omega_j = \varnothing \quad \forall i \neq j \quad \text{y} \quad \overline{\Omega} = \bigcup_{i=1}^{E} \overline{\Omega}_i, \tag{295}$$

y al conjunto

$$\Gamma = \bigcup_{i=1}^{E} \Gamma_i, \quad \text{si } \Gamma_i = \partial\Omega_i \backslash \partial\Omega \tag{296}$$

lo llamaremos la frontera interior del dominio Ω, denotamos por H al diámetro $H_i = Diam(\Omega_i)$ de cada Ω_i que satisface $Diam(\Omega_i) \leq H$ para cada $i =$

$1, 2, ..., E$, además, cada subdominio Ω_i es descompuesto en una mallado fino $\mathcal{T}_h$ de K subdominios mediante una triangulación Ω_e de modo que esta sea conforme, denotamos por h al diámetro $h_i = Diam(\Omega_e)$ de cada Ω_e que satisface $Diam(\Omega_e) \leq h$ para cada $e = 1, 2, ..., K$ de cada $i = 1, 2, ..., E$.

Un ejemplo de un dominio Ω y su descomposición en subdominios Ω_i y cada Ω_i a su vez descompuesto en Ω_e subdominios como se muestra en la figura:

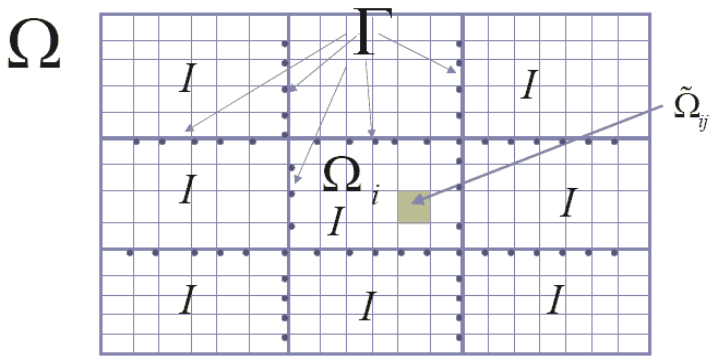

Figura 2: Dominio Ω descompuesto en una partición gruesa de 3×3 y cada subdominio Ω_i en una partición fina de 7×5.

Sin perdida de generalidad tomemos $g = 0$ en $\partial\Omega$, notemos que siempre es posible poner el problema de la Ec. (294) como uno con condiciones de frontera Dirichlet que se nulifiquen mediante la adecuada manipulación del término del lado derecho de la ecuación.

Denotemos por $V^h(\Omega_i)$ al espacio de funciones definidas por pedazos que son generadas por el método de elemento finito estándar las cuales se nulifican en $\partial\Omega_i \cap \partial\Omega$.

La aproximación por elementos finitos de un problema elíptico que es continuo a través de Γ se denota por $V^h(\Omega)$.

Si consideramos el dominio Ω particionado en dos subdominios Ω_1 y Ω_2, la formulación dada por la Ec. (294), es equivalente al siguiente problema acoplado

$$\begin{aligned} -\Delta u_1 &= f \quad \text{en } \Omega_1 \\ u_1 &= 0 \quad \text{en } \partial\Omega_1 \backslash \Gamma \\ u_1 &= u_2 \quad \text{en } \Gamma \\ \frac{\partial u_1}{\partial n_1} &= \frac{\partial u_2}{\partial n_2} \quad \text{en } \Gamma \\ -\Delta u_2 &= f \quad \text{en } \Omega_2 \\ u_2 &= 0 \quad \text{en } \partial\Omega_2 \backslash \Gamma \end{aligned} \tag{297}$$

aquí u_i es la restricción de u a Ω_i y $\underline{n_i}$ es el vector normal a Ω_i. La condición sobre la interfase Γ es llamada las condiciones de transmisibilidad y ellas son

equivalentes a la igualdad de cualquier combinación lineal independiente de trazas de funciones y sus derivadas normales, a las derivadas normales en muchos ámbitos se les conoce también como flujo.

Considerando una triangulación del dominio y una aproximación de la Ec. (294) por del método de elemento finito ver sección (10.3). Tal aproximación da origen a un sistema lineal

$$\underline{\underline{A}}\,\underline{u} = \underline{b} \tag{298}$$

donde la matriz $\underline{\underline{A}}$ es simétrica y positiva definida, la cual para una malla de diámetro h, típicamente tiene un número de condicionamiento sobre el orden de $1/h^2$. Aquí

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{A}}^1_{II} & 0 & \underline{\underline{A}}^1_{I\Gamma} \\ 0 & \underline{\underline{A}}^2_{II} & \underline{\underline{A}}^2_{I\Gamma} \\ \underline{\underline{A}}^1_{\Gamma I} & \underline{\underline{A}}^2_{\Gamma I} & \underline{\underline{A}}_{\Gamma\Gamma} \end{bmatrix}, \underline{u} = \begin{bmatrix} \underline{u_I}_1 \\ \underline{u_I}_2 \\ \underline{u_\Gamma} \end{bmatrix} \text{ y } \underline{b} = \begin{bmatrix} \underline{b_I}_1 \\ \underline{b_I}_2 \\ \underline{b_\Gamma} \end{bmatrix} \tag{299}$$

se ha particionado en los grados de libertad de los nodos interiores de Ω_1 y Ω_2 y los nodos de la frontera interior Γ. La matriz de rigidez $\underline{\underline{A}}$ y el vector de carga $\underline{f}$ se obtienen por subensamble de las correspondientes contribuciones de ambos subdominios, i.e.

$$\underline{\underline{A}}_i = \begin{bmatrix} \underline{\underline{A}}^i_{II} & \underline{\underline{A}}^i_{I\Gamma} \\ \underline{\underline{A}}^i_{\Gamma I} & \underline{\underline{A}}^i_{\Gamma\Gamma} \end{bmatrix} \text{ y } \underline{b}_i = \begin{bmatrix} \underline{b_I}_i \\ \underline{b_\Gamma}_i \end{bmatrix} \text{ con } i = 1, 2 \tag{300}$$

son las matrices y vectores de rigidez para el problema de Poisson con condiciones de frontera Dirichlet sobre $\partial\Omega_i \backslash \Gamma$ y condiciones de Neumann sobre Γ, así tenemos que

$$\underline{\underline{A}}_{\Gamma\Gamma} = \underline{\underline{A}}^1_{\Gamma\Gamma} + \underline{\underline{A}}^2_{\Gamma\Gamma} \text{ y } \underline{b_\Gamma}_i = \underline{b_\Gamma}_1 + \underline{b_\Gamma}_2. \tag{301}$$

De la Ec. (297), se buscará una aproximación a las condiciones de transmisión, mediante una aproximación a las derivadas normales sobre Γ. Suponiendo la existencia de la solución exacta local $\underline{u}_i$, su derivada normal puede ser definida como una funcional lineal usando la formula de Green Ec. (115). Así, si ϕ_j es una base nodal de funciones para los nodos de Γ, se tiene de la Ec. (297) la siguiente expresión

$$\int_\Gamma \frac{\partial \underline{u}_i}{\partial \underline{n}_i} \phi_j \, d\underline{s} = \int_{\Omega_i} \left(\Delta \underline{u}_i \phi_j + \nabla \underline{u}_i \cdot \nabla \phi_j \right) d\underline{x} = \int_{\Omega_i} \left(-\underline{f} \phi_j + \nabla \underline{u}_i \cdot \nabla \phi_j \right) d\underline{x}. \tag{302}$$

Una aproximación λ_i de la funcional representante de la derivada normal puede ser encontrada mediante el reemplazo de la solución exacta $\underline{u}_i$ en el lado derecho de la ecuación anterior con la aproximación por elementos finitos. Corriendo j sobre todos los nodos de Γ y usando la definición de la matriz de carga local, se introduce la expresión

$$\lambda_i = \underline{\underline{A}}^i_{\Gamma I} \underline{u_I}_i + \underline{\underline{A}}^i_{\Gamma\Gamma} \underline{u_\Gamma}_i - \underline{b_\Gamma}_i. \tag{303}$$

Notemos que esta expresión coincide con el residual correspondiente a los nodos sobre Γ del problema de Poisson con condiciones de Neumann sobre Γ.

Usando la Ec. (303), se aproxima la Ec. (297) mediante

$$\begin{aligned} \underline{\underline{A}}^1_{II}\underline{u_I}_1 + \underline{\underline{A}}^1_{I\Gamma}\underline{u_\Gamma}_1 &= \underline{b_I}_1 \\ \underline{u_\Gamma}_1 &= \underline{u_\Gamma}_2 = \underline{u_\Gamma} \\ \left(\underline{\underline{A}}^1_{\Gamma I}\underline{u_I}_1 + \underline{\underline{A}}^1_{\Gamma\Gamma}\underline{u_\Gamma}_1 - \underline{b_\Gamma}_1\right) &= -\left(\underline{\underline{A}}^2_{\Gamma I}\underline{u_I}_2 + \underline{\underline{A}}^2_{\Gamma\Gamma}\underline{u_\Gamma}_2 - \underline{b_\Gamma}_2\right) = \underline{\lambda_\Gamma} \\ \underline{\underline{A}}^2_{II}\underline{u_I}_1 + \underline{\underline{A}}^2_{I\Gamma}\underline{u_\Gamma}_2 &= \underline{b_I}_2 \end{aligned} \tag{304}$$

nótese que la primera y última ecuación es la discretización del problema de Poisson para las funciones interiores $\underline{u_I}_i$ con condiciones de frontera tipo Dirichlet que se nulifican sobre $\partial\Omega_i\backslash\Gamma$ y es igual al valor común $\underline{u_\Gamma}$ sobre Γ. Alternativamente, la primera y tercera ecuación provee una discretización del problema de Poisson en Ω_1 para la función local $\underline{u}_1$ con datos de frontera Neumann igual a $\underline{\lambda_\Gamma}$ y que se nulifique sobre $\partial\Omega_1\backslash\Gamma$, de forma análoga se formula un problema con datos de frontera en Ω_2 con las ecuaciones tres y cuatro.

Es posible también obtener una ecuación para la traza de la solución exacta sobre Γ trabajando directamente con el problema continuo Ec. (303) el correspondiente operador es llamado Steklov-Poincaré. El complemento de Schur sección (10.4) es una aproximación de la ecuación de Steklov-Poincaré, determinado directamente por la aproximación mediante el método de elemento finito, particularmente, por la aproximación de la derivada normal Ec. (303).

5.1.1. Una Ecuación para el Flujo Usando el Complemento de Schur

Consideremos el sistema de ecuaciones lineales dado por la Ec. (298), donde la matriz $\underline{\underline{A}}$, $\underline{u}$ y $\underline{b}$ se definen como en Ec. (299). En el primer paso de la gran mayoría de los métodos iterativos de descomposición de dominio, los nodos interiores $\underline{u_I}_i$ desconocidos en los subdominios son eliminadados. Esto corresponde a un factorización por bloques de la matriz de la Ec. (299) en

$$\underline{\underline{A}} = \begin{bmatrix} \underline{\underline{I}} & 0 & 0 \\ 0 & \underline{\underline{I}} & 0 \\ \underline{\underline{A}}^1_{\Gamma I}\left(\underline{\underline{A}}^1_{II}\right)^{-1} & \underline{\underline{A}}^2_{\Gamma I}\left(\underline{\underline{A}}^2_{II}\right)^{-1} & \underline{\underline{I}} \end{bmatrix} \begin{bmatrix} \underline{\underline{A}}^1_{II} & 0 & \underline{\underline{A}}^1_{I\Gamma} \\ 0 & \underline{\underline{A}}^2_{II} & \underline{\underline{A}}^2_{I\Gamma} \\ 0 & 0 & \underline{\underline{S}} \end{bmatrix} \tag{305}$$

resultando el sistema lineal

$$\begin{pmatrix} \underline{\underline{A}}^1_{II} & 0 & \underline{\underline{A}}^1_{I\Gamma} \\ 0 & \underline{\underline{A}}^2_{II} & \underline{\underline{A}}^2_{I\Gamma} \\ 0 & 0 & \underline{\underline{S}} \end{pmatrix} \underline{u} = \begin{bmatrix} \underline{b_I}_1 \\ \underline{b_I}_2 \\ \underline{g_\Gamma} \end{bmatrix}. \tag{306}$$

Aquí $\underline{\underline{I}}$ es la matriz identidad y

$$\underline{\underline{S}} = \underline{\underline{A}}_{\Gamma\Gamma} - \underline{\underline{A}}^1_{\Gamma I}\left(\underline{\underline{A}}^1_{II}\right)^{-1}\underline{\underline{A}}^1_{I\Gamma} - \underline{\underline{A}}^2_{\Gamma I}\left(\underline{\underline{A}}^2_{II}\right)^{-1}\underline{\underline{A}}^2_{I\Gamma} \tag{307}$$

es la matriz del complemento de Schur véase sección (10.4) relativo a las incógnitas sobre Γ. También definimos el complemento de Schur local como

$$\underline{\underline{S}}^i = \underline{\underline{A}}^i_{\Gamma\Gamma} - \underline{\underline{A}}^i_{\Gamma I}\left(\underline{\underline{A}}^i_{II}\right)^{-1}\underline{\underline{A}}^i_{I\Gamma} \tag{308}$$

con $i = 1, 2$. Pudiendo encontrar el complemento de Schur para $\underline{u_\Gamma}$ mediante el sistema

$$\underline{\underline{S}}\,\underline{u_\Gamma} = \underline{g_\Gamma} \tag{309}$$

donde

$$\begin{aligned}
\underline{\underline{S}} &= \underline{\underline{S}}^1 + \underline{\underline{S}}^2 \\
\underline{g_\Gamma} &= \underline{g_\Gamma}_1 + \underline{g_\Gamma}_2 \\
\underline{g_\Gamma}_1 &= \left(\underline{b_\Gamma}_1 - \underline{\underline{A}}^1_{\Gamma I}\left(\underline{\underline{A}}^1_{II}\right)^{-1}\underline{b_I}_1\right) \\
\underline{g_\Gamma}_2 &= \left(\underline{b_\Gamma}_2 - \underline{\underline{A}}^2_{\Gamma I}\left(\underline{\underline{A}}^2_{II}\right)^{-1}\underline{b_I}_2\right)
\end{aligned} \tag{310}$$

nótese que una vez encontrado $\underline{u_\Gamma}$ mediante la resolución del sistema de la Ec. (309) los nodos interiores pueden ser encontrados usando

$$\underline{u_I}_i = \left(\underline{\underline{A}}^i_{II}\right)^{-1}\left(\underline{b_I}_i - \underline{\underline{A}}^i_{I\Gamma}\underline{u_\Gamma}\right). \tag{311}$$

Para derivar una ecuación para la derivada normal $\underline{\lambda_\Gamma}$ sobre Γ se usa los valores $\underline{\lambda_\Gamma} = \underline{\lambda_\Gamma}_1 = -\underline{\lambda_\Gamma}_2$ de frontera interior que son desconocidos tercera ecuación de la Ec. (304) y resolver el sistema local de problemas de Neumann para encontrar $\underline{u}_1$ y $\underline{u}_2$, i.e.

$$\begin{bmatrix} \underline{\underline{A}}^i_{II} & \underline{\underline{A}}^i_{I\Gamma} \\ \underline{\underline{A}}^i_{\Gamma I} & \underline{\underline{A}}^i_{\Gamma\Gamma} \end{bmatrix} \begin{bmatrix} \underline{u_I}_i \\ \underline{u_\Gamma}_i \end{bmatrix} = \begin{bmatrix} \underline{b_I}_i \\ \underline{b_\Gamma}_i + \underline{\lambda_\Gamma} \end{bmatrix}, \text{ con } i = 1, 2 \tag{312}$$

usando factorización por bloques de las matrices locales, se tiene

$$\underline{u}^\Gamma_i = \left(\underline{\underline{S}}^i\right)^{-1}\left(\underline{g_\Gamma}_i + \underline{\lambda_\Gamma}_i\right) \tag{313}$$

donde $\underline{g_\Gamma}_i$ es dado como en la Ec. (310). Usando la segunda ecuación del problema de la Ec. (304) la cual crea $\underline{u_\Gamma}_1$ y $\underline{u_\Gamma}_2$ iguales, encontramos la ecuación para el flujo, y esta dada por

$$\underline{\underline{F}}\,\underline{\lambda_\Gamma} = \underline{d_\Gamma} \tag{314}$$

con

$$\begin{aligned}
\underline{\underline{F}} &= \left(\underline{\underline{S}}^1\right)^{-1} + \left(\underline{\underline{S}}^2\right)^{-1} \\
\underline{d_\Gamma} &= \underline{d_\Gamma}_1 + \underline{d_\Gamma}_2 \\
\underline{d_\Gamma}_1 &= \left(\underline{\underline{S}}^1\right)^{-1}\underline{g_\Gamma}_1 \\
\underline{d_\Gamma}_2 &= \left(\underline{\underline{S}}^2\right)^{-1}\underline{g_\Gamma}_2.
\end{aligned} \tag{315}$$

5.1.2. Extensión Armónica Discreta

El espacio de funciones armónicas discretas es un subespacio importante, relacionado directamente con el complemento de Schur y de los valores de los nodos sobre Γ.

Denotaremos por $\mathcal{H}_i\left(\underline{u_\Gamma}\right)$ a la proyección de los nodos $\underline{u_\Gamma} \in \Gamma$ a los nodos interiores $\underline{u_I}_i$ pertenecientes al subdominio Ω_i.

Definición 44 *Una función $\underline{u}_i$ definida sobre Ω_i es llamada armónica discreta sobre Ω_i si*

$$\underline{\underline{A}}^i_{II}\underline{u_I}_i + \underline{\underline{A}}^i_{I\Gamma}\underline{u_\Gamma}_i = 0. \tag{316}$$

Notemos que $\underline{u}_i = \mathcal{H}_i\left(\underline{u_\Gamma}\right)$ es completamente definido por sus valores sobre $\partial\Omega_i \cap \Gamma$ y es ortogonal -en el producto interior $a(\cdot,\cdot)$ – al espacio $V^h \cap H^1_0(\Omega_i)$.

El espacio $\widetilde{V^h} \subset V^h(\Omega)$ global de funciones armónicas discretas consistentes de funciones armónicas discretas sobre cada subdominio Ω_i. Una función $\underline{u}$ pertenece a $\widetilde{V^h}$ si y sólo si

$$\underline{\underline{A}}^i_{II}\underline{u_I} + \underline{\underline{A}}^i_{I\Gamma}\underline{u_\Gamma} = 0 \tag{317}$$

y es completamente definida por sus valores sobre Γ. El espacio $\widetilde{V^h}$ es ortogonal -en el producto interior $a(\cdot,\cdot)$- a todos los espacios interiores

$$V^h \cap H^1_0(\Omega_i), \text{ con } i = 1, ..., E. \tag{318}$$

Denotaremos la extensión armónica discreta por pedazos de $\underline{u_\Gamma}$ por $\mathcal{H}\left(\underline{u_\Gamma}\right)$.

5.2. One-Level FETI

5.2.1. Algoritmos en Dos Subdominios

En esta subsección se mostrarán dos algoritmos uno corresponde al Neumann-Neumann y otro al Dirichlet-Dirichlet, este último conocido como FETI.

El Algoritmo Neumann-Neumann Si consideramos el dominio Ω particionado en dos subdominios Ω_1 y Ω_2, la formulación dada por la Ec. (294), puede ser expresada como un problema iterativo en el se inicia con un valor supuesto $\underline{u_\Gamma}_0$ el primer paso en este algoritmo consiste en resolver un problema con condiciones de Frontera Dirichlet en cada Ω_i con datos $\underline{u_\Gamma}_0$ sobre Γ, para después resolver un problema en cada subdominio con condiciones de frontera Neumann sobre Γ eligiendo como la derivada normal la diferencia de la solución de los dos problemas con condiciones de frontera Dirichlet. Los valores sobre Γ de la solución de dichos problemas Neumann es empleado para corregir el valor inicial $\underline{u_\Gamma}_0$ y encontrar el nuevo valor de $\underline{u_\Gamma}_1$, y así continuar de manera iterativa

para $n \geq 0$, y el algoritmo queda expresado como

$$(D_i) \left\{ \begin{array}{ll} -\Delta \underline{u}_i^{n+1/2} = f & , \text{ en } \Omega_i \\ \underline{u}_i^{n+1/2} = 0 & , \text{ en } \partial\Omega_i \backslash \Gamma \\ \underline{u}_i^{n+1/2} = \underline{u_\Gamma}^n & , \text{ en } \Gamma \end{array} \right\}, \text{ con } i = 1, 2 \tag{319}$$

$$(N_i) \left\{ \begin{array}{ll} -\Delta \underline{v}_i^{n+1} = 0 & , \text{en } \Omega_i \\ \underline{v}_i^{n+1} = 0 & , \text{en } \partial\Omega_i \backslash \Gamma \\ \frac{\partial v^{n+1}}{\partial \underline{n}_i} = \frac{\partial u_1^{n+1/2}}{\partial \underline{n}_1} + \frac{\partial u_2^{n+1/2}}{\partial \underline{n}_2} & , \text{ en } \Gamma \end{array} \right\}, \text{ con } i = 1, 2$$

$$\underline{u_\Gamma}^{n+1} = \underline{u_\Gamma}^n - \theta \left(\underline{v}_1^{n+1} + \underline{v}_2^{n+1} \right) \qquad \text{en } \Gamma$$

para algún adecuado $\theta \in (0, \theta_{máx})$. Usando una aproximación a la derivada normal, se derivará un método iterativo para el problema discreto. Definiendo los vectores $\underline{w}_i = \underline{u}_{I_i}$ y $\underline{r}_i = \underline{v}_{I_i}$ entonces se tiene que

$$(D_i) \left\{ \underline{\underline{A}}_{II}^i \underline{w}_i^{n+1/2} + \underline{\underline{A}}_{I\Gamma}^i \underline{u_\Gamma}^n = \underline{b}_{I_i}, \text{ con } i = 1, 2 \right\} \tag{320}$$

$$(N_i) \left\{ \left[\begin{array}{cc} \underline{\underline{A}}_{II}^i & \underline{\underline{A}}_{I\Gamma}^i \\ \underline{\underline{A}}_{\Gamma I}^i & \underline{\underline{A}}_{\Gamma\Gamma}^i \end{array} \right] \left[\begin{array}{c} \underline{w}_i^{n+1} \\ \underline{r}_i^{n+1} \end{array} \right] = \left[\begin{array}{c} 0 \\ \underline{t_\Gamma} \end{array} \right] \right\}, \text{ con } i = 1, 2$$

$$\underline{u_\Gamma}^{n+1} = \underline{u_\Gamma}^n - \theta \left(\underline{\eta}_1^{n+1} + \underline{\eta}_2^{n+1} \right)$$

donde el residual $\underline{t_\Gamma}$ es definido como

$$\underline{t_\Gamma} = \left(\underline{\underline{A}}_{\Gamma I}^1 \underline{w}_1^{n+1/2} + \underline{\underline{A}}_{\Gamma\Gamma}^1 \underline{u_\Gamma}^n - \underline{b_\Gamma}_1 \right) + \left(\underline{\underline{A}}_{\Gamma I}^2 \underline{w}_2^{n+1/2} + \underline{\underline{A}}_{\Gamma\Gamma}^2 \underline{u_\Gamma}^n - \underline{b_\Gamma}_2 \right) \tag{321}$$

en vista de la tercera ecuación de la Ec. (304).

Eliminando $\underline{w}_i^{n+1/2}, \underline{r}_i^{n+1/2}$ de la Ec. (320) entonces (D_i) queda dado por

$$\underline{t_\Gamma} = -\left(\underline{g_\Gamma} - \underline{\underline{S}} \underline{u_\Gamma}^n \right), \tag{322}$$

el cual muestra que la diferencia $\underline{t_\Gamma}$ del flujo local es igual a menos el residual del sistema del complemento de Schur. Usando factorización por bloques de las matrices locales $\underline{\underline{A}}_i$, los problemas (N_i), quedan en términos de

$$\underline{\eta}_i^{n+1} = \left(\underline{\underline{S}}^i \right)^{-1} \underline{t_\Gamma} = -\left(\underline{\underline{S}}^i \right)^{-1} \left(\underline{g_\Gamma} - \underline{\underline{S}} \underline{u_\Gamma}^n \right). \tag{323}$$

Por lo tanto, encontramos

$$\underline{u_\Gamma}^{n+1} - \underline{u_\Gamma}^n = \theta \left(\left(\underline{\underline{S}}^1 \right)^{-1} + \left(\underline{\underline{S}}^2 \right)^{-1} \right) \left(\underline{g_\Gamma} - \underline{\underline{S}} \underline{u_\Gamma}^n \right) \tag{324}$$

lo cual muestra que el algoritmo Neumann-Neumann es también un sistema iterativo precondicionado de Richardson para el sistema del complemento de Schur, con precondicionador $\left(\underline{\underline{S}}^1 \right)^{-1} + \left(\underline{\underline{S}}^2 \right)^{-1}$. La matriz precondicionada es

$$\underline{\underline{FS}} = \left(\left(\underline{\underline{S}}^1 \right)^{-1} + \left(\underline{\underline{S}}^2 \right)^{-1} \right) \underline{\underline{S}} = \left(\left(\underline{\underline{S}}^1 \right)^{-1} + \left(\underline{\underline{S}}^2 \right)^{-1} \right) \left(\underline{\underline{S}}^1 + \underline{\underline{S}}^2 \right). \tag{325}$$

La aplicación de este algoritmo implica la solución de dos problemas con condición de frontera tipo Dirichlet y dos problemas con condición de frontera tipo Neumann con datos sobre Γ.

El Algoritmo Dirichlet-Dirichlet Consideremos aquí el dual del algoritmo Neumann-Neumann este se le conoce como Finite Element Tearing and Interconnecting (FETI) o algoritmo Dirichlet-Dirichlet. Si consideramos el dominio Ω particionado en dos subdominios Ω_1 y Ω_2, la formulación dada por la Ec. (294), puede ser expresada como un problema iterativo en el se inicia con un valor supuesto $\underline{\lambda_\Gamma}_0$ del flujo sobre Γ ver Ec. (314) - aquí $\underline{\lambda_\Gamma}$ es una aproximación a la derivada normal en la dirección $\underline{n}_i-$, el primer paso en este algoritmo consiste en resolver un problema con condiciones de Frontera Neumann en cada Ω_i con datos $\underline{\lambda_\Gamma}_0$ sobre Γ, para después resolver un problema en cada subdominio con condiciones de frontera Dirichlet sobre Γ eligiendo como la derivada normal la diferencia de la solución de los dos problemas con condiciones de frontera Neumann. Los valores sobre Γ de la solución de dichos problemas Neumann es empleado para corregir el valor inicial $\underline{\lambda_\Gamma}_0$ y encontrar el nuevo valor de $\underline{\lambda_\Gamma}_1$, y así continuar de manera iterativa para $n \geq 0$, y el algoritmo queda expresado como

$$(N_i) \left\{ \begin{array}{ll} -\Delta \underline{u}_i^{n+1/2} = f & , \text{ en } \Omega_i \\ \underline{u}_i^{n+1/2} = 0 & , \text{ en } \partial\Omega_i \backslash \Gamma \\ \frac{\partial \underline{u}_i^{n+1/2}}{\partial \underline{n}_i} = \underline{\lambda_\Gamma}_i^n & , \text{ en } \Gamma \end{array} \right\}, \text{ con } i = 1, 2 \tag{326}$$

$$(D_i) \left\{ \begin{array}{ll} -\Delta \underline{v}_i^{n+1} = 0 & , \text{ en } \Omega_i \\ \underline{v}_i^{n+1} = 0 & , \text{ en } \partial\Omega_i \backslash \Gamma \\ v^{n+1} = u_1^{n+1/2} - u_2^{n+1/2} & , \text{ en } \Gamma \end{array} \right\}, \text{ con } i = 1, 2$$

$$\underline{\lambda_\Gamma}^{n+1} = \underline{\lambda_\Gamma}^n - \theta \left(\frac{\partial \underline{v}_1^{n+1}}{\partial \underline{n}_1} + \frac{\partial \underline{v}_2^{n+1}}{\partial \underline{n}_2} \right) \qquad \text{en } \Gamma$$

para algún adecuado $\theta \in (0, \theta_{máx})$. Usando una aproximación a la derivada normal, se derivará un método iterativo para el problema discreto. Definiendo los vectores $\underline{w}_i = \underline{u}_{I_i}$ y $\underline{r}_i = \underline{v}_{I_i}$ entonces se tiene que

$$(N_i) \left\{ \left[\begin{array}{cc} \underline{\underline{A}}_{II}^i & \underline{\underline{A}}_{I\Gamma}^i \\ \underline{\underline{A}}_{\Gamma I}^i & \underline{\underline{A}}_{\Gamma\Gamma}^i \end{array} \right] \left[\begin{array}{c} \underline{w}_i^{n+1} \\ \underline{r}_i^{n+1} \end{array} \right] = \left[\begin{array}{c} \underline{b}_{I_i} \\ \underline{b}_{\Gamma_i} + \underline{\lambda_\Gamma}_i^n \end{array} \right] \right\}, \text{ con } i = 1, 2 \tag{327}$$

$$(D_i) \left\{ \underline{\underline{A}}_{II}^i \underline{r}_i^{n+1/2} + \underline{\underline{A}}_{I\Gamma}^i \underline{t_\Gamma}^n = 0, \text{ con } i = 1, 2 \right\}$$

$$\underline{\lambda_\Gamma}_i^{n+1} = \underline{\lambda_\Gamma}_i^n - \theta \left(\underline{\eta}_1^{n+1} + \underline{\eta}_2^{n+1} \right)$$

donde el residual $\underline{t_\Gamma}$ es definido como

$$\underline{t_\Gamma} = \underline{\gamma}_1^{n+1} - \underline{\gamma}_2^{n+1} \tag{328}$$

y el flujo $\underline{\eta}_i^{n+1}$ por

$$\underline{\eta}_i^{n+1} = \underline{\underline{A}}_{\Gamma I}^i \underline{r}_i^{n+1/2} + \underline{\underline{A}}_{\Gamma\Gamma}^i \underline{t_\Gamma} \tag{329}$$

conforme a la Ec (303).

Eliminando $\underline{w}_i^{n+1/2}, \underline{\gamma}_i^{n+1/2}$ y $\underline{r}_i^{n+1/2}$ de la Ec. (327), usando factorización por bloques de las matrices locales $\underline{\underline{A}}^i$, los problemas (N_i), quedan en términos de

$$\underline{t_\Gamma} = -\left(\underline{d_\Gamma} - \underline{\underline{F}}\underline{\lambda_\Gamma}^n\right), \tag{330}$$

el cual muestra que la diferencia $\underline{t_\Gamma}$ del flujo local es igual a menos el residual del sistema de la Ec. (314). Los problemas (D_i) quedan en términos de

$$\underline{\eta}_i^{n+1} = \underline{\underline{S}}^i \underline{t_\Gamma} = -\underline{\underline{S}}^i \left(\underline{d_\Gamma} - \underline{\underline{F}}\underline{\lambda_\Gamma}^n\right). \tag{331}$$

Por lo tanto, encontramos

$$\underline{\lambda_\Gamma}^{n+1} - \underline{\lambda_\Gamma}^n = \theta \left(\underline{\underline{S}}_1 + \underline{\underline{S}}_2\right) \left(\underline{d_\Gamma} - \underline{\underline{F}}\underline{\lambda_\Gamma}^n\right) \tag{332}$$

lo cual muestra que el algoritmo Dirichlet-Dirichlet es también un sistema iterativo precondicionado de Richardson para el sistema del complemento de Schur, con precondicionador $\underline{\underline{S}}_1 + \underline{\underline{S}}_2$. La matriz precondicionada es

$$\underline{\underline{SF}} = \underline{\underline{S}} \left(\left(\underline{\underline{S}}^1\right)^{-1} + \left(\underline{\underline{S}}^2\right)^{-1}\right) = \left(\underline{\underline{S}}^1 + \underline{\underline{S}}^2\right) \left(\left(\underline{\underline{S}}^1\right)^{-1} + \left(\underline{\underline{S}}^2\right)^{-1}\right) \tag{333}$$

la aplicación de este algoritmo implica la solución de dos problemas con condición de frontera tipo Neumann y dos problemas con condición de frontera tipo Dirichlet con datos sobre Γ.

5.2.2. Algoritmos en Múltiples Subdominios

Sea $\Omega \subset \mathbb{R}^n$ un dominio, y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios del dominio Ω, i.e. se asume que:

1.- Ω_α, para $\alpha = 1, ..., E$ es un subdominio de Ω,

2.- $\Omega_\alpha \bigcap \Omega_\beta = \varnothing$, siempre que $\alpha \neq \beta$.

3.- $\Omega \subset \bigcup_{\alpha=1}^{E} \overline{\Omega_\alpha}$.

La notación $\partial\Omega$ y $\partial\Omega_\alpha$, $\alpha = 1, ..., E$ es tomada de la frontera del dominio Ω y la frontera del subdominio Ω_α respectivamente, claramente

$$\partial\Omega \subset \bigcup_{\alpha=1}^{E} \partial\Omega_\alpha. \tag{334}$$

Adicionalmente definimos $\Gamma_i = \partial\Omega_i \bigcap \partial\Omega$, a la frontera interior como

$$\Gamma = \bigcup_i \Gamma_i. \tag{335}$$

Notemos que Γ y Γ_i son conjuntos abiertos. Un problema acoplado como en la Ec. (297) puede ser resuelto hallando las condiciones de transmisión impuestas

a lo largo de cada borde $\partial\Omega_i \bigcap \partial\Omega_j$.El sistema lineal dado por la Ec. (298) puede ser escrito como

$$\begin{bmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Gamma} \\ \underline{\underline{A}}_{\Gamma I} & \underline{\underline{A}}_{\Gamma\Gamma} \end{bmatrix} \begin{bmatrix} \underline{u}_I \\ \underline{u}_\Gamma \end{bmatrix} = \begin{bmatrix} \underline{b}_I \\ \underline{b}_\Gamma \end{bmatrix} \tag{336}$$

el cual ha sido particionado en los grados de libertad de los nodos interiores de los subdominios y los nodos de frontera sobre Γ. La matriz de carga y el lado derecho fueron obtenidos por subensamble de los correspondientes componentes relativos a los subdominios según la Ec. (300).

Las incógnitas en el interior de los subdominios $\underline{u}_I$ pueden ser eliminadas por eliminación Gaussiana en bloques y el sistema lineal resultante es

$$\begin{bmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Gamma} \\ 0 & \underline{\underline{S}} \end{bmatrix} \begin{bmatrix} \underline{u}_I \\ \underline{u}_\Gamma \end{bmatrix} = \begin{bmatrix} \underline{b}_I \\ \underline{g}_\Gamma \end{bmatrix}. \tag{337}$$

Como antes, el complemento de Schur y el vector $\underline{g}_\Gamma$ puede ser encontrado por subensamble de las contribuciones locales, ver sección (10.4). Primero definimos una familia de operadores de restricción, dado un vector de la cantidad de grados de libertad que el vector $\underline{u}_\Gamma$ sobre la interfase Γ, definimos la restricción $\underline{\underline{R}}^i(\underline{u}_\Gamma)$ como el vector de grados de libertad de $\underline{u}_\Gamma$ sobre Γ_i. Aquí $\underline{\underline{R}}^i$ es una matriz rectangular de ceros y unos. Y para cada subdominio Ω_i los grados de libertad de los nodos de la frontera interior Γ_i de Ω_i como en la Ec. (300), teniendo

$$\begin{aligned} \underline{\underline{S}} &= \sum_{i=1}^{E} \left(\underline{\underline{R}}^i\right)^T \underline{\underline{S}}^i \underline{\underline{R}}^i \\ \underline{g}_\Gamma &= \sum_{i=1}^{E} \left(\underline{\underline{R}}^i\right)^T \left(\underline{b}_{\Gamma_i} - \underline{\underline{A}}^i_{\Gamma I} \left(\underline{\underline{A}}^i_{II}\right)^{-1} \underline{b}_{I_i}\right) \end{aligned} \tag{338}$$

donde el complemento de Schur local es definido como en la Ec. (308) y $\left(\underline{\underline{R}}^i\right)^T$ es el transpuesto de $\underline{\underline{R}}^i$.

El Algoritmo Neumann-Neumann Examinando el método Neumann-Neumann para dos subdominios, en particular la Ec. (325), entonces la generalización al caso de múltiples subdominios queda dada en términos de

$$\underline{\underline{S}}_{NN} = \sum_{i=1}^{E} \left(\underline{\underline{R}}^i\right)^T \left(\underline{\underline{S}}^i\right)^{-1} \underline{\underline{R}}^i \underline{\underline{S}}. \tag{339}$$

Notemos que la aplicación de este operador a un vector implica la solución en cada subdominio Ω_i, de un problema con condiciones de frontera Dirichlet y un problema con condiciones de frontera Neumann sobre $\partial\Omega_i \cap \Gamma$. Así también, todos los subdominios que no tocan $\partial\Omega$, $\underline{\underline{S}}^i$ es no singular y $\left(\underline{\underline{S}}^i\right)^{-1}$ puede entenderse como una seudo inversa o una inversa de un problema regularizado.

Adicionalmente, el algoritmo Neumann-Neumann también puede ser definido a nivel continuo usando la Ec. (319), con $i = 1, ..., E$ con condiciones de frontera

Neumann para el problema N_i sobre cada cara $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$. La nueva iterada $\underline{u_\Gamma}^{n+1}$ en los nodos de la interfase, se construye por la corrección en todos los subdominios que tienen nodos sobre esas fronteras.

El Algoritmo Dirichlet-Dirichlet Aquí trataremos la versión más general del algoritmo Dirichlet-Dirichlet mejor conocido cómo método One-Level FETI.

Primeramente consideremos a $\Omega \subset \mathbb{R}^n$ un dominio, y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios del dominio Ω, además sea Γ_i la frontera de interior del subdominio Ω_i y Γ la frontera interior del dominio Ω_i.

Asumiremos que las discontinuidades en la ecuación diferencial parcial -si existen- estarán alineadas con las fronteras de los subdominios, tal que en cada subdominio Ω_i, el coeficiente $\rho(x)$ de la ecuación tenga un valor constante, sin perdida de generalidad se asumirá $\rho_i > 0$. Además, denotaremos a W_i como el espacio de trazas de Ω_i, es decir

$$W_i = W^h\left(\partial\Omega_i \cap \Gamma\right), \text{ con } i = 1, ..., E \tag{340}$$

también denotaremos W como el espacio producto del espacio de las trazas, es decir

$$W = \prod_{i=1}^{E} W_i \tag{341}$$

y la extensión armónica discreta Sec(5.1.2) por pedazos de $\underline{u_\Gamma}$ por $\mathcal{H}\left(\underline{u_\Gamma}\right)$.

Así, en lo que resta de esta sección, se trabajara casi exclusivamente con funciones en el espacio de trazas W_i y cuando sea conveniente, se considerarán como un elemento representante de las funciones armónicas discretas en Ω_i, de tal forma que $\underline{w} \in W$, $\mathcal{H}\left(\underline{w}\right)$ denotará la extensión por pedazos de la armónica discreta sobre todo el subdominio Ω_i, entenderemos $\mathcal{H}\left(\underline{w}\right)$ como un elemento en el espacio producto W con componentes $\mathcal{H}_i\left(\underline{w}_i\right)$.

Reformulando el problema definido por la Ec. (294) a uno reducido a la interfase Γ por medio de elementos finitos, como un problema de minimización con restricciones impuestas por lo requerimientos de continuidad en Γ queda como:

Encontrar $\underline{u} \in W$ tal que

$$\left.\begin{array}{c} J(\underline{u}) = \frac{1}{2}\left\langle \underline{\underline{S}}\underline{u}, \underline{u}\right\rangle - \left\langle \underline{f}, \underline{u}\right\rangle \to \text{mín} \\ \underline{\underline{B}}\underline{u} = 0 \end{array}\right\} \tag{342}$$

donde la matriz por bloques $\underline{\underline{S}}$ es formada por las matrices $\underline{\underline{S}}_i$ Ec. (308) del complemento de Schur en el $i-$ ésimo subdominio, el vector por bloques $\underline{u}$ es formado por los vectores $\underline{u}_i$ solución de la frontera interior en cada $i-$ ésimo subdominio y el vector $\underline{f}$ es formado por los vectores $\underline{f}_i$ de la frontera interior

en cada $i-$ ésimo subdominio, i.e.

$$\underline{\underline{S}} = \begin{bmatrix} \underline{\underline{S}}^1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \underline{\underline{S}}^E \end{bmatrix}, \underline{u} = \begin{bmatrix} \underline{u}_1 \\ \vdots \\ \vdots \\ \underline{u}_E \end{bmatrix}, \underline{f} = \begin{bmatrix} \underline{f}_1 \\ \vdots \\ \vdots \\ \underline{f}_E \end{bmatrix} \tag{343}$$

y la matriz $\underline{\underline{B}}$ es formada por las matrices $\underline{\underline{B}}_i$ en cada $i-$ ésimo subdominio tal que la solución asociada a más de un subdominio coincida, i.e.

$$\underline{\underline{B}} = \begin{bmatrix} \underline{\underline{B}}^1, & \cdots & \cdots, & \underline{\underline{B}}^E \end{bmatrix} \tag{344}$$

es construida de $\{0, 1, -1\}$ tal que los valores de la solución $, \underline{u}$ asociada a más de un subdominio coincida cuando $\underline{\underline{B}}\underline{u} = 0$, donde la elección de $\underline{\underline{B}}$ no es única.

Nótese que un mismo nodo en la frontera interior pertenece a dos o más subdominios, por ello es necesario algún mecanismo para asegurar que la solución asociada a más de un subdominio coincida.

El problema (342) es soluble de manera única ya que el

$$Kernel\left(\underline{\underline{S}}\right) \cap Kernel\left(\underline{\underline{B}}\right) = \{0\} \tag{345}$$

lo cual indica que $\underline{\underline{S}}$ es invertible sobre el $Kernel\left(\underline{\underline{B}}\right)$.

Pero introduciendo un vector de multiplicadores de Lagrange $\underline{\lambda}$ para imponer las restricciones $\underline{\underline{B}}\underline{u} = 0$, obtenemos una formulación silla de la Ec. (342):

Encontrar $(\underline{u}, \underline{\lambda}) \in W \times U$ tal que

$$\begin{cases} \underline{\underline{S}}\underline{u} + \underline{\underline{B}}^T\underline{\lambda} = \underline{f} \\ \underline{\underline{B}}\underline{u} = \underline{0} \end{cases} \tag{346}$$

la solución $\underline{\lambda}$ de la Ec. (346) es única salvo la adición de un elemento del $Kernel\left(\underline{\underline{B}}^T\right)$. El espacio de multiplicadores de Lagrange U, es por lo tanto elegido como el $Rango\left(\underline{\underline{B}}\right)$. Este espacio puede ser entendido como el espacio de las funciones de salto en W.

Definición 45 *Decimos que un subdominio Ω_i es un subdominio flotante si la intersección de este con la frontera del dominio $\partial\Omega$ es vacía.*

También usamos a la matriz $\underline{\underline{R}}$ construida de todos los espacios nulos de los elementos de $\underline{\underline{S}}$, cuyos elementos están asociados cada subdominio de manera individual

$$\underline{\underline{R}} = \begin{bmatrix} \underline{\underline{R}}^1 & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \underline{\underline{R}}^E \end{bmatrix} \tag{347}$$

tal que el $Rango\left(\underline{\underline{R}}\right) = Kernel\left(\underline{\underline{S}}\right)$. De hecho, sólo los subdominio flotantes contribuyen, i.e. el subdominio que intersecta a $\partial\Omega$ no contribuye al $Kernel\left(\underline{\underline{S}}\right)$, y por tanto esas columnas de $\underline{\underline{R}}$ son nulas.

Una solución $\underline{u}$ a la primera ecuación de la Ec. (346) existe si y sólo si $\underline{f} - \underline{\underline{B}}^T\underline{\lambda} \in Rango\left(\underline{\underline{S}}\right)$, esta restricción permite la introducción de un operador proyección P. Obteniendo

$$\underline{u} = \underline{\underline{S}}^{\dagger}\left(\underline{f} - \underline{\underline{B}}^T\underline{\lambda}\right) - \underline{\underline{R}}\underline{\alpha} \text{ si } \left(\underline{f} - \underline{\underline{B}}^T\underline{\lambda}\right) \perp Kernel\left(\underline{\underline{S}}\right) \tag{348}$$

donde $\underline{\underline{S}}^{\dagger}$ es una pseudo-inversa de $\underline{\underline{S}}$, notemos que $\underline{\alpha}$ puede ser fácilmente determinada una vez encontrada $\underline{\lambda}$.

Sustituyendo la expresión para $\underline{u}$ dentro de la segunda ecuación de la Ec. (346) obtenemos

$$\underline{\underline{BS}}^{\dagger}\underline{\underline{B}}^T\underline{\lambda} = \underline{\underline{BS}}^{\dagger}\underline{f} - \underline{\underline{BR}}\underline{\alpha} \tag{349}$$

así, se obtiene el sistema

$$\left\{ \begin{array}{c} \underline{\underline{BS}}^{\dagger}\underline{\underline{B}}^T\underline{\lambda} + \underline{\underline{BR}}\underline{\alpha} = \underline{\underline{BS}}^{\dagger}\underline{f} \\ \underline{\underline{BR}}^T\underline{\lambda} = \underline{\underline{R}}^T\underline{f} \end{array} \right. \tag{350}$$

donde la primera ecuación se obtiene de la segunda ecuación de Ec. (346) al sustituir $\underline{u}$ de la Ec. (348) y la segunda se obtiene de la primera Ec. (346) al sustituir también $\underline{u}$ de la Ec. (348) y usando el hecho que $\left(\underline{f} - \underline{\underline{B}}^T\underline{\lambda}\right) \perp Kernel\left(\underline{\underline{S}}\right)$. Esta última ecuación puede escribirse más compactamente como

$$\left\{ \begin{array}{c} \underline{\underline{F}}\underline{\lambda} + \underline{\underline{G}}\underline{\alpha} = \underline{d} \\ \underline{\underline{G}}^T\underline{\lambda} = \underline{e} \end{array} \right. \tag{351}$$

donde $\underline{\underline{F}} = \underline{\underline{BS}}^{\dagger}\underline{\underline{B}}^T, \underline{\underline{G}} = \underline{\underline{BR}}, \underline{d} = \underline{\underline{BS}}^{\dagger}\underline{f}$ y $\underline{e} = \underline{\underline{R}}^T\underline{f}$.

Introduciendo una matriz simétrica y positiva definida $\underline{\underline{Q}}$ y un producto interior $\left\langle \underline{\lambda}, \underline{\mu} \right\rangle = \left\langle \underline{\lambda}, \underline{\underline{Q}}\underline{\mu} \right\rangle$ sobre $U = Rango(\underline{\underline{B}})$, como antes $\langle \cdot, \cdot \rangle$ es el producto estándar del espacio L^2. Sea

$$\underline{\underline{P}}^T = \underline{\underline{I}} - \underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{Q}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\underline{\underline{Q}} \tag{352}$$

la proyección de U sobre el subespacio de multiplicadores de Lagrange que son Q-ortogonales al $Rango\left(\underline{\underline{G}}\right)$; también definimos

$$\underline{\underline{P}} = \underline{\underline{I}} - \underline{\underline{Q}}\underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{Q}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T \tag{353}$$

como la proyección de U sobre el $Kernel\left(G^T\right)$, esta proyección es ortogonal en el Q^{-1} producto interior, es decir, el producto interior definido por $\left\langle \underline{\lambda}, \underline{\underline{Q}}^{-1}\underline{\mu} \right\rangle$.

Notemos que si $\underline{\underline{Q}} = \underline{\underline{I}}$, entonces $\underline{\underline{P}}^T = \underline{\underline{I}} - \underline{\underline{BR}} \left(\left(\underline{\underline{BR}} \right)^T \underline{\underline{I}} \left(\underline{\underline{BR}} \right) \right)^{-1} \left(\underline{\underline{BR}} \right)^T \underline{\underline{I}}$, desarrollando

$$\begin{aligned} \underline{\underline{P}}^T &= \underline{\underline{I}} - \underline{\underline{BR}} \left(\underline{\underline{BR}} \right)^{-T} \left(\underline{\underline{BR}} \right)^{-1} \left(\underline{\underline{BR}} \right)^T \\ &= \underline{\underline{I}} - \underline{\underline{BRR}}^{-1} \underline{\underline{B}}^{-1} \underline{\underline{B}}^{-1} \underline{\underline{R}}^{-1} \underline{\underline{RB}} \\ &= \underline{\underline{I}} - \underline{\underline{I}}_{Kernel(S)} \end{aligned} \tag{354}$$

así, $\underline{\underline{P}}^T \underline{v} = \underline{v} - Proy_{Kernel(S)}(\underline{v})$.

Aplicando P^T al sistema dado por la Ec. (351) obtenemos

$$\left\{ \begin{array}{c} P^T \underline{\underline{F}} \underline{\lambda} = P^T \underline{d} \\ P^T \underline{\underline{G}}^T \underline{\lambda} = \underline{e} \end{array} \right. \tag{355}$$

ya que $P^T \underline{\underline{G}} \underline{\alpha} = \underline{\underline{BR}} \underline{\alpha} - Proy_{Kernel(S)}(\underline{\underline{BR}} \underline{\alpha}) = 0$.

Multiplicando la Ec. (349) por $\left(\underline{\underline{G}}^T \underline{\underline{Q}} \underline{\underline{G}} \right)^{-1} \underline{\underline{G}}^T \underline{\underline{Q}}$ encontramos que

$$\alpha = \left(\underline{\underline{G}}^T \underline{\underline{Q}} \underline{\underline{G}} \right)^{-1} \underline{\underline{G}}^T \underline{\underline{Q}} \left(\underline{d} - \underline{\underline{F}} \underline{\lambda} \right) \tag{356}$$

el cual queda totalmente determinado por los valores de λ. Notemos que los operadores P y P^T representa solamente la parte global del precondicionador.

Introduciendo ahora los subespacios

$$\begin{aligned} V &= \left\{ \underline{\lambda} \in U \mid \left\langle \underline{\lambda}, \underline{\underline{B}} \underline{z} \right\rangle = 0, \text{ con } \underline{z} \in Kernel \left(\underline{\underline{S}} \right) \right\} \\ &= Kernel \left(\underline{\underline{G}}^T \right) = Rango \left(\underline{\underline{P}} \right) \end{aligned} \tag{357}$$

y

$$\begin{aligned} V' &= \left\{ \underline{\mu} \in U \mid \left\langle \underline{\mu}, \underline{\underline{B}} \underline{z} \right\rangle_Q = 0, \text{ con } \underline{z} \in Kernel \left(\underline{\underline{S}} \right) \right\} \\ &= Rango \left(\underline{\underline{P}}^T \right) \end{aligned} \tag{358}$$

donde el espacio V' es isomorfo al dual del espacio V.

El método One-Level FETI es un método de Gradiente Conjugado precondicionado -ver sección (9.2.1)- en el espacio V, aplicado a

$$\underline{\underline{P}}^T \underline{\underline{F}} \underline{\lambda} = \underline{\underline{P}}^T \underline{d}, \underline{\lambda} \in \underline{\lambda_0} + V \tag{359}$$

con condición inicial aproximada $\underline{\lambda_0}$ escogido tal que $\underline{\underline{G}} \underline{\lambda_0} = \underline{e}$.

El precondicionador más básico para FETI al tomar $\underline{\underline{Q}} = \underline{\underline{I}}$, es de la forma

$$\underline{\underline{M}}^{-1} = \underline{\underline{BSB}}^T = \sum_{i=1}^{E} \underline{\underline{B}}^i \underline{\underline{S}}^i \left(\underline{\underline{B}}^i \right)^T \tag{360}$$

otra variante es

$$\underline{\underline{PM}}^{-1}\underline{\underline{P}}^{T}\underline{\underline{F}}\underline{\lambda} = \underline{\underline{PM}}^{-1}\underline{\underline{P}}^{T}\underline{d}, \underline{\lambda} \in \underline{\lambda_0} + V \tag{361}$$

notemos que en esta variante, para $\underline{\lambda} \in V$, $\underline{\underline{PM}}^{-1}\underline{\underline{P}}^{T}\underline{\underline{F}}\underline{\lambda} = \underline{\underline{PM}}^{-1}\underline{\underline{P}}^{T}\underline{\underline{P}}^{T}\underline{\underline{FP}}\underline{\lambda}$, y esto puede ser visto como el producto de dos matrices simétricas, nótese que estrictamente $\underline{\underline{M}}^{-1}$ no tiene una inversa por eso usaremos $\hat{\underline{\underline{M}}}^{-1}$ que si la tiene y es definido como

$$\begin{aligned} \hat{\underline{\underline{M}}}^{-1} &= \left(\underline{\underline{BD}}^{-1}\underline{\underline{B}}^{T}\right)^{-1}\underline{\underline{BD}}^{-1}\underline{\underline{SD}}^{-1}\underline{\underline{B}}^{T}\left(\underline{\underline{BD}}^{-1}\underline{\underline{B}}^{T}\right)^{-1} \\ &= \left(\underline{\underline{BD}}^{-1}\underline{\underline{B}}^{T}\right)^{-1}\sum_{i}^{E}\underline{\underline{B}}^{i}\left(\underline{\underline{D}}^{i}\right)^{-1}\underline{\underline{S}}^{i}\left(\underline{\underline{D}}^{i}\right)^{-1}\left(\underline{\underline{B}}^{i}\right)^{T}\left(\underline{\underline{BD}}^{-1}\underline{\underline{B}}^{T}\right)^{-1} \end{aligned} \tag{362}$$

donde

$$\underline{\underline{D}} = \begin{bmatrix} \underline{\underline{D}}^{1} & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \underline{\underline{D}}^{E} \end{bmatrix} \tag{363}$$

en la cual cada $\underline{\underline{D}}^{i}$ es la matriz diagonal con los elementos $\delta_i^{\dagger}(x)$ correspondientes a los puntos $x \in \partial\Omega_{i,h} \cap \Gamma_h$.

Entonces el método One-Level FETI queda en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$\underline{\lambda_0} = \underline{\underline{QG}}\left(\underline{\underline{G}}^{T}\underline{\underline{QG}}\right)^{-1}\underline{e} + \underline{\mu}, \underline{\mu} \in Rango\left(\underline{\underline{P}}\right)$
$\underline{r_0} = \underline{d} - \underline{\underline{F}}\underline{\lambda}^{0}$
$\beta^{0} = 0$
$\underline{p}^{1} = 0$

2.- Itera $k = 1, 2...$ hasta converger

$\underline{q}^{k-1} = \underline{\underline{P}}^{T}\underline{r}^{k-1}$
$\underline{z}^{k-1} = \hat{\underline{\underline{M}}}^{-1}\underline{q}^{k-1}$
$\underline{y}^{k-1} = \underline{\underline{P}}\underline{z}^{k-1}$
$\beta^{k} = \frac{\left\langle \underline{y}^{k-1}, q^{k-1}\right\rangle}{\left\langle \underline{y}^{k-2}, \underline{q}^{k-2}\right\rangle}$
$\underline{p}^{k} = \underline{y}^{k-1} + \beta^{k}\underline{p}^{k-1}$
$\alpha^{k} = \frac{\left\langle \underline{y}^{k-1}, q^{k-1}\right\rangle}{\left\langle \underline{p}^{k}, \underline{\underline{F}}\underline{q}^{k}\right\rangle}$
$\underline{\lambda}^{k} = \underline{\lambda}^{k-1} + \alpha^{k}\underline{p}^{k}$
$\underline{r}^{k} = \underline{r}^{k-1} - \alpha^{k}\underline{\underline{F}}\underline{p}^{k}$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u}$ mediante

$$\underline{u} = \underline{\underline{S}}^{\dagger} \left(\underline{f} - \underline{\underline{B}}^{T} \underline{\lambda} \right) \tag{364}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

Como se ha mencionado a lo largo de esta sección, el algoritmo One-Level FETI es determinado por la elección de $\underline{\underline{Q}}$ y $\underline{\underline{\hat{M}}}^{-1}$. Para la elección $\underline{\underline{Q}} = \underline{\underline{\hat{M}}}^{-1}$, cada paso correspondiente al método de Gradiente Conjugado supone una aplicación de $\underline{\underline{P}}^{T}$ y uno de $\underline{\underline{P}}$, la solución de un problema con condiciones de frontera Dirichlet sobre los subdominios es necesario para la aplicación de $\underline{\underline{\hat{M}}}^{-1}$ y la solución de un problema con condiciones de frontera Neumann sobre los subdominios es necesario para la aplicación de $\underline{\underline{F}}$ en el cálculo del nuevo residual. Entonces la aplicación de $\underline{\underline{P}}^{T}$ y $\underline{\underline{P}}$ implica dos adicionales aplicaciones de $\underline{\underline{Q}} = \underline{\underline{\hat{M}}}^{-1}$ y la solución de dos problemas sobre la partición gruesa, esto es un total de un problema con condiciones de frontera tipo Neumann y tres problemas con condiciones de frontera tipo Dirichlet sobre cada subdominio y dos problemas sobre la partición gruesa en cada paso de la iteración.

5.2.3. El Algoritmo One-Level FETI Simplificado

Aquí trataremos la versión particular del algoritmo Dirichlet-Dirichlet en el cual $\underline{\underline{R}} = \underline{0}$, entonces reformulando el problema definido por la Ec. (294) a uno reducido a la interfase Γ por medio de elementos finitos, como un problema de minimización con restricciones impuestas por lo requerimientos de continuidad en Γ queda como:

Encontrar $\underline{u} \in W$ tal que

$$\left.\begin{array}{c} J(\underline{u}) = \frac{1}{2} \left\langle \underline{\underline{S}}\underline{u}, \underline{u} \right\rangle - \left\langle \underline{f}, \underline{u} \right\rangle \rightarrow \text{mín} \\ \underline{\underline{B}}\underline{u} = 0 \end{array}\right\} \tag{365}$$

donde la matriz por bloques $\underline{\underline{S}}$ es formada por las matrices $\underline{\underline{S}}^{i}$ Ec. (308) del complemento de Schur en el $i-$ ésimo subdominio, el vector por bloques $\underline{u}$ es formado por los vectores $\underline{u}_i$ solución de la frontera interior en cada $i-$ ésimo subdominio y el vector $\underline{f}$ es formado por los vectores $\underline{f}_i$ de la frontera interior en cada $i-$ ésimo subdominio, i.e.

$$\underline{\underline{S}} = \begin{bmatrix} \underline{\underline{S}}^{1} & 0 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & \cdots & \underline{\underline{S}}^{E} \end{bmatrix}, \underline{u} = \begin{bmatrix} \underline{u}_1 \\ \vdots \\ \vdots \\ \underline{u}_E \end{bmatrix}, \underline{f} = \begin{bmatrix} \underline{f}_1 \\ \vdots \\ \vdots \\ \underline{f}_E \end{bmatrix} \tag{366}$$

y la matriz $\underline{\underline{B}}$ es formada por las matrices $\underline{\underline{B}}^i$ en cada $i-$ ésimo subdominio tal que la solución asociada a más de un subdominio coincida, i.e.

$$\underline{\underline{B}} = \left[\begin{array}{ccc} \underline{\underline{B}}^1, & \cdots & \cdots, \quad \underline{\underline{B}}^E \end{array} \right] \tag{367}$$

es construida de $\{0, 1, -1\}$ tal que los valores de la solución $, \underline{u}$ asociada a más de un subdominio coincida cuando $\underline{\underline{B}}\underline{u} = 0$, donde la elección de $\underline{\underline{B}}$ no es única.

Nótese que un mismo nodo en la frontera interior pertenece a dos o más subdominios, por ello es necesario algún mecanismo para asegurar que la solución asociada a más de un subdominio coincida.

El problema (365) es soluble de manera única ya que el

$$Kernel\left(\underline{\underline{S}}\right) \cap Kernel\left(\underline{\underline{B}}\right) = \{0\} \tag{368}$$

lo cual indica que $\underline{\underline{S}}$ es invertible sobre el $Kernel\left(\underline{\underline{B}}\right)$.

Pero introduciendo un vector de multiplicadores de Lagrange $\underline{\lambda}$ para imponer las restricciones $\underline{\underline{B}}\underline{u} = 0$, obtenemos una formulación silla de la Ec. (365):

Encontrar $(\underline{u}, \underline{\lambda}) \in W \times U$ tal que

$$\left\{ \begin{array}{c} \underline{\underline{S}}\underline{u} + \underline{\underline{B}}^T\underline{\lambda} = \underline{f} \\ \underline{\underline{B}}\underline{u} = \underline{0} \end{array} \right. \tag{369}$$

la solución $\underline{\lambda}$ de la Ec. (369) es única salvo la adición de un elemento del $Kernel\left(\underline{\underline{B}}^T\right)$. El espacio de multiplicadores de Lagrange U, es por lo tanto elegido como el $Rango\left(\underline{\underline{B}}\right)$. Este espacio puede ser entendido como el espacio de las funciones de salto en W.

Una solución $\underline{u}$ a la primera ecuación de la Ec. (369) existe si y sólo si $\underline{f} - \underline{\underline{B}}^T\underline{\lambda} \in Rango\left(\underline{\underline{S}}\right)$, sustituyendo la expresión para $\underline{u}$ dentro de la segunda ecuación de la Ec. (369) obtenemos

$$\underline{\underline{B}}\underline{\underline{S}}^\dagger\underline{\underline{B}}^T\underline{\lambda} = \underline{\underline{B}}\underline{\underline{S}}^\dagger\underline{f} \tag{370}$$

donde $\underline{\underline{S}}^\dagger$ es la inversa de $\underline{\underline{S}}$ y tomado como precondicionador -el más básico- para FETI

$$\underline{\underline{M}}^{-1} = \underline{\underline{BSB}}^T = \sum_{i=1}^{E} \underline{\underline{B}}^i\underline{\underline{S}}^i\left(\underline{\underline{B}}^i\right)^T. \tag{371}$$

Entonces el método one-level FETI simplificado queda en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$\underline{\lambda_0} = \underline{0}$
$\underline{r_0} = \left(\underline{\underline{B}}\underline{\underline{S}}^\dagger\underline{f}\right) - \left(\underline{\underline{B}}\underline{\underline{S}}^\dagger\underline{\underline{B}}^T\right)\underline{\lambda}^0$
$\beta^0 = 0$
$\underline{p}^1 = 0$

2.- Itera $k = 1, 2...$ hasta converger

$$\begin{aligned}
&\underline{q}^{k-1} = \underline{r}^{k-1} \\
&\underline{z}^{k-1} = \left(\underline{\underline{B}}\underline{\underline{S}}\underline{\underline{B}}^T\right)^{-1} \underline{q}^{k-1} \\
&\underline{y}^{k-1} = \underline{z}^{k-1} \\
&\beta^k = \frac{\left\langle \underline{y}^{k-1}, q^{k-1} \right\rangle}{\left\langle \underline{y}^{k-2}, \underline{q}^{k-2} \right\rangle} \\
&\underline{p}^k = \underline{y}^{k-1} + \beta^k \underline{p}^{k-1} \\
&\alpha^k = \frac{\left\langle \underline{y}^{k-1}, q^{k-1} \right\rangle}{\left\langle \underline{p}^k, (\underline{\underline{B}}\underline{\underline{S}}^\dagger \underline{\underline{B}}^T) \underline{q}^k \right\rangle} \\
&\underline{\lambda}^k = \underline{\lambda}^{k-1} + \alpha^k \underline{p}^k \\
&\underline{r}^k = \underline{r}^{k-1} - \alpha^k \left(\underline{\underline{B}}\underline{\underline{S}}^\dagger \underline{\underline{B}}^T\right) \underline{p}^k
\end{aligned}$$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u}$ mediante

$$\underline{u} = \underline{\underline{S}}^\dagger \left(\underline{f} - \underline{\underline{B}}^T \underline{\lambda}\right) \tag{372}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

5.3. Dual-Primal FETI

El método dual-primal FETI (FETI-DP) fue introducido posteriormente que el método One-Level FETI, siendo esto una gran contribución a la teoría para la resolución de problemas elípticos de segundo y cuarto orden. Este método se basa en hacer cumplir un número relativamente pequeño de restricciones de continuidad a través de la frontera interior en cada paso de las iteraciones en comparación con el método de One-Level FETI.

Primeramente consideremos a $\Omega \subset \mathbb{R}^n$ un dominio, y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios del dominio Ω, además sea Γ_i la frontera de interior del subdominio Ω_i y Γ la frontera interior del dominio Ω_i.

Asumiremos que las discontinuidades en la ecuación diferencial parcial -si existen- estarán alineadas con las fronteras de los subdominios, tal que en cada subdominio Ω_i, el coeficiente $\rho(x)$ de la ecuación tenga un valor constante, sin perdida de generalidad se asumirá $\rho_i > 0$. Además, denotaremos a W_i como el espacio de trazas de Ω_i, es decir

$$W_i = W^h \left(\partial\Omega_i \cap \Gamma\right), \text{ con } i = 1, ..., E \tag{373}$$

también denotaremos W como el espacio producto del espacio de las trazas, es decir

$$W = \prod_{i=1}^{E} W_i \tag{374}$$

y la extensión armónica discreta (5.1.2) por pedazos de $\underline{u_\Gamma}$ por $\mathcal{H}\left(\underline{u_\Gamma}\right)$.

Así, en lo que resta de esta sección, se trabajara casi exclusivamente con funciones en el espacio de trazas W_i y cuando sea conveniente, se considerarán como un elemento representante de las funciones armónicas discretas en Ω_i, de tal forma que $\underline{w} \in W$, $\mathcal{H}_i\left(\underline{w}\right)$ denotará la extensión por pedazos de la armónica discreta sobre todo el subdominio Ω_i, entenderemos $\mathcal{H}\left(\underline{w}\right)$ como un elemento en el espacio producto W con componentes $\mathcal{H}_i\left(\underline{w}_i\right)$.

La aproximación por medio de elementos finitos estándar al problema elíptico es continua a través de la frontera interior Γ y denotamos al correspondiente subespacio de W por $\hat{W}$. En este método se usan subespacios intermedios $\tilde{W}$ de W, de tal manera que el complemento de Schur usado en los cálculos será estrictamente positivo definido.

Denotamos $\tilde{W}^h(\Omega)$ como un subespacio de $\prod_{i=1}^{E} W_i^h(\Omega_i)$ el cual es igual a $\tilde{W}$ cuando es restringido a la frontera interior Γ. Adicionalmente se introducen dos subespacios $\hat{W}_\Pi, \tilde{W}_\Delta \subset \hat{W}$ correspondiendo a la parte primal y dual del espacio $\tilde{W}$, además $\hat{W} = \hat{W}_\Pi \oplus \tilde{W}_\Delta$. Relacionamos al espacio dual $\tilde{W}_\Delta$ con los saltos en la frontera interior Γ y los multiplicadores de Lagrange son introducidos para eliminar tales saltos.

En el método FETI-DP se expresará al complemento de Schur $\underline{\underline{\tilde{S}}}$ relacionado con el espacio dual $\tilde{W}_\Delta$, así, en esta sección, $\tilde{W}$ consiste de funciones en W que toman el mismo valor en los vértices del subdominio y puede escribirse como

$$\tilde{W} = \hat{W}_\Pi \oplus \tilde{W}_\Delta. \tag{375}$$

Aquí $\hat{W}_\Pi \subset \hat{W}$ es el espacio de funciones con la frontera interior Γ_h continua que se nulifican en todos los nodos sobre Γ excepto en los vértices del subdominio Ω_i, con $i = 1, ..., E$, y $\tilde{W}_\Delta$ es la suma directa de los subespacios locales $\tilde{W}_{\Delta,i}$. i.e. $\tilde{W}_\Delta = \oplus \tilde{W}_{\Delta,i}$ donde $\tilde{W}_{\Delta,i} \subset W_i$ y consiste de las funciones locales sobre $\partial\Omega_i$ que se nulifican en los vértices de Ω_i.

El continuo de los grados de libertad asociados con los vértices de cada subdominio y con el subespacio $\hat{W}_\Pi$ es llamado primal (Π), mientras aquellos -potenciales discontinuidades a través de Γ- asociados con los subespacios $\tilde{W}_{\Delta,i}$ y con el interior de la frontera de cada subdominio Ω_i es llamado dual (Δ).

Además consideramos la familia de funciones de peso $\delta_i \in W_i$, las cuales están asociadas con cada $\partial\Omega_i$ y definidas para $\gamma \in [\frac{1}{2}, \infty)$ por la suma de contribuciones de Ω_i con los vecinos pertinentes, así definimos

$$\delta_i(x) = \frac{\sum_{j \in N_x} \rho_j^\gamma}{\rho_i^\gamma}, \qquad x \in \partial\Omega_{i,h} \cap \Gamma_h \tag{376}$$

donde N_x es el conjunto de índices j de las subregiones tal que $x \in \partial\Omega_{i,h}$, la pseudo-inversa $\delta_i^\dagger$ es definida por

$$\delta_i^\dagger(x) = \left(\delta_i(x)\right)^{-1}, \qquad x \in \partial\Omega_{i,h} \cap \Gamma_h. \tag{377}$$

Sea $\underline{\underline{\tilde{A}}}$ la matriz de carga, la cual es obtenida por la restricción

$$\underline{\underline{\tilde{A}}} = diag\left\{\underline{\underline{A}}^1, ..., \underline{\underline{A}}^E\right\} \tag{378}$$

de $\prod_i^E W^h(\Omega_i)$ a $\tilde{W}^h(\Omega)$, donde $\underline{\underline{A}}^i$ es la matriz de carga generada por el método de subestructuración en el subdominio i, notemos que $\underline{\underline{\tilde{A}}}$ no es una matriz diagonal en bloques ya que ahora están acoplados los distintos subdominios que tienen un vértice en común. Particionando $\underline{\underline{\tilde{A}}}$ como

$$\underline{\underline{\tilde{A}}} = \begin{bmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Pi} & \underline{\underline{A}}_{I\Delta} \\ \left(\underline{\underline{A}}_{I\Pi}\right)^T & \underline{\underline{A}}_{\Pi\Pi} & \underline{\underline{A}}_{\Pi\Delta} \\ \left(\underline{\underline{A}}_{I\Delta}\right)^T & \left(\underline{\underline{A}}_{\Pi\Delta}\right)^T & \underline{\underline{A}}_{\Delta\Delta} \end{bmatrix} \tag{379}$$

$$\underline{\tilde{f}} = \begin{bmatrix} \underline{f_I} \\ \underline{f_\Pi} \\ \underline{f_\Delta} \end{bmatrix} \tag{380}$$

donde el superíndice I se refiere a los grados de libertad asociados a los nodos internos de los subdominios Ω_i, Π se refiere a los asociados con los vértices de los subdominios Ω_i y Δ a los asociados al interior de las caras de la frontera de los subdominios Ω_i.

Notemos que $\underline{\underline{A}}_{II}$ y $\underline{\underline{A}}_{\Delta\Delta}$ son matrices diagonales por bloques y cada bloque corresponde a un dominio individual Ω_i y cada no-cero de $\underline{\underline{A}}_{I\Delta}$ representa un acoplamiento entre los grados de libertad asociados a un subdominio. $\underline{\underline{\tilde{A}}}$ es obtenida por el ensamble parcial de las contribuciones locales asociadas con cada subdominio Ω_i.

Eliminado las variables I y Π, entonces el complemento de Schur asociado a los grados de libertad del conjunto Δ, del interior de las caras de las fronteras $\partial\Omega_i$, queda como

$$\underline{\underline{\tilde{S}}} = \underline{\underline{A}}_{\Delta\Delta} - \begin{bmatrix} \left(\underline{\underline{A}}_{I\Delta}\right)^T & \left(\underline{\underline{A}}_{\Pi\Delta}\right)^T \end{bmatrix} \begin{bmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Pi} \\ \left(\underline{\underline{A}}_{I\Pi}\right)^T & \underline{\underline{A}}_{\Pi\Pi} \end{bmatrix}^{-1} \begin{bmatrix} \underline{\underline{A}}_{I\Delta} \\ \underline{\underline{A}}_{\Pi\Delta} \end{bmatrix} \tag{381}$$

$$\underline{\tilde{f}_\Delta} = \underline{f_\Delta} - \begin{bmatrix} \left(\underline{\underline{A}}_{I\Delta}\right)^T & \left(\underline{\underline{A}}_{\Pi\Delta}\right)^T \end{bmatrix} \begin{bmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Pi} \\ \left(\underline{\underline{A}}_{I\Pi}\right)^T & \underline{\underline{A}}_{\Pi\Pi} \end{bmatrix}^{-1} \begin{bmatrix} \underline{f_I} \\ \underline{f_\Pi} \end{bmatrix} \tag{382}$$

también obtenemos una reducción del lado derecho $\underline{\tilde{f}}_\Delta$ del vector de carga asociado con los subdominios individuales. Denotamos por $\underline{u_\Delta} \in \tilde{W}_\Delta$ el vector de grados de libertad asociado a las caras de los subdominios.

Reformulando el problema definido por la Ec. (294) a uno reducido a un segundo subespacio $\tilde{W}_\Delta$ como un problema de minimización con restricciones impuestas por lo requerimientos de continuidad en Γ queda como:

Encontrar $\underline{u_\Delta} \in \tilde{W}$ tal que

$$\left.\begin{array}{c} J(\underline{u_\Delta}) = \frac{1}{2}\left\langle \underline{\underline{\tilde{S}}}\,\underline{u_\Delta}, \underline{u_\Delta} \right\rangle - \left\langle \underline{\tilde{f}_\Delta}, \underline{u_\Delta} \right\rangle \rightarrow \text{mín} \\ \underline{\underline{B_\Delta}}\,\underline{u_\Delta} = 0 \end{array}\right\} \tag{383}$$

la matriz $\underline{\underline{B}}_\Delta$ es construida de $\{0, 1, -1\}$ tal que los valores de la solución , $\underline{u_\Delta}$ asociada a más de un subdominio coincida cuando $\underline{\underline{B}}_\Delta \underline{u_\Delta} = 0$, donde la elección de $\underline{\underline{B}}_\Delta$ no es única.

Pero introduciendo un vector de multiplicadores de Lagrange $\underline{\lambda} \in V = Rango\left(\underline{\underline{B}}_\Delta\right)$ para imponer las restricciones $\underline{\underline{B}}\underline{u} = 0$, obtenemos una formulación silla de la Ec. (383). Eliminando el subvector $\underline{u}_\Delta$, y obteniendo el siguiente sistema de multiplicadores de Lagrange

$$\underline{\underline{F}}\underline{\lambda} = \underline{d} \tag{384}$$

con

$$\underline{\underline{F}} = \underline{\underline{B}}_\Delta \left(\tilde{\underline{\underline{S}}}\right)^{-1} \left(\underline{\underline{B}}_\Delta\right)^T, \qquad \underline{d} = \underline{\underline{B}}_\Delta \left(\tilde{\underline{\underline{S}}}\right)^{-1} \underline{\tilde{f}_\Delta}. \tag{385}$$

Una vez $\underline{\lambda}$ encontrada, podemos resolver hacia atrás y obtener

$$\underline{u_\Delta} = \left(\tilde{\underline{\underline{S}}}\right)^{-1} \left(\underline{\tilde{f}_\Delta} - \left(\underline{\underline{B}}_\Delta\right)^T \underline{\lambda}\right) \in \tilde{W}_\Delta. \tag{386}$$

Los valores de la solución en el interior de los subdominios $\underline{u_I}$ y en los vértices de los subdominios $\underline{u_\Pi}$ son obtenidos como un subproducto cuando se resuelve el sistema lineal con la matriz por bloques dada por la Ec. (381).

Introduciendo una matriz de escalamiento diagonal $\underline{\underline{D}}^i_\Delta$, donde cada uno de los elementos de la diagonal corresponde a un multiplicador de Lagrange que fuerzan la continuidad entre los valores de los nodos de algunas $\underline{u_i} \in W^i$ y $\underline{u_j} \in W^j$ en algún punto $\underline{x} \in \Gamma_h$ y esta dado por $\delta^\dagger_i(x)$. También se define un escalamiento del salto por medio del operador

$$\underline{\underline{B}}_{D,\Delta} = \left[\begin{array}{cccc} \underline{\underline{D}}^1_\Delta \underline{\underline{B}}^1_\Delta & , ..., & \underline{\underline{D}}^E_\Delta \underline{\underline{B}}^E_\Delta \end{array}\right] \tag{387}$$

aquí, como antes, el bloque $\underline{\underline{B}}^i_\Delta$ es obtenido por extracción de columnas de $\underline{\underline{B}}_\Delta$ asociadas con el espacio local W_i.

Resolviendo el sistema dual dado por la Ec. (384) usando el método de Gradiente Conjugado - ver sección (9.2.1)- con el precondicionador

$$\begin{aligned} \underline{\underline{M}}^{-1} &= \underline{\underline{B}}_{D,\Delta} \underline{\underline{S}}_\Delta \left(\underline{\underline{B}}_{D,\Delta}\right)^T \\ &= \sum_{i=1}^{E} \underline{\underline{D}}^i_\Delta \underline{\underline{B}}^i_\Delta \underline{\underline{S}}^i_\Delta \left(\underline{\underline{B}}^i_\Delta\right)^T \underline{\underline{D}}^i_\Delta \end{aligned}$$

donde $\underline{\underline{S}}^i_\Delta$ es la restricción del complemento de Schur local $\underline{\underline{S}}^i$ a $\tilde{W}_{\Delta,i} \subset W^i$.

El método FETI-DP es un método de Gradiente Conjugado precondicionado para resolver el sistema precondicionado

$$\underline{\underline{M}}^{-1}\underline{\underline{F}}\underline{\lambda} = \underline{\underline{M}}^{-1}\underline{d} \tag{388}$$

quedando en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$\underline{r}^0 = \underline{d} - \underline{\underline{F}}\underline{\lambda}^0$
$\beta^1 = 0$
$\underline{p}^1 = \underline{z}^0$

2.- Itera $k = 1, 2...$ hasta converger

$\underline{z}^{k-1} = \underline{\underline{M}}^{-1}\underline{r}^{k-1}$
$\beta^k = \frac{\langle \underline{z}^{k-1}, r^{k-1} \rangle}{\langle \underline{z}^{k-2}, \underline{r}^{k-2} \rangle}$
$\underline{p}^k = \underline{z}^{k-1} + \beta^k \underline{p}^{k-1}$
$\alpha^k = \frac{\langle \underline{z}^{k-1}, r^{k-1} \rangle}{\langle \underline{p}^k, \underline{\underline{F}}\underline{q}^k \rangle}$
$\underline{\lambda}^k = \underline{\lambda}^{k-1} + \alpha^k \underline{p}^k$
$\underline{r}^k = \underline{r}^{k-1} - \alpha^k \underline{\underline{F}}\underline{p}^k$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u_\Delta}$ mediante

$$\underline{u_\Delta} = \left(\underline{\underline{\tilde{S}}}\right)^{-1}\left(\underline{\tilde{f}_\Delta} - \left(\underline{\underline{B}}_\Delta\right)^T \underline{\lambda}\right) \tag{389}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

El método FETI-DP presenta las siguientes ventajas:

- El algoritmo no requiere de la caracterización del kernel de los problemas locales con condiciones de frontera Neumann. Adicionalmente, la imposición de adicionales restricciones en cada iteración siempre crea problemas locales no singulares y al mismo tiempo proporciona un subyacente problema grueso global.
- El algoritmo no requiere la introducción de matrices de escalabilidad Q.
- El algoritmo en el método de Gradiente Conjugado puede usar un valor inicial arbitrario $\underline{\lambda_0}$.

5.4. Variantes para la Implementación Numérica

En esta sección describiremos como hacer el cálculo en donde estén involucradas las matrices $\underline{\underline{S}}$ y $\underline{\underline{S}}^{-1}$ (ya que estas matrices son virtuales, ya que todo queda en función de las matrices locales $\underline{\underline{S}}^{\alpha}$), el cálculo de los nodos interiores y otras variantes de la implementación numérica que ofrece mejoras computacionales al modelo.

5.4.1. Cálculo de la Matriz $\underline{\underline{S}}$

La matriz $\underline{\underline{S}}$ definida por

$$\underline{\underline{S}} = \underline{\underline{A}}_{\Delta\Delta} - \underline{\underline{A}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \tag{390}$$

es formada por $\underline{\underline{S}} = \sum_{\alpha=1}^{E} \underline{\underline{S}}^{\alpha}$, donde $\underline{\underline{S}}^{\alpha}$ esta formada por el complemento de Schur local

$$\underline{\underline{S}}^{\alpha} = \underline{\underline{A}}^{\alpha}_{\Delta\Delta} - \underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta}. \tag{391}$$

Así que, las matrices locales $\underline{\underline{S}}^{\alpha}$ y $\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1}$ no se construyen, ya que estas serian matrices densas y su construcción es computacionalmente muy costosa, y como sólo nos interesa el producto $\underline{\underline{S}}\underline{y_{\Gamma}}$, o más precisamente $\left[\sum_{\alpha=1}^{E} \underline{\underline{S}}^{\alpha} \right] \underline{y_{\Gamma}}$, entonces si llamamos $\underline{y_{\Gamma}}^{\alpha}$ al vector correspondiente al subdominio α, entonces tendremos

$$\underline{\tilde{u}_{\Gamma}}^{\alpha} = \left(\underline{\underline{A}}^{\alpha}_{\Delta\Delta} - \underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \right) \underline{y_{\Gamma}}^{\alpha}. \tag{392}$$

Para evaluar eficientemente esta expresión, realizamos las siguientes operaciones equivalentes

$$\begin{aligned} \underline{x1} &= \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \underline{y_{\Gamma}}^{\alpha} \\ \underline{x2} &= \left(\underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \right) \underline{y_{\Gamma}}^{\alpha} \\ \underline{\tilde{u}_{\Gamma}}^{\alpha} &= \underline{x1} - \underline{x2} \end{aligned} \tag{393}$$

la primera y tercera expresión no tienen ningún problema en su evaluación, para la segunda expresión tendremos que hacer

$$\underline{x3} = \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \underline{y_{\Gamma}}^{\alpha} \tag{394}$$

con este resultado intermedio deberíamos calcular

$$\underline{x4} = \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{x3} \tag{395}$$

pero como no contamos con $\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}$, entonces multiplicamos la expresión por $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}$ obteniendo

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{x4} = \underline{\underline{A}}^{\alpha}_{\Pi\Pi}\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{x3} \tag{396}$$

al simplificar, tenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{x4} = \underline{x3}. \tag{397}$$

Esta última expresión puede ser resuelta usando Factorización LU o Gradiente Conjugado (cada una de estas opciones tiene ventajas y desventajas que deben ser evaluadas al momento de implementar el código para un problema particular). Una vez obtenido $\underline{x4}$, podremos calcular

$$\underline{x2} = \underline{\underline{A}}^{\alpha}_{\Delta\Pi}\underline{x4} \tag{398}$$

así

$$\underline{\tilde{u}_{\Gamma}}^{\alpha} = \underline{x1} - \underline{x2} \tag{399}$$

completando la secuencia de operaciones necesaria para obtener $\underline{\underline{S}}_{\alpha}\underline{y_{\Gamma}}^{\alpha}$.

5.4.2. Cálculo de los Nodos Interiores

La evaluación de

$$\underline{u}_{\Pi} = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_{\Delta} \tag{400}$$

involucra nuevamente cálculos locales de la expresión

$$\underline{u_I}^{\alpha} = -\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}^{\alpha}_{\Pi\Delta}\underline{u_{\Gamma}}^{\alpha} \tag{401}$$

en esta está nuevamente involucrado $\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}$, por ello deberemos de usar el siguiente procedimiento para evaluar eficientemente esta expresión, realizando las operaciones equivalentes

$$\begin{aligned} \underline{x4} &= \underline{\underline{A}}^{\alpha}_{\Pi\Delta}\underline{u_{\Gamma}}^{\alpha} \\ \underline{u_I}^{\alpha} &= \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{x4} \end{aligned} \tag{402}$$

multiplicando por $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}$ a la última expresión, obtenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{u_I}^{\alpha} = \underline{\underline{A}}^{\alpha}_{\Pi\Pi}\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{x4} \tag{403}$$

simplificando, tenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{u_I}^{\alpha} = \underline{x4} \tag{404}$$

esta última expresión puede ser resuelta usando Factorización LU o Gradiente Conjugado.

Como se observo, para resolver el sistema $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{x} = \underline{b}$ podemos usar Factorización LU, Gradiente Conjugado o cualquier otro método para resolver sistemas lineales, pero deberá de usarse aquel que proporcione la mayor velocidad en el cálculo o que consuma la menor cantidad de memoria (ambas condicionantes son mutuamente excluyentes), por ello la decisión de que método usar deberá de tomarse al momento de tener que resolver un problema particular en un equipo dado y básicamente el condicionante es el tamaño del la matriz $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}$.

Para usar el método de Factorización LU, se deberá primeramente de factorizar la matriz bandada $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}$ en una matriz $\underline{\underline{LU}}$, la cual es bandada pero incrementa el tamaño de la banda a más del doble, pero esta operación sólo se deberá de realizar una vez por cada subdominio, y para solucionar los diversos sistemas lineales $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{x} = \underline{b}$ sólo será necesario evaluar los sistemas

$$\begin{aligned} \underline{\underline{L}}\underline{y} &= \underline{b} \\ \underline{\underline{U}}\underline{x} &= \underline{y} \end{aligned} \tag{405}$$

en donde $\underline{y}$ es un vector auxiliar. Esto proporciona una manera muy eficiente de evaluar el sistema lineal pero el consumo en memoria para un problema particular puede ser excesivo.

Por ello, si el problema involucra una gran cantidad de nodos interiores y el equipo en el que se implantará la ejecución del programa tiene una cantidad de memoria muy limitada, es recomendable usar el método de Gradiente Conjugado, este consume una cantidad de memoria adicional muy pequeña y el tiempo de ejecución se optimiza versus la Factorización LU.

De esta forma, es posible adaptar el código para tomar en cuenta la implementación de este en un equipo de cómputo en particular y poder sacar el máximo provecho al método de Subestructuración en la resolución de problemas elípticos de gran envergadura.

En lo que resta del presente trabajo, se asume que el método empleado para resolver $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{x} = \underline{b}$ en sus respectivas variantes necesarias para evitar el cálculo de $\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}$ es el método de Gradiente Conjugado, logrando así el máximo desempeño en velocidad en tiempo de ejecución.

5.4.3. Cálculo de la Matriz $\underline{\underline{S}}^{-1}$

En los algoritmos desarrollados interviene el cálculo de $\underline{\underline{S}}^{-1}$, dado que la matriz $\underline{\underline{S}}$ no se construye, entonces la matriz $\underline{\underline{S}}^{-1}$ tampoco es necesaria construirla, en lugar de ello se procede de la siguiente manera. Se asume que en las operaciones anteriores al producto de $\underline{\underline{S}}^{-1}$, se ha obtenido un vector, supongamos que es $\underline{v}$, entonces para hacer

$$\underline{u} = \underline{\underline{S}}^{-1}\underline{v} \tag{406}$$

se procedemos a multiplicar por $\underline{\underline{S}}$ a la ecuación anterior

$$\underline{\underline{S}}\underline{u} = \underline{\underline{S}}\underline{\underline{S}}^{-1}\underline{v} \tag{407}$$

obteniendo

$$\underline{\underline{S}}\underline{u} = \underline{v}. \tag{408}$$

Es decir, mediante algún proceso iterativo (usando CGM o factorización de LU) resolveremos el sistema anterior, de tal forma que en cada iteración de $\underline{u}^i$ se procede como se indico en la sección del cálculo de $\underline{\underline{S}}$, resolviendo $\underline{u}^{i+1} = \underline{\underline{S}}\underline{u}^i$.

5.4.4. Implementación de la Matriz $\underline{\underline{J}}$

Primeramente indicaremos una forma de construir la matriz $\underline{\underline{B}}$ y como construir a la matriz $\underline{\underline{J}}$ véase [**?**], la cual en nuestras pruebas presenta mejores resultados en la implementación computacional que la matriz $\underline{\underline{B}}$, para ello recordando, la matriz $\underline{\underline{B}}$ es formada por las matrices $\underline{\underline{B}}^i$ en cada $i-$ ésimo subdominio tal que la solución asociada a más de un subdominio coincida, i.e.

$$\underline{\underline{B}} = \left[\begin{array}{cccc} \underline{\underline{B}}^1, & \cdots & \cdots, & \underline{\underline{B}}^E \end{array} \right] \tag{409}$$

es construida de $\{0, 1, -1\}$ tal que los valores de la solución ,$\underline{u}$ asociada a más de un subdominio coincida cuando $\underline{\underline{B}}\underline{u} = 0$, donde la elección de $\underline{\underline{B}}$ no es única.

Estructura de $\underline{\underline{B}}^{(q)}$, donde q es un nodo de multiplicidad 2 queda como

$$\underline{\underline{B}}^{(q)} = \left[\begin{array}{cc} 1 & -1 \\ 1 & -1 \end{array} \right] \tag{410}$$

y la estructura de $\underline{\underline{B}}^{(q)}$, donde q es un nodo de multiplicidad 4 queda como

$$\underline{\underline{B}}^{(q)} = \left[\begin{array}{cccc} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{array} \right]. \tag{411}$$

Por otro lado, la matriz $\underline{\underline{j}}$ es formada por las matrices $\underline{\underline{j}}_i$ en cada $i-$ ésimo subdominio tal que la solución $\underline{u}$ asociada a más de un subdominio coincida, i.e.

$$\underline{\underline{j}} = \left[\begin{array}{cccc} \underline{\underline{j}}^1, & \cdots & \cdots, & \underline{\underline{j}}^E \end{array} \right] \text{ cy } \qquad \underline{\underline{j}}\underline{u} = 0. \tag{412}$$

Para ello, primeramente definimos dos matrices $\underline{\underline{a}}$ y $\underline{\underline{j}}$ con la propiedad de que $\underline{\underline{I}} = \underline{\underline{a}} + \underline{\underline{j}}$, donde $\underline{\underline{a}}$ y $\underline{\underline{j}}$ son ambas simétricas, no-negativas e indenpotentes, donde además $\underline{\underline{a}}\underline{\underline{j}} = \underline{\underline{j}}\underline{\underline{a}} = \underline{\underline{0}}$, mediante la matriz local de promedio definida como $\underline{\underline{a}}^{(q)} = \frac{1}{|Z|} \sum_{q \in Z} \underline{u}(q)$ donde $|Z|$ es la multiplicidad de los nodos primales q, y la matriz local de salto definida como $\underline{\underline{j}}^{(q)} = \underline{\underline{I}} - \underline{\underline{a}}^{(q)}$.

Así, la estructura de $\underline{\underline{a}}^{(q)}$ y $\underline{\underline{j}}^{(q)}$, donde q es un nodo de multiplicidad 2 queda como

$$\underline{\underline{a}}^{(q)} = \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{array} \right] \qquad y \qquad \underline{\underline{j}}^{(q)} = \left[\begin{array}{cc} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{array} \right] \tag{413}$$

y la estructura de $\underline{\underline{a}}^{(q)}$ y $\underline{\underline{j}}^{(q)}$, donde q es un nodo de multiplicidad 4 queda como

$$\underline{\underline{a}}^{(q)} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \qquad y \qquad \underline{\underline{j}}^{(q)} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix}. \quad (414)$$

Más concretamente supóngase que se tiene un dominio que es particionado en 2x2 y cada subdominio en 2x2, usando FETI One-Level tenemos que hay 4 subdominios, en cada subdominio se tiene 4 elementos, 9 nodos, 1 nodo interior y 3 nodos de frontera interior. En total se tienen 4 nodos interiores, 5 nodos en la frontera interior y 12 multiplicadores de Lagrange.

Suponiendo que se construyera las matrices globales, entonces la estructura de las matrices $\underline{\underline{B}}$, $\underline{\underline{a}}$ y $\underline{\underline{j}}$ seria la siguiente:

$$\underline{\underline{B}} = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix},$$

$$\underline{\underline{a}} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}$$

y

$$\underline{\underline{j}} = \begin{bmatrix} \frac{1}{2} & 0 & 0 & \frac{-1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{4} & 0 & 0 & \frac{-1}{4} & 0 & \frac{-1}{4} & 0 & 0 & \frac{-1}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{-1}{2} & 0 & 0 & 0 & 0 \\ \frac{-1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{-1}{4} & 0 & 0 & \frac{3}{4} & 0 & \frac{-1}{4} & 0 & 0 & \frac{-1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{-1}{2} & 0 \\ 0 & \frac{-1}{4} & 0 & 0 & \frac{-1}{4} & 0 & \frac{3}{4} & 0 & 0 & \frac{-1}{4} & 0 & 0 \\ 0 & 0 & \frac{-1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{-1}{2} \\ 0 & \frac{-1}{4} & 0 & 0 & \frac{-1}{4} & 0 & \frac{-1}{4} & 0 & 0 & \frac{3}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{-1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{-1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

FETI One-Level En el caso de FETI One-Level, al ser la matriz $\underline{\underline{j}}$ simétrica y positiva definida, la cual cumple también la restricción dada por la Ec. (342) ya que $\underline{\underline{j}}\underline{u} = 0$, entonces la formulación silla de dicha ecuación, se reduce a:

Encontrar $(\underline{u}, \underline{\lambda}) \in W \times U$ tal que

$$\begin{cases} \underline{\underline{S}}\underline{u} + \underline{\underline{j}}\underline{\lambda} = \underline{f} \\ \underline{\underline{j}}\underline{u} = \underline{0} \end{cases} \tag{415}$$

la solución $\underline{\lambda}$ de la Ec. (415) es única salvo la adición de un elemento del $Kernel\left(\underline{\underline{j}}\right)$. El espacio de multiplicadores de Lagrange U, es por lo tanto elegido como el $Rango\left(\underline{\underline{j}}\right)$. Este espacio puede ser entendido como el espacio de las funciones de salto en W.

El método One-Level FETI es un método de Gradiente Conjugado precondicionado -ver sección (9.2.1)- en el espacio V, aplicado a

$$\underline{\underline{P}}^T \underline{\underline{F}}\underline{\lambda} = \underline{\underline{P}}^T \underline{d}, \qquad \underline{\lambda} \in \underline{\lambda_0} + V \tag{416}$$

con condición inicial aproximada $\underline{\lambda_0}$ escogido tal que $\underline{\underline{G}}\underline{\lambda_0} = \underline{e}$. Con el precondicionador más básico para FETI al tomar $\underline{\underline{Q}} = \underline{\underline{I}}$, tiene la forma

$$\underline{\underline{M}}^{-1} = \underline{\underline{j}}\underline{\underline{S}}\underline{\underline{j}} = \sum_{i=1}^{E} \underline{\underline{j}}^i \underline{\underline{S}}^i \underline{\underline{j}}^i \tag{417}$$

donde

$$\underline{\underline{F}} = \underline{\underline{j}}\underline{\underline{S}}^{\dagger}\underline{\underline{j}}^T, \qquad \underline{\underline{G}} = \underline{\underline{j}}\underline{\underline{R}}, \qquad \underline{d} = \underline{\underline{j}}\underline{\underline{S}}^{\dagger}\underline{f}, \qquad \underline{e} = \underline{\underline{R}}^T\underline{f} \tag{418}$$

y

$$\underline{\underline{P}}^T = \underline{\underline{I}} - \underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{Q}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T\underline{\underline{Q}} \tag{419}$$

$$\underline{\underline{P}} = \underline{\underline{I}} - \underline{\underline{Q}}\underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{Q}}\underline{\underline{G}}\right)^{-1}\underline{\underline{G}}^T. \tag{420}$$

Entonces el método One-Level FETI usando la matriz $\underline{\underline{j}}$ en vez de la matriz $\underline{\underline{B}}$ queda en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$$\underline{\lambda_0} = \underline{\underline{Q}}\underline{\underline{G}}\left(\underline{\underline{G}}^T\underline{\underline{Q}}\underline{\underline{G}}\right)^{-1}\underline{e} + \underline{\mu}, \underline{\mu} \in Rango\left(\underline{\underline{P}}\right)$$
$$\underline{r_0} = \underline{d} - \underline{\underline{F}}\underline{\lambda}^0$$
$$\beta^0 = 0$$
$$\underline{p}^1 = 0$$

2.- Itera $k = 1, 2...$ hasta converger

$$\underline{q}^{k-1} = \underline{\underline{P}}^T\underline{r}^{k-1}$$
$$\underline{z}^{k-1} = \hat{\underline{\underline{M}}}^{-1}\underline{q}^{k-1}$$
$$\underline{y}^{k-1} = \underline{\underline{P}}\underline{z}^{k-1}$$
$$\beta^k = \frac{\left\langle \underline{y}^{k-1}, q^{k-1}\right\rangle}{\left\langle \underline{y}^{k-2}, \underline{q}^{k-2}\right\rangle}$$
$$\underline{p}^k = \underline{y}^{k-1} + \beta^k\underline{p}^{k-1}$$
$$\alpha^k = \frac{\left\langle \underline{y}^{k-1}, q^{k-1}\right\rangle}{\left\langle \underline{p}^k, \underline{\underline{F}}\underline{q}^k\right\rangle}$$
$$\underline{\lambda}^k = \underline{\lambda}^{k-1} + \alpha^k\underline{p}^k$$
$$\underline{r}^k = \underline{r}^{k-1} - \alpha^k\underline{\underline{F}}\underline{p}^k$$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u}$ mediante

$$\underline{u} = \underline{\underline{S}}^{\dagger}\left(\underline{f} - \underline{\underline{j}}\underline{\lambda}\right) \tag{421}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

FETI One-Level Simplificado En el caso de FETI One-Level simplificado se toma a $\underline{\underline{R}} = \underline{\underline{0}}$, se reduce a:

El método One-Level FETI es un método de Gradiente Conjugado precondicionado -ver sección (9.2.1)- en el espacio V, aplicado a

$$\underline{\underline{F}}\underline{\lambda} = \underline{d}, \tag{422}$$

con el precondicionador

$$\underline{\underline{M}}^{-1} = \underline{\underline{j}}\underline{\underline{S}}\underline{\underline{j}} = \sum_{i=1}^{E}\underline{\underline{j}}^i\underline{\underline{S}}^i\underline{\underline{j}}^i \tag{423}$$

donde

$$\underline{\underline{F}} = \underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{\underline{j}}^{T}, \qquad \underline{d} = \underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{f}. \tag{424}$$

Entonces el método One-Level FETI simplificado queda en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$$\begin{aligned}
&\underline{\lambda_0} = \underline{0}\\
&\underline{r_0} = \left(\underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{f}\right) - \left(\underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{\underline{j}}\right)\underline{\lambda}^{0}\\
&\beta^{0} = 0\\
&\underline{p}^{1} = 0
\end{aligned}$$

2.- Itera $k = 1, 2...$ hasta converger

$$\begin{aligned}
&\underline{q}^{k-1} = \underline{r}^{k-1}\\
&\underline{z}^{k-1} = \left(\underline{\underline{j}}\,\underline{\underline{S}}\,\underline{\underline{j}}\right)^{-1}\underline{q}^{k-1}\\
&\underline{y}^{k-1} = \underline{z}^{k-1}\\
&\beta^{k} = \frac{\left\langle \underline{y}^{k-1}, \underline{q}^{k-1}\right\rangle}{\left\langle \underline{y}^{k-2}, \underline{q}^{k-2}\right\rangle}\\
&\underline{p}^{k} = \underline{y}^{k-1} + \beta^{k}\underline{p}^{k-1}\\
&\alpha^{k} = \frac{\left\langle \underline{y}^{k-1}, \underline{q}^{k-1}\right\rangle}{\left\langle \underline{p}^{k}, \left(\underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{\underline{j}}\right)\underline{q}^{k}\right\rangle}\\
&\underline{\lambda}^{k} = \underline{\lambda}^{k-1} + \alpha^{k}\underline{p}^{k}\\
&\underline{r}^{k} = \underline{r}^{k-1} - \alpha^{k}\left(\underline{\underline{j}}\,\underline{\underline{S}}^{\dagger}\,\underline{\underline{j}}\right)\underline{p}^{k}
\end{aligned}$$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u}$ mediante

$$\underline{u} = \underline{\underline{S}}^{\dagger}\left(\underline{f} - \underline{\underline{j}}\,\underline{\lambda}\right) \tag{425}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

FETI Dual-Primal En el caso de Dual-Primal, al ser la matriz $\underline{\underline{j}}$ simétrica y positiva definida, la cual cumple también la restricción dada por la Ec. (383) ya que $\underline{\underline{j}}\,\underline{u} = 0$, entonces la formulación silla de dicha ecuación, se reduce a:

Encontrar $\underline{u_{\Delta}} \in \tilde{W}$ tal que

$$\left.\begin{array}{c} J(\underline{u_{\Delta}}) = \frac{1}{2}\left\langle \underline{\tilde{S}}\,\underline{u_{\Delta}}, \underline{u_{\Delta}}\right\rangle - \left\langle \underline{\tilde{f}_{\Delta}}, \underline{u_{\Delta}}\right\rangle \rightarrow \text{mín} \\ \underline{\underline{j}}_{\Delta}\,\underline{u_{\Delta}} = 0 \end{array}\right\} \tag{426}$$

la matriz $\underline{\underline{j_\Delta}}$ es construida tal que los valores de la solución , $\underline{u_\Delta}$ asociada a más de un subdominio coincida cuando $\underline{\underline{j}}_\Delta \underline{u_\Delta} = 0$.

Así, el método FETI-DP es un método de Gradiente Conjugado precondicionado para resolver el sistema precondicionado

$$\underline{\underline{M}}^{-1} \underline{\underline{F}} \underline{\lambda} = \underline{\underline{M}}^{-1} \underline{d} \tag{427}$$

con el precondicionador

$$\begin{aligned} \underline{\underline{M}}^{-1} &= \underline{\underline{j}}_{D,\Delta} \underline{\underline{S}}_\Delta \underline{\underline{j}}_{D,\Delta} \\ &= \sum_{i=1}^{E} \underline{\underline{D}}^i_\Delta \underline{\underline{j}}^i_\Delta \underline{\underline{S}}^i_\Delta \underline{\underline{j}}^i_\Delta \underline{\underline{D}}^i_\Delta \end{aligned} \tag{428}$$

donde $\underline{\underline{S}}^i_\Delta$ es la restricción del complemento de Schur local $\underline{\underline{S}}^i$ a $\tilde{W}_{\Delta,i} \subset W^i$ y

$$\underline{\underline{F}} = \underline{\underline{j}}_\Delta \left(\underline{\underline{\tilde{S}}} \right)^{-1} \underline{\underline{j}}_\Delta, \qquad \underline{d} = \underline{\underline{j}}_\Delta \left(\underline{\underline{\tilde{S}}} \right)^{-1} \underline{\tilde{f}_\Delta} \tag{429}$$

$$\underline{\underline{j}}_{D,\Delta} = \left[\begin{array}{cccc} \underline{\underline{D}}^1_\Delta \underline{\underline{j}}^1_{D,\Delta} & , ..., & \underline{\underline{D}}^E_\Delta \underline{\underline{j}}^E_{D,\Delta} \end{array} \right] \tag{430}$$

$$\underline{\underline{\tilde{S}}} = \underline{\underline{A}}_{\Delta\Delta} - \left[\begin{array}{cc} \left(\underline{\underline{A}}_{I\Delta} \right)^T & \left(\underline{\underline{A}}_{\Pi\Delta} \right)^T \end{array} \right] \left[\begin{array}{cc} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Pi} \\ \left(\underline{\underline{A}}_{I\Pi} \right)^T & \underline{\underline{A}}_{\Pi\Pi} \end{array} \right]^{-1} \left[\begin{array}{c} \underline{\underline{A}}_{I\Delta} \\ \underline{\underline{A}}_{\Pi\Delta} \end{array} \right] \tag{431}$$

$$\underline{\tilde{f}_\Delta} = \underline{f_\Delta} - \left[\begin{array}{cc} \left(\underline{\underline{A}}_{I\Delta} \right)^T & \left(\underline{\underline{A}}_{\Pi\Delta} \right)^T \end{array} \right] \left[\begin{array}{cc} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\Pi} \\ \left(\underline{\underline{A}}_{I\Pi} \right)^T & \underline{\underline{A}}_{\Pi\Pi} \end{array} \right]^{-1} \left[\begin{array}{c} \underline{f_I} \\ \underline{f_\Pi} \end{array} \right] \tag{432}$$

El método FETI-DP queda en términos del método de Gradiente Conjugado precondicionado como a continuación se muestra:

1.- Inicializa

$\underline{r}^0 = \underline{d} - \underline{\underline{F}} \underline{\lambda}^0$
$\beta^1 = 0$
$\underline{p}^1 = \underline{z}^0$

2.- Itera $k = 1, 2...$ hasta converger

$\underline{z}^{k-1} = \underline{\underline{M}}^{-1} \underline{r}^{k-1}$
$\beta^k = \frac{\left\langle \underline{z}^{k-1}, r^{k-1} \right\rangle}{\left\langle \underline{z}^{k-2}, \underline{r}^{k-2} \right\rangle}$
$\underline{p}^k = \underline{z}^{k-1} + \beta^k \underline{p}^{k-1}$
$\alpha^k = \frac{\left\langle \underline{z}^{k-1}, r^{k-1} \right\rangle}{\left\langle p^k, \underline{\underline{F}} q^k \right\rangle}$
$\underline{\lambda}^k = \underline{\lambda}^{k-1} + \alpha^k \underline{p}^k$
$\underline{r}^k = \underline{r}^{k-1} - \alpha^k \underline{\underline{F}} p^k$

Así, una vez calculando el multiplicador de Lagrange $\underline{\lambda}$ obtenemos la solución en los nodos de la frontera interior $\underline{u_\Delta}$ mediante

$$\underline{u_\Delta} = \left(\underline{\underline{\tilde{S}}}\right)^{-1} \left(\underline{\tilde{f}_\Delta} - \left(\underline{\underline{B}}_\Delta\right)^T \underline{\lambda}\right) \tag{433}$$

y para obtener la solución en los nodos interiores de cada subdominio se recurre a la aplicación de la Ec (691) del método de subestructuración o complemento de Schur, solucionando así el problema.

5.5. 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++ usando 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 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.

La implementación de los métodos a priori, requieren de más trabajo tanto en la face de construcción como en la parte de su aplicación, la gran ventaja de este tipo de precondicionadores es que pueden ser óptimos, 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.

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:

$$\begin{aligned} -\nabla^2 u + k^2 u &= f_\Omega \text{ en } \Omega \\ u &= g_{\partial\Omega} \text{ en } \partial\Omega. \end{aligned} \tag{434}$$

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.

En particular consideramos el problema con Ω definido en:

$$\Omega = [0,1] \times [0,1] \tag{435}$$

donde

$$f_\Omega = \exp(xy) \quad \text{y} \quad g_{\partial\Omega} = 0. \tag{436}$$

En todos los cálculos de los métodos numéricos usados para resolver el sistema lineal algebraico asociado se usó una tolerancia mínima de 1×10^{-5}.

A partir de la formulación del método de elemento finito visto en la sección (2.1.2), 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

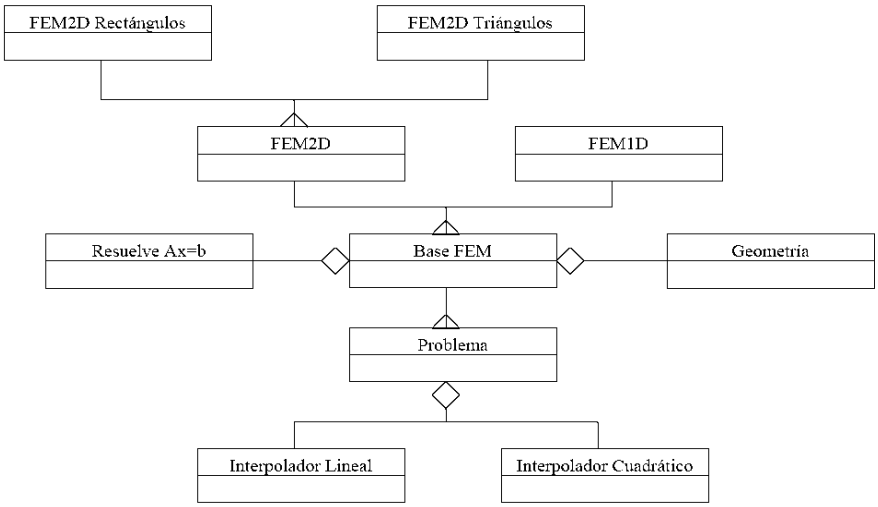

Figura 3: Jerarquía de clases para el método de elemento finito

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.

Por otro lado, la computación en paralelo es una técnica que nos permite distribuir una gran carga computacional entre muchos procesadores. Y es bien sabido que una de las mayores dificultades del procesamiento en paralelo es la coordinación de las actividades de los diferentes procesadores y el intercambio de información entre los mismos.

Para hacer una adecuada coordinación de actividades entre los diferentes procesadores, el programa que soporta el método de subestructuración paralelo, usa la misma jerarquía de clases que el método de subestructuración secuencial. Este se desarrolló para usar el esquema maestro-esclavo, de forma tal que el nodo maestro mediante la agregación de un objeto de la clase de *Geometría* genere la descomposición gruesa del dominio y los nodos esclavos creen un conjunto de

objetos *FEM2D Rectángulos* para que en estos objetos se genere la participación fina y mediante el paso de mensajes (vía MPI) puedan comunicarse los nodos esclavos con el nodo maestro.

La implementación computacional que se desarrolló tiene una jerarquía de clases en la cual se agregan las clases *FEM2D Rectángulos* y *Geometría*, además de heredar a la clase *Problema*. De esta forma se rehusó todo el código desarrollado para *FEM2D Rectángulos*, la jerarquía queda como:

> La clase *DDM2D* realiza la partición gruesa del dominio mediante la clase *Geometría* y controla la partición de cada subdominio mediante un objeto de la clase de *FEM2D Rectángulos* generando la partición fina del dominio. La resolución de los nodos de la frontera interior se hace mediante el método de gradiente conjugado, necesaria para resolver los nodos internos de cada subdominio.

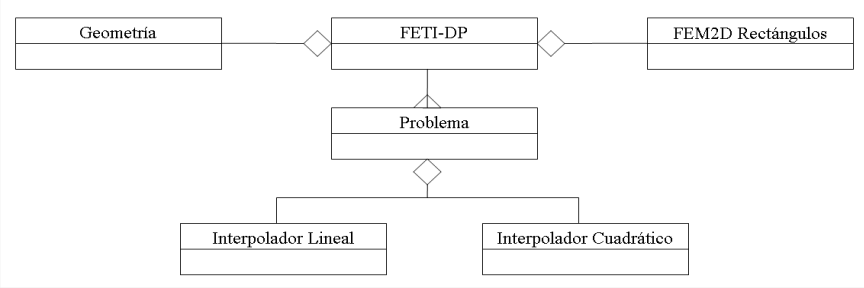

Figura 4: Jerarquía de clases para el método de subestructuración secuencial

Así, el dominio Ω es descompuesto en una descomposición gruesa de $n \times m$ subdominios y cada subdominio Ω_i se parte en $p \times q$ subdominios, generando la participación fina del dominio como se muestra en la figura:

Realizando las siguientes tareas:

> A) El nodo maestro genera la descomposición gruesa del dominio (supongamos particionado en $n \times m$ subdominios) mediante la agregación de un objeto de la clase *Geometría*, esta geometría es pasada a los nodos esclavos.
>
> B) Con esa geometría se construyen los objetos *FEM2D Rectángulos* (uno por cada subdominio), donde cada subdominio es particionado (supongamos en $p \times q$ subdominios). Cada objeto de *FEM2D Rectángulos* genera la geometría solicitada, regresando las coordenadas de los nodos de frontera del subdominio correspondiente al nodo maestro.

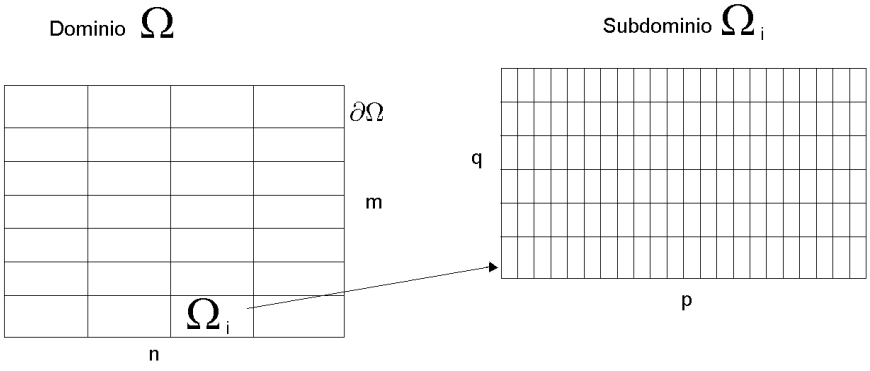

Figura 5: Descomposición del dominio Ω en $E = n \times m$ subdominios y cada subdominio Ω_i en $p \times q$ subdominios

C) Con estas coordenadas, el nodo maestro conoce a los nodos de la frontera interior (son estos los que resuelve el método de descomposición de dominio). Las coordenadas de los nodos de la frontera interior se dan a conocer a los objetos *FEM2D Rectángulos* en los nodos esclavos, transmitiendo sólo aquellos que están en su subdominio.

D) Después de conocer los nodos de la frontera interior, cada objeto *FEM2D Rectángulos* calcula las matrices

$$\underline{\underline{A}}^i_{II}, \underline{\underline{A}}^i_{I\Pi}, \underline{\underline{A}}^i_{I\Delta}, \underline{\underline{A}}^i_{\Pi\Pi}, \underline{\underline{A}}^i_{\Pi\Delta} \text{ y } \underline{\underline{A}}_{\Delta\Delta}$$

necesarias para construir el complemento de Schur local

$$\underline{\underline{\tilde{S}}}^i = \underline{\underline{A}}^i_{\Delta\Delta} - \begin{bmatrix} \left(\underline{\underline{A}}^i_{I\Delta}\right)^T & \left(\underline{\underline{A}}^i_{\Pi\Delta}\right)^T \end{bmatrix} \begin{bmatrix} \underline{\underline{A}}^i_{II} & \underline{\underline{A}}^i_{I\Pi} \\ \left(\underline{\underline{A}}^i_{I\Pi}\right)^T & \underline{\underline{A}}^i_{\Pi\Pi} \end{bmatrix}^{-1} \begin{bmatrix} \underline{\underline{A}}^i_{I\Delta} \\ \underline{\underline{A}}^i_{\Pi\Delta} \end{bmatrix}$$

$$\underline{\tilde{f}_\Delta}^i = \underline{f_\Delta}^i - \begin{bmatrix} \left(\underline{\underline{A}}^i_{I\Delta}\right)^T & \left(\underline{\underline{A}}^i_{\Pi\Delta}\right)^T \end{bmatrix} \begin{bmatrix} \underline{\underline{A}}^i_{II} & \underline{\underline{A}}^i_{I\Pi} \\ \left(\underline{\underline{A}}^i_{I\Pi}\right)^T & \underline{\underline{A}}^i_{\Pi\Pi} \end{bmatrix}^{-1} \begin{bmatrix} \underline{f_I}^i \\ \underline{f_\Pi}^i \end{bmatrix}$$

sin realizar comunicación alguna. Al terminar de calcular las matrices se avisa al nodo maestro de la finalización de los cálculos.

E) Mediante la comunicación de vectores del tamaño del número de nodos de la frontera interior entre el nodo maestro y los objetos *FEM2D Rectángulos*, se prepara todo lo necesario para empezar el método de gradiente conjugado y resolver el sistema lineal virtual $\underline{\underline{M}}^{-1}\underline{\underline{F}}\underline{\lambda} = \underline{\underline{M}}^{-1}\underline{d}$.

F) Para usar el método de gradiente conjugado, se transmite un vector del tamaño del número de nodos de la frontera interior para

que en cada objeto se realicen las operaciones pertinentes y resolver así el sistema algebraico asociado, esta comunicación se realiza de ida y vuelta entre el nodo maestro y los objetos *FEM2D Rectángulos* tantas veces como iteraciones haga el método. Resolviendo con esto los nodos de la frontera interior $\underline{u_{\Sigma}}_i$.

G) Al término de las iteraciones se pasa la solución $\underline{\lambda}^i$ de los nodos de la frontera interior que pertenecen a cada subdominio dentro de cada objeto *FEM2D Rectángulos* para que se resuelvan los nodos interiores $\underline{u_{\Delta}}^i = \left(\underline{\underline{\tilde{S}}}^i\right)^{-1}\left(\underline{\tilde{f}_{\Delta}}^i - \left(\underline{\underline{B}}^i_{\Delta}\right)^T \underline{\lambda}^i\right)$, sin realizar comunicación alguna en el proceso, al concluir se avisa al nodo maestro de ello.

I) El nodo maestro mediante un último mensaje avisa que se concluya el programa, terminado así el esquema maestro-esclavo.

Del algoritmo descrito anteriormente hay que destacar la sincronía entre el nodo maestro y los objetos *FEM2D Rectángulos* contenidos en los nodos esclavos, esto es patente en las actividades realizadas en los incisos A, B y C, estas consumen una parte no significativa del tiempo de cálculo.

Una parte importante del tiempo de cálculo es consumida en la generación de las matrices locales descritas en el inciso D que se realizan de forma independiente en cada nodo esclavo, esta es muy sensible a la discretización particular del dominio usado en el problema.

Los incisos E y F del algoritmo consumen la mayor parte del tiempo total del ejecución al resolver el sistema lineal que dará la solución a los nodos de la frontera interior. La resolución de los nodos interiores planteada en el inciso G consume muy poco tiempo de ejecución, ya que sólo se realiza una serie de cálculos locales previa transmisión del vector que contiene la solución a los nodos de la frontera interior.

Este algoritmo es altamente paralelizable ya que los nodos esclavos están la mayor parte del tiempo ocupados y la fracción serial del algoritmo esta principalmente en las actividades que realiza el nodo maestro, estas nunca podrán ser eliminadas del todo pero consumirán menos tiempo del algoritmo conforme se haga más fina la malla en la descomposición del dominio.

Por ejemplo, para resolver la Ec. (434), usando 3072×3072 nodos podemos tomar alguna de las siguientes descomposiciones:

Descomposición	Subdominios	Nodos Interiores	Elementos Subdominio	Total Nodos Subdominio	Nodos Desconocidos Subdominio
8×8 y 384×384	64	147456	148225	146689	9388096
16×16 y 192×192	96	36864	37249	36481	9339136
32×32 y 96×96	1024	9216	9409	9025	9241600
64×64 y 48×48	4096	2304	2401	2409	9048064
128×128 y 24×24	16384	576	625	529	8667136

Cada una de las descomposiciones genera un problema distinto. Usando el equipo secuencial a 2,8 GHz y evaluando el desempeño del método de subestructuración secuencial se obtuvieron los siguientes resultados:

Partición	Nodos Frontera Interior	Iteraciones Subestructuración	Tiempo Subestructuración
8×8 y 384×384	42945	4	18071 seg.
16×16 y 192×192	91905	3	4751 seg.
32×32 y 96×96	189441	2	911 seg.
64×64 y 48×48	382977	1	781 seg.
128×128 y 24×24	76395	1	3130 seg.

y para el método FETI-DP secuencial se obtuvieron los siguientes resultados:

Partición	Nodos Frontera Interior	Iteraciones FETI-DP	Tiempo FETI-DP
8×8 y 384×384	42945	2	14685 seg.
16×16 y 192×192	91905	2	3985 seg.
32×32 y 96×96	189441	1	777 seg.
64×64 y 48×48	382977	1	673 seg.
128×128 y 24×24	76395	1	2977 seg.

Nótese que aún en un solo procesador es posible encontrar una descomposición que desminuya los tiempos de ejecución (la descomposición de 64×64 y 48×48 concluye en 673 seg. versus los 781 seg. en el caso del algoritmo de subestructuración*)*.

Notemos también que en la última descomposición, en lugar de disminuir el tiempo de ejecución este aumenta, esto se debe a que se construyen muchos objetos *FEM2D Rectángulos* (76395 en este caso), con los cuales hay que hacer comunicación resultando muy costoso computacionalmente.

Por otro lado, para la implementación paralela, la descomposición adecuada del dominio para tener un buen balanceo de cargas se logra cuando se descompone en $n \times m$ nodos en la partición gruesa, generándose $n * m$ subdominios y si se trabaja con P procesadores (1 para el nodo maestro y $P-1$ para los nodos esclavos), entonces el balance de cargas adecuado será cuando $(P-1) \mid (n * m)$. Así, los siguientes tiempos fueron obtenidos al usar 1,2,3,5,9 y 17 procesadores.

Usando para los cálculos en un procesador el equipo secuencial y para la parte paralela el cluster homogéneo a 2,8 GHz resolviendo por el método de gradiente conjugado, la solución para una partición 64×64 y 48×48 se encontró la solución en 1 iteración en los siguientes tiempos:

Partición	CPUs	Tiempo Total
64×64 y 48×48	1	673 seg.
64×64 y 48×48	2	820 seg.
64×64 y 48×48	3	415 seg.
64×64 y 48×48	5	286 seg.
64×64 y 48×48	9	222 seg.
64×64 y 48×48	17	190 seg.

Las métricas de desempeño son las siguientes

Procesadores	Tiempo	Factor de Aceleración	Eficiencia	Fracción Serial
1	673 seg.			
2	820 seg.	0.8207	0.41036	1.43684
3	409 seg	1.6216	0.54056	0.42496
5	399 seg.	2.3531	0.47062	0.28120
9	353 seg.	3.0315	0.33683	0.24609
17	330 seg	3.5421	0.20835	0.23746

Estos resultados pueden ser apreciados mejor de manera gráfica como se muestra a continuación:

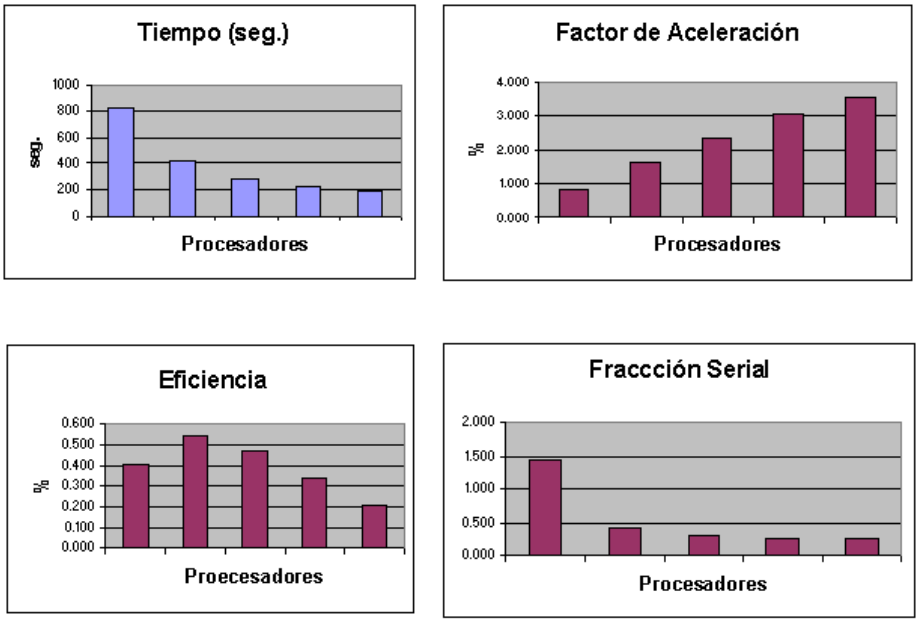

Figura 6: Métricas de desempeño mostrando sólo cuando las cargas están bien balanceadas (2, 3, 5, 9 y 17 procesadores).

En cuanto a las métricas de desempeño, obtenemos que el factor de aceleración en el caso ideal debería de aumentar de forma lineal al aumento del número de procesadores, que en nuestro caso no es lineal pero cumple bien este hecho si están balanceadas las cargas de trabajo.

El valor de la eficiencia deberá ser cercano a uno cuando el hardware es usado de manera eficiente, como es en nuestro caso cuando se tiene un procesador por cada subdominio.

Y en la fracción serial su valor debiera de tender a cero en el caso ideal, siendo este nuestro caso si están balanceadas las cargas de trabajo, de aquí se

puede concluir que la granularidad del problema es gruesa, es decir, no existe una sobrecarga en los procesos de comunicación siendo el cluster una buena herramienta de trabajo para este tipo de problemas.

Finalmente las posibles mejoras de eficiencia para el método de subestructuración en paralelo para disminuir el tiempo de ejecución pueden ser:

- Balanceo de cargas de trabajo homogéneo.
- Al compilar los códigos usar directivas de optimización.
- Usar bibliotecas que optimizan las operaciones en el manejo de los elementos de la matriz usando punteros en las matrices densas o bandadas.
- El cálculo de las matrices que participan en el complemento de Schur pueden ser obtenidas en paralelo.

6. Funciones Definidas por Tramos

Sea $\Omega \subset \mathbb{R}^n$ un dominio, y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios Ω_i sin traslape del dominio Ω -también conocida como malla gruesa $\mathcal{T}_H$. Un ejemplo de un dominio Ω y su descomposición en subdominios Ω_i y cada Ω_i a su vez descompuesto en Ω_e subdominios se muestra en la figura:

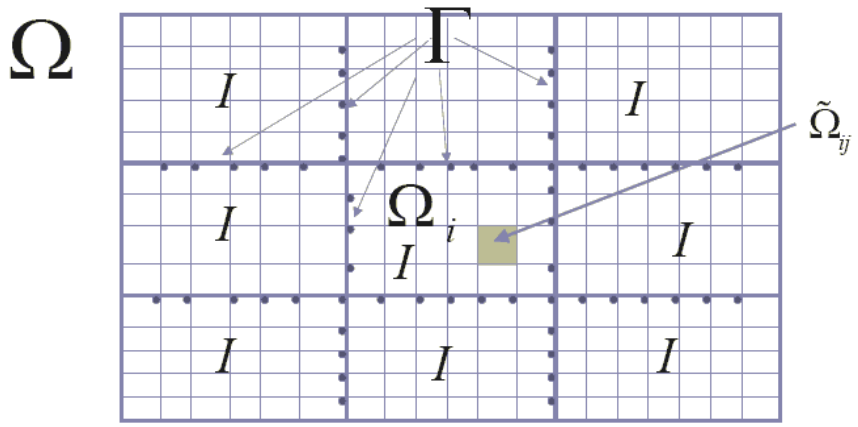

Figura 7: El dominio Ω, su frontera externa $\partial\Omega$ y la frontera interna Γ.

Se asume que:

1.- Ω_i, para $i = 1, ..., E$ es un subdominio de Ω,

2.- $\Omega_i \bigcap \Omega_j = \varnothing$, siempre que $i \neq j$.

3.- $\Omega \subset \bigcup_{i=1}^{E} \overline{\Omega_i}$.

La notación $\partial\Omega$ y $\partial\Omega_i$, $i = 1, ..., E$ es tomada de la frontera del dominio Ω y la frontera del subdominio Ω_i respectivamente, claramente

$$\partial\Omega \subset \bigcup_{i=1}^{E} \partial\Omega_i. \tag{437}$$

Adicionalmente definimos a la frontera interior como

$$\Gamma = \bigcup_{i \neq j} \partial\Omega_i \bigcap \partial\Omega_j \tag{438}$$

y a $\partial\Omega$ como la frontera exterior del dominio Ω, denotamos por H al diámetro $H_i = Diam(\Omega_i)$ de cada Ω_i que satisface $Diam(\Omega_i) \leq H$ para cada $i =$

$1, 2, ..., E$, además, cada subdominio Ω_i es descompuesto en una mallado fino $\mathcal{T}_h$ de K subdominios mediante una triangulación Ω_e de modo que esta sea conforme, denotamos por h al diámetro $h_i = Diam(\Omega_e)$ de cada Ω_e que satisface $Diam(\Omega_e) \leq h$ para cada $e = 1, 2, ..., K$ de cada $i = 1, 2, ..., E$.

También asumimos que salvo en un conjunto de medida cero sobre Γ se define un único vector normal denotado por $\underline{n}$ cuyo sentido se elige arbitrariamente y el lado positivo de Γ es definido hacia el sentido positivo del vector normal.

Sea $D(\Omega)$ un espacio lineal de funciones definidas en Ω, entonces

Definición 46 *Identificamos como una sola función, a dos funciones $u, w \in D(\Omega)$ cuyo dominio de definición está contenido en Ω, cuando se satisface la condición de que el conjunto de puntos en los cuales $u \neq w$ tiene medida de Lebesgue cero.*

Dada una partición $\Pi = \{\Omega_1, ..., \Omega_E\}$ del dominio Ω, entonces

Definición 47 *Una función definida por pedazos es una sucesión de funciones $\{w_1, ..., w_E\}$, tal que para cada $i = 1, ..., E$, la función w_i esta definida en Ω_i. Dada una función w definida en Ω esta tiene una única función definida por pedazos $\{w_1, ..., w_E\}$, tal que*

$$w_i = w\,|_{\Omega_i} \quad con\ i = 1, ..., E \tag{439}$$

donde $w\,|_{\Omega_i}$ es la restricción de w a Ω_i.

Esto establece una correspondencia biunívoca entre las funciones definidas en Ω y las funciones definidas por pedazos.

Definición 48 *Identificaremos a la función w definida en casi todos lados salvo un conjunto de medida cero en Ω y la correspondiente sucesión $\{w_1, ..., w_E\}$.*

Así, dada una función w contenida en Ω, la sucesión $\{w_1, ..., w_E\}$ será referida como la representación definida por pedazos de w y las funciones $w_i, i = 1, ..., E$ como componentes locales de w. También, establecemos una correspondencia biunívoca entre los espacios $\{D(\Omega_1), ..., D(\Omega_E)\}$ y $\hat{D}(\Omega)$ en Ω.

Definición 49 *Dada una familia $\{D(\Omega_1), ..., D(\Omega_E)\}$ de espacios lineales definidos en $\Omega_1, ..., \Omega_E$ respectivamente, definimos el espacio lineal $\hat{D}(\Omega)$ contenido en Ω por*

$$\hat{D}(\Omega) \equiv D(\Omega_1) \oplus ... \oplus D(\Omega_E). \tag{440}$$

Teorema 50 *Sea $\{w_1, ..., w_E\}$ la representación en pedazos de cualquier función w, entonces $w \in \hat{D}(\Omega)$ si y sólo si $w_i \in D(\Omega_i)$ para todo $i = 1, ..., E$.*

Definición 51 *El espacio lineal de funciones definidas por pedazos $\hat{D}(\Omega)$ por el espacio cuyos elementos son la restricción a Ω_i de las funciones pertenecientes a $D(\Omega_i)$.*

En cuyo caso la función de $D(\Omega)$ a $\hat{D}(\Omega)$ la cual asocia a cada $w \in D(\Omega)$ una representación en pedazos de

$$\{w_1, ..., w_E\} \in D(\Omega_1) \oplus ... \oplus D(\Omega_E) \tag{441}$$

es una biyección la cual es referida como la inmersión natural de $D(\Omega)$ sobre $D(\Omega_1) \oplus ... \oplus D(\Omega_E)$. En lo sucesivo identificaremos los dos espacios lineales y escribiremos

$$D(\Omega) \subset D(\Omega_1) \oplus ... \oplus D(\Omega_E). \tag{442}$$

En vista de las definiciones anteriores, para cada función $v \in \hat{D}(\Omega)$, existe una sucesión de funciones $\{v_1, ..., v_E\}$ tal que $v_i = v\mid_{\Omega_i}; i = 1, ..., E$, donde $v\mid_{\Omega_i}$ es la restricción de v a Ω_i.

Puesto que una función $v \in \hat{D}(\Omega)$ se forma por las restricciones de funciones definidas de manera independiente en cada subdominio Ω_i, esta es en general totalmente discontinua en la frontera interior Γ, así, las funciones que pertenecen a este espacio pueden tener discontinuidades de salto finitas tanto en el valor de la función como en el valor de sus derivadas normales en Γ.

Cuando se califica a una función como continua, se entiende que la función es continua en su valor, sin asumir nada acerca de la continuidad en sus derivadas. Aunque es claro que la función es continua en todo Ω, esto es, en cada subdominio Ω_i y en la frontera interior Γ, se pondrá énfasis en la continuidad através de Γ. Cuando se califica a una función como totalmente continua, se entiende que la función es continua tanto en su valor como en sus derivadas normales através de Γ. Cuando se califica a una función como totalmente discontinua, se entiende que la función puede presentar discontinuidades tanto en su valor como en sus derivadas normales a través de Γ.

Sea $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$ donde $\partial\Omega_i$ y $\partial\Omega_j$ son las fronteras de dos subregiones adyacentes, entonces definimos como la traza a la restricción de v^i a Γ_{ij}.

Pero como Γ_{ij}, para dos subregiones vecinas hay dos trazas definidas una que corresponde a v^i y otra a v^j, entonces se requiere introducir la siguiente notación para poderlas distinguir entre si:

$$v_+ \equiv Tr(v^i) \tag{443}$$

cuando Ω_i cae del lado positivo de Γ_{ij} y

$$v_- \equiv Tr(v^j) \tag{444}$$

en caso contrario. Aquí $Tr(v)$ designa al operador traza de la función v. En general $v_+ \neq v_-$ ya que se trabaja con espacios de funciones definidas por tramos.

Observación 52 *Notemos que al considerar una función w en Ω, su definición en Γ es innecesaria, ya que la medida de Lebesgue de Γ es cero. Si la traza de*

w_α es definida en casi todos lados salvo un conjunto de medida cero sobre $\partial\Omega_\alpha$ para $\alpha = 1, ..., E$, entonces tal traza es también definida en Γ. En particular, si la traza de w_α esta definida sobre $\partial\Omega_\alpha$ para cada $\alpha = 1, ..., E$, entonces ellas definen dos funciones definidas en casi todos lados salvo un conjunto de medida cero sobre Γ, denotadas por (w_+, w_-) correspondientes a los lados de trazas positivas y negativas de Γ respectivamente.

Definición 53 *El salto de v sobre Γ de funciones definidas por pedazos como*

$$[[w]] \equiv w_+ - w_- \tag{445}$$

y el promedio como

$$\dot{w} \equiv \frac{1}{2}(w_+ + w_-) \tag{446}$$

respectivamente.

Observemos que tanto el promedio de una función $\dot{w}$, como el producto $[w]\cdot\underline{n}$ no depende de como se elija el sentido del vector normal unitario $\underline{n}$ en Γ, además las siguientes identidades se satisfacen

$$w_+ = \dot{w} + \frac{1}{2}[[w]] \text{ y } w_- = \dot{w} - \frac{1}{2}[[w]]. \tag{447}$$

La discontinuidad de una función en la frontera interior Γ se puede expresar ya sea especificando los valores de sus trazas en Γ, o bien, especificando los valores de su promedio y de su salto en Γ. Además, si la función v es continua a través de Γ se tiene que

$$v = v_+ = v_- = \dot{v} \qquad \text{y que } [[v]] = 0.$$

6.1. Espacios de Sobolev de Funciones Definidas por Tramos

Dada una familia de espacios lineales $\{D(\Omega_1), ..., D(\Omega_E)\}$, tal que $D(\Omega_i)$, para cada $i = 1, 2, ...E$, es un espacio lineal de funciones definido en casi todos lados salvo un conjunto de medida cero en Ω_i, se puede considerar el espacio

$$D(\Omega) = D(\Omega_1) \oplus ... \oplus D(\Omega_E) \tag{448}$$

entonces, los elementos de $D(\Omega)$ son funciones definidas por tramos, $(w_1, ..., w_E)$, con $w_i \in D(\Omega_i)$, $i = 1, 2, ..., E$.

Definición 54 *El espacio de Sobolev de orden $p \geq 0$ para funciones definidas por tramos esta dado por*

$$\hat{H}^p(\Omega, \Pi) = H^p(\Omega_1) \oplus ... \oplus H^p(\Omega_E) \tag{449}$$

aquí, $H^p(\Omega_i)$ es el espacio de Sobolev de orden p, de funciones definidas en Ω_i. Cada función $w \in \hat{H}^p(\Omega)$ es una sucesión, $w \equiv (w_1, ..., w_E)$,con $w_i \in H^p(\Omega_i)$, $i = 1, 2, ..., E$.

Observemos que cuando $w \in H^p(\Omega)$, entonces las restricción, w_i de w a Ω_i tiene la propiedad que $w_i \in H^p(\Omega_i)$. Por lo tanto

$$H^p(\Omega) \subset \hat{H}^p(\Omega) \tag{450}$$

Para $p > 0$, esta es una inclusión propia. Sin embargo para $p = 0$, $H^0(\Omega) \equiv \hat{H}^0(\Omega) \equiv L^2(\Omega)$. Además

$$H^0(\Omega) \equiv \hat{H}^0(\Omega) \supset \hat{H}^p(\Omega) \qquad \forall p \geq 0 \tag{451}$$

Aquí las funciones definidas en Ω han sido identificadas con sus representaciones por partes. Todos los espacios $\hat{H}^p(\Omega)$, para $p = 0, 1, 2, ...$, están hechos de funciones las cuales pertenecen a $H^0(\Omega) \equiv L^2(\Omega)$.

Teorema 55 *Para cada $p \geq 0$, una función $\hat{u} = (u_1, ..., u_E) \in H^0(\Omega)$ pertenecen a $\hat{H}^p(\Omega)$ si y sólo si la norma*

$$||\hat{v}||_{p,\Omega,\Pi} = \left(\sum_{i=1}^{E} ||v_i||_{p,\Omega_i}^2 \right)^{\frac{1}{2}} \tag{452}$$

está bien definida.

Cuando $\hat{H}^p(\Omega)$ es equipada con esta norma el correspondiente producto interior $(\cdot, \cdot)$, se convierte en un espacio de Hilbert.

Observación 56 *Las siguientes propiedades se satisfacen:*

1.- Cuando $w \in H^p(\Omega)$, entonces la restricción de w a Ω_α, w_α tiene la propiedad de que $w_\alpha \in H^p(\Omega_p)$. Por lo tanto

$$H^p(\Omega) \subset \hat{H}^p(\Omega) \tag{453}$$

2.- Cuando $u \in \hat{H}^1(\Omega)$ entonces

$$[[u]] = 0 \text{ sobre } \Gamma \Leftrightarrow u \in H^1(\Omega) \tag{454}$$

3.- Cuando $u \in \hat{H}^2(\Omega)$ entonces

$$[[u]] = \left[\left[\frac{\partial u}{\partial n} \right] \right] = 0 \text{ sobre } \Gamma \Leftrightarrow u \in H^2(\Omega). \tag{455}$$

La identidad

$$\begin{aligned} \sum_{\alpha=1}^{E} \int_{\partial\Omega_\alpha} u_\alpha w_\alpha n_i dx &= \int_{\partial\Omega} uwn_i dx - \int_{\Gamma} (uw) n_i dx \\ &= \int_{\partial\Omega} uwn_i dx - \int_{\Gamma} \left(\dot{u} [[w]] + \dot{w} [[u]] \right) n_i dx \end{aligned} \tag{456}$$

puede ser fácilmente verificada. Aquí n_i es cualquier componente del vector normal unitario.

6.2. Fórmulas Green-Herrera

Sea Ω un dominio y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios del dominio Ω. Sea una ecuación diferencial en forma general

$$\mathcal{L}u = \mathcal{L}u_\Omega \equiv f_\Omega, \text{ en } \Omega_i, i = 1, ..., E \tag{457}$$

con condiciones de frontera

$$B_j u = B_j u_\partial \equiv g_\partial, \text{ en } \partial\Omega \tag{458}$$

y saltos prescritos

$$J_k u = J_k u_\Gamma \equiv j_\Gamma, \text{ en } \Gamma \tag{459}$$

donde B_j y J_k son k operadores diferenciales. Aquí, $u_\Omega = \left(u_\Omega^1, ..., u_\Omega^E\right)$, u_∂, y u_Γ son funciones dadas en $\widehat{D}_1(\Omega)$, que definen los datos del problema. De manera tal que tenemos un problema bien planteado, es decir, se garantiza la existencia y la unicidad de la solución. El problema enunciado se denomina *Problema con Valores en la Frontera con Saltos Prescritos.*

Si $u \in \widehat{D}_1(\Omega)$, entonces la ecuación diferencial $\mathcal{L}u$ está definida en el interior de cada Ω_i para $i = 1, ..., E$. De igual forma, si $w \in \widehat{D}_2(\Omega)$ entonces $\mathcal{L}^* w$ está definida en el interior de cada Ω_i, para $i = 1, ..., E$. Ambos operadores diferenciales podrían no estar definidas en $\Gamma \cup \partial\Omega$.

Y como por definición del operador diferencial $\mathcal{L}$ y su operador diferencial adjunto formal $\mathcal{L}^*$ satisfacen la condición

$$w\mathcal{L}u - u\mathcal{L}^* w = \nabla \cdot \underline{\mathfrak{D}}(u, w) \tag{460}$$

donde $\underline{\mathfrak{D}}(u, w)$ es una función bilineal definida en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ apropiada para el operador $\mathcal{L}$. Además asumimos que existen funcionales bilineales $\mathcal{B}(u, w), \mathcal{C}(w, u), \mathcal{J}(u, w)$ y $\mathcal{K}(w, u)$ donde las primeras dos están definidas en $\partial\Omega$ y las dos últimas sobre Γ, tal que

$$\underline{\mathfrak{D}}(u, w) \cdot \underline{n} = \mathcal{B}(u, w) - \mathcal{C}(w, u) \qquad \text{en } \partial\Omega \tag{461}$$

y

$$-\left[\left[\underline{\mathfrak{D}}(u, w)\right]\right] \cdot \underline{n} = \mathcal{J}(u, w) - \mathcal{K}(w, u) \qquad \text{en } \Gamma \tag{462}$$

generalmente, las definiciones de $\mathcal{B}$ y $\mathcal{J}$ depende de las condiciones de frontera y de criterios de suavidad del problema en particular de que se trate, si los coeficientes del operador diferencial son continuos, las formulas de Herrera para $\mathcal{J}$ y $\mathcal{K}$ satisfacen

$$\left[\left[\underline{\mathfrak{D}}(u, w)\right]\right] = \underline{\mathfrak{D}}(\dot{u}, [[w]]) + \underline{\mathfrak{D}}([[u]], \dot{w}) \tag{463}$$

i.e.

$$\mathcal{J}(u, w) \equiv -\underline{\mathfrak{D}}([[u]], \dot{w}) \cdot \underline{n} \text{ y } \mathcal{K}(w, u) \equiv \underline{\mathfrak{D}}(\dot{u}, [[w]]) \cdot \underline{n}. \tag{464}$$

Si integramos la ecuación (460) cada Ω_i para $i = 1, 2, ...E$, y se considera $\overline{\Omega} = \bigcup_{i=1}^{E} \overline{\Omega}_i$, se tiene que

$$\sum_{i=1}^{E} \int_{\Omega_i} \left(w\mathcal{L}u - u\mathcal{L}^* w \right) d\underline{x} = \sum_{i=1}^{E} \int_{\Omega_i} \nabla \cdot \underline{\mathfrak{D}}\left(u, w\right) d\underline{x} \tag{465}$$

aplicando la teorema generalizado de la divergencia

$$\int_{\Omega} \nabla \cdot \underline{\mathfrak{D}}\left(u, w\right) d\underline{x} = \int_{\partial\Omega} \underline{\mathfrak{D}}\left(u, w\right) \cdot \underline{n}_{\partial\Omega} d\underline{x} - \int_{\Gamma} \left[\left[\underline{\mathfrak{D}}\left(u, w\right)\right]\right] \cdot \underline{n}_{\Gamma} d\underline{x} \tag{466}$$

en el lado derecho de la Ec. (465) obtenemos

$$\sum_{i=1}^{E} \int_{\Omega_i} w\mathcal{L}u d\underline{x} - \sum_{i=1}^{E} \int_{\Omega_i} u\mathcal{L}^* w d\underline{x} = \int_{\Omega} \underline{\mathfrak{D}}\left(u, w\right) \cdot \underline{n} d\underline{x} - \int_{\Gamma} \left[\left[\underline{\mathfrak{D}}\left(u, w\right)\right]\right] \cdot \underline{n} d\underline{x} \tag{467}$$

desarrollando el algebra de saltos en el segundo sumando del lado derecho de la ecuación anterior se tiene

$$\begin{aligned} & \sum_{i=1}^{E} \int_{\Omega_i} w\mathcal{L}u d\underline{x} - \sum_{i=1}^{E} \int_{\Omega_i} u\mathcal{L}^* w d\underline{x} \qquad (468) \\ = & \int_{\Omega} \underline{\mathcal{D}}\left(u, w\right) \cdot \underline{n} d\underline{x} - \int_{\Gamma} \left[\left[\underline{\mathfrak{D}}\left(\dot{u}, \left[w\right]\right)\right]\right] \cdot \underline{n} d\underline{x} - \int_{\Gamma} \left[\left[\underline{\mathfrak{D}}\left(\left[\left[u\right]\right], \dot{w}\right)\right]\right] \cdot \underline{n} d\underline{x}. \end{aligned}$$

Ahora, para poder usar las formulas de Green se introducen las siguientes funcionales bilineales reales definidas en $\widehat{D}_1\left(\Omega\right) \times \widehat{D}_2\left(\Omega\right)$. Sean las funcionales bilineales $\mathcal{B}\left(u, w\right), \mathcal{C}^*\left(u, w\right), \mathcal{J}\left(u, w\right)$ y $\mathcal{K}^*\left(u, w\right)$ tales que producen las siguientes descomposiciones

$$\underline{\mathfrak{D}}\left(u, w\right) \cdot \underline{n} = \mathcal{B}\left(u, w\right) - \mathcal{C}^*\left(u, w\right) \qquad \text{en } \partial\Omega \tag{469}$$

$$\underline{\mathfrak{D}}\left(\dot{u}, \left[\left[w\right]\right]\right) \cdot \underline{n} \equiv \mathcal{K}^*\left(u, w\right) \qquad \text{en } \Gamma \tag{470}$$

$$-\underline{\mathfrak{D}}\left(\left[\left[u\right]\right], \dot{w}\right) \cdot \underline{n} \equiv \mathcal{J}\left(u, w\right) \qquad \text{en } \Gamma \tag{471}$$

donde $\mathcal{C}^*\left(u, w\right)$ es el transpuesto de la funcional bilineal $\mathcal{C}\left(w, u\right)$ y se define como

$$\mathcal{C}^*\left(u, w\right) \equiv \mathcal{C}\left(w, u\right) \tag{472}$$

de igual manera para

$$\mathcal{K}^*\left(u, w\right) \equiv \mathcal{K}\left(w, u\right). \tag{473}$$

La funcional bilineal $\mathcal{B}\left(u, w\right)$ está en función de los valores de frontera (condiciones de frontera y condiciones iniciales), mientras que la funcional $\mathcal{C}^*\left(u, w\right)$ involucra los valores desconocidos en $\partial\Omega$ (información desconocida).

Por otra parte, la funcional $\mathcal{K}^*(u,w)$ involucra los valores relacionados con los promedios $\dot{u}$ en Γ mientras que la funcional $\mathcal{J}(u,w)$ involucra los valores relacionados con los saltos $[[u]]$ en Γ.

Adicionalmente, consideramos las funcionales bilineales reales definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$:

$$\mathcal{P}(u,w) \equiv w\mathcal{L}u, \qquad \text{en } \Omega_i \text{ para } i = 1, ..., E \tag{474}$$

$$\mathcal{J}^*(u,w) \equiv w\mathcal{L}^*u, \qquad \text{en } \Omega_i \text{ para } i = 1, ..., E \tag{475}$$

Sustituyendo las Ecs. (469) a (475) en Ec. (468) y reordenando los términos, se obtiene la fórmula de Green-Herrara

$$\begin{aligned} &\sum_{i=1}^{E} \int_{\Omega_i} \mathcal{P}(u,w) d\underline{x} - \int_{\partial\Omega} \mathcal{B}(u,w) d\underline{x} - \int_{\Gamma} \mathcal{J}(u,w) d\underline{x} \qquad (476) \\ = \quad &\sum_{i=1}^{E} \int_{\Omega_i} \mathcal{J}^*(u,w) d\underline{x} - \int_{\partial\Omega} \mathcal{C}^*(u,w) d\underline{x} - \int_{\Gamma} \mathcal{K}^*(u,w) d\underline{x}. \end{aligned}$$

Ahora, sean las funcionales bilineales reales

$$\langle \mathcal{P}u, w\rangle, \langle \mathcal{B}u, w\rangle, \langle \mathcal{J}u, w\rangle, \langle Q^*u, w\rangle, \langle \mathcal{C}^*u, w\rangle \text{ y } \langle \mathcal{K}^*u, w\rangle \tag{477}$$

definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ tales que

$$\langle \mathcal{P}u, w\rangle \equiv \sum_{i=1}^{E} \int_{\Omega_i} \mathcal{P}(u,w) d\underline{x} = \sum_{i=1}^{E} \int_{\Omega_i} w\mathcal{L}u d\underline{x} \qquad \text{en } \Omega_i \text{ para } i = 1, ..., E \tag{478}$$

$$\langle \mathcal{B}u, w\rangle \equiv \int_{\partial\Omega} \mathcal{B}(u,w) d\underline{x} \qquad \text{en } \partial\Omega \tag{479}$$

$$\langle \mathcal{J}u, w\rangle \equiv \int_{\Gamma} \mathcal{J}(u,w) d\underline{x} \qquad \text{en } \Gamma \tag{480}$$

$$\langle Q^*u, w\rangle \equiv \sum_{i=1}^{E} \int_{\Omega_i} \mathcal{J}^*(u,w) d\underline{x} = \sum_{i=1}^{E} \int_{\Omega_i} w\mathcal{L}^*u d\underline{x} \qquad \text{en } \Omega_i \text{ para } i = 1, ..., E \tag{481}$$

$$\langle \mathcal{C}^*u, w\rangle \equiv \int_{\partial\Omega} \mathcal{C}^*(u,w) d\underline{x} \qquad \text{en } \partial\Omega \tag{482}$$

$$\langle \mathcal{K}^*u, w\rangle \equiv \int_{\Gamma} \mathcal{K}^*(u,w) d\underline{x} \qquad \text{en } \Gamma. \tag{483}$$

Si se sustituye las Ecs. (478) a (483) en (476), se tiene

$$\langle \mathcal{P}u, w\rangle - \langle \mathcal{B}u, w\rangle - \langle \mathcal{J}u, w\rangle = \langle Q^*u, w\rangle - \langle \mathcal{C}^*u, w\rangle - \langle \mathcal{K}^*u, w\rangle \tag{484}$$

que se satisface para todo $(u,w) \in \widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$, o bien, por la propiedad de linealidad

$$\langle (\mathcal{P} - \mathcal{B} - \mathcal{J})\, u, w \rangle = \langle (Q^* - \mathcal{C}^* - \mathcal{K}^*)\, u, w \rangle \tag{485}$$

o de manera compacta

$$\mathcal{P} - \mathcal{B} - \mathcal{J} = Q^* - \mathcal{C}^* - \mathcal{K}^* \tag{486}$$

esta expresión representa la fórmula Green-Herrera para operadores en campos discontinuos.

Las funcionales bilineales

$$\langle \mathcal{P}u, w \rangle, \langle \mathcal{B}u, w \rangle, \langle \mathcal{J}u, w \rangle, \langle Q^*u, w \rangle, \langle \mathcal{C}^*u, w \rangle \text{ y } \langle \mathcal{K}^*u, w \rangle \tag{487}$$

también se pueden considerar como operadores funcionales lineales definidos en $\widehat{D}_1(\Omega)$ y valuadas en $\widehat{D_2^*}(\Omega)$, i.e., valuados en el dual algebraico de $\widehat{D}_2(\Omega)$. Por ejemplo, $\mathcal{P} : \widehat{D}_1(\Omega) \to \widehat{D_2^*}(\Omega)$ donde $\mathcal{P}u \in \widehat{D_2^*}(\Omega)$ y $u \in \widehat{D}_1(\Omega)$. De igual modo, los transpuestos de las funcionales bilineales

$$\langle \mathcal{P}^*u, w \rangle, \langle \mathcal{B}^*u, w \rangle, \langle \mathcal{J}^*u, w \rangle, \langle Qu, w \rangle, \langle \mathcal{C}u, w \rangle \text{ y } \langle \mathcal{K}u, w \rangle \tag{488}$$

se pueden considerar como operadores funcionales lineales definidos en $\widehat{D}_2(\Omega)$ y valuados en $\widehat{D_1^*}(\Omega)$, i.e., valuados en el espacio dual algebraico de $\widehat{D}_1(\Omega)$, por ejemplo $\mathcal{P}^* : \widehat{D}_2(\Omega) \to \widehat{D_1^*}(\Omega)$ donde $\mathcal{P}w \in \widehat{D_1^*}(\Omega)$ y $w \in \widehat{D}_2(\Omega)$.

Finalmente, sean las funciones $u_\partial \in \widehat{D}_1(\Omega)$ y $u_\Gamma \in \widehat{D}_1(\Omega)$. Entonces las funcionales lineales $g_\partial \in \widehat{D_2^*}(\Omega)$ y $j_\Gamma \in \widehat{D_2^*}(\Omega)$ se define como

$$g_\partial(w) \equiv \langle \mathcal{B}u_\partial, w \rangle \text{ para todo } w \in \widehat{D}_2(\Omega) \tag{489}$$

$$j_\Gamma(w) \equiv \langle \mathcal{J}u_\Gamma, w \rangle \text{ para todo } w \in \widehat{D}_2(\Omega) \tag{490}$$

o brevemente:

$$g_\partial j_\Gamma \mathcal{B}u_\partial \tag{491}$$

$$j_\Gamma \equiv \mathcal{J}u_\Gamma. \tag{492}$$

Formula de Green La formula de Green es un caso particular de las formulas de Green-Herrera cuando $\langle \mathcal{J}u, w \rangle = 0$ y $\langle \mathcal{K}^*u, w \rangle = 0$, lo cual implica la continuidad de la función u y de sus derivadas normales a través de Γ.

Supóngase que las condiciones de salto en Γ son nulas, entonces al aplicar el teorema de la divergencia (466) en el lado derecho de Ec. (465) se obtienen

$$\sum_{i=1}^{E} \int_{\Omega_i} w\mathcal{L}u - \sum_{i=1}^{E} \int_{\Omega_i} u\mathcal{L}^*w d\underline{x} = \int_{\partial\Omega} \underline{\mathfrak{D}}(u,w) \cdot \underline{n} d\underline{x}. \tag{493}$$

Si se compara con la Ec. (467) se observa que $[\underline{\mathfrak{D}}(u,w)] = 0$. Ahora, si se introduce las funcionales bilineales previamente definidas se tiene

$$\sum_{i=1}^{E}\int_{\Omega_i}\mathcal{P}(u,w)d\underline{x} - \sum_{i=1}^{E}\int_{\Omega_i}\mathcal{J}^*(u,w)d\underline{x} = \int_{\partial\Omega}\left(\mathcal{B}(u,w) - \mathcal{C}^*(u,w)\right)d\underline{x} \tag{494}$$

de esta forma se obtiene

$$\langle \mathcal{P}u, w\rangle - \langle \mathcal{B}u, w\rangle = \langle Q^*u, w\rangle - \langle \mathcal{C}^*u, w\rangle \tag{495}$$

y más compactamente

$$\mathcal{P} - \mathcal{B} = Q^* - \mathcal{C}^* \tag{496}$$

para todo $(u,w) \in \widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$. Esta última formula, es la llamada formula de Green convencional.

6.3. Formulaciones Variacionales con Valor en la Frontera con Saltos Prescritos

Definición 57 *Se dice que las condiciones de frontera $g_\partial \in \widehat{D_2^*}(\Omega)$ en $\partial\Omega$ y las condiciones de salto prescrito $j_\Gamma \in \widehat{D_2^*}(\Omega)$ son condiciones compatibles cuando existe una función $u_{\partial\Gamma}\widehat{D}_1(\Omega)$ tal que*

$$\begin{aligned} g_\partial &= \mathcal{B}u_{\partial\Gamma}, \ \textit{en } \partial\Omega \\ j_\Gamma &= \mathcal{J}u_{\partial\Gamma}, \ \textit{en } \Gamma. \end{aligned} \tag{497}$$

Es decir, que ambas condiciones se pueden derivar de una única función, de modo que realmente no imponen condiciones contradictorias.

Definición 58 *El problema de contorno con valores en la frontera con saltos prescritos (BVPJ) consiste en buscar una función $u \in \widehat{D}_1(\Omega)$, tal que satisfaga:*

1) El operador

$$\mathcal{L}u = \mathcal{L}u_\Omega = f_\Omega, \ \textit{en } \Omega \tag{498}$$

2) Condiciones de frontera

$$\mathcal{B}(u,\cdot) = \mathcal{B}(u_\partial,\cdot) = g_\partial, \ \textit{en } \partial\Omega \tag{499}$$

3) Saltos prescritos

$$\mathcal{J}(u,\cdot) = \mathcal{J}(u_\Gamma,\cdot) = j_\Gamma, \ \textit{en } \Gamma \tag{500}$$

donde u_Ω, u_∂, y u_Γ son funciones dadas en $\widehat{D}_1(\Omega)$, que definen los datos del problema, además las condiciones de frontera en $\partial\Omega$ y las condiciones de salto en Γ deben de ser compatibles.

Además se usa la convención de que la Ec. (**??**) que se satisface solamente en los puntos interiores de cada una de las subregiones $\Omega_i, i = 1, ..., E$.

Si definimos las siguientes funcionales f, g y j que pertenecen a $\widehat{D}_2^* (\Omega)$ como

$$\langle f, w \rangle \equiv \langle \mathcal{P} u_\Omega, w \rangle , \text{ para todo } w \in \widehat{D}_2 (\Omega) \tag{501}$$

$$\langle g, w \rangle \equiv \langle \mathcal{B} u_\partial, w \rangle , \text{ para todo } w \in \widehat{D}_2 (\Omega) \tag{502}$$

$$\langle j, w \rangle \equiv \langle \mathcal{J} u_\Gamma, w \rangle , \text{ para todo } w \in \widehat{D}_2 (\Omega) \tag{503}$$

entonces

Definición 59 *Una formulación débil del problema con valores en la frontera con saltos prescritos se puede escribir como*

$$\mathcal{P} u = f; \qquad \mathcal{B} u = g; \qquad \mathcal{J} u = j. \tag{504}$$

Definición 60 *Llamaremos solución de un problema con valores en la frontera con saltos prescritos, a una función $u \in \widehat{D}_1 (\Omega)$ que satisfaga la formulación débil del problema con valores en la frontera con saltos prescritos.*

En lo sucesivo se supondrá que existe almenos una solución del problema con valores en la frontera con saltos prescritos además la Ec. (504) es equivalente a la siguiente ecuación simple:

$$(\mathcal{P} - \mathcal{B} - \mathcal{J}) u = f - g - j. \tag{505}$$

Una condición suficiente para que la Ec. (505) sea equivalente a la Ec. (504) es que $\mathcal{B}$ y $\mathcal{J}$ sean operadores de frontera para $\mathcal{P} : \widehat{D}_1 (\Omega) \to \widehat{D}_2^* (\Omega)$.

Si escribimos la Ec. (505) de manera más explicita resulta

$$\langle (\mathcal{P} - \mathcal{B} - \mathcal{J}) u, w \rangle = \langle f - g - j, w \rangle \qquad \text{para todo } w \in \widehat{D}_2^* (\Omega) \tag{506}$$

la cual representa la formulación variacional del problema con valores en la frontera con saltos prescritos y nos referiremos a ella como la ecuación variacional en términos de los datos del problema.

Haciendo uso de la fórmula Green-Herrera Ec. (486) se obtiene la siguiente formulación variacional equivalente en términos de la información complementaria:

$$\langle (Q - \mathcal{C} - \mathcal{K})^* u, w \rangle = \langle f - g - j, w \rangle \qquad \text{para todo } w \in \widehat{D}_2 (\Omega) \tag{507}$$

cuando se aplica el método de residuos pesados, la solución aproximada $u \in \widehat{D}_1 (\Omega)$ satisface

$$\langle (Q - \mathcal{C} - \mathcal{K})^* \tilde{u}, w^\alpha \rangle = \langle f - g - j, w^\alpha \rangle \qquad \text{con } \alpha = 1, ..., E \tag{508}$$

donde $\left\{ w^1, ..., w^E \right\} \subset \widehat{D}_2 (\Omega)$ es un sistema de funciones de peso.

Como la solución exacta satisface la Ec. (507) entonces se cumple que

$$\langle (Q - \mathcal{C} - \mathcal{K})^* (u - \tilde{u}), w^\alpha \rangle = 0 \qquad \text{con } \alpha = 1, ..., E \tag{509}$$

este resultado puede ser usado para analizar la información acerca de la solución exacta que está contenida en la solución aproximada.

7. Método de Trefftz

En este capítulo se considerarán problemas con valor en la frontera (VBVP) de la forma

$$\begin{aligned} \mathcal{L}u &= f \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{510}$$

donde

$$\mathcal{L}u = -\Delta u \tag{511}$$

como un caso particular del operador elíptico definido por la Ec. (43) de orden dos.Consideremos el problema dado por la Ec. (510) en el dominio $\Omega \subset \mathbb{R}^n$ y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición o descomposición en subdominios del dominio Ω, i.e. se asume que:

1.- Ω_i, para $i = 1, ..., E$ es un subdominio de Ω,

2.- $\Omega_i \bigcap \Omega_j = \varnothing$, siempre que $i \neq j$.

3.- $\Omega \subset \bigcup_{i=1}^{E} \overline{\Omega_i}$.

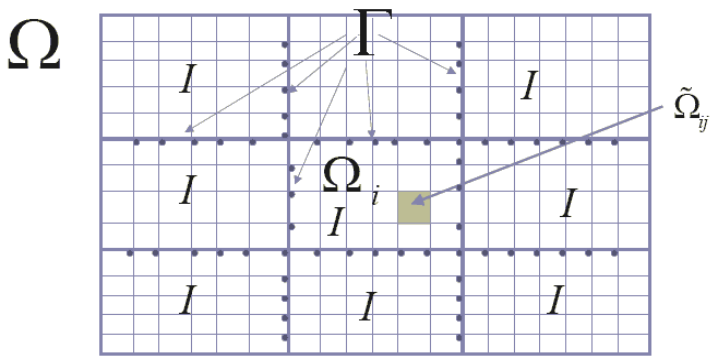

Figura 8: El dominio Ω, su frontera externa $\partial\Omega$ y la frontera interna Γ.

La notación $\partial\Omega$ y $\partial\Omega_i$, $i = 1, ..., E$ es tomada de la frontera del dominio Ω y la frontera del subdominio Ω_i respectivamente, claramente

$$\partial\Omega \subset \bigcup_{i=1}^{E} \partial\Omega_i. \tag{512}$$

Adicionalmente definimos a la frontera interior como

$$\Gamma = \bigcup_{i \neq j} \partial\Omega_i \bigcap \partial\Omega_j \tag{513}$$

y a $\partial\Omega$ como la frontera exterior del dominio Ω.

También asumimos que salvo en un conjunto de medida cero sobre Γ se define un único vector normal denotado por $\underline{n}$ cuyo sentido se elige arbitrariamente y el lado positivo de Γ es definido hacia el sentido positivo del vector normal.

El método Trefftz-Herrera es un método -al igual que en los métodos de Descomposición de Dominio-, permite construir la solución global definida en todo el dominio, resolviendo exclusivamente problemas de contorno locales formulados en cada uno de los subdominios de la partición. La estrategia general para alcanzar dicho propósito consiste en recabar cierta información de la solución pero únicamente en la frontera interior Γ de la partición, que sea la suficiente para definir problemas de contorno independientes y bien planteados en cada uno de los subdominios, cuyas soluciones individuales sean precisamente las restricciones correspondientes de la solución global. Para esto, se elige de antemano cierta información objetivo de la solución en que posea esta propiedad, la cual se denota como información buscada.

Existen dos grandes categorías de métodos para recabar dicha información buscada: los métodos directos y los métodos indirectos. Los métodos directos utilizan soluciones locales del operador diferencial original para establecer las condiciones de compatibilidad que debe aportar la información buscada; mientras que los métodos indirectos utilizan para tal fin el operador diferencial adjunto. Entonces, a partir de estas condiciones de compatibilidad se deriva una matriz del sistema global asociada con el problema.

Los métodos directos se derivan de una forma muy general de las condiciones de continuidad de Poincaré-Steklov, mientras que los métodos indirectos se derivan de una teoría desarrollada por Herrera, la cual se relaciona con la metodología Trefftz, llamada con frecuencia teoría Trefftz-Herrera. Existen diversas maneras de implementar cada uno de estos métodos. Una de ellas, se basa en el uso de una clase especial de funciones de base y de peso, llamadas genéricamente funciones óptimas.

En los métodos de localización indirectos se desarrolla y se aplica un sistema de funciones de peso especializadas que tiene la propiedad de capturar la información buscada de la solución en la frontera interior exclusivamente. La idea de construir tales funciones óptimas de peso surge del hecho de que en el método de residuos pesados, la información acerca de la solución exacta que contiene la solución aproximada, depende del sistema de funciones de peso que se aplica. Para utilizar esta dependencia en la construcción de las funciones óptimas de peso, se requiere de un procedimiento de análisis. Las herramientas básicas de este análisis son las fórmulas de Green-Herrera, las cuales se pueden aplicar aún cuando las funciones de base y de peso son completamente discontinuas, cosa que no se puede hacer en los espacios de Sovolev estándares ni con los operadores definidos en ellos. Entonces, las fórmulas de Green-Herrera se aplican para diseñar las funciones óptimas de peso adecuadamente, las cuales tienen entre otras propiedades las de ser soluciones locales a la ecuación diferencial homogénea asociada con el operador diferencial adjunto. Las funciones óptimas de peso se utilizan para derivar las condiciones de compatibilidad de las cuales

se obtiene la información buscada.

Por otro lado, en los métodos de localización directos se introduce un espacio lineal cuyos elementos, llamados funciones óptimas de base, tiene la siguiente propiedad: una función óptima de base satisface las condiciones de continuidad de Poincaré-Steklov si y solo si contienen la información buscada. Cuando se utiliza la aproximación directa, la obtención de la información buscada se transforma en la construcción de la función óptima de base que cumple con las condiciones de Poincaré-Steklov, de la cual se deriva la matriz del sistema global. Las funciones óptimas de base son soluciones locales de la ecuación diferencial homogénea asociada con el operador diferencial original del problema considerado.

Como se mencionó anteriormente, el término de funciones óptimas denota un conjunto de funciones que incluye tanto a las funciones óptimas de base como a las funciones óptimas de peso, de modo que el primero se compone de la suma directa de los dos últimos.

7.1. Conceptos Básicos

Trabajaremos con los espacios de Sobolev para definir los espacios $\widehat{D}_{\alpha}(\Omega)$, ya que estos son apropiados para manejar funciones discontinuas definidas por tramos, en donde las trazas de las funciones y de sus derivadas normales en las fronteras de los subdominios $\partial\Omega_i$ siempre están definidas y en consecuencia, los saltos y los promedios de éstas. De esta forma, se garantiza la existencia y la unicidad de la solución para los problemas de contorno con saltos prescritos y con coeficientes discontinuos, en donde la solución también puede ser discontinua.

La ubicación dentro del dominio Ω tanto de los saltos prescritos como de las discontinuidades de los coeficientes del operador diferencial determinan la partición Π, de forma que solamente haya éstos sobre la frontera interior Γ. De esta forma, la solución solamente puede presentar discontinuidades en Γ.

Usando las definiciones del capítulo de Funciones Definidas por Tramos, entonces los espacios $\widehat{D}_1(\Omega)$ y $\widehat{D}_2(\Omega)$ corresponden a los espacios de las funciones de base y de las funciones de peso respectivamente. Las funciones de base construyen la solución de la ecuación diferencial; mientras que las funciones de peso o funciones de prueba ponderan a la ecuación diferencial.

En esta sección se trabaja con ecuaciones diferenciales parciales elípticas de segundo orden, en consecuencia, se utilizaran para definir los espacios de las funciones base y funciones de peso los espacios de Sobolev por tramos $\hat{H}^s(\Omega)$ de orden $s = 2$, es decir

$$\widehat{D}_1(\Omega) \subset \hat{H}^2(\Omega) \text{ y } \widehat{D}_2(\Omega) \subset \hat{H}^2(\Omega). \tag{514}$$

Definición 61 *Decimos que un conjunto de funciones $\mathcal{E} \subset \widehat{D}_2(\Omega)$ es un conjunto de funciones TH-Completo para $\mathcal{P} : \widehat{D}_1(\Omega) \to \widehat{D}_2^*(\Omega)$ cuando*

$$\langle \mathcal{P}u, w \rangle = 0, \text{ para todo } w \in \mathcal{E} \tag{515}$$

es decir

$$\mathcal{P}u = 0. \tag{516}$$

Definición 62 *Decimos que un operador* $\mathcal{B} : \widehat{D}_1(\Omega) \to \widehat{D_2^*}(\Omega)$ *es un operador de frontera para* $\mathcal{P} : \widehat{D}_1(\Omega) \to \widehat{D_2^*}(\Omega)$, *cuando* $\mathcal{N}_{\mathcal{B}^*} \subset \widehat{D}_2(\Omega)$ *es un conjunto de funciones TH-Completo para* $\mathcal{P}$, *es decir*

$$\langle \mathcal{P}u, w\rangle = 0, \text{para todo } w \in \mathcal{N}_{\mathcal{B}^*}. \tag{517}$$

Teorema 63 *Sean los operadores* $\mathcal{B}$ *y* $\mathcal{J}$ *operadores de frontera para el operador* $\mathcal{P}$ *y además, sean los espacios* $\mathcal{N}_{\mathcal{P}^*}, \mathcal{N}_{\mathcal{B}^*}$ *y* $\mathcal{N}_{\mathcal{J}^*}$ *espacios TH-Completos para los operadores* $\mathcal{P}$, $\mathcal{B}$ *y* $\mathcal{J}$ *respectivamente: Entonces, las ecuaciones*

$$\mathcal{P}u = f; \qquad \mathcal{B}u = g; \qquad \mathcal{J}u = j \tag{518}$$

son equivalentes a la ecuación

$$\langle (\mathcal{P} - \mathcal{B} - \mathcal{J})u, w\rangle = \langle f - g - j, w\rangle. \tag{519}$$

En virtud de las fórmulas de Green-Herrera, se tienen dos formulaciones débiles. La primera, la ecuación

$$(\mathcal{P} - \mathcal{B} - \mathcal{J})u = f - g - j \tag{520}$$

referida como formulación variacional en términos de los datos del problema y la segunda llamada formulación variacional en términos de la información complementaria

$$(Q - \mathcal{C} - \mathcal{K})^* u = f - g - j \tag{521}$$

de hecho, si la función $u \in \widehat{D}_1(\Omega)$ es la solución del BVPJ se dice que los términos $\mathcal{P}u$, $\mathcal{B}u$ y $\mathcal{J}u$ son los datos del problema, mientras que los términos Q^*u, $\mathcal{C}^*u$ y $\mathcal{K}^*u$ son la información complementaria.

Si se tiene un BVPJ homogéneo con condiciones homogéneas, las siguientes observaciones son útiles.

Corolario 64 *Sean los operadores* $\mathcal{B}$ *y* $\mathcal{J}$ *operadores de frontera para el operador* $\mathcal{P}$ *y además, sean los espacios* $\mathcal{N}_{\mathcal{P}^*}, \mathcal{N}_{\mathcal{B}^*}$ *y* $\mathcal{N}_{\mathcal{J}^*}$ *espacios TH-Completos para los operadores* $\mathcal{P}$, $\mathcal{B}$ *y* $\mathcal{J}$ *respectivamente. Entonces se tiene que*

$$\mathcal{P}u = 0; \qquad \mathcal{B}u = 0; \qquad \mathcal{J}u = 0 \Leftrightarrow \langle (\mathcal{P} - \mathcal{B} - \mathcal{J})u, w\rangle = 0. \tag{522}$$

Teorema 65 *Sean los operadores* $\mathcal{C}$ *y* $\mathcal{K}$ *operadores de frontera para el operador* Q *y además, sean los espacios* $\mathcal{N}_Q, \mathcal{N}_C$ *y* $\mathcal{N}_{\mathcal{K}}$ *espacios TH-Completos para los operadores* $Q, \mathcal{C}$ *y* $\mathcal{K}$ *respectivamente. Entonces se tiene que*

$$Q^*u = 0; \qquad C^*u = 0; \qquad \mathcal{K}^*u = 0 \Leftrightarrow \langle (Q - \mathcal{C} - \mathcal{K})^* u, w\rangle = 0. \tag{523}$$

7.1.1. Condiciones de Poincaré-Steklov

Sea el siguiente problema de Poisson con solución única $u \in C^1(\Omega)$ sujeto a condiciones de frontera homogéneas tipo Dirichlet

$$\begin{aligned} -\Delta u &= f_\Omega \quad \text{en } \Omega \\ u &= 0 \quad \text{en } \partial\Omega \end{aligned} \tag{524}$$

introduciendo la siguiente partición $\Pi = \{\Omega_1, \Omega_2\}$. Entonces, este problema se puede formular de manera equivalente en múltiples subdominios como

$$\begin{aligned} -\Delta u_1 &= f_\Omega \quad \text{en } \Omega_1 \\ u_1 &= 0 \quad \text{en } \partial\Omega \cap \partial\Omega_1 \end{aligned} \tag{525}$$

$$\begin{aligned} -\Delta u_2 &= f_\Omega \quad \text{en } \Omega_2 \\ u_2 &= 0 \quad \text{en } \partial\Omega \cap \partial\Omega_2 \end{aligned} \tag{526}$$

$$\begin{aligned} u_1 &= u_2 \quad \text{en } \Gamma \\ \frac{\partial u_1}{\partial n} &= \frac{\partial u_2}{\partial n} \quad \text{en } \Gamma \end{aligned} \tag{527}$$

donde u_i es la restricción de la solución u en Ω_i y $\underline{n}$ es un vector normal a $\partial\Omega_i$. Las ecuaciones (527) representan las condiciones de transmisión en Γ.

Cuando se aplica algún método de descomposición de dominio a una ecuación diferencial, en general se tiene que resolver un problema de condiciones de transmisión en Γ. En particular, esta ecuación de transmisión se puede representar por medio del operador de Steklov-Poincaré. Para tal efecto, considerese los siguientes problemas tipo Dirichlet

$$\begin{aligned} -\Delta w_i &= f_\Omega \quad \text{en } \Omega_i \\ w_i &= 0 \quad \text{en } \partial\Omega \cap \partial\Omega_i \\ w_i &= \lambda \quad \text{en } \Gamma \end{aligned} \tag{528}$$

para $i = 1, 2$, donde λ es el valor desconocido de u en Γ. La solución w_i se puede escribir como

$$w_i = u_i^H + u_i^P \tag{529}$$

donde u_i y u_p son la solución de los siguientes problemas

$$\begin{aligned} -\Delta u_i^H &= 0 \quad \text{en } \Omega_i \\ u_i^H &= 0 \quad \text{en } \partial\Omega \cap \partial\Omega_i \\ u_i^H &= \lambda \quad \text{en } \Gamma \end{aligned} \tag{530}$$

y

$$\begin{aligned} -\Delta u_i^P &= f_\Omega \quad \text{en } \Omega_i \\ u_i^P &= 0 \quad \text{en } \partial\Omega \cap \partial\Omega_i \\ u_i^P &= 0 \quad \text{en } \Gamma \end{aligned} \tag{531}$$

de aquí que la función u_i^H sea una extensión armónica de λ en Ω_i, la cual se denotará como $\mathcal{H}_i\lambda$; además $\mathcal{H} = \mathcal{H}_1 + \mathcal{H}_2$. Por otro lado, la función u_i^P se denotará como $\mathcal{G}_i\lambda$; además $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2$.

En consecuencia, comparando las ecuaciones del problema (525) a (527) con el problema (528) se tiene que

$$w_i = u_i, i = 1, 2 \text{ si y sólo si } \frac{\partial w_1}{\partial n} = \frac{\partial w_2}{\partial n} \quad \text{en } \Gamma \tag{532}$$

esta última condición equivale a que λ satisfaga la ecuación de transmisión de Steklov-Poincaré

$$\mathcal{S}\lambda = \mathcal{X} \quad \text{en } \Gamma \tag{533}$$

donde $\mathcal{S}$ es el operador de Steklov-Poincaré y se define como

$$\begin{aligned} \mathcal{S}\lambda &\equiv \frac{\partial}{\partial n}\mathcal{H}_1\lambda - \frac{\partial}{\partial n}\mathcal{H}_2\lambda \equiv \left[\left[\frac{\partial}{\partial n}\mathcal{H}\lambda\right]\right] \\ \mathcal{X} &\equiv \frac{\partial}{\partial n}\mathcal{G}_1 f_\Omega - \frac{\partial}{\partial n}\mathcal{G}_2 f_\Omega \equiv \left[\left[\frac{\partial}{\partial n}\mathcal{G} f_\Omega\right]\right] \end{aligned} \tag{534}$$

finalmente, el operador inverso del operador Steklov-Poincaré $\mathcal{S}^{-1}$, se le llama operador de Poincaré-Steklov.

7.2. Método Indirecto de Trefftz-Herrera

Partiendo de la formulación variacional en términos de la información complementaria de un BVPJ

$$\langle (Q - \mathcal{C} - \mathcal{K})^* u, w\rangle = \langle f - g - j, w\rangle \,, \text{ para todo } w \in \widehat{D}_2(\Omega) \tag{535}$$

de la información relacionada con la solución $u \in \widehat{D}_1(\Omega)$ en el interior de los subdominios Ω_i de la partición Π esta dada por el término Q^*u, en la frontera exterior $\partial\Omega$ por el término $\mathcal{C}^*u$ y en la frontera interior Γ por el término $\mathcal{K}^*u$.

El método indirecto de Trefftz-Herrera se carateriza por construir un espacio de funciones de peso especializado para capturar cierta información de la solución en las fronteras de los subdominios, dicho espacio se llama espacio de funciones óptimas de peso $\widehat{O}_T \in \widehat{D}_2(\Omega)$. La información que se busca de la solución en la frontera interior puede ser el promedio de la función, el promedio de sus derivadas o una combinación de éstos. De hecho, esa información buscada determina los espacios nulos relacionados con los operadores $Q, \mathcal{C}$ y $\mathcal{K}$ cuando se aplica el método de residuos pesados. El análisis para construir las funciones óptimas de peso se detalla a continuación.

primero notemos que la información buscada de la solución sólo está en $\Gamma \cup \partial\Omega$ y no en el interior de los subdominios Ω_i. Entonces, para obtener esa información se requiere eliminar el término Q^*u de la Ec. (535) aplicando el método de residuos pesados, la condición anterior se logra construyendo funciones de peso tales que satisfaga la condición $Qw = 0$ en el interior de cada subdominio Ω_i por separado, y de este modo, se introduce el espacio nulo de Q el cual es $N_Q = \left\{ w \in \widehat{D}_2(\Omega) \mid Qw = 0 \text{ en cada } \Omega_i \right\}$.

Definición 66 *Definimos al nucleo de* Q*, como* $w \in \widehat{D}_2(\Omega)$ *tales que* $Qw = 0$ *en cada* Ω_i*, i.e.*

$$N_Q \equiv \left\{ w \in \widehat{D}_2(\Omega) \mid Qw = 0 \text{ en cada } \Omega_i \right\}. \tag{536}$$

En este método se pretende recabar suficiente información de la solución en $\Gamma \cup \partial\Omega$ para proponer problemas de contorno locales bien planteados en cada subdominio Ω_i, i.e., no es necesario recabar toda la información posible de *la* solución en $\Gamma \cup \partial\Omega$. Esto induce la descomposición del operador $\mathcal{K}$ en dos partes. Una parte se refiere a la información buscada de la solución en Γ y otra parte se refiere a la información redundante o no buscada de la solución en Γ.

Definición 67 *Sea la descomposición* $\{S_k, R_k\}$ *de la funcional bilineal* $\mathcal{K}$ *tal que*

$$\mathcal{K} \equiv S_k + R_k \tag{537}$$

donde el operador S_k *se toma de tal forma que* $S_k^* u$ *sea precisamente la información buscada de la solución en* Γ*.*

Con el objetivo de formalizar la descomposición del operador $\mathcal{K}$ se introduce lo siguiente: Sean $S_k^*(u, w)$ y $R_k^*(u, w)$ funcionales bilineales reales definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ tales que producen la siguiente descomposición

$$\mathcal{K}^*(u, w) \equiv S_k^*(u, w) + R_k^*(u, w) \text{ para todo } w \in \Gamma \tag{538}$$

además sean $\langle S_k^* u, w\rangle$ y $\langle R_k^* u, w\rangle$ funcionales bilineales reales definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ tales que

$$\begin{aligned} \langle S_k^* u, w\rangle &\equiv \int_\Gamma S_k^*(u, w)\, d\underline{x} \text{ en } \Gamma \\ \langle R_k^* u, w\rangle &\equiv \int_\Gamma R_k^*(u, w)\, d\underline{x} \text{ en } \Gamma. \end{aligned} \tag{539}$$

El operador R_k se asocia con la información redundante o no buscada de la solución en Γ. Así, una vez considerada la descomposición del operador $\mathcal{K}$, cuando se requiere eliminar el término $R_k^* u$ de la Ec. (535). Aplicando el método de residuos pesados, la condición anterior se logra construyendo funciones de peso tales que satisfagan la condición $R_k^* u = 0$ en Γ y de este modo, se introduce el espacio nulo de R_k en cual es $N_{R_k} = \left\{ w \in \widehat{D}_2(\Omega) \mid R_k w = 0 \text{ en } \Gamma \right\}$.

Definición 68 *Definimos al nucleo de* R_k *como* $w \in \widehat{D}_2(\Omega)$ *tales que* $R_k w = 0$ *en* Γ*,i.e*

$$N_{R_k} \equiv \left\{ w \in \widehat{D}_2(\Omega) \mid R_k w = 0 \text{ en } \Gamma \right\}. \tag{540}$$

Definición 69 *Sea un BVPJ con solución única* $u \in \widehat{D}_1(\Omega)$, *sean las funcionales bilineales* S_k *y* R_k *tales que* $\mathcal{K} = S_k + R_k$, *además sea* $\hat{u} \in \widehat{D}_1(\Omega)$ *tal que*

$$S_k^* \hat{u} = S_k^* u \text{ en } \Gamma \tag{541}$$

entonces decimos que $\hat{u}$ *contiene la información buscada de la solución en* Γ.

Por otro lado, notemos que para una ecuación diferencial elíptica de segundo orden se necesita condiciones de frontera en todo $\partial\Omega$ para que el problema de contorno esté bien planteado. Lo anterior significa que, en general, no se busca información en $\partial\Omega$. Entonces se requiere eliminar el término $\mathcal{C}_k^*$ de (535). Aplicando el método de residuos pesados, la condición anterior se logra construyendo funciones de peso tales que satisfagan la condición $\mathcal{C}w = 0$ en $\partial\Omega$ y de este modo, se introduce el espacio nulo de $\mathcal{C}$ el cual es $N_{\mathcal{C}} = \left\{ w \in \widehat{D}_2(\Omega) \mid \mathcal{C}w = 0 \text{ en } \partial\Omega \right\}$.

Definición 70 *Definimos el nulo de* $\mathcal{C}$ *como* $w \in \widehat{D}_2(\Omega)$ *tales que* $\mathcal{C}w = 0$ *en* $\partial\Omega$, *i.e.*

$$N_{\mathcal{C}} \equiv \left\{ w \in \widehat{D}_2(\Omega) \mid \mathcal{C}w = 0 \text{ en } \partial\Omega \right\}. \tag{542}$$

Definición 71 *Definimos al espacio de funciones óptimas de peso* $\widehat{O}_T$ *como*

$$\widehat{O}_T \equiv N_Q \cap N_{\mathcal{C}} \cap N_{R_k} \subset \widehat{D}_2(\Omega) \tag{543}$$

i.e.

$$\langle (Q - \mathcal{C} - R_k)^* u, w \rangle = 0, \text{ para todo } w \in N_Q \cap N_{\mathcal{C}} \cap N_{R_k}. \tag{544}$$

Teorema 72 *Método indirecto de Trefftz-Herrera*

Sea el BVPJ $(\mathcal{P} - \mathcal{B} - \mathcal{J})u = f - g - j$, *con los datos* $f = \mathcal{P}u_\Omega, g = \mathcal{B}u_\partial$ *y* $j = \mathcal{J}u_\Gamma$ *donde* $u_\Omega \in \widehat{D}_1(\Omega), u_\partial \in \widehat{D}_1(\Omega)$ *y* $u_\Gamma \in \widehat{D}_1(\Omega)$. *Además sea* $\widehat{O}_T$ *el espacio de funciones óptimas de peso TH-Completo para* $S_k^* : \widehat{D}_1(\Omega) \times \widehat{D_2^*}(\Omega)$ *y supongase que el BVPJ tiene una única solución* $u \in \widehat{D}_1(\Omega)$. *Entonces* $\hat{u} \in \widehat{D}_1(\Omega)$ *contiene la información buscada, i.e.* $S_k^* \hat{u} = S_k^* u$, *si y sólo si*

$$-\langle S_k^* \hat{u}, w \rangle = \langle f - g - j, w \rangle, \text{ para toda } w \in \widehat{O}_T. \tag{545}$$

Por último, puesto que el operador Q se relaciona con el operador diferencial adjunto y el espacio de funciones base se toma igual que el espacio de funciones óptimas de peso, el método Trefftz-Herrera solamente es capaz de obtener información de la solución en la fortera interior Γ y no es capaz de obtener información de la solución en el interior de los subdominios Ω_i. Para encontrar la solución en el interior de los subdonios Ω_i se necesita un procedimiento llamado interpolación óptima.

El procedimiento llamado interpolación óptiman consiste en extender la información buscada de la solución en Γ al interior de cada ubdominio Ω_i de la siguiente manera. Una vez encontrada la información buscada de la solución en Γ, junto con las condiciones de frontera $\mathcal{B}u = \mathcal{B}u_\partial$ en $\partial\Omega$ y las condiciones de

salto $\mathcal{J}u = \mathcal{J}u_\Gamma$ en Γ, se obtienen problemas de contorno locales bien planteados e independientes, en cada Ω_i. La solución de estos problemas de contorno locales extiende la información buscada de la solución en Γ al interior de los subdominios Ω_i.

7.3. Método directo de Steklov-Poincaré

Partiendo de la formulación variacional en términos de los datos de un BVPJ

$$\langle(\mathcal{P}-\mathcal{B}-\mathcal{J})u, w\rangle = \langle f-g-j, w\rangle\,, \text{ para todo } w \in \widehat{D}_2(\Omega) \tag{546}$$

donde la información relacionada con la solución $u \in \widehat{D}_1(\Omega)$ en el interior de los subdominios Ω_i de la partición Π está dada por el término $\mathcal{P}u$, en la frontera exterior $\partial\Omega$ por el término $\mathcal{B}u$ y en la frontera interior Γ por el término $\mathcal{J}u$.

El método directo de Steklov-Poincaré se caracteriza por construir un espacio de funciones de base especializada paracontener cierta información de la solución en las fronteras de los subdominios, dicho espacio se llamara funciones óptimas de base, donde la información que se busca de la solución en la frontera interior es el promedio de la función. Una propiedad relevante es que las funciones óptimas de base contienen la información buscada si y sólo si satisfacen las condiciones de continuidad de Poincaré-Steklov en la frontera interior Γ. De hecho, la condición anterior determina los espacios nulos relacionados con los operadores $\mathcal{P}, \mathcal{B}$ y $\mathcal{J}$. El ánalisis para construir las funciones óptimas de base se detalla a continuación.

Primero, se requiere que las funciones óptimas de base satisfagan las condiciones de continuidad de Poincaré-Steklov en la frontera interior Γ. Esto induce la descomposición del operador $\mathcal{J}$ en dos partes. Una parte se refiere precisamente a dichas condiciones de continuidad en Γ, mientras que la otra parte se refiere a condiciones de continuidad redundantes en Γ.

Definición 73 *Sea la descomposición $\{S_j, R_j\}$ de la funcional bilineal $\mathcal{J}$ tal que*

$$\mathcal{J} \equiv S_j + R_j \tag{547}$$

donde el operador S_j se toma de tal forma que $S_j v$ sea precisamente las condiciones de continuidad de Poincaré-Steklov en Γ.

Con el objetivo de formalizar la descomposición del operador $\mathcal{J}$ se introduce lo siguiente. Sean $S_j(u,w)$ y $R_j(u,w)$ funciones bilineales reales definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ tales que producen la siguiente descomposición

$$\mathcal{J} \equiv S_j(u,w) + R_j(u,w) \text{ para toda } \underline{x} \in \Gamma, \tag{548}$$

luego, sean $\langle S_j u, w\rangle$ y $\langle R_j u, w\rangle$ funcionales bilineales reales definidas en $\widehat{D}_1(\Omega) \times \widehat{D}_2(\Omega)$ tales que

$$\begin{aligned} \langle S_j u, w\rangle &= \int_\Gamma S_j(u,w)\, d\underline{x} \text{ en } \Gamma \\ \langle R_j u, w\rangle &= \int_\Gamma R_j(u,w)\, d\underline{x} \text{ en } \Gamma \end{aligned} \tag{549}$$

El operador R_j se asocia con las condiciones de continuidad redundantes en Γ. También la descomposición del operador $\mathcal{J}$ induce una descomposición correspondiente pero en los datos del problema, especificamente en $j = \mathcal{J}u_\Gamma$, de modo que

$$\begin{aligned} j_s + j_r &= j = \mathcal{J}u_\Gamma \\ j_s &= S_j u_\Gamma \\ j_R &= R_j u_\Gamma. \end{aligned} \tag{550}$$

La estrategia general del método directo de Steklov-Poincaré consiste en lo siguiente. La solución $u \in \widehat{D}_1(\Omega)$ se conforma de la suma de dos funciones. Una función óptima de base $v \in \widehat{O}_B$ la cual cumple con las condiciones de continuidad de Poincaré-Steklov en Γ y que contiene la información buscada de la solución en Γ;y una función auxilar $u_p \in \widehat{D}_1(\Omega)$ la cual cumple las condiciones de continuidad redundantes en Γ y que no contiene en absoluto la información buscada. Además, la función auxiliar u_p se construye con las soluciones particulares locales de cada uno de los subdominios de la partición, satisfaciendo las condiciones de frontera y las condiciones de saltos preescritos asociadas con las condiciones redundantes de continuidad en Γ. Ciertamente la función auxiliar u_p no es una funciòn óptima de base.

De este modo, la solución $u \in \widehat{D}_1(\Omega)$ se puede escribir como $u = v + u_p$, donde $u_p \in \widehat{D}_1(\Omega)$ y $v \in \widehat{O}_B$, lo cual implica que la Ec. (546) se transforma en

$$\langle (\mathcal{P} - \mathcal{B} - \mathcal{J})v, w\rangle = \langle f - g - j, w\rangle - \langle (\mathcal{P} - \mathcal{B} - \mathcal{J})u_p, w\rangle \tag{551}$$

para toda $w \in \widehat{D}_2(\Omega)$. En cierto modo se puede decir que la función auxiliar $u_p$es una solución particular, construida resolviendo problemas de contorno locales, mientras que la función óptima de base v es una solución homogénea, construidad resolviendo un problema de contorno global y cuya utilidad es la de acoplar las mensionadas soluciones locales al contrarrestar sus discontinuidades que presentan en Γ introducidas por la forma en que se construyen.

Definición 74 *Sea la función auxiliar $u_p \in \widehat{D}_1(\Omega)$ tal que*

$$\langle (\mathcal{P} - \mathcal{B} - R_j)u_p, w\rangle = \langle f - g - j_R, w\rangle \text{ para toda } w \in \widehat{D}_2(\Omega) \tag{552}$$

y

$$S_k^* u_p = 0 \tag{553}$$

donde el operador S_k se utiliza para establecer la información buscada y se definio en la Ec.(537).

Nótese que si $S_k^* u_p = 0$ entonces $S_k^* u = S_k^*(v + u_p) = S_k^* v$, lo cual significa que efectivamente la función óptima de base contiene completamente la información buscada de la solución en Γ.

Para el caso de una ecuación diferencial elíptica de segundo orden la solución particular $u_p \in \widehat{D}_1(\Omega)$ se construye como una función que satisface por

separado los problemas de contorno no homogéneos locales en cada subdominio Ω_i. Esta función también satisface las condiciones de frontera en $\partial\Omega$ y las condiciones de saltos prescritos pero únicamente de la función en Γ, imponiéndose que su promedio sea cero en Γ. Sin embargo, puesto que dicha función se obtiene a partir de problemas locales independientes, aunque sí es posible construirla satisfaciendo condiciones de continuidad de la función en las fronteras de los subdominios, no lo es satisfaciendo simultaneamente condiciones de continuidad en sus derivadas normales. De aquí que la función óptima de base $v \in \widehat{O}_B$, con el objetivo de acoplar soluciones locales particulares, debe contribuir a satisfacer estas últimas condiciones de continuidad. Lo anterior se logra gracias a que la función óptima de base contrarresta el salto en las derivadas normales en Γ que representa la función auxiliar u_p. Esto último constituye precisamente las condiciones de continuidad de Poincaré-Steklov.

Además, para el caso elíptico, la información buscada de la solución en Γ es el promedio de la función. Nótese que la función auxiliar u_p satisface las condiciones de salto de la función en Γ y su promedio es cero en Γ. En consecuencia, la función óptima de base v es continua en Γ y es precisamente el promedio de la solución en Γ, i.e.

$$\dot{u} \mid_{\Gamma} = (\dot{u}_p + \dot{v}) \mid_{\Gamma} = \dot{v} \mid_{\Gamma} = v \mid_{\Gamma} . \tag{554}$$

A continuación, se plantea el espacio de funciones óptimas de base. Puesto que la función auxiliar u_p ya satisface la ecuación diferencial homogénea, entonces las funciones óptimas de base se construyen de tal modo que satisfagan la condición $\mathcal{P}v = 0$ en el interior de cada subdominio Ω_i por separado y de este modo, se introduce el espacio nulo de $\mathcal{P}$ el cual es $N_{\mathcal{P}} = \left\{ v \in \widehat{D}_1(\Omega) \mid \mathcal{P}v = 0 \text{ en cada } \Omega_i \right\}$ y puesto que la función auxiliar $u_p$ya satisfizo las condiciones de frontera, entonces las funciones óptimas de base se construyen de tal modo que satisfagan la condición $\mathcal{B}v = 0$en $\partial\Omega$ y de este modo se introduce el espacio nulo de $\mathcal{B}$ el cual es $N_{\mathcal{B}} = \left\{ v \in \widehat{D}_1(\Omega) \mid \mathcal{B}v = 0 \text{ en cada } \partial\Omega \right\}$ .Por último, la descomposición del operador $\mathcal{J}$ introduce su espacio nulo de R_j, el cual es $N_{R_j} = \left\{ v \in \widehat{D}_1(\Omega) \mid R_j v = 0 \text{ en cada } \Gamma \right\}$.

Definición 75 *Definimos a los espacios nulos de $\mathcal{P}$, $\mathcal{B}$ y R_j como*

$$\begin{aligned} N_{\mathcal{P}} &= \left\{ v \in \widehat{D}_1(\Omega) \mid \mathcal{P}v = 0 \textit{ en cada } \Omega_i \right\} \\ N_{\mathcal{P}} &= \left\{ v \in \widehat{D}_1(\Omega) \mid \mathcal{B}v = 0 \textit{ en cada } \partial\Omega \right\} \\ N_{R_j} &= \left\{ v \in \widehat{D}_1(\Omega) \mid R_j v = 0 \textit{ en cada } \Gamma \right\} \end{aligned} \tag{555}$$

respectivamente.

Definición 76 *Sea el espacio de las funciones óptimas de base $\widehat{O}_B$ definido como*

$$\widehat{O}_B \equiv N_{\mathcal{P}} \cap N_{\mathcal{P}} \cap N_{R_j} \subset \widehat{D}_1(\Omega) \tag{556}$$

lo cual implica que

$$\langle(\mathcal{P}-\mathcal{B}-R_j)u_p, w\rangle = 0 \text{ para toda } v \in N_{\mathcal{P}} \cap N_{\mathcal{P}} \cap N_{R_j}. \tag{557}$$

Así, el espacio de funciones de peso $\widehat{D}_2(\Omega)$ lo tomamos igual que el espacio de funciones de base $\widehat{D}_1(\Omega)$ y denotaremos a $\hat{v} \in \widehat{O}_B$ como una función óptima de base que contiene la información buscada de la solución $u \in \widehat{D}_1(\Omega)$, i.e. $S_k^* \hat{v} = S_k^* u$.

Como conclusión, el siguiente teorema establece las condiciones que se debe cumplir para que la función de base $\hat{v}$ contenga la información buscada de la solución u en Γ.

Teorema 77 *Sea el BVPJ* $(\mathcal{P}-\mathcal{B}-\mathcal{J})u = f - g - j$ *con los datos* $f = \mathcal{P}u_\Omega, g = \mathcal{B}u_\partial$ *y* $j = \mathcal{J}u_\Gamma$ *donde* $u_\Omega \in \widehat{D}_1(\Omega), u_\partial \in \widehat{D}_1(\Omega)$ *y* $u_\Gamma \in \widehat{D}_1(\Omega)$. *Además sea* $\widehat{O}_B$ *el espacio de funciones óptimas de base TH-Completo para* S_j : $\widehat{D}_1(\Omega) \times \widehat{D_2^*}(\Omega)$ *y supongase que el BVPJ tiene una única solución* $u \in \widehat{D}_1(\Omega)$ *tal que* $u = \hat{v} + u_p$, *donde* $\hat{v} \in \widehat{O}_B$ *y* $u_p \in \widehat{D}_1(\Omega)$ *cumple con la Ec.(551) y Ec.(552). Entonces* $\hat{v} \in \widehat{O}_B$ *contiene la información buscada, i.e.* $S_k^* \hat{v} = S_k^* u$, *si y sólo si*

$$-\langle S_j \hat{v}, w\rangle = \langle S_j u_p, w\rangle - \langle J_s, w\rangle, \text{ para toda } w \in \widehat{O}_B. \tag{558}$$

Finalmente, a diferencia del método indirecto de Trefftz-Herrera, el método directo de Steklov-Poincaré no requiere del procedimiento de interpolación óptima para extender la información buscada de la solución en Γ al interior de los subdominios Ω_i de la partición Π, ya que el operador $\mathcal{P}$ se relaciona con el operador diferencial original.

8. 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.

8.1. Nociones de Algebra Lineal

A continuación detallaremos algunos resultados de álgebra lineal, las demostraciones de los siguientes resultados puede ser consultada en [20].

Definición 78 *Sea V un espacio vectorial y sea $f(\cdot) : V \to \mathbb{R}$, f es llamada funcional lineal si satisface la condición*

$$f(\alpha v + \beta w) = \alpha f(v) + \beta f(w) \quad \forall v, w \in V \quad y \quad \alpha, \beta \in \mathbb{R}. \tag{559}$$

Definición 79 *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 80 *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 81 *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*

$$N(D) = \{f \in V^* \mid f(v) = 0 \quad \forall v \in D\}. \tag{560}$$

Teorema 82 *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 83 *Si $V = L \oplus M$ (suma directa) entonces $V^* = N(L) \oplus N(M)$.*

Teorema 84 *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*

$$T^*(w^*)(u) = w^*(Tu). \tag{561}$$

Teorema 85 *Si H es un espacio completo con producto interior, entonces $H^* = H$.*

Definición 86 *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*

$$\langle Tu, v\rangle = \langle u, T^* v\rangle . \tag{562}$$

Definición 87 *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*

$$\begin{aligned} \tau(\alpha x + \beta y, z) &= \alpha\tau(x, z) + \beta\tau(y, z) \\ \tau(x, \alpha y + \beta z) &= \alpha\tau(x, y) + \beta\tau(x, z) . \end{aligned} \tag{563}$$

Definición 88 *Si $\tau(\cdot,\cdot)$ es una forma bilineal sobre V, entonces la función $q_\tau(\cdot): V \to \mathbb{R}$ definida por*

$$q_\tau(x) = \tau(x, x) \quad \forall x \in V \tag{564}$$

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 89 *Sea $V \subset \mathbb{R}^n$ un subespacio, $P \in \mathbb{R}^n \times \mathbb{R}^n$*

8.2. 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 [22] y [3].

Definición 90 *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 91 *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 92 *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*

$$\mu\left(\bigcup_{n=1}^{\infty}\psi_n\right)=\sum_{n=1}^{\infty}\mu\left(\psi_n\right). \tag{565}$$

Teorema 93 *Existe una función de medida* μ *sobre el conjunto de Borel de* $\mathbb{R}$ *llamada la medida de Lebesge que satisface* $\mu\left([a,b]\right)=b-a$.

Definición 94 *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 95 *Sea* $E\subset\Omega$ *un conjunto, se dice que el conjunto* E *tiene medida cero si* $\mu\left(E\right)=0$.

Teorema 96 *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 97 *(Fubini)*
Si $f\left(x,y\right)$ *es medible en* $X\times Y$ *entonces*

$$\int\limits_{X\times Y}f\left(x,y\right)d\mu=\int\limits_{X}\int\limits_{Y}f\left(x,y\right)d\beta d\alpha=\int\limits_{Y}\int\limits_{X}f\left(x,y\right)d\alpha d\beta \tag{566}$$

en el sentido de que cualquiera de las integrales existe y son iguales.

Teorema 98 *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 99 *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 100 *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.

8.3. Espacios L^p

Las definiciones y material adicional puede ser consultada en [12], [18] y [3].

Definición 101 *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*

$$\int_{\Omega}\left|f\right|d\underline{x}<\infty. \tag{567}$$

Definición 102 *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*

$$\int_{\Omega} |u(\underline{x})|^p \, d\underline{x} \tag{568}$$

es integrable.

Al espacio $L^2(\Omega)$ se le llama cuadrado integrable.

Definición 103 *La norma $L^2(\Omega)$ se define como*

$$\|u\|_{L^2(\Omega)} = \left(\int_{\Omega} |u(\underline{x})|^2 \, d\underline{x} \right)^{\frac{1}{2}} < \infty \tag{569}$$

y el producto interior en la norma $L^2(\Omega)$ como

$$\langle u, v \rangle_{L^2(\Omega)} = \int_{\Omega} u(\underline{x}) v(\underline{x}) d\underline{x}. \tag{570}$$

Definición 104 *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,*

$$L^\infty(\Omega) = \{u \mid |u(x)| \leq k\} \tag{571}$$

definida en casi todo Ω, para algún $k \in \mathbb{R}$.

8.4. 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 [12], [18] y [3]

Definición 105 *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 106 *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 107 *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*

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} ... \partial x_n^{\alpha_n}} \tag{572}$$

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

Definición 108 *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 109 *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 110 *Una distribución sobre un dominio $\Omega \subset \mathbb{R}^n$ es toda funcional lineal continua sobre $\mathcal{D}(\Omega)$.*

Definición 111 *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 112 *Un función $f(\cdot)$ es llamada localmente integrable, si para todo subconjunto compacto $K \subset \Omega$ se tiene*

$$\int_K |f(x)|\, dx < \infty. \tag{573}$$

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

$$\langle F, \phi \rangle = \int_\Omega f\phi dx \tag{574}$$

con $\phi \in \mathcal{D}(\Omega)$.

Si el soporte de ϕ es $K \subset \Omega$, entonces

$$|\langle F, \phi \rangle| = \left| \int_\Omega f\phi dx \right| = \left| \int_K f\phi dx \right| \leq \sup_{x \in K} |\phi| \int_\Omega |f(x)|\, dx \tag{575}$$

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 113 *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

$$\langle (uf), \phi \rangle = \langle f, u\phi \rangle \tag{576}$$

para toda $\phi \in \mathcal{D}(\Omega)$. Notemos que la anterior ecuación es una generalización de la identidad

$$\int_{\Omega} [u(x) f(x)] \phi(x) dx = \int_{\Omega} f(x) [u(x) \phi(x)] dx \tag{577}$$

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 114 *La versión clásica del teorema de Green es dada por la identidad*

$$\int_{\Omega} u \frac{\partial v}{\partial x_i} d\underline{x} = \int_{\partial\Omega} uvn_i d\underline{s} - \int_{\Omega} v \frac{\partial u}{\partial x_i} d\underline{x} \tag{578}$$

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. (578) en una dimensión se obtiene usando la formula de integración por partes, quedando como

$$\int_a^b uv'dx = [uv]\,|_a^b - \int_a^b vu'dx, \quad u, v \in C^1[a, b] \tag{579}$$

como un caso particular de la Ec. (578).

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. (578) y con $|\alpha| = m$, entonces se puede mostrar que:

Teorema 115 *Otra versión del teorema de Green es dado por*

$$\int_{\Omega} (D^\alpha u) v d\underline{x} = (-1)^{|\alpha|} \int_{\Omega} u D^\alpha v d\underline{x} + \int_{\partial\Omega} h(u, v) d\underline{s} \tag{580}$$

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. (580) por ϕ perteneciente a $\mathcal{D}(\Omega)$ y como $\phi = 0$ en la frontera $\partial\Omega$ tenemos

$$\int_{\Omega} (D^\alpha u) \phi d\underline{x} = (-1)^{|\alpha|} \int_{\Omega} u D^\alpha \phi d\underline{x} \tag{581}$$

ya que u es m-veces continuamente diferenciable, esta genera una distribución denotada por u, tal que

$$\langle u, \phi \rangle = \int_{\Omega} u \phi d\underline{x} \tag{582}$$

o, como $D^{\alpha}\phi$ también pertenece a $\mathcal{D}(\Omega)$, entonces

$$\langle u, D^{\alpha}\phi\rangle = \int_{\Omega} uD^{\alpha}\phi d\underline{x} \tag{583}$$

además, $D^{\alpha}u$ es continua, así que es posible generar una distribución regular denotada por $D^{\alpha}u$ satisfaciendo

$$\langle D^{\alpha}u, \phi\rangle = \int_{\Omega} \left(D^{\alpha}u\right)\phi d\underline{x} \tag{584}$$

entonces la Ec. (581) puede reescribirse como

$$\langle D^{\alpha}u, \phi\rangle = (-1)^{|\alpha|}\langle u, D^{\alpha}\phi\rangle, \quad \forall\phi \in D(\Omega). \tag{585}$$

Definición 116 *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

$$\langle u, \phi\rangle = \int_{\Omega} u\phi dx \tag{586}$$

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

$$\langle D^{\alpha}u, \phi\rangle = (-1)^{|\alpha|}\langle u, D^{\alpha}\phi\rangle, \quad \forall\phi \in D(\Omega). \tag{587}$$

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

$$\langle D^{\alpha}u, \phi\rangle = \int_{\Omega} D^{\alpha}u(x)\,\phi(x)\,d\underline{x} \tag{588}$$

y se sigue que la función u y $D^{\alpha}u$ están relacionadas por

$$\int_{\Omega} D^{\alpha}u(x)\,\phi(x)\,d\underline{x} = (-1)^{|m|}\int_{\Omega} u(x)\,D^{\alpha}\phi(x)\,d\underline{x} \tag{589}$$

para $|\alpha| \leq m$.

Definición 117 *Llamamos a la función (o más precisamente, a la equivalencia de clases de funciones) $D^{\alpha}u$ obtenida en la Ec. (589), 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 $|\alpha| \leq m$.

9. Apéndice B

9.1. 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

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{590}$$

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.

9.1.1. 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

$$\sum_{j=1}^{N} a_{ij}^{(0)} x_j = a_{i,n+1}^{(0)}, \quad i = 1, 2, ..., N \tag{591}$$

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)

$$x_i + \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j = a_{i,n+1}^{(i)}, \quad i = 1, 2, ..., N \tag{592}$$

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

$$\begin{aligned} x_N &= a_{N,N+1}^{(N)}; \\ i &= N-1, N-2, ..., 1 \\ x_i &= a_{i,N+1}^{(i)} - \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j. \end{aligned} \tag{593}$$

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

$$\underline{\underline{L}}\underline{y} = \underline{b} \quad \text{y} \quad \underline{\underline{U}}\underline{u} = \underline{y}. \tag{594}$$

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.

9.1.2. 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 [15] y [24].

Un método iterativo en el cual se resuelve el sistema lineal

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{595}$$

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

$$\underline{u}^k = \underline{\underline{T}}\underline{u}^{k-1} + \underline{c} \quad \forall k = 1, 2, 3, ... \tag{596}$$

La convergencia a la solución la garantiza el siguiente teorema cuya solución puede verse en [25].

Teorema 118 *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

$$cond(\underline{\underline{A}}) = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} \geq 1 \tag{597}$$

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 (595) 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

$$\underline{u} = \underline{\underline{B}}\underline{u} + \underline{d} \tag{598}$$

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

$$x_i = \sum_{j=1}^{n} b_{ij} x_j^{(k-1)} + d_i \tag{599}$$

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. (595) se puede reescribir como

$$\left(\underline{\underline{I}} - \underline{\underline{B}}\right) \underline{u} = \underline{b}. \tag{600}$$

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

$$\underline{u}^{m+1} = \underline{\underline{B}}\underline{u}^m + \underline{k}. \tag{601}$$

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. (595) por $D^{-1} = \left[diag(\underline{\underline{A}})\right]^{-1}$ obteniendo

$$\underline{\underline{B}} = \underline{\underline{I}} - \underline{\underline{D}}^{-1}\underline{\underline{A}}, \quad \underline{k} = \underline{\underline{D}}^{-1}\underline{b}. \tag{602}$$

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. (601) puede ser escrita como

$$u_i^{m+1} = \sum_{j<i} b_{ij} u_j^{m+1} + \sum_{j>i} b_{ij} u_j^m + k_i. \tag{603}$$

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

$$\underline{u}^{m+1} = \underline{\underline{E}}\underline{u}^{m+1} + \underline{\underline{F}}\underline{u}^m + \underline{k} \tag{604}$$

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

$$\underline{u}^{m+1} - \underline{u}^{m} = \underline{b} - \underline{\underline{A}}\underline{u}^{m} \tag{605}$$

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

$$\underline{u}^{m+1} = \underline{u}^{m} + \omega\left(\underline{b} - \underline{\underline{A}}\underline{u}^{m}\right). \tag{606}$$

El método de Richardson se define como

$$\underline{u}^{m+1} = \left(\underline{\underline{I}} - \omega\underline{\underline{A}}\right)\underline{u}^{m} + \omega\underline{b} \tag{607}$$

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

$$u_i^{m+1} = (1-\omega)\,u_i^{m} + \omega\left[\sum_{j=1}^{i-1} b_{ij}u_j^{m+1} + \sum_{j=i+1}^{N} b_{ij}u_j^{m} + k_i\right] \tag{608}$$

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.

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 119 *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*

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

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

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{609}$$

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:

$$\begin{aligned} \beta^{k+1} &= \frac{\underline{\underline{A}}\underline{p}^k \cdot \underline{r}^k}{\underline{\underline{A}}\underline{p}^k \cdot \underline{p}^k} \\ \underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1}\underline{p}^k \\ \alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{\underline{A}}\underline{p}^{k+1} \cdot \underline{p}^{k+1}} \end{aligned} \tag{610}$$

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

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 $\|\underline{u}\|^2_{\underline{\underline{A}}} = \underline{u} \cdot \underline{\underline{A}}\underline{u}$ entonces

$$\left\|\underline{u} - \underline{u}^k\right\|_{\underline{\underline{A}}} \leq \left\|\underline{u} - \underline{u}^0\right\|_{\underline{\underline{A}}} \left[\frac{1 - \sqrt{Cond(\underline{\underline{A}})}}{1 + \sqrt{Cond(\underline{\underline{A}})}}\right]^{2k}. \tag{611}$$

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 120 *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*

$$\frac{\|e^n\|}{\|e^0\|} \leq \frac{2}{\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^n + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-n}\right]} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \tag{612}$$

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

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \simeq 1 - \frac{2}{\sqrt{\kappa}} \tag{613}$$

tal que

$$\|\underline{e}^n\|_{\underline{\underline{A}}} \simeq \left\|\underline{e}^0\right\|_{\underline{\underline{A}}} \exp\left(-2\frac{n}{\sqrt{\kappa}}\right) \tag{614}$$

de lo anterior podemos esperar un espectro de convergencia del orden de $O(\sqrt{\kappa})$ iteraciones, para mayor referencia ver [25].

Definición 121 *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.*

9.2. 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 [17].

¿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

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1}\underline{b} \tag{615}$$

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

$$Cond(\underline{\underline{C}}^{-1}\underline{\underline{A}}) < Cond(\underline{\underline{A}}), \tag{616}$$

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

$$\underline{\underline{C}}^{-1} \simeq \underline{\underline{A}}^{-1} \tag{617}$$

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.

9.2.1. Gradiente Conjugado Precondicionado

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

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{618}$$

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

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{C}}^{-1}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{619}$$

entonces la Ec. (609) se pude escribir como

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1} b \tag{620}$$

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

$$\begin{aligned} \left\langle \underline{u}, \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right\rangle &= \underline{u} \cdot \underline{\underline{C}} \left(\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right) \\ &= \underline{u} \cdot \underline{\underline{A}}\underline{v} \end{aligned} \tag{621}$$

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

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

$$\langle \underline{u}, \underline{v} \rangle = \underline{u} \cdot \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \tag{622}$$

por ello las Ecs. (610[1]) y (610[3]), se convierten en

$$\alpha^{k+1} = \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \tag{623}$$

y

$$\beta^{k+1} = \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \tag{624}$$

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:

$$\begin{aligned} \beta^{k+1} &= \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \\ \underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1}\underline{p}^k \\ \alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \\ \underline{u}^{k+1} &= \underline{u}^k + \alpha^{k+1}\underline{p}^{k+1} \\ \underline{r}^{k+1} &= \underline{\underline{C}}^{-1}\underline{r}^k - \alpha^{k+1}\underline{\underline{A}}\underline{p}^{k+1}. \end{aligned} \tag{625}$$

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 122 *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.

9.2.2. 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 $\underline{\underline{A}}$

$$C_{ij} = \begin{cases} A_{ij} \quad si \quad i = j \\ \\ 0 \quad si \quad i \neq j. \end{cases} \tag{626}$$

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

$$\underline{\underline{A}} = \underline{\underline{D}} + \underline{\underline{L}} + \underline{\underline{L}}^T \tag{627}$$

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

$$\underline{\underline{C}}(\omega) = \frac{1}{2-w}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)\left(\frac{1}{\omega}\underline{\underline{D}}\right)^{-1}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)^T \tag{628}$$

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 $i, j > k$:

$$S_{ij} = \begin{cases} A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} & \text{Si} \quad (i,j) \in S \\ \\ A_{ij} & \text{Si} \quad (i,j) \notin S. \end{cases} \tag{629}$$

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

$$A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} \neq 0 \tag{630}$$

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

$$\underline{\underline{C}}_0 = \arg mín_{\underline{\underline{C}} \in \mathcal{S}} \left\| \underline{\underline{AC}} - \underline{\underline{I}} \right\|. \tag{631}$$

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

$$\left\| \underline{\underline{AC}}_0 - \underline{\underline{I}} \right\| = \varepsilon < 1. \tag{632}$$

Existen dos razones para esto, primero, la ecuación (632) 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

$$r_k = \underline{\underline{AC}}_k - \underline{e}_k \tag{633}$$

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

$$\left\| \underline{\underline{AC}}_k - \underline{e}_k \right\|_2^2 = 1 - \sum_{l=1}^{p_k} \frac{\left[\det\left(\underline{\underline{D}}_l^k\right)\right]^2}{\det\left(\underline{\underline{G}}_{l-2}^k\right) \det\left(\underline{\underline{G}}_l^k\right)} \tag{634}$$

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 (631) 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

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

$$\underline{m}_k = \sum_{l=1}^{p_k} \frac{\det\left(\underline{\underline{D}}_l^k\right)}{\det\left(\underline{\underline{G}}_{l-2}^k\right) \det\left(\underline{\underline{G}}_l^k\right)} \tilde{\underline{m}}_l \tag{635}$$

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.

9.2.3. 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

$$(\underline{\underline{A}} + \underline{\underline{A}}^T)/2 \tag{636}$$

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

$$c_1 \leq \frac{\underline{x}^T \underline{\underline{A}} \underline{x}}{\underline{x}^T \underline{\underline{B}} \underline{x}} \leq c_2 \quad \forall \underline{x} \tag{637}$$

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 [11], [5], [4] y [2].

Definición 123 *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.

10. Apéndice C

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

$$\begin{aligned} \mathcal{L}u &= f_\Omega \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{638}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{639}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, como un caso particular del operador elíptico definido por la Ec. (43) 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

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{640}$$

aplicando el teorema de Green (115) obtenemos la Ec. (50), que podemos reescribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{641}$$

Definiendo el operador bilineal

$$a(u,v) = \int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} \tag{642}$$

y la funcional lineal

$$l(v) = \langle f, v\rangle = \int_\Omega v f_\Omega d\underline{x} \tag{643}$$

podemos reescribir el problema dado por la Ec. (638) de orden 2 en forma variacional, haciendo uso de la forma bilineal $a(\cdot,\cdot)$ y la funcional lineal $l(\cdot)$.

10.1. 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 124 *Una familia de triangulaciones $\mathcal{T}_h$ es llamada de forma-regular si existe una constante independiente de h, tal que*

$$h_K \leq C\rho_K, \ \ con\ K \in \mathcal{T}_h,$$

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 125 *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*

$$h_K \leq Ch, \ \ con\ K \in \mathcal{T}_h.$$

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.

10.2. 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.

10.3. Método de Elemento Finito Usando Discretización de Rectángulos

Para resolver la Ec. (638), 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. (189) queda escrita como

$$\int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy = \int_\Omega f_\Omega \phi_j dxdy \tag{644}$$

donde

$$\begin{aligned} K_{ij} &= \int_\Omega \left(a \nabla \phi_i \cdot \nabla \phi_j + c \phi_i \phi_j \right) dxdy \qquad (645) \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \nabla \phi_i^{(e)} \cdot \nabla \phi_j^{(e)} + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \\ &= \sum_{e=1}^{E} \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \end{aligned}$$

y el lado derecho como

$$\begin{aligned} F_j &= \int_\Omega f_\Omega \phi_j dxdy \\ &= \sum_{e=1}^{E} \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy. \end{aligned} \tag{646}$$

Para cada Ω_e de Ω, la submatriz de integrales (matriz de carga local)

$$K_{ij} = \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{647}$$

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)

$$F_j = \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \tag{648}$$

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. (638) 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

$$\int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \tag{649}$$

y

$$\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy, \tag{650}$$

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

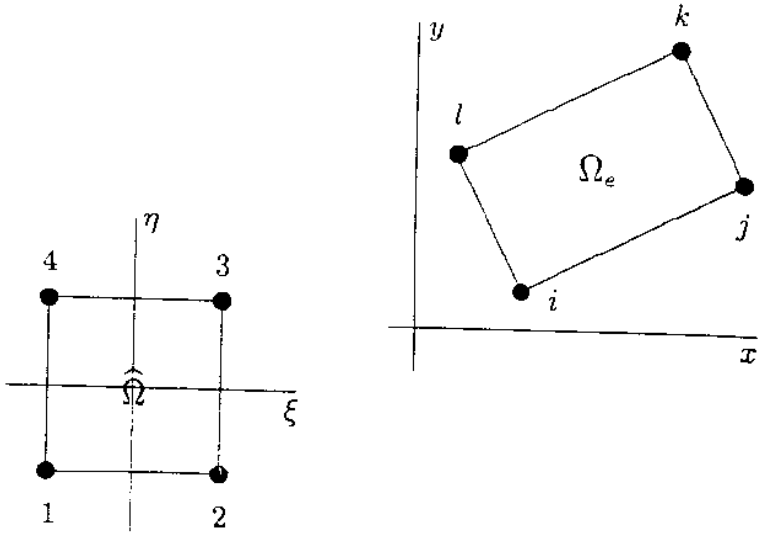

Figura 9:

mediante la transformación $f(x,y) = \underline{\underline{T}}(\varepsilon,\eta)+\underline{b}$, quedando dicha transformación como

$$\begin{aligned} x &= \frac{x_2^{(e)} - x_1^{(e)}}{2}\varepsilon + \frac{y_2^{(e)} - y_1^{(e)}}{2}\eta \\ y &= \frac{x_4^{(e)} - x_1^{(e)}}{2}\varepsilon + \frac{y_4^{(e)} - y_1^{(e)}}{2}\eta \end{aligned} \tag{651}$$

en la cual la matriz $\underline{\underline{T}}$ está dada por

$$\underline{\underline{T}} = \begin{pmatrix} \frac{x_2^{(e)} - x_1^{(e)}}{2} & \frac{y_2^{(e)} - y_1^{(e)}}{2} \\ \frac{x_4^{(e)} - x_1^{(e)}}{2} & \frac{y_4^{(e)} - y_1^{(e)}}{2} \end{pmatrix} \tag{652}$$

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

$$\begin{aligned}
\varepsilon &= \frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2} \left[\frac{y - b_2}{\left(\frac{x_4^{(e)} - x_1^{(e)}}{2} \right) \left(\frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2}}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \right)} \right]}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \\
\eta &= \frac{y - b_2}{\left(\frac{x_4^{(e)} - x_1^{(e)}}{2} \right) \left(\frac{x - b_1 - \frac{y_2^{(e)} - y_1^{(e)}}{2}}{\frac{x_2^{(e)} - x_1^{(e)}}{2}} \right) + \frac{y_4^{(e)} - y_1^{(e)}}{2}}.
\end{aligned} \tag{653}$$

Entonces las $\phi_i^{(e)}$ quedan definidas en términos de $\hat{\phi}_i$ como

$$\begin{aligned}
\hat{\phi}_1(\varepsilon, \eta) &= \frac{1}{4}(1 - \varepsilon)(1 - \eta) \\
\hat{\phi}_2(\varepsilon, \eta) &= \frac{1}{4}(1 + \varepsilon)(1 - \eta) \\
\hat{\phi}_3(\varepsilon, \eta) &= \frac{1}{4}(1 + \varepsilon)(1 + \eta) \\
\hat{\phi}_4(\varepsilon, \eta) &= \frac{1}{4}(1 - \varepsilon)(1 + \eta)
\end{aligned} \tag{654}$$

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. (651), entonces se tendrian las siguientes integrales

$$\begin{aligned}
K_{ij}^{(e)} &= \int_{\Omega_e} \left(a \left[\frac{\partial \phi_i^{(e)}}{\partial x} \frac{\partial \phi_j^{(e)}}{\partial x} + \frac{\partial \phi_i^{(e)}}{\partial y} \frac{\partial \phi_j^{(e)}}{\partial y} \right] + c \phi_i^{(e)} \phi_j^{(e)} \right) dxdy \\
&= \int_{\hat{\Omega}} \left(\left[a \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial x} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial x} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial x} \right) + \right. \right. \\
&\quad \left. \left. \left(\frac{\partial \hat{\phi}_i}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_i}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \left(\frac{\partial \hat{\phi}_j}{\partial \varepsilon} \frac{\partial \varepsilon}{\partial y} + \frac{\partial \hat{\phi}_j}{\partial \eta} \frac{\partial \eta}{\partial y} \right) \right] + c \hat{\phi}_i \hat{\phi}_j \right) |J| \, d\varepsilon d\eta
\end{aligned} \tag{655}$$

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. (652). Para resolver $\int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy$ en cada Ω_e se genera las integrales

$$\begin{aligned}
F_j^{(e)} &= \int_{\Omega_e} f_\Omega \phi_j^{(e)} dxdy \\
&= \int_{\hat{\Omega}} f_\Omega \hat{\phi}_j \, |J| \, d\varepsilon d\eta
\end{aligned} \tag{656}$$

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. (654) obtenemos

$$\begin{array}{ll} \frac{\partial \phi_1}{\partial \varepsilon} = -\frac{1}{4}(1-\eta) & \frac{\partial \phi_1}{\partial \eta} = -\frac{1}{4}(1-\varepsilon) \\ \frac{\partial \phi_2}{\partial \varepsilon} = \frac{1}{4}(1-\eta) & \frac{\partial \phi_2}{\partial \eta} = -\frac{1}{4}(1+\varepsilon) \\ \frac{\partial \phi_3}{\partial \varepsilon} = \frac{1}{4}(1+\eta) & \frac{\partial \phi_3}{\partial \eta} = \frac{1}{4}(1+\varepsilon) \\ \frac{\partial \phi_4}{\partial \varepsilon} = -\frac{1}{4}(1+\eta) & \frac{\partial \phi_4}{\partial \eta} = \frac{1}{4}(1-\varepsilon) \end{array} \tag{657}$$

y también

$$\begin{array}{ll} \frac{\partial \varepsilon}{\partial x} = \begin{pmatrix} \frac{y_4^{(e)} - y_1^{(e)}}{2\det T} \end{pmatrix} & \frac{\partial \varepsilon}{\partial y} = \begin{pmatrix} \frac{x_4^{(e)} - x_1^{(e)}}{2\det T} \end{pmatrix} \\ \frac{\partial \eta}{\partial x} = \begin{pmatrix} \frac{y_2^{(e)} - y_1^{(e)}}{2\det T} \end{pmatrix} & \frac{\partial \eta}{\partial y} = \begin{pmatrix} \frac{x_2^{(e)} - x_1^{(e)}}{2\det T} \end{pmatrix} \end{array} \tag{658}$$

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.

10.4. Método de Descomposición de Dominio de Subestructuración

La solución numérica por los esquemas tradicionales de discretización tipo elemento finito y diferencias finitas generan una discretización del problema, la cual es usada para generar un sistema de ecuaciones algebraicos $\underline{\underline{A}}\underline{u} = \underline{b}$. Este sistema algebraico en general es de gran tamaño para problemas reales, al ser estos algoritmos secuenciales su implantación suele hacerse en equipos secuenciales y por ello no es posible resolver muchos problemas que involucren el uso de una gran cantidad de memoria, actualmente para tratar de subsanar dicha limitante, se usa equipo paralelo para soportar algoritmos secuenciales, haciendo ineficiente su implantación en dichos equipos.

Los métodos de descomposición de dominio son un paradigma natural usado por la comunidad de modeladores. Los sistemas físicos son descompuestos en dos o más subdominios contiguos basados en consideraciones fenomenológicas. Estas descomposiciones basadas en dominios físicos son reflejadas en la ingeniería de software del código correspondiente.

Los métodos de descomposición de dominio permiten tratar los problemas de tamaño considerable, empleando algoritmos paralelos en computadoras secuenciales y/o paralelas. Esto es posible ya que cualquier método de descomposición de dominio se basa en la suposición de que dado un dominio computacional Ω, este se puede particionar -triangular- en E subdominios $\Omega_i, i = 1, 2, ..., E$ entre los cuales no existe traslape. Entonces el problema es reformulado en términos de cada subdominio (empleando algún método del tipo elemento finito) obteniendo una familia de subproblemas de tamaño reducido independientes en principio entre si, que están acoplados a través de la solución en la interfaz de los subdominios que es desconocida.

En esta sección se considerarán problemas con valor en la frontera (VBVP) de la forma

$$\begin{aligned} \mathcal{L}u &= f \quad \text{en } \Omega \\ u &= g \quad \text{en } \partial\Omega \end{aligned} \tag{659}$$

donde

$$\mathcal{L}u = -\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu \tag{660}$$

con $\underline{\underline{a}}$ una matriz positiva definida, simétrica y $c \geq 0$, como un caso particular del operador elíptico definido por la Ec. (43) 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.

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

$$-v\left(\nabla \cdot \underline{\underline{a}} \cdot \nabla u + cu\right) = v f_\Omega \tag{661}$$

aplicando el teorema de Green (115) obtenemos la Ec. (50), que podemos reescribir como

$$\int_\Omega \left(\nabla v \cdot \underline{\underline{a}} \cdot \nabla u + cuv\right) d\underline{x} = \int_\Omega v f_\Omega d\underline{x}. \tag{662}$$

Definiendo el operador bilineal

$$a\left(u,v\right)=\int_{\Omega}\left(\nabla v\cdot\underline{\underline{a}}\cdot\nabla u+cuv\right)d\underline{x} \tag{663}$$

y la funcional lineal

$$l(v)=\langle f,v\rangle=\int_{\Omega}vf_{\Omega}d\underline{x} \tag{664}$$

podemos reescribir el problema dado por la Ec. (638) de orden 2 en forma variacional, haciendo uso de la forma bilineal $a\left(\cdot,\cdot\right)$ y la funcional lineal $l\left(\cdot\right)$.

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. (189) queda escrita como

$$\int_{\Omega}\left(a\nabla\phi_i\cdot\nabla\phi_j+c\phi_i\phi_j\right)dxdy=\int_{\Omega}f_{\Omega}\phi_jdxdy \tag{665}$$

donde

$$\begin{aligned}K_{ij}&=\int_{\Omega}\left(a\nabla\phi_i\cdot\nabla\phi_j+c\phi_i\phi_j\right)dxdy \qquad (666)\\&=\sum_{e=1}^{E}\int_{\Omega_e}\left(a\nabla\phi_i^{(e)}\cdot\nabla\phi_j^{(e)}+c\phi_i^{(e)}\phi_j^{(e)}\right)dxdy\\&=\sum_{e=1}^{E}\int_{\Omega_e}\left(a\left[\frac{\partial\phi_i^{(e)}}{\partial x}\frac{\partial\phi_j^{(e)}}{\partial x}+\frac{\partial\phi_i^{(e)}}{\partial y}\frac{\partial\phi_j^{(e)}}{\partial y}\right]+c\phi_i^{(e)}\phi_j^{(e)}\right)dxdy\end{aligned}$$

y el lado derecho como

$$\begin{aligned}F_j&=\int_{\Omega}f_{\Omega}\phi_jdxdy \qquad (667)\\&=\sum_{e=1}^{E}\int_{\Omega_e}f_{\Omega}\phi_j^{(e)}dxdy.\end{aligned}$$

Consideremos el problema dado por la Ec. (659) en el dominio Ω, el cual es subdividido en E subdominios Ω_i, $i=1,2,...,E$ sin traslape, también conocida como malla gruesa $\mathcal{T}_H$, es decir

$$\Omega_i\cap\Omega_j=\varnothing\quad\forall i\neq j\quad\text{y}\quad\overline{\Omega}=\bigcup_{i=1}^{E}\overline{\Omega}_i, \tag{668}$$

y al conjunto

$$\Gamma=\bigcup_{i=1}^{E}\Gamma_i,\quad\text{si }\Sigma_i=\partial\Omega_i\backslash\partial\Omega \tag{669}$$

lo llamaremos la frontera interior del dominio Ω, denotamos por H al diámetro $H_i = Diam(\Omega_i)$ de cada Ω_i que satisface $Diam(\Omega_i) \leq H$ para cada $i = 1, 2, ..., E$, además, cada subdominio Ω_i es descompuesto en una mallado fino $\mathcal{T}_h$ de K subdominios mediante una triangulación Ω_e de modo que esta sea conforme, denotamos por h al diámetro $h_i = Diam(\Omega_e)$ de cada Ω_e que satisface $Diam(\Omega_e) \leq h$ para cada $e = 1, 2, ..., K$ de cada $i = 1, 2, ..., E$.

Un ejemplo de un dominio Ω y su descomposición en subdominios Ω_i y cada Ω_i a su vez descompuesto en Ω_e subdominios se muestra en la figura:

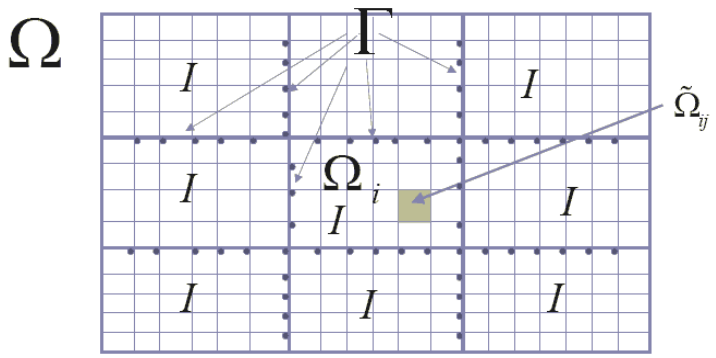

Figura 10: Dominio Ω descompuesto en subdominios Ω_i, con $i = 1, 2, ..., 9$.

Sin perdida de generalidad tomemos $g = 0$ en $\partial\Omega$, notemos que siempre es posible poner el problema de la Ec. (659) como uno con condiciones de frontera Dirichlet que se nulifiquen mediante la adecuada manipulación del término del lado derecho de la ecuación.

Primeramente sea $D \subset H_0^1(\Omega)$ un espacio lineal de funciones de dimensión finita N, en el cual esté definido un producto interior denotado para cada $u, v \in D$ por

$$u \cdot v = \langle u, v \rangle \tag{670}$$

Considerando la existencia de los subconjuntos linealmente indepandientes

$$\mathcal{B} \subset \tilde{D}, \mathcal{B}_I \subset \tilde{D}_I, \mathcal{B}_\Gamma \subset \tilde{D}_\Gamma$$
$$\mathcal{B}_\Gamma \subset \tilde{D}_\Gamma, \mathcal{B}_{\Gamma J} \subset \tilde{D}_{\Gamma 1}, \mathcal{B}_{\Gamma M} \subset \tilde{D}_{\Gamma 2}$$

los cuales satisfacen

$$\mathcal{B} = \mathcal{B}_I \cup \mathcal{B}_\Gamma \text{ y } \bar{\mathcal{B}}_\Gamma = \mathcal{B}_{\Gamma J} \cup \mathcal{B}_{\Gamma M}$$

el espacio generado por cada uno de los subconjuntos $\mathcal{B}_\Gamma$ y $\bar{\mathcal{B}}_\Gamma$ es $\tilde{D}_\Gamma$, sin embargo distinguimos la propiedad de que los miembros de $\mathcal{B}_\Gamma$ tienen soporte local.

Definimos las bases

$$\mathcal{B}_I = \left\{ w_I^1, ..., w_I^{\bar{N}_I} \right\}, \mathcal{B}_{\Gamma M} = \left\{ w_M^1, ..., w_M^{\bar{N}_\Gamma} \right\} \text{ y } \mathcal{B}_{\Gamma J} = \left\{ w_J^1, ..., w_J^{\bar{N}_\Gamma} \right\}$$

de las funcionales lineales ϕ_i en Ω.

Entonces definiendo para toda $\delta = 1, ..., K$, la matriz de $N_\delta \times N_\delta$

$$\underline{\underline{A}}_\delta^{II} \equiv \left[\left\langle w_I^i, w_I^j \right\rangle \right] \tag{671}$$

que sólo esta definida en cada subespacio (subdominio Ω_δ). Entonces, la matriz virtual $\underline{\underline{A}}^{II}$ es dada por la matriz diagonal de la forma

$$\underline{\underline{A}}^{II} \equiv \begin{bmatrix} \underline{\underline{A}}_1^{II} & & & \\ & \underline{\underline{A}}_2^{II} & & \\ & & \ddots & \\ & & & \underline{\underline{A}}_E^{II} \end{bmatrix} \tag{672}$$

donde el resto de la matriz fuera de la diagonal en bloques es cero.

De forma similar definimos

$$\underline{\underline{A}}_\delta^{I\Gamma} \equiv \left[\left\langle w_I^i, w_\Gamma^\alpha \right\rangle \right], \quad \underline{\underline{A}}_\delta^{\Gamma I} \equiv \left[\left\langle w_\Gamma^\alpha, w_I^i \right\rangle \right] \tag{673}$$

y

$$\underline{\underline{A}}_\delta^{\Gamma\Gamma} \equiv \left[\langle w_\Gamma^\alpha, w_\Gamma^\alpha \rangle \right] \tag{674}$$

para toda $\delta = 1, ..., K$, obsérvese que como $\bar{\mathcal{B}}_\Gamma = \mathcal{B}_{\Gamma J} \cup \mathcal{B}_{\Gamma M}$ entonces

$$\underline{\underline{A}}_\delta^{\Gamma\Gamma} = \left[\langle w_\Gamma^\alpha, w_\Gamma^\alpha \rangle \right] = \left[\left\langle w_{\Gamma_J}^\alpha, w_{\Gamma_J}^\alpha \right\rangle \right] + \left[\left\langle w_{\Gamma_M}^\alpha, w_{\Gamma_M}^\alpha \right\rangle \right]$$

también que $\underline{\underline{A}}_\delta^{I\Gamma} = \left(\underline{\underline{A}}_\delta^{\Gamma I} \right)^T$. Entonces las matrices virtuales $\underline{\underline{A}}^{\Gamma I}, \underline{\underline{A}}^{\Gamma I}$ y $\underline{\underline{A}}^{\Gamma\Gamma}$ quedarán definidas como

$$\underline{\underline{A}}^{I\Gamma} \equiv \begin{bmatrix} \underline{\underline{A}}_1^{I\Gamma} \\ \underline{\underline{A}}_2^{I\Gamma} \\ \vdots \\ \underline{\underline{A}}_E^{I\Gamma} \end{bmatrix} \tag{675}$$

$$\underline{\underline{A}}^{\Gamma I} \equiv \begin{bmatrix} \underline{\underline{A}}_1^{\Gamma I} & \underline{\underline{A}}_2^{\Gamma I} & \cdots & \underline{\underline{A}}_E^{\Gamma I} \end{bmatrix} \tag{676}$$

y

$$\underline{\underline{A}}^{\Gamma\Gamma} \equiv \left[\sum_{i=1}^{E} \underline{\underline{A}}_i^{\Gamma\Gamma} \right] \tag{677}$$

donde $\left[\sum_{i=1}^{E} \underline{\underline{A}}_i^{\Gamma\Gamma} \right]$ es construida sumando las $\underline{\underline{A}}_i^{\Gamma\Gamma}$ según el orden de los nodos globales versus los nodos locales.

También consideremos al vector $\underline{u} \equiv (u_1, .., u_E)$ el cual puede ser escrito como $\underline{u} = (\underline{u_I}, \underline{u_\Gamma})$ donde $\underline{u_I} = (u_1, ..., _{N_I})$ y $\underline{u_\Gamma} = (u_1, ..., _{N_\Gamma})$.

Así, el sistema virtual

$$\begin{aligned} \underline{\underline{A}}^{II}\underline{u_I} + \underline{\underline{A}}^{I\Gamma}\underline{u_\Gamma} &= \underline{b_I} \\ \underline{\underline{A}}^{\Gamma I}\underline{u_I} + \underline{\underline{A}}^{\Gamma\Gamma}\underline{u_\Gamma} &= \underline{b_\Gamma} \end{aligned} \tag{678}$$

quedando expresado como

$$\begin{bmatrix} \underline{\underline{A}}_1^{II} & & \\ & \ddots & \\ & & \underline{\underline{A}}_E^{II} \end{bmatrix} \begin{bmatrix} \underline{u_I}_1 \\ \vdots \\ \underline{u_I}_E \end{bmatrix} + \begin{bmatrix} \underline{\underline{A}}_1^{I\Gamma} \\ \vdots \\ \underline{\underline{A}}_E^{I\Gamma} \end{bmatrix} \begin{bmatrix} \underline{u_\Gamma}_1 \\ \vdots \\ \underline{u_\Gamma}_E \end{bmatrix} = \begin{bmatrix} \underline{b_I}_1 \\ \vdots \\ \underline{b_I}_E \end{bmatrix}$$

$$\begin{bmatrix} \underline{\underline{A}}_1^{\Gamma I} & \cdots & \underline{\underline{A}}_E^{\Gamma I} \end{bmatrix} \begin{bmatrix} \underline{u_I}_1 \\ \vdots \\ \underline{u_I}_E \end{bmatrix} + \begin{bmatrix} \underline{\underline{A}}^{\Gamma\Gamma} \end{bmatrix} \begin{bmatrix} \underline{u_\Gamma}_1 \\ \vdots \\ \underline{u_\Gamma}_E \end{bmatrix} = \begin{bmatrix} \underline{b_\Gamma}_1 \\ \vdots \\ \underline{b_\Gamma}_E \end{bmatrix}$$

o más compactamente como $\underline{\underline{A}}\,\underline{u} = \underline{b}$, notemos que las matrices $\underline{\underline{A}}_i^{\Gamma\Gamma}, \underline{\underline{A}}_i^{\Gamma I}, \underline{\underline{A}}_i^{I\Gamma}$ y $\underline{\underline{A}}_i^{II}$ son matrices bandadas.

Si ahora despejamos $\underline{u_I}$ del la primera ecuación del sistema dado por la Ec. (678) obtenemos

$$\underline{u_I} = \left(\underline{\underline{A}}^{II}\right)^{-1}\left(\underline{b_I} - \underline{\underline{A}}^{I\Gamma}\underline{u_\Gamma}\right)$$

si sustituimos $\underline{u_I}$ en la segunda ecuación del sistema dado por la Ec. (678) entonces tenemos

$$\left(\underline{\underline{A}}^{\Gamma\Gamma} - \underline{\underline{A}}^{\Gamma I}\left(\underline{\underline{A}}^{II}\right)^{-1}\underline{\underline{A}}^{I\Gamma}\right)\underline{u_\Gamma} = \underline{b_\Gamma} - \underline{\underline{A}}^{\Gamma I}\left(\underline{\underline{A}}^{II}\right)^{-1}\underline{b_I} \tag{679}$$

en la cual los nodos interiores no figuran en la ecuación y todo queda en función de los nodos de la frontera interior $\underline{u_\Gamma}$.

A la matriz formada por $\underline{\underline{A}}^{\Gamma\Gamma} - \underline{\underline{A}}^{\Gamma I}\left(\underline{\underline{A}}^{II}\right)^{-1}\underline{\underline{A}}^{I\Gamma}$ se le conoce como el complemento de Schur global y se le denota como

$$\underline{\underline{S}} = \underline{\underline{A}}^{\Gamma\Gamma} - \underline{\underline{A}}^{\Gamma I}\left(\underline{\underline{A}}^{II}\right)^{-1}\underline{\underline{A}}^{I\Gamma}. \tag{680}$$

En nuestro caso, como estamos planteado todo en términos de subdominios Ω_i, con $i = 1, ..., E$, entonces las matrices $\underline{\underline{A}}_i^{\Gamma\Gamma}, \underline{\underline{A}}_i^{\Gamma I}, \underline{\underline{A}}_i^{I\Gamma}$ y $\underline{\underline{A}}_i^{II}$ quedan definidas de manera local, así que procedemos a definir el complemento de Schur local como

$$\underline{\underline{S}}_i = \underline{\underline{A}}_i^{\Gamma\Gamma} - \underline{\underline{A}}_i^{\Gamma I}\left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{\underline{A}}_i^{I\Gamma} \tag{681}$$

adicionalmente definimos

$$\underline{b}_i = \underline{b_\Gamma}_i - \underline{\underline{A}}_i^{\Gamma I}\left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{b_I}_i.$$

El sistema dado por la Ec. (679) lo escribimos como

$$\underline{\underline{S}}\,\underline{u_\Gamma} = \underline{b} \tag{682}$$

y queda definido de manera virtual a partir de

$$\left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right] \underline{u_\Gamma} = \left[\sum_{i=1}^{E} \underline{b}_i\right] \tag{683}$$

donde $\left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right]$ y $\left[\sum_{i=1}^{E} \underline{b}_i\right]$ podrían ser construida sumando las S_i y b_i respectivamente según el orden de los nodos globales versus los nodos locales.

El sistema lineal virtual obtenido de esta forma (682) se resuelve eficientemente usando el método de gradiente conjugado visto en la sección (9.1.2), para ello no es necesario construir la matriz $\underline{\underline{S}}$ con las contribuciones de cada S_i co-rrespondientes al subdominio i, lo que hacemos es pasar a cada subdominio el vector $\underline{u_\Gamma}^i$ correspondiente a la i-ésima iteración del método de gradiente conjugado para que en cada subdominio se evalué $\underline{\tilde{u}_\Gamma}^i = \underline{\underline{S}}_i \underline{u_\Gamma}^i$ localmente y con el resultado se forma el vector $\underline{\tilde{u}_\Gamma} = \sum_{i=1}^{E} \underline{\tilde{u}_\Gamma}^i$ y se continué con los demás pasos del método. Esto es ideal para una implementación en paralelo del método de gradiente conjugado.

Observación 126 *Notemos que el normalmente las matrices locales* $\underline{\underline{S}}_i$ *y* $\left(\underline{\underline{A}}_i^{II}\right)^{-1}$ *no se construyen, ya que estas serian matrices densas y su construcción es computacionalmente muy costosa. Y como sólo nos interesa el producto* $\underline{\underline{S}y_\Gamma}$*, o más precisamente* $\left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right] \underline{y_\Gamma}$*, entonces si llamamos* $\underline{y_\Gamma}_i$ *al vector correspondiente al subdominio* i*, entonces tendremos*

$$\underline{z} = \left(\underline{\underline{A}}^{\Gamma\Gamma} - \underline{\underline{A}}^{\Gamma I} \left(\underline{\underline{A}}^{II}\right)^{-1} \underline{\underline{A}}^{I\Gamma}\right) \underline{y_\Gamma}. \tag{684}$$

Para evaluar eficientemente esta expresión, realizamos las siguientes operaciones equivalentes

$$\begin{aligned} \underline{x1} &= \underline{\underline{A}}^{\Gamma\Gamma} \underline{y_\Gamma}_i \\ \underline{x2} &= \left(\underline{\underline{A}}^{\Gamma I} \left(\underline{\underline{A}}^{II}\right)^{-1} \underline{\underline{A}}^{I\Gamma}\right) \underline{y_\Gamma}_i \\ \underline{z} &= \underline{x1} - \underline{x2} \end{aligned} \tag{685}$$

la primera y tercera expresión no tienen ningún problema en su evaluación, para la segunda expresión tendremos que hacer

$$\underline{x3} = \underline{\underline{A}}^{I\Gamma} \underline{y_\Gamma}_i \tag{686}$$

con este resultado intermedio, deberíamos calcular

$$\underline{x4} = \left(\underline{\underline{A}}^{II}\right)^{-1} \underline{x3} \tag{687}$$

pero como no contamos con $\left(\underline{\underline{A}}^{II}\right)^{-1}$*, entonces multiplicamos la expresión por* $\underline{\underline{A}}^{II}$ *obteniendo*

$$\underline{\underline{A}}^{II} \underline{x4} = \underline{\underline{A}}^{II} \left(\underline{\underline{A}}^{II}\right)^{-1} \underline{x3} \tag{688}$$

al simplificar, tenemos

$$\underline{\underline{A}}^{II}\underline{x4} = \underline{x3}. \tag{689}$$

Esta última expresión puede ser resuelta usando Factorización LU o Gradiente Conjugado (cada una de estas opciones tiene ventajas y desventajas que deben ser evaluadas al momento de implementar el código para un problema particular). Una vez obtenido $\underline{x3}$*, podremos calcular*

$$\underline{x2} = \underline{\underline{A}}^{\Gamma I}\underline{x3} \tag{690}$$

completando la secuencia de operaciones necesaria para obtener $\underline{\underline{S}}_i\underline{y_\Gamma}$.

Una vez resuelto el sistema de la Ec. (683) en el que hemos encontrado la solución para los nodos de la frontera interior $\underline{u_\Gamma}$, entonces debemos resolver localmente los $\underline{u_I}_i$ correspondientes a los nodos interiores para cada subespacio Ω_i, para esto empleamos

$$\underline{u_I}_i = \left(\underline{\underline{A}}_i^{II}\right)^{-1}\left(\underline{b_I}_i - \underline{\underline{A}}_i^{I\Gamma}\underline{u_\Gamma}_i\right) \tag{691}$$

para cada $i = 1, 2, ..., E$, quedando así resuelto el problema $\underline{\underline{A}}\underline{u} = \underline{b}$ tanto en los nodos interiores $\underline{u_I}_i$ como en los de la frontera interior $\underline{u_\Gamma}_i$ correspondientes a cada subespacio Ω_i.

Observación 127 *En la evaluación de* $\underline{u_I}_i = \left(\underline{\underline{A}}_i^{II}\right)^{-1}\left(\underline{b_I}_i - \underline{\underline{A}}_i^{I\Gamma}\underline{u_\Gamma}_i\right)$*, esta nuevamente involucrado* $\left(\underline{\underline{A}}_i^{II}\right)^{-1}$*, por ello deberemos de usar el siguiente procedimiento para evaluar eficientemente esta expresión, realizando las siguientes operaciones equivalentes*

$$\begin{aligned} \underline{x4} &= \underline{b_I}_i - \underline{\underline{A}}_i^{I\Gamma}\underline{u_\Gamma}_i \\ \underline{u_I}_i &= \left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{x4} \end{aligned} \tag{692}$$

multiplicando por $\underline{\underline{A}}_i^{II}$ *a la última expresión, obtenemos*

$$\underline{\underline{A}}_i^{II}\underline{u_I}_i = \underline{\underline{A}}_i^{II}\left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{x4} \tag{693}$$

simplificando, tenemos

$$\underline{\underline{A}}_i^{II}\underline{u_I}_i = \underline{x4} \tag{694}$$

esta última expresión puede ser resuelta usando Factorización LU o Gradiente Conjugado.

Como se indico en las dos últimas observaciones, para resolver el sistema $\underline{\underline{A}}_i^{II}\underline{x} = \underline{b}$ podemos usar Factorización LU, Gradiente Conjugado o cualquier otro método para resolver sistemas lineales, pero deberá de usarse aquel que proporcione la mayor velocidad en el cálculo o que consuma la menor cantidad

de memoria (ambas condicionantes son mutuamente excluyentes), por ello la decisión de que método usar deberá de tomarse al momento de tener que resolver un problema particular en un equipo dado.

Para usar el método de Factorización LU, se deberá primeramente de factorizar la matriz bandada $\underline{\underline{A}}_i^{II}$ en una matriz $\underline{\underline{LU}}$ que es densa (del orden de $(N_\delta)^2$ por ello consume una gran cantidad de memoria) pero esta operación sólo se deberá de realizar una vez por cada subdominio, y para solucionar los diversos sistemas lineales $\underline{\underline{A}}_i^{II}\underline{x} = \underline{b}$ sólo será necesario evaluar los sistemas

$$\begin{aligned} \underline{\underline{L}}\underline{y} &= \underline{b} \\ \underline{\underline{U}}\underline{x} &= \underline{y} \end{aligned} \tag{695}$$

en donde $\underline{y}$ es un vector auxiliar. Esto proporciona una manera muy eficiente de evaluar el sistema lineal pero el consumo en memoria para un problema particular puede ser excesivo.

Por ello si el problema involucra una gran cantidad de nodos interiores y el equipo en el que se implantará la ejecución del programa tiene una cantidad de memoria muy limitada, es recomendable usar el método de Gradiente Conjugado, este consume una cantidad de memoria adicional muy pequeña y pese a que no es tan eficiente (dos o tres veces mas lento) como la Factorización LU, si proporciona un buen desempeño versus el obtenido al realizar el cálculo directo de $\left(\underline{\underline{A}}_i^{II}\right)^{-1}$.

De esta forma, es posible adaptar el código para tomar en cuenta la implementación de este en un equipo de cómputo en particular y poder sacar el máximo provecho al método de Subestructuración en la resolución de problemas elípticos de gran envergadura.

En lo que resta del presente trabajo, se asume que el método empleado para resolver $\underline{\underline{A}}_i^{II}\underline{x} = \underline{b}$ en sus respectivas variantes necesarias para evitar el cálculo de $\left(\underline{\underline{A}}_i^{II}\right)^{-1}$ es la Factorización LU, logrando así el máximo desempeño en velocidad en tiempo de ejecución.

El número de condicionamiento del complemento de Schur sin precondicionamiento puede ser estimado, para ello:

Definición 128 *Introducimos una norma-L^2 equivalente sobre Γ mediante*

$$\left\|\underline{u_\Gamma}\right\|_\Gamma^2 = \sum_{i=1}^{E} \left\|\underline{u_\Gamma}\right\|_{L^2(\partial\Omega_i)}^2 .$$

Teorema 129 *Sea $\underline{u_\Gamma}$ la traza de funciones de elemento finito en V^h sobre Γ, asumimos que los coeficientes de la ecuación diferencial parcial $\rho_i = 1$, $i = 1, 2, ..., E$, y que la malla fina $\mathcal{T}_h$ y la malla gruesa $\mathcal{T}_H$ sea cuasi-uniforme. Entonces existen dos constantes positivas c y C, independientes de h y H, tal que*

$$cH \left\|\underline{u_\Gamma}\right\|_\Gamma^2 \leq s(\underline{u_\Gamma}, \underline{u_\Gamma}) \leq Ch^{-1} \left\|\underline{u_\Gamma}\right\|_\Gamma^2$$

de este modo

$$\kappa = cond(\underline{\underline{S}}) \leq \frac{C}{Hh}.$$

10.5. 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:

$$\begin{aligned} -\nabla^2 u &= f_\Omega \text{ en } \Omega \\ u &= g_{\partial\Omega} \text{ en } \partial\Omega. \end{aligned} \tag{696}$$

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:

$$\Omega = [-1, 1] \times [0, 1] \tag{697}$$

donde

$$f_\Omega = 2n^2\pi^2 \sin(n\pi x) * \sin(n\pi y) \quad \text{y} \quad g_{\partial\Omega} = 0 \tag{698}$$

cuya solución es

$$u(x, y) = \sin(n\pi x) * \sin(n\pi y). \tag{699}$$

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:

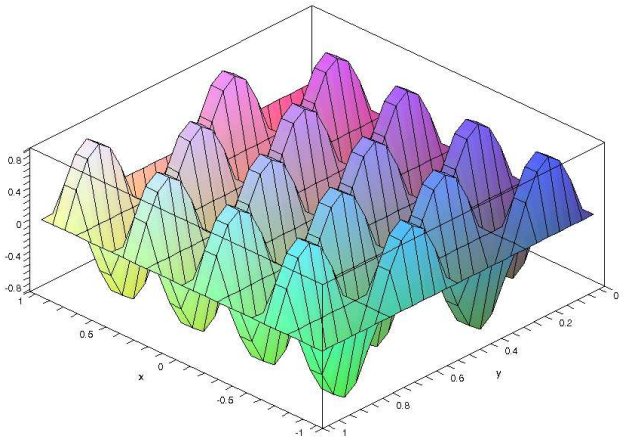

Figura 11: Solución a la ecuación de Poisson para n=4.

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.

10.5.1. Método del Elemento Finito Secuencial

A partir de la formulación del método de elemento finito visto en la sección (2.1.2), la implementación computacional que se desarrolló tiene la jerarquía de clases siguiente:

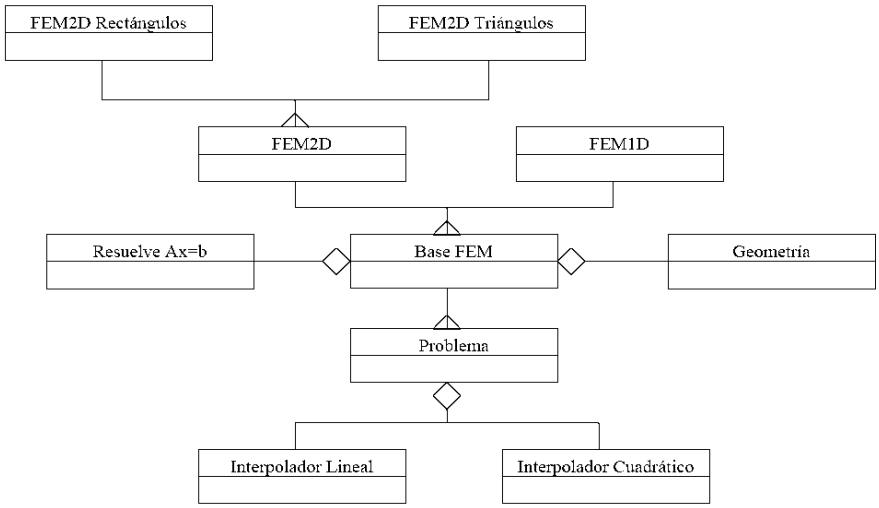

Figura 12: Jerarquía de clases para el método de elemento finito

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:

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.

10.5.2. Método de Subestructuración Secuencial

A partir de la formulación del método de subestructuración visto en la sección (**??**) se generan las matrices locales $\underline{\underline{A}}_i^{II}, \underline{\underline{A}}_i^{I\Sigma}, \underline{\underline{A}}_i^{\Sigma I}$ y $\underline{\underline{A}}_i^{\Sigma\Sigma}$ y con ellas se construyen $\underline{\underline{S}}_i = \underline{\underline{A}}_i^{\Sigma\Sigma} - \underline{\underline{A}}_i^{\Sigma I}\left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{\underline{A}}_i^{I\Sigma}$ y $\underline{b}_i = \underline{\underline{A}}_i^{\Sigma I}\left(\underline{\underline{A}}_i^{II}\right)^{-1}\underline{b}_{I_i}$ que son problemas locales a cada subdominio Ω_i, con $i = 1, 2, ..., E$. Generando de manera virtual el sistema lineal $\underline{\underline{S}}\underline{u_\Sigma} = \underline{b}$ a partir de

$$\left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right] \underline{u_\Sigma} = \left[\sum_{i=1}^{E} \underline{b}_i\right] \tag{700}$$

donde $\underline{\underline{S}} = \left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right]$ y $\underline{b} = \left[\sum_{i=1}^{E} \underline{b}_i\right]$ podría ser construida sumando las $\underline{\underline{S}}_i$ y $\underline{b}_i$ respectivamente según el orden de los nodos globales versus los nodos locales a cada subdominio.

El sistema lineal virtual resultante

$$\underline{\underline{S}}\underline{u_\Sigma} = \underline{b} \tag{701}$$

es resuelto usando el método de gradiente conjugado visto en la sección (9.1.2), para ello no es necesario construir la matriz $\underline{\underline{S}}$ con las contribuciones de cada S_i correspondientes al subdominio i. Lo que hacemos es pasar a cada subdominio el vector $\underline{u_\Sigma}^i$ correspondiente a la i-ésima iteración del método de gradiente conjugado para que en cada subdominio se evalué $\underline{\tilde{u}_\Sigma}^i = \underline{\underline{S}}_i \underline{u_\Sigma}^i$ localmente y con el resultado se forma el vector $\underline{\tilde{u}_\Sigma} = \sum_{i=1}^{E} \underline{\tilde{u}_\Sigma}^i$ y se continué con los demás pasos del método.

La implementación computacional que se desarrolló tiene una jerarquía de clases en la cual se agregan las clases *FEM2D Rectángulos* y *Geometría*, además de heredar a la clase *Problema*. De esta forma se rehusó todo el código desarrollado para *FEM2D Rectángulos*, la jerarquía queda como:

> La clase *DDM2D* realiza la partición gruesa del dominio mediante la clase *Geometría* y controla la partición de cada subdominio mediante un objeto de la clase de *FEM2D Rectángulos* generando la partición fina del dominio. La resolución de los nodos de la frontera interior se hace mediante el método de gradiente conjugado, necesaria para resolver los nodos internos de cada subdominio.

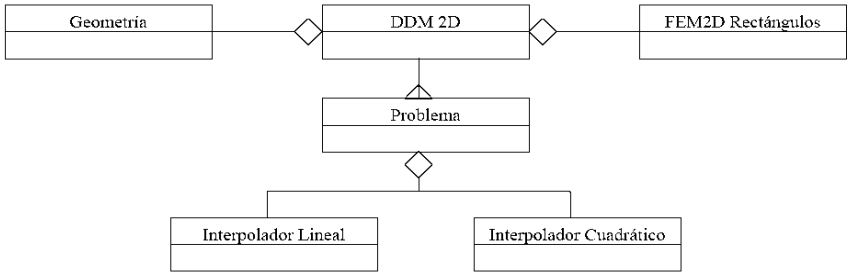

Figura 13: Jerarquía de clases para el método de subestructuración secuencial

Así, el dominio Ω es descompuesto en una descomposición gruesa de $n \times m$ subdominios y cada subdominio Ω_i se parte en $p \times q$ subdominios, generando la participación fina del dominio como se muestra en la figura:

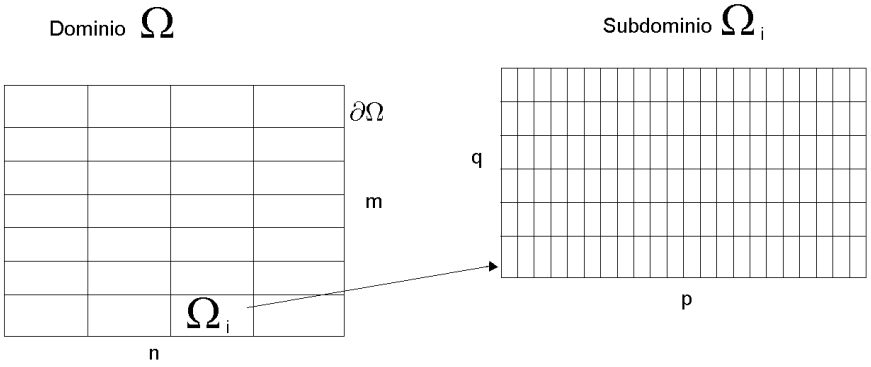

Figura 14: Descomposición del dominio Ω en $E = n \times m$ subdominios y cada subdominio Ω_i en $p \times q$ subdominios

El método de descomposición de dominio se implementó realizando las siguientes tareas:

A) La clase *DDM2D* genera la descomposición gruesa del dominio mediante la agregación de un objeto de la clase *Geometría* (supongamos particionado en $n \times m$ subdominios, generando $s = n * m$ subdominios Ω_i, $i = 1, 2, ..., E$).

B) Con esa geometría se construyen los objetos de *FEM2D Rectángulos* (uno por cada subdominio Ω_i), donde cada subdominio es particionado (supongamos en $p \times q$ subdominios) y regresando las coordenadas de los nodos de frontera del subdominio correspondiente a la clase DDM2D.

C) Con estas coordenadas, la clase *DDM2D* conoce a los nodos de la frontera interior (son estos los que resuelve el método de descomposición de dominio). Las coordenadas de los nodos de la frontera interior se dan a conocer a los objetos *FEM2D Rectángulos*, transmitiendo sólo aquellos que están en su subdominio.

D) Después de conocer los nodos de la frontera interior, cada objeto *FEM2D Rectángulos* calcula las matrices $\underline{\underline{A}}_i^{\Sigma\Sigma}, \underline{\underline{A}}_i^{\Sigma I}, \underline{\underline{A}}_i^{I\Sigma}$ y $\underline{\underline{A}}_i^{II}$ necesarias para construir el complemento de Schur local $\underline{\underline{S}}_i = \underline{\underline{A}}_i^{\Sigma\Sigma} - \underline{\underline{A}}_i^{\Sigma I} \left(\underline{\underline{A}}_i^{II}\right)^{-1} \underline{\underline{A}}_i^{I\Sigma}$ sin realizar comunicación alguna. Al terminar de calcular las matrices se avisa a la clase *DDM2D* de la finalización de los cálculos.

E) Mediante la comunicación de vectores del tamaño del número de nodos de la frontera interior entre la clase *DDM2D* y los objetos *FEM2D Rectángulos*, se prepara todo lo necesario para empezar el método de gradiente conjugado y resolver el sistema lineal virtual $\left[\sum_{i=1}^{E} \underline{\underline{S}}_i\right] \underline{u_\Sigma} = \left[\sum_{i=1}^{E} \underline{b}_i\right]$.

F) Para usar el método de gradiente conjugado, se transmite un vector del tamaño del número de nodos de la frontera interior para que en cada objeto se realicen las operaciones pertinentes y resolver así el sistema algebraico asociado, esta comunicación se realiza de ida y vuelta entre la clase *DDM2D* y los objetos *FEM2D Rectángulos* tantas veces como iteraciones haga el método. Resolviendo con esto los nodos de la frontera interior $\underline{u_{\Sigma_i}}$.

G) Al término de las iteraciones se pasa la solución $\underline{u_{\Sigma_i}}$ de los nodos de la frontera interior que pertenecen a cada subdominio dentro de cada objeto *FEM2D Rectángulos* para que se resuelvan los nodos interiores $\underline{u_{I_i}} = \left(\underline{\underline{A}}_i^{II}\right)^{-1} \left(\underline{b_{I_i}} - \underline{\underline{A}}_i^{I\Sigma} \underline{u_{\Sigma_i}}\right)$, sin realizar comunicación alguna en el proceso, al concluir se avisa a la clase *DDM2D* de ello.

I) La clase *DDM2D* mediante un último mensaje avisa que se concluya el programa, terminado así el esquema maestro-esclavo secuencial.

Por ejemplo, para resolver la Ec. (**??**), usando 513×513 nodos (igual al ejemplo de *FEM2D Rectángulos* secuencial), podemos tomar alguna de las siguientes descomposiciones:

Descomposición	Nodos Interiores	Subdominios	Elementos Subdominio	Total Nodos Subdominio	Nodos Desconocidos Subdominio
2x2 y 256x256	260100	4	65536	66049	65025
4x4 y 128x128	258064	16	16384	16641	16129
8x8 y 64x64	254016	64	4096	4225	3969
16x16 y 32x32	246016	256	1024	1089	961
32x32 y 16x16	230400	1024	256	289	225

Cada una de las descomposiciones genera un problema distinto. Usando el equipo secuencial y evaluando el desempeño del método de subestructuración secuencial se obtuvieron los siguientes resultados:

Partición	Nodos Frontera Interior	Iteraciones	Tiempo Total
2x2 y 256x256	1021	139	5708 seg.
4x4 y 128x128	3057	159	2934 seg.
8x8 y 64x64	7105	204	1729 seg.
16x16 y 32x32	15105	264	1077 seg.
32x32 y 16x16	30721	325	1128 seg

Nótese que aún en un solo procesador es posible encontrar una descomposición que desminuya los tiempos de ejecución (la descomposición de 2x2 y 256x256 concluye en 5708 seg. versus los 6388 seg. en el caso de *FEM2D Rectángulos)*, ello es debido a que al descomponer el dominio en múltiples subdominios, la complejidad del problema es también disminuida y esto se ve reflejado en la disminución del tiempo de ejecución.

En la última descomposición, en lugar de disminuir el tiempo de ejecución este aumenta, esto se debe a que se construyen muchos objetos *FEM2D Rectángulos* (1024 en este caso), con los cuales hay que hacer comunicación resultando muy costoso computacionalmente.

Finalmente las posibles mejoras de eficiencia para el método de subestructuración secuencial para disminuir el tiempo de ejecución son las mismas que en el caso del método de elemento finito pero además se tienen que:

- Encontrar la descomposición pertinente entre las posibles descomposiciones que consuma el menor tiempo de cálculo.

10.6. Análisis de Convergencia

11. Apéndice D

11.1. El Cómputo en Paralelo

Los sistemas de cómputo con procesamiento en paralelo surgen de la necesidad de resolver problemas complejos en un tiempo razonable, utilizando las ventajas de memoria, velocidad de los procesadores, formas de interconexión de estos y distribución de la tarea, a los que en su conjunto denominamos arquitectura en paralelo. Entenderemos por una arquitectura en paralelo a un conjunto de procesadores interconectados capaces de cooperar en la solución de un problema.

Así, para resolver un problema en particular, se usa una o combinación de múltiples arquitecturas (topologías), ya que cada una ofrece ventajas y desventajas que tienen que ser sopesadas antes de implementar la solución del problema en una arquitectura en particular. También es necesario conocer los problemas a los que se enfrenta un desarrollador de programas que se desean correr en paralelo, como son: el partir eficientemente un problema en múltiples tareas y como distribuir estas según la arquitectura en particular con que se trabaje.

11.1.1. Arquitecturas de Software y Hardware

En esta sección se explican en detalle las dos clasificaciones de computadoras más conocidas en la actualidad. La primera clasificación, es la clasificación clásica de Flynn en dónde se tienen en cuenta sistemas con uno o varios procesadores, la segunda clasificación es moderna en la que sólo tienen en cuenta los sistemas con más de un procesador.

El objetivo de esta sección es presentar de una forma clara los tipos de clasificación que existen en la actualidad desde el punto de vista de distintos autores, así como cuáles son las ventajas e inconvenientes que cada uno ostenta, ya que es común que al resolver un problema particular se usen una o más arquitecturas de hardware interconectadas generalmente por red.

Clasificación de Flynn Clasificación clásica de arquitecturas de computadoras que hace alusión a sistemas con uno o varios procesadores, Flynn la publicó por primera vez en 1966 y por segunda vez en 1970.

Esta taxonomía se basa en el flujo que siguen los datos dentro de la máquina y de las instrucciones sobre esos datos. Se define como flujo de instrucciones al conjunto de instrucciones secuenciales que son ejecutadas por un único procesador y como flujo de datos al flujo secuencial de datos requeridos por el flujo de instrucciones.

Con estas consideraciones, Flynn clasifica los sistemas en cuatro categorías:

Single Instruction stream, Single Data stream (SISD) Los sistemas de este tipo se caracterizan por tener un único flujo de instrucciones sobre un único flujo de datos, es decir, se ejecuta una instrucción detrás de otra. Este es el concepto de arquitectura serie de Von Neumann donde, en cualquier

momento, sólo se ejecuta una única instrucción, un ejemplo de estos sistemas son las máquinas secuenciales convencionales.

Single Instruction stream, Multiple Data stream (SIMD) Estos sistemas tienen un único flujo de instrucciones que operan sobre múltiples flujos de datos. Ejemplos de estos sistemas los tenemos en las máquinas vectoriales con hardware escalar y vectorial.

El procesamiento es síncrono, la ejecución de las instrucciones sigue siendo secuencial como en el caso anterior, todos los elementos realizan una misma instrucción pero sobre una gran cantidad de datos. Por este motivo existirá concurrencia de operación, es decir, esta clasificación es el origen de la máquina paralela.

El funcionamiento de este tipo de sistemas es el siguiente. La unidad de control manda una misma instrucción a todas las unidades de proceso (ALUs). Las unidades de proceso operan sobre datos diferentes pero con la misma instrucción recibida.

Existen dos alternativas distintas que aparecen después de realizarse esta clasificación:

- Arquitectura Vectorial con segmentación, una CPU única particionada en unidades funcionales independientes trabajando sobre flujos de datos concretos.
- Arquitectura Matricial (matriz de procesadores), varias ALUs idénticas a las que el procesador da instrucciones, asigna una única instrucción pero trabajando sobre diferentes partes del programa.

Multiple Instruction stream, Single Data stream (MISD) Sistemas con múltiples instrucciones que operan sobre un único flujo de datos. Este tipo de sistemas no ha tenido implementación hasta hace poco tiempo. Los sistemas MISD se contemplan de dos maneras distintas:

- Varias instrucciones operando simultáneamente sobre un único dato.
- Varias instrucciones operando sobre un dato que se va convirtiendo en un resultado que será la entrada para la siguiente etapa. Se trabaja de forma segmentada, todas las unidades de proceso pueden trabajar de forma concurrente.

Multiple Instruction stream, Multiple Data stream (MIMD) Sistemas con un flujo de múltiples instrucciones que operan sobre múltiples datos. Estos sistemas empezaron a utilizarse antes de la década de los 80s. Son sistemas con memoria compartida que permiten ejecutar varios procesos simultáneamente (sistema multiprocesador).

Cuando las unidades de proceso reciben datos de una memoria no compartida estos sistemas reciben el nombre de MULTIPLE SISD (MSISD). En

arquitecturas con varias unidades de control (MISD Y MIMD), existe otro nivel superior con una unidad de control que se encarga de controlar todas las unidades de control del sistema (ejemplo de estos sistemas son las máquinas paralelas actuales).

11.1.2. Categorías de Computadoras Paralelas

Clasificación moderna que hace alusión única y exclusivamente a los sistemas que tienen más de un procesador (i.e máquinas paralelas). Existen dos tipos de sistemas teniendo en cuenta su acoplamiento:

- Los sistemas fuertemente acoplados son aquellos en los que los procesadores dependen unos de otros.
- Los sistemas débilmente acoplados son aquellos en los que existe poca interacción entre los diferentes procesadores que forman el sistema.

Atendiendo a esta y a otras características, la clasificación moderna divide a los sistemas en dos tipos: Sistemas multiprocesador (fuertemente acoplados) y sistemas multicomputadoras (débilmente acoplados).

Multiprocesadores o Equipo Paralelo de Memoria Compartida Un multiprocesador puede verse como una computadora paralela compuesta por varios procesadores interconectados que comparten un mismo sistema de memoria.

Los sistemas multiprocesadores son arquitecturas MIMD con memoria compartida. Tienen un único espacio de direcciones para todos los procesadores y los mecanismos de comunicación se basan en el paso de mensajes desde el punto de vista del programador.

Dado que los multiprocesadores comparten diferentes módulos de memoria, pudiendo acceder a un mismo módulo varios procesadores, a los multiprocesadores también se les llama sistemas de memoria compartida.

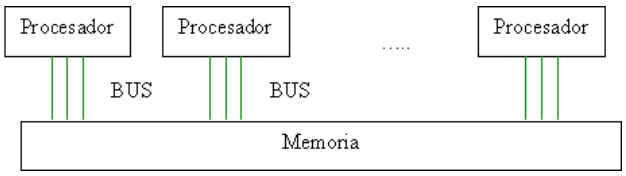

Figura 15: Arquitectura de una computadora paralela con memoria compartida

Para hacer uso de la memoria compartida por más de un procesador, se requiere hacer uso de técnicas de semáforos que mantienen la integridad de la memoria; esta arquitectura no puede crecer mucho en el número de procesadores interconectados por la saturación rápida del bus o del medio de interconexión.

Dependiendo de la forma en que los procesadores comparten la memoria, se clasifican en sistemas multiprocesador UMA, NUMA, COMA y Pipeline.

Uniform Memory Access (UMA) Sistema multiprocesador con acceso uniforme a memoria. La memoria física es uniformemente compartida por todos los procesadores, esto quiere decir que todos los procesadores tienen el mismo tiempo de acceso a todas las palabras de la memoria. Cada procesador tiene su propia caché privada y también se comparten los periféricos.

Los multiprocesadores son sistemas fuertemente acoplados (tightly-coupled), dado el alto grado de compartición de los recursos (hardware o software) y el alto nivel de interacción entre procesadores, lo que hace que un procesador dependa de lo que hace otro.

El sistema de interconexión debe ser rápido y puede ser de uno de los siguientes tipos: bus común, red crossbar y red multietapa. Este modelo es conveniente para aplicaciones de propósito general y de tiempo compartido por varios usuarios, existen dos categorías de sistemas UMA.

- Sistema Simétrico

  Cuando todos los procesadores tienen el mismo tiempo de acceso a todos los componentes del sistema (incluidos los periféricos), reciben el nombre de sistemas multiprocesador simétrico. Los procesadores tienen el mismo dominio (prioridad) sobre los periféricos y cada procesador tiene la misma capacidad para procesar.

- Sistema Asimétrico

  Los sistemas multiprocesador asimétrico, son sistemas con procesadores maestros y procesadores esclavos, en donde sólo los primeros pueden ejecutar aplicaciones y dónde en tiempo de acceso para diferentes procesadores no es el mismo. Los procesadores esclavos (attached) ejecutan código usuario bajo la supervisión del maestro, por lo tanto cuando una aplicación es ejecutada en un procesador maestro dispondrá de una cierta prioridad.

Non Uniform Memory Access (NUMA) Un sistema multiprocesador NUMA es un sistema de memoria compartida donde el tiempo de acceso varía según donde se encuentre localizado el acceso.

El acceso a memoria, por tanto, no es uniforme para diferentes procesadores, existen memorias locales asociadas a cada procesador y estos pueden acceder a datos de su memoria local de una manera más rápida que a las memorias de otros procesadores, debido a que primero debe aceptarse dicho acceso por el procesador del que depende el módulo de memoria local.

Todas las memorias locales conforman la memoria global compartida y físicamente distribuida y accesible por todos los procesadores.

Cache Only Memory Access (COMA) Los sistemas COMA son un caso especial de los sistemas NUMA. Este tipo de sistemas no ha tenido mucha trascendencia, al igual que los sistemas SIMD.

Las memorias distribuidas son memorias cachés, por este motivo es un sistema muy restringido en cuanto a la capacidad de memoria global. No hay jerarquía de memoria en cada módulo procesador. Todas las cachés forman un mismo espacio global de direcciones. El acceso a las cachés remotas se realiza a través de los directorios distribuidos de las cachés.

Dependiendo de la red de interconexión utilizada, se pueden utilizar jerarquías en los directorios para ayudar a la localización de copias de bloques de caché.

Procesador Vectorial Pipeline En la actualidad es común encontrar en un solo procesador los denominados Pipeline o Procesador Vectorial Pipeline del tipo MISD. En estos procesadores los vectores fluyen a través de las unidades aritméticas Pipeline.

Las unidades constan de una cascada de etapas de procesamiento compuestas de circuitos que efectúan operaciones aritméticas o lógicas sobre el flujo de datos que pasan a través de ellas, las etapas están separadas por registros de alta velocidad usados para guardar resultados intermedios. Así la información que fluye entre las etapas adyacentes está bajo el control de un reloj que se aplica a todos los registros simultáneamente.

Multicomputadoras o Equipo Paralelo de Memoria Distribuida Los sistemas multicomputadoras se pueden ver como una computadora paralela en el cual cada procesador tiene su propia memoria local. En estos sistemas la memoria se encuentra distribuida y no compartida como en los sistemas multiprocesador. Los procesadores se comunican a través de paso de mensajes, ya que éstos sólo tienen acceso directo a su memoria local y no a las memorias del resto de los procesadores.

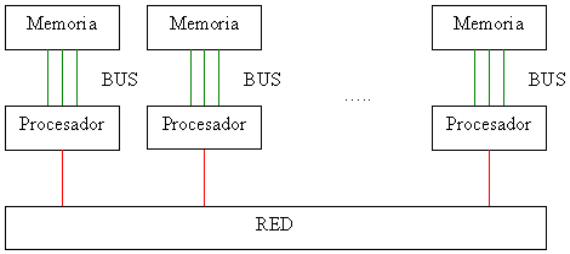

Figura 16: Arquitectura de una computadora paralela con memoria distribuida

La transferencia de los datos se realiza a través de la red de interconexión que conecta un subconjunto de procesadores con otro subconjunto. La transferencia de unos procesadores a otros se realiza por múltiples transferencias entre procesadores conectados dependiendo del establecimiento de dicha red.

Dado que la memoria está distribuida entre los diferentes elementos de proceso, estos sistemas reciben el nombre de distribuidos. Por otra parte, estos sistemas son débilmente acoplados, ya que los módulos funcionan de forma casi independiente unos de otros. Este tipo de memoria distribuida es de acceso lento por ser peticiones a través de la red, pero es una forma muy efectiva de tener acceso a un gran volumen de memoria.

Equipo Paralelo de Memoria Compartida-Distribuida La tendencia actual en las máquinas paralelas es de aprovechar las facilidades de programación que ofrecen los ambientes de memoria compartida y la escalabilidad de las ambientes de memoria distribuida. En este modelo se conectan entre si módulos de multiprocesadores, pero se mantiene la visión global de la memoria a pesar de que es distribuida.

Clusters El desarrollo de sistemas operativos y compiladores del dominio público (Linux y software GNU), estándares para el pase de mensajes (MPI), conexión universal a periféricos (PCI), etc. han hecho posible tomar ventaja de los económicos recursos computacionales de producción masiva (CPU, discos, redes).

La principal desventaja que presenta a los proveedores de multicomputadoras es que deben satisfacer una amplia gama de usuarios, es decir, deben ser generales. Esto aumenta los costos de diseños y producción de equipos, así como los costos de desarrollo de software que va con ellos: sistema operativo, compiladores y aplicaciones. Todos estos costos deben ser añadidos cuando se hace una venta. Por supuesto alguien que sólo necesita procesadores y un mecanismo de pase de mensajes no debería pagar por todos estos añadidos que nunca usará. Estos usuarios son los que están impulsando el uso de clusters principalmente de computadoras personales (PC), cuya arquitectura se muestra a continuación:

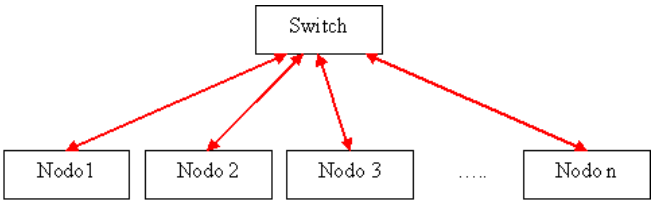

Figura 17: Arquitectura de un cluster

Los cluster se pueden clasificar en dos tipos según sus características físicas:

- Cluster homogéneo si todos los procesadores y/o nodos participantes en el equipo paralelo son iguales en capacidad de cómputo (en la cual es permitido variar la cantidad de memoria o disco duro en cada procesador).

- Cluster heterogéneo es aquel en que al menos uno de los procesadores y/o nodos participantes en el equipo paralelo son de distinta capacidad de cómputo.

Los cluster pueden formarse de diversos equipos; los más comunes son los de computadoras personales, pero es creciente el uso de computadoras multiprocesador de más de un procesador de memoria compartida interconectados por red con los demás nodos del mismo tipo, incluso el uso de computadoras multiprocesador de procesadores vectoriales Pipeline. Los cluster armados con la configuración anterior tienen grandes ventajas para procesamiento paralelo:

• La reciente explosión en redes implica que la mayoría de los componentes necesarios para construir un cluster son vendidos en altos volúmenes y por lo tanto son económicos. Ahorros adicionales se pueden obtener debido a que sólo se necesitará una tarjeta de vídeo, un monitor y un teclado por cluster. El mercado de los multiprocesadores es más reducido y más costoso.

• Remplazar un componente defectuoso en un cluster es relativamente trivial comparado con hacerlo en un multiprocesador, permitiendo una mayor disponibilidad de clusters cuidadosamente diseñados.

Desventajas del uso de clusters de computadoras personales para procesamiento paralelo:

• Con raras excepciones, los equipos de redes generales producidos masivamente no están diseñados para procesamiento paralelo y típicamente su latencia es alta y los anchos de banda pequeños comparados con multiprocesadores. Dado que los clusters explotan tecnología que sea económica, los enlaces en el sistema no son veloces implicando que la comunicación entre componentes debe pasar por un proceso de protocolos de negociación lentos, incrementando seriamente la latencia. En muchos y en el mejor de los casos (debido a costos) se recurre a una red tipo Fast Ethernet restringimiento la escalabilidad del cluster.

• Hay poco soporte de software para manejar un cluster como un sistema integrado.

• Los procesadores no son tan eficientes como los procesadores usados en los multiprocesadores para manejar múltiples usuarios y/o procesos. Esto hace que el rendimiento de los clusters se degrade con relativamente pocos usuarios y/o procesos.

• Muchas aplicaciones importantes disponibles en multiprocesadores y optimizadas para ciertas arquitecturas, no lo están en clusters.

Sin lugar a duda los clusters presentan una alternativa importante para varios problemas particulares, no sólo por su economía, si no también porque

pueden ser diseñados y ajustados para ciertas aplicaciones. Las aplicaciones que pueden sacar provecho de clusters son en donde el grado de comunicación entre procesos es de bajo a medio.

`Tipos de Cluster`

Básicamente existen tres tipo de clusters, cada uno de ellos ofrece ventajas y desventajas, el tipo más adecuado para el cómputo científico es del de alto-rendimiento, pero existen aplicaciones científicas que pueden usar más de un tipo al mismo tiempo.

- Alta-disponibilidad (Fail-over o High-Availability): este tipo de cluster esta diseñado para mantener uno o varios servicios disponibles incluso a costa de rendimiento, ya que su función principal es que el servicio jamás tenga interrupciones como por ejemplo un servicio de bases de datos.
- Alto-rendimiento (HPC o High Performance Computing): este tipo de cluster está diseñado para obtener el máximo rendimiento de la aplicación utilizada incluso a costa de la disponibilidad del sistema, es decir el cluster puede sufrir caídas, este tipo de configuración está orientada a procesos que requieran mucha capacidad de cálculo.
- Balanceo de Carga (Load-balancing): este tipo de cluster esta diseñado para balancear la carga de trabajo entre varios servidores, lo que permite tener, por ejemplo, un servicio de cálculo intensivo multiusuarios que detecte tiempos muertos del proceso de un usuario para ejecutar en dichos tiempos procesos de otros usuarios.

Grids Son cúmulos (grupo de clusters) de arquitecturas en paralelo interconectados por red, los cuales distribuyen tareas entre los clusters que lo forman, estos pueden ser homogéneos o heterogéneos en cuanto a los nodos componentes del cúmulo. Este tipo de arquitecturas trata de distribuir cargas de trabajo acorde a las características internas de cada cluster y las necesidades propias de cada problema, esto se hace a dos niveles, una en la parte de programación conjuntamente con el balance de cargas y otra en la parte de hardware que tiene que ver con las características de cada arquitectura que conforman al cúmulo.

11.2. Métricas de Desempeño

Las métricas de desempeño del procesamiento de alguna tarea en paralelo es un factor importante para medir la eficiencia y consumo de recursos al resolver una tarea con un número determinado de procesadores y recursos relacionados de la interconexión de éstos.

Entre las métricas para medir desempeño en las cuales como premisa se mantiene fijo el tamaño del problema, destacan las siguientes: Factor de aceleración, eficiencia y fracción serial. Cada una de ellas mide algo en particular

y sólo la combinación de estas dan un panorama general del desempeño del procesamiento en paralelo de un problema en particular en una arquitectura determinada al ser comparada con otras.

Factor de Aceleración (o Speed-Up) Se define como el cociente del tiempo que se tarda en completar el cómputo de la tarea usando un sólo procesador entre el tiempo que necesita para realizarlo en p procesadores trabajando en paralelo

$$s = \frac{T(1)}{T(p)} \tag{702}$$

en ambos casos se asume que se usará el mejor algoritmo tanto para un solo procesador como para p procesadores.

Esta métrica en el caso ideal debería de aumentar de forma lineal al aumento del número de procesadores.

Eficiencia Se define como el cociente del tiempo que se tarda en completar el cómputo de la tarea usando un solo procesador entre el número de procesadores multiplicado por el tiempo que necesita para realizarlo en p procesadores trabajando en paralelo

$$e = \frac{T(1)}{pT(p)} = \frac{s}{p}. \tag{703}$$

Este valor será cercano a la unidad cuando el hardware se esté usando de manera eficiente, en caso contrario el hardware será desaprovechado.

Fracción serial Se define como el cociente del tiempo que se tarda en completar el cómputo de la parte secuencial de una tarea entre el tiempo que se tarda el completar el cómputo de la tarea usando un solo procesador

$$f = \frac{T_s}{T(1)} \tag{704}$$

pero usando la ley de Amdahl

$$T(p) = T_s + \frac{T_p}{p}$$

y rescribiéndola en términos de factor de aceleración, obtenemos la forma operativa del cálculo de la fracción serial que adquiere la forma siguiente

$$f = \frac{\frac{1}{s} - \frac{1}{p}}{1 - \frac{1}{p}}. \tag{705}$$

Esta métrica permite ver las inconsistencias en el balance de cargas, ya que su valor debiera de tender a cero en el caso ideal, por ello un incremento en el valor de f es un aviso de granularidad fina con la correspondiente sobrecarga en los procesos de comunicación.

11.3. Cómputo Paralelo para Sistemas Continuos

Como se mostró en los capítulos anteriores, la solución de los sistemas continuos usando ecuaciones diferenciales parciales genera un alto consumo de memoria e involucran un amplio tiempo de procesamiento; por ello nos interesa trabajar en computadoras que nos puedan satisfacer estas demandas.

Actualmente, en muchos centros de cómputo es una práctica común usar directivas de compilación en equipos paralelos sobre programas escritos de forma secuencial, con la esperanza que sean puestos por el compilador como programas paralelos. Esto en la gran mayoría de los casos genera códigos poco eficientes, pese a que corren en equipos paralelos y pueden usar toda la memoria compartida de dichos equipos, el algoritmo ejecutado continua siendo secuencial en la gran mayoría del código.

Si la arquitectura paralela donde se implemente el programa es UMA de acceso simétrico, los datos serán accesados a una velocidad de memoria constante. En caso contrario, al acceder a un conjunto de datos es común que una parte de estos sean locales a un procesador (con un acceso del orden de nano segundos), pero el resto de los datos deberán de ser accesados mediante red (con acceso del orden de mili segundos), siendo esto muy costoso en tiempo de procesamiento.

Por ello, si usamos métodos de descomposición de dominio es posible hacer que el sistema algebraico asociado pueda distribuirse en la memoria local de múltiples computadoras y que para encontrar la solución al problema se requiera poca comunicación entre los procesadores.

Por lo anterior, si se cuenta con computadoras con memoria compartida o que tengan interconexión por bus, salvo en casos particulares no será posible explotar éstas características eficientemente. Pero en la medida en que se adecuen los programas para usar bibliotecas y compiladores acordes a las características del equipo disponible (algunos de ellos sólo existen de manera comercial) la eficiencia aumentará de manera importante.

La alternativa más adecuada (en costo y flexibilidad), es trabajar con computadoras de escritorio interconectadas por red que pueden usarse de manera cooperativa para resolver nuestro problema. Los gastos en la interconexión de los equipos son mínimos (sólo el switch y una tarjeta de red por equipo y cables para su conexión). Por ello los clusters y los grids son en principio una buena opción para la resolución de este tipo de problemas.

Esquema de Paralelización Maestro-Esclavo La implementación de los métodos de descomposición de dominio que se trabajarán será mediante el esquema Maestro-Esclavo (Farmer) en el lenguaje de programación C++ bajo la interfaz de paso de mensajes MPI trabajando en un cluster Linux Debian.

Donde tomando en cuenta la implementación en estrella del cluster, el modelo de paralelismo de MPI y las necesidades propias de comunicación del programa, el nodo maestro tendrá comunicación sólo con cada nodo esclavo y no existirá comunicación entre los nodos esclavos, esto reducirá las comunicaciones y optimizará el paso de mensajes.

El esquema de paralelización maestro-esclavo, permite sincronizar por parte

del nodo maestro las tareas que se realizan en paralelo usando varios nodos esclavos, éste modelo puede ser explotado de manera eficiente si existe poca comunicación entre el maestro y el esclavo y los tiempos consumidos en realizar las tareas asignadas son mayores que los períodos involucrados en las comunicaciones para la asignación de dichas tareas. De esta manera se garantiza que la mayoría de los procesadores estarán trabajando de manera continua y existirán pocos tiempos muertos.

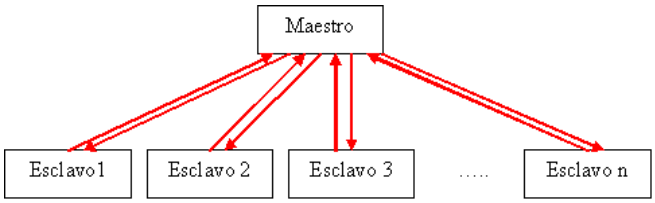

Figura 18: Esquema del maestro-esclavo

Un factor limitante en este esquema es que el nodo maestro deberá de atender todas las peticiones hechas por cada uno de los nodos esclavos, esto toma especial relevancia cuando todos o casi todos los nodos esclavos compiten por ser atendidos por el nodo maestro.

Se recomienda implementar este esquema en un cluster heterogéneo en donde el nodo maestro sea más poderoso computacionalmente que los nodos esclavos. Si a éste esquema se le agrega una red de alta velocidad y de baja latencia, se le permitirá operar al cluster en las mejores condiciones posibles, pero este esquema se verá degradado al aumentar el número de nodos esclavos inexorablemente.

Pero hay que ser cuidadosos en cuanto al número de nodos esclavos que se usan en la implementación en tiempo de ejecución versus el rendimiento general del sistema al aumentar estos, algunas observaciones posibles son:

- El esquema maestro-esclavo programado en C++ y usando MPI lanza P procesos (uno para el nodo maestro y $P-1$ para los nodos esclavos), estos en principio corren en un solo procesador pero pueden ser lanzados en múltiples procesadores usando una directiva de ejecución, de esta manera es posible que en una sola maquina se programe, depure y sea puesto a punto el código usando mallas pequeñas (del orden de cientos de nodos) y cuando este listo puede mandarse a producción en un cluster.

- El esquema maestro-esclavo no es eficiente si sólo se usan dos procesadores (uno para el nodo maestro y otro para el nodo esclavo), ya que el nodo maestro en general no realiza los cálculos pesados y su principal función será la de distribuir tareas; los cálculos serán delegados al nodo esclavo. En el caso que nos interesa implementar, el método de descomposición de dominio adolece de este problema.

Paso de Mensajes Usando MPI Para poder intercomunicar al nodo maestro con cada uno de los nodos esclavos se usa la interfaz de paso de mensajes (MPI), una biblioteca de comunicación para procesamiento en paralelo. MPI ha sido desarrollado como un estándar para el paso de mensajes y operaciones relacionadas.

Este enfoque es adoptado por usuarios e implementadores de bibliotecas, en la cual se proveen a los programas de procesamiento en paralelo de portabilidad y herramientas necesarias para desarrollar aplicaciones que puedan usar el cómputo paralelo de alto desempeño.

El modelo de paso de mensajes posibilita a un conjunto de procesos que tienen solo memoria local la comunicación con otros procesos (usando Bus o red) mediante el envío y recepción de mensajes. Por definición el paso de mensajes posibilita transferir datos de la memoria local de un proceso a la memoria local de cualquier otro proceso que lo requiera.

En el modelo de paso de mensajes para equipos paralelos, los procesos se ejecutan en paralelo, teniendo direcciones de memoria separada para cada proceso, la comunicación ocurre cuando una porción de la dirección de memoria de un proceso es copiada mediante el envío de un mensaje dentro de otro proceso en la memoria local mediante la recepción del mismo.

Las operaciones de envío y recepción de mensajes es cooperativa y ocurre sólo cuando el primer proceso ejecuta una operación de envío y el segundo proceso ejecuta una operación de recepción, los argumentos base de estas funciones son:

- Para el que envía, la dirección de los datos a transmitir y el proceso destino al cual los datos se enviarán.
- Para el que recibe, debe de tener la dirección de memoria donde se pondrán los datos recibidos, junto con la dirección del proceso del que los envío.

Es decir:
Send(dir, lg, td, dest, etiq, com)

$\{dir, lg, td\}$ describe cuántas ocurrencias *lg* de elementos del tipo de dato *td* se transmitirán empezando en la dirección de memoria *dir*.

$\{des, etiq, com\}$ describe el identificador *etq* de destino *des* asociado con la comunicación *com*.

Recv(dir, mlg, td, fuent, etiq, com, st)

$\{dir, lg, td\}$ describe cuántas ocurrencias *lg* de elementos del tipo de dato *td* se transmitirán empezando en la dirección de memoria *dir*.

$\{fuent, etiq, com, est\}$ describe el identificador *etq* de la fuente *fuent* asociado con la comunicación *com* y el estado *st*.

El conjunto básico de directivas (en nuestro caso sólo se usan estas) en C++ de MPI son:

MPI::Init	Inicializa al MPI
MPI::COMM_WORLD.Get_size	Busca el número de procesos existentes
MPI::COMM_WORLD.Get_rank	Busca el identificador del proceso
MPI::COMM_WORLD.Send	Envía un mensaje
MPI::COMM_WORLD.Recv	Recibe un mensaje
MPI::Finalize	Termina al MPI

Estructura del Programa Maestro-Esclavo La estructura del programa se realizo para que el nodo maestro mande trabajos de manera síncrona a los nodos esclavos. Cuando los nodos esclavos terminan la tarea asignada, avisan al nodo maestro para que se le asigne otra tarea (estas tareas son acordes a la etapa correspondiente del método de descomposición de dominio ejecutándose en un instante dado). En la medida de lo posible se trata de mandar paquetes de datos a cada nodo esclavo y que estos regresen también paquetes al nodo maestro, a manera de reducir las comunicaciones al mínimo y tratar de mantener siempre ocupados a los nodos esclavos para evitar los tiempos muertos, logrando con ello una granularidad gruesa, ideal para trabajar con clusters.

La estructura básica del programa bajo el esquema maestro-esclavo codificada en C++ y usando MPI es:

```
main(int argc, char *argv[])
{
    MPI::Init(argc,argv);
    ME_id = MPI::COMM_WORLD.Get_rank();
    MP_np = MPI::COMM_WORLD.Get_size();
    if (ME_id == 0) {
        // Operaciones del Maestro
    } else {
        // Operaciones del esclavo con identificador ME_id
    }
    MPI::Finalize();
}
```

En este único programa se deberá de codificar todas las tareas necesarias para el nodo maestro y cada uno de los nodos esclavos, así como las formas de intercomunicación entre ellos usando como distintivo de los distintos procesos a la variable *ME_id*. Para más detalles de esta forma de programación y otras funciones de MPI ver [21] y [6].

Los factores limitantes para el esquema maestro-esclavo pueden ser de dos tipos, los inherentes al propio esquema maestro-esclavo y al método de descomposición de dominio:

• El esquema de paralelización maestro-esclavo presupone contar con un nodo maestro lo suficientemente poderoso para atender simultáneamente las tareas síncronas del método de descomposición de dominio, ya que este distribuye tareas acorde al número de subdominios, estas si son balanceadas ocasionaran que todos los procesadores esclavos terminen al mismo tiempo y el nodo maestro tendrá que atender múltiples comunicaciones simultáneamente, degradando su rendimiento al aumentar el número de subdominios.

• Al ser síncrono el método de descomposición de dominio, si un nodo esclavo acaba la tarea asignada y avisa al nodo maestro, este no podrá asignarle otra tarea hasta que todos los nodos esclavos concluyan la suya.

Para los factores limitantes inherente al propio esquema maestro-esclavo, es posible implementar algunas operaciones del nodo maestro en paralelo, ya sea usando equipos multiprocesador o en más de un nodo distintos a los nodos esclavos.

Para la parte inherente al método de descomposición de dominio, la parte medular la da el balanceo de cargas. Es decir que cada nodo esclavo tenga una carga de trabajo igual al resto de los nodos. Este balanceo de cargas puede no ser homogéneo por dos razones:

- Al tener P procesadores en el equipo paralelo, la descomposición del dominio no sea la adecuada.
- Si se tiene una descomposición particular, esta se implemente en un número de procesadores inadecuado.

Cualquiera de las dos razones generarán desbalanceo de la carga en los nodos esclavos, ocasionando una perdida de eficiencia en el procesamiento de un problema bajo una descomposición particular en una configuración del equipo paralelo especifica, es por esto que en algunos casos al aumentar el número de procesadores que resuelvan la tarea no se aprecia una disminución del de tiempo de procesamiento.

El número de procesadores P que se usen para resolver un dominio Ω y tener buen balance de cargas puede ser conocido si aplicamos el siguiente procedimiento: Si el dominio Ω se descompone en $n \times m$ subdominios (la partición gruesa), entonces se generarán $s = n * m$ subdominios Ω_i, en este caso, se tiene un buen balanceo de cargas si $(P-1) \mid s$. La partición fina se obtiene al descomponer a cada subdominio Ω_i en $p \times q$ subdominios.

Como ejemplo, supongamos que deseamos resolver el dominio Ω usando 81×81 nodos ($nodos = n*p+1$ y $nodos = m*q+1$), de manera inmediata nos surgen las siguientes preguntas: ¿cuales son las posibles descomposiciones validas? y ¿en cuantos procesadores se pueden resolver cada descomposición?. Para este ejemplo, sin hacer la tabla exhaustiva obtenemos:

Partición	Subdominios	Procesadores
1x2 y 80x40	2	2,3
1x4 y 80x20	5	2,5
1x5 y 80x16	6	2,6
2x1 y 40x80	2	2,3
2x2 y 40x40	4	2,3,5
2x4 y 40x20	8	2,3,5,9
2x5 y 40x16	10	2,3,6,11
2x8 y 40x10	16	2,3,5,9,17
4x1 y 20x80	4	2,3,5
4x2 y 20x40	8	2,3,5,9
4x4 y 20x20	16	2,3,5,9,17
4x5 y 20x16	20	2,3,5,6,11,21
5x1 y 16x80	5	2,6
5x2 y 16x40	10	2,3,6,11
5x4 y 16x20	20	2,3,5,6,11,21
5x5 y 16x16	25	2,6,26

De esta tabla es posible seleccionar (para este ejemplo en particular), las descomposiciones que se adecuen a las necesidades particulares del equipo con que se cuente. Sin embargo hay que tomar en cuenta siempre el número de nodos por subdominio de la partición fina, ya que un número de nodos muy grande puede que exceda la cantidad de memoria que tiene el nodo esclavo y un número pequeño estaría infrautilizando el poder computacional de los nodos esclavos. De las particiones seleccionadas se pueden hacer corridas de prueba para evaluar su rendimiento, hasta encontrar la que menor tiempo de ejecución consuma, maximizando así la eficiencia del equipo paralelo.

Programación Paralela en Multihilos En una computadora, sea secuencial o paralela, para aprovechar las capacidades crecientes del procesador, el sistema operativo divide su tiempo de procesamiento entre los distintos procesos, de forma tal que para poder ejecutar a un proceso, el kernel les asigna a cada proceso una prioridad y con ello una fracción del tiempo total de procesamiento, de forma tal que se pueda atender a todos y cada uno de los procesos de manera eficiente.

En particular, en la programación en paralelo usando MPI, cada proceso (que eventualmente puede estar en distinto procesador) se lanza como una copia del programa con datos privados y un identificador del proceso único, de tal forma que cada proceso sólo puede compartir datos con otro proceso mediante paso de mensajes.

Esta forma de lanzar procesos por cada tarea que se desee hacer en paralelo es costosa, por llevar cada una de ellas todo una gama de subprocesos para poderle asignar recursos por parte del sistema operativo. Una forma más eficiente de hacerlo es que un proceso pueda generar bloques de subprocesos que puedan ser ejecutados como parte del proceso (como subtareas), así en el tiempo asignado

se pueden atender a más de un subproceso de manera más eficiente, esto es conocido como programación multihilos.

Los hilos realizarán las distintas tareas necesarias en un proceso. Para hacer que los procesos funcionen de esta manera, se utilizan distintas técnicas que le indican kernel cuales son las partes del proceso que pueden ejecutarse simultáneamente y el procesador asignará una fracción de tiempo exclusivo al hilo del tiempo total asignado al proceso.

Los datos pertenecientes al proceso pasan a ser compartidos por los subprocesos lanzados en cada hilo y mediante una técnica de semáforos el kernel mantiene la integridad de estos. Esta técnica de programación puede ser muy eficiente si no se abusa de este recurso, permitiendo un nivel más de paralelización en cada procesador. Esta forma de paralelización no es exclusiva de equipos multiprocesadores o multicomputadoras, ya que pueden ser implementados a nivel de sistema operativo.

12. 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] 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.

[12] J. L. Lions & E. Magenes; *Non-Homogeneous Bonduary Value Problems and Applications Vol. I*, Springer-Verlag Berlin Heidelber New York 1972.

[13] K. Hutter & K. Jöhnk; *Continuum Methods of Physical Modeling.* Springer-Verlag Berlin Heidelber New York 2004.

[14] L. F. Pavarino, A. Toselli; *Recent Developments in Domain Decomposition Methods*. Springer, 2003.

[15] M.B. Allen III, I. Herrera & G. F. Pinder; *Numerical Modeling in Science And Engineering.* John Wiley & Sons, Inc . 1988.

[16] M. Diaz; *Desarrollo del Método de Colocación Trefftz-Herrera Aplicación a Problemas de Transporte en las Geociencias.* Tesis Doctoral, Instituto de Geofísica, UNAM, 2001.

[17] 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.

[18] P.G. Ciarlet, J. L. Lions; *Handbook of Numerical Analysis, Vol. II.* North-Holland, 1991.

[19] R. L. Burden y J. D. Faires; *Análisis Numérico.* Math Learning, 7 ed. 2004.

[20] S. Friedberg, A. Insel, and L. Spence; *Linear Algebra*, 4th Edition, Prentice Hall, Inc. 2003.

[21] W. Gropp, E. Lusk, A. Skjellem, *Using MPI, Portable Parallel Programming Whit the Message Passing Interface.* Scientific and Engineering Computation Series, 2ed, 1999.

[22] W. Rudin; *Principles of Mathematical Analysis.* McGraw-Hill International Editions, 1976.

[23] X. O. Olivella, C. A. de Sacribar; *Mecánica de Medios Continuos para Ingenieros.* Ediciones UPC, 2000.

[24] Y. Saad; *Iterative Methods for Sparse Linear Systems.* SIAM, 2 ed. 2000.

[25] Y. Skiba; *Métodos y Esquemas Numéricos, un Análisis Computacional.* UNAM, 2005.