Trabajo de Investigación A: Unificación y Simplificación de los Métodos Dual-Primal de Descomposición de Dominio

Antonio Carrillo Ledesma

9 de diciembre de 2008

Índice

1. Métodos de Funciones Discontinuas Definidas por Tramos **5**
1.1. Algortimos a Nivel Continuo 6
1.1.1. Algoritmo Neumann-Neumann 6
1.1.2. Algoritmo Dirichlet-Dirichlet 8
1.2. Discretización Axiomática 9
1.3. Esquema General 12
1.4. Procedimiento para Evaluar la Transformación de Componentes . 18
1.5. Métodos Dual-Primal 21

2. Unificación y Simplificación de los Métodos Dual-Primal de Descomposición de Dominio **23**
2.1. Espacio Dual-Primal 23
2.1.1. Espacio de Vectores 24
2.1.2. La Inmersión Natural 28
2.1.3. La Formulación Matricial Discontinua Libre de Multiplicadores de Lagrange 29
2.1.4. Construcción de la Matriz $\underline{\underline{A}}$ 31
2.2. Fórmula de Green-Herrera para Matrices 33
2.3. El Operador de Steklov-Poincaré 34
2.4. Formulación en Término de la Matriz de Complemento de Schur 36
2.4.1. Relación con la Formulación de Multiplicadores de Lagrange 37
2.4.2. Espacio de Vectores Armónicos 38
2.5. Métodos basados en la Formulación Matricial Discontinua 39
2.5.1. Métodos Single-Trip 40
2.5.2. Métodos Round-Trip 42
2.5.3. Caso Cuando Todos los Nodos Primales son Interiores . . 44
2.5.4. Caso Cuando no Todos los Nodos Primales son Interiores 47
2.6. Problemas del tipo del Operador Laplaciano 48
2.6.1. Cálculo de la inversa de $\underline{\underline{M}}$ 50

2.6.2. Cálculo de $\underline{\underline{I}}_S^C$, $\left(\underline{\underline{j}}^w\right)^{-1}$ y $\underline{\underline{k}}^w$ 52

3. Implementación Computacional de los Métodos Round-Trip 54
3.1. Discretización del Espacio 54
3.2. Construcción de la Solución Particular 57
3.3. Métodos Single-Trip 57
3.3.1. El Enfoque Dirichlet 58
3.3.2. El Enfoque Neumann 59
3.4. Métodos Round-Trip 59
3.4.1. El enfoque Dirichlet-Dirichlet 60
3.4.2. El enfoque Neumann-Neumann 61
3.5. Consideraciones Computacionales 61
3.5.1. Cálculo de la Matriz $\underline{\underline{S}}$ 62
3.5.2. Cálculo de los Nodos Interiores 63
3.5.3. Cálculo de la Matriz $\underline{\underline{S}}^{-1}$ 64

4. Apéndice A 65
4.1. El Cómputo en Paralelo 65
4.1.1. Arquitecturas de Software y Hardware 65
4.1.2. Categorías de Computadoras Paralelas 67
4.2. Métricas de Desempeño 72
4.3. Cómputo Paralelo para Sistemas Continuos 74

5. Apéndice B 81
5.1. Solución de Grandes Sistemas de Ecuaciones 81
5.1.1. Métodos Directos 81
5.1.2. Métodos Iterativos 83
5.2. Precondicionadores 88
5.2.1. Gradiente Conjugado Precondicionado 90
5.2.2. Precondicionador a Posteriori 92
5.2.3. Precondicionador a Priori 95

6. Bibliografía 98

Introducción

Una característica del método de elemento finito (FEM) y de muchos otros métodos numéricos para ecuaciones diferenciales parciales es el uso -después de que la partición del domino del problema ha sido introducida- de las funciones de base y funciones de peso definidas por tramos, i.e. ellas son definidas separadamente en cada uno de los subdominios de la partición. Pero como las funciones son definidas independientemente en cada uno de los subdominios de la partición y, sobre la frontera común de dos subdominios, los límites de uno y otro lado no necesariamente coinciden.

Por lo consiguiente, una teoría general y sistemática del método de elemento finito debe de ser reformulada en espacios de funciones en los cuales las funciones de peso y base puedan ser totalmente discontinuas a través de la frontera interior. Dicha teoría puede ser incluida en el método de Galerkin discontinuo (dG) y nos permite movernos suavemente sin interrupciones del método de elemento finito estándar -basado en funciones continuas definidas por tramos- al método de Galerkin discontinuo.

La mayoría de los métodos FEM que existen en el presente usan funciones definidas por tramos con cierto grado de continuidad; típicamente, para una ecuación elíptica de segundo orden las funciones son tomadas del espacio de Sobolev $H^1(\Omega)$, en el cual las funciones son continuas con posibles discontinuidades en la derivada de primer orden [2].

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 y son de los métodos de dominios ajenos mejor conocidos y más probados para la resolución de ecuaciones diferenciales parciales elípticas. Ellos son métodos iterativos de subestructuración que genera precondicionadores a priori dentro del mismo método 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, haciéndolo eficiente.

El método FETI usa lo que se denomina nodos duales-primales y dependiendo del número de estos y su interacción entre subdominios optimiza el de-

sempeño del método, ya que genera un mejor precondicionador a priori para el problema, pudiendo ser este casi no dependiente del tamaño de la malla. Una cosa particular que tienen este método es que usan multiplicadores de Lagrange y que requieren el generar varias matrices auxiliares para forzar que las matrices que intervienen en el método CGM sean positivas definidas aun si el problema original es tipo silla.

El método en el que se trabaja (desarrollado por el Dr. Herrera y sus colaboradores), que permitirá unificar conceptos de métodos numéricos para la resolución de ecuaciones diferenciales parciales elípticas mostrando cómo problemas de valores en la frontera con saltos preescritos pueden formularse como un problema con valores en la frontera estándar. Derivando y mostrando como construir las funciones de peso y de prueba de forma tal que sean funciones definidas por tramos, las cuales sean totalmente discontinuas y no usen a los multiplicadores de Lagrange. Y en el cual no es necesario la construcción de matrices auxiliares para que las matrices resultantes sean positivas definidas si el problema original es tipo silla, esto genera un ahorro importante en memoria, tiempo de ejecución y genera un método robusto y flexible.

Así, cuando las ecuaciones diferenciales parciales son formuladas en funciones discontinuas definidas por tramos, los problemas bien planteados son problemas de valor en la frontera con saltos prescritos (BVPJ), en los cuales las condiciones de frontera son complementadas por adecuadas condiciones de salto, que satisfacen los cruces en la frontera interior asociados con la partición del dominio. Un resultado mostrado en el trabajo de Herrera [1], muestra que para problemas elípticos de orden 2m, con $m \geq 1$, el BVPJ satisface existencia si y sólo si el problema estándar de valor en la frontera suave lo satisface. Por lo tanto, este resultado esencialmente reduce el problema de establecer condiciones para la existencia de la solución para el problema de valores en la frontera con saltos preescritos a un problema con valores en la frontera estándar.

El trabajo se centra en desarrollar un método libre de multiplicadores de Lagrange que permita trabajar problemas elípticos y parabólicos, tanto lineales como no lineales y generar esquemas de precondicionamiento a priori para este método, el método desarrollado se comparará con los métodos más usado y con mejor desempeño computacional para las aplicaciones que nos interesan.

1. Métodos de Funciones Discontinuas Definidas por Tramos

Consideremos la ecuación de Poisson en un dominio Ω y $\Pi = \{\Omega_1, ..., \Omega_E\}$ una descomposición en subdominios del dominio Ω y asumiendo condiciones de frontera tipo Dirichlet igual a cero, i.e.

$$\begin{aligned} -\Delta \bar{u} &= f_\Omega \text{ en } \Omega \\ \bar{u} &= 0 \text{ sobre } \partial\Omega \end{aligned} \tag{1}$$

el espacio $\overline{D}$, donde la solución $\bar{u}$ es buscada, se define como

$$\overline{D} = \left\{ v \in H^2(\Omega) \mid traza\ v = 0 \text{ sobre } \partial\Omega \right\} \tag{2}$$

otro espacio que usaremos en lo sucesivo es

$$\widetilde{D} = \left\{ v \in \hat{H}^2(\Omega, \Pi) \mid traza\ v = 0 \text{ sobre } \partial\Omega \right\}. \tag{3}$$

Cuando el dato f_Ω es tal que la solución $\bar{u}$ pertenece a $\overline{D}$, entonces este problema es equivalente a el siguiente problema con valores en la frontera con saltos preescritos (BVPJ):

Encontrar $\tilde{u} \in \tilde{D}$ tal que

$$\begin{aligned} -\Delta \tilde{u} &= f_\Omega \text{ en } \Omega_\alpha, \alpha = 1, ..., E \\ [[\tilde{u}]] &= \left[\left[\frac{\partial \tilde{u}}{\partial n}\right]\right] = 0 \text{ en } \Gamma. \end{aligned} \tag{4}$$

aquí las condiciones de frontera no aparecen ya que estas fueron incorporadas en la definición del espacio $\tilde{D}$, más precisamente, $\tilde{u} \in \tilde{D}$ satisface la ecuación (5) si y sólo si $\tilde{u} = u$.

Se desarrollarán dos maneras de aproximar el problema dado por la Ec. (4):

A.- Una manera es introducir una función auxiliar $\tilde{u}_p \in \tilde{D}$ que satisfaga

$$\begin{aligned} -\Delta \tilde{u}_p &= f_\Omega \text{ en } \Omega_\alpha, \alpha = 1, ..., E \\ [[\tilde{u}_p]] &= 0 \text{ y } \widehat{\dot{\tilde{u}}_p} = 0 \text{ sobre } \Gamma \end{aligned} \tag{5}$$

entonces, si $u \equiv \tilde{u} - \tilde{u}_p$ obtenemos para $u \in \tilde{D}$ las ecuaciones

$$\begin{aligned} -\Delta u &= 0 \text{ en } \Omega_\alpha, \alpha = 1, ..., E \\ [[u]] &= 0 \text{ y } \left[\left[\frac{\partial u}{\partial n}\right]\right] = -\left[\left[\frac{\partial \tilde{u}_p}{\partial n}\right]\right] \text{ en } \Gamma. \end{aligned} \tag{6}$$

B.- Otra opción es remplazar la ecuación (5) por

$$\begin{aligned} -\Delta \tilde{u}_p &= f_\Omega \text{ en } \Omega_\alpha, \alpha = 1, ..., E \\ \left[\left[\frac{\partial \tilde{u}_p}{\partial n}\right]\right] &= 0 \text{ y } \widehat{\frac{\partial \tilde{u}_p}{\partial n}} = 0 \text{ sobre } \Gamma \end{aligned} \tag{7}$$

en cuyo caso

$$\begin{aligned} -\Delta u &= 0 \text{ en } \Omega_\alpha, \alpha = 1, ..., E \\ [[u]] &= -[[\tilde{u}_p]] \text{ y } \left[\left[\frac{\partial u}{\partial n}\right]\right] = 0 \text{ en } \Gamma. \end{aligned} \tag{8}$$

La primera aproximación da lugar al algoritmo Neumann-Neumann y la segunda aproximación da lugar al algoritmo Dirichlet-Dirichlet. Independientemente de la aproximación elegida, se están buscando funciones en el espacio lineal

$$D \equiv \left\{ u \in \tilde{D} \mid -\Delta u = 0 \text{ en } \Omega_\alpha; \alpha = 1, 2, ..., E \right\}. \tag{9}$$

1.1. Algortimos a Nivel Continuo

En esta sección, los algoritmos mostrados en la sección anterior, serán presentados a nivel discreto en una manera en que pueden ser aplicados a cualquier número de dimensiones y a cualquier número de particiones del subdominio Ω, incluyendo particiones con vértices y operadores diferenciales que sean positivos pero no positivos definidos. Aprovechando la ventaja del tipo de sistema discreto que se obtiene se usará en el algoritmo el método para resolución de sistemas lineales Gradiente Conjugado, ver sección (5.1.2).

1.1.1. Algoritmo Neumann-Neumann

Construir:

1.- $\tilde{u}_p \in \tilde{D}$ que satisfaga

$$\begin{aligned} -\Delta \tilde{u}_p &= f_\Omega \text{ en } \Omega_\alpha; \alpha = 1, ..., E \\ [[\tilde{u}_p]] &= 0 \text{ y } \widehat{\tilde{u}_p} = 0 \text{ en } \Gamma \end{aligned} \tag{10}$$

$\tilde{u} \in \tilde{D}$ que satisfaga

$$\begin{aligned} -\Delta \tilde{u} &= f_\Omega \text{ en } \Omega_\alpha; \alpha = 1, ..., E \\ [[\tilde{u}]] &= \left[\left[\frac{\partial \tilde{u}}{\partial n}\right]\right] = 0 \text{ en } \Gamma \end{aligned} \tag{11}$$

y definimos $u = \tilde{u} - \tilde{u}_p$, donde $u = u_{21} \oplus u_{22}$.

2.- Construir $u_{21} \in D$, tal que

$$\left[\left[\frac{\partial u_{21}}{\partial n}\right]\right] = -\left[\left[\frac{\partial \tilde{u}_p}{\partial n}\right]\right] \text{ y } \dot{\widehat{\frac{\partial u_{21}}{\partial n}}} = 0 \text{ en } \Gamma \tag{12}$$

3.- Definiendo $r^0 \in D$ tal que

$$\left[\left[r^0\right]\right] = 0 \text{ y } \dot{\widehat{r^0}} = \dot{\widehat{u_{21}}} \text{ en } \Gamma \tag{13}$$

sea $p^0 = r^0$ y $u^0 = 0$.

Usando el método de Gradiente Conjugado, entonces para $n = 0, 1, 2, ...$

4.- Construir $\psi^n \in D$ tal que

$$\left[\left[\frac{\partial \psi^n}{\partial n}\right]\right] = 0 \text{ y } \dot{\widehat{\frac{\partial \psi^n}{\partial n}}} = \dot{\widehat{\frac{\partial p^n}{\partial n}}} \text{ en } \Gamma \tag{14}$$

5.-

$$\alpha^n = \frac{p^n \cdot p^n}{p^n \cdot p^n + \psi^n \cdot \psi^n} \tag{15}$$

6.-

$$u^{n+1} = u^n + \alpha^n p^n \tag{16}$$

7.- Además, construir $q^n \in D$ tal que

$$[[q^n]] = 0 \text{ y } \dot{\widehat{q^n}} = \dot{\widehat{\psi^n}} \text{ en } \Gamma \tag{17}$$

8.-

$$r^{n+1} = r^n - \alpha^n q^n \tag{18}$$

9.-

$$\beta^n = \frac{r^{n+1} \cdot r^{n+1}}{r^n \cdot r^n} \tag{19}$$

10.-

$$p^{n+1} = r^{n+1} + \beta^n p^n \tag{20}$$

11.-

$$n = n + 1, \tag{21}$$

y regresar a 4.

1.1.2. Algoritmo Dirichlet-Dirichlet

Construir:

1.- $\tilde{u}_p \in \tilde{D}$ que satisfaga

$$\begin{aligned} -\Delta\tilde{u}_p &= f_\Omega \text{ en } \Omega_\alpha; \alpha = 1, ..., E \\ \left[\left[\frac{\tilde{u}_p}{\partial n}\right]\right] &= 0 \text{ y } \dot{\widehat{\frac{\tilde{u}_p}{\partial n}}} = 0 \text{ en } \Gamma \end{aligned} \tag{22}$$

$\tilde{u} \in \tilde{D}$ que satisfaga

$$\begin{aligned} -\Delta\tilde{u} &= f_\Omega \text{ en } \Omega_\alpha; \alpha = 1, ..., E \\ [[u]] &= -[[\tilde{u}_p]] \text{ y } \left[\left[\frac{\partial u}{\partial n}\right]\right] = 0 \text{ en } \Gamma. \end{aligned} \tag{23}$$

y definimos $u = \tilde{u} - \tilde{u}_p$, donde $u = u_{11} \oplus u_{12}$.

2.- Construir $u_{11} \in D$, tal que

$$[[u_{11}]] = -[[\tilde{u}_p]] \text{ y } \dot{\widehat{u_{11}}} = 0 \text{ en } \Gamma \tag{24}$$

3.- Definiendo $r^0 \in D$ tal que

$$\left[\left[\frac{\partial r^0}{\partial n}\right]\right] = 0 \text{ y } \dot{\widehat{\frac{\partial r^0}{\partial n}}} = \dot{\widehat{\frac{\partial u_{11}}{\partial n}}} \text{ en } \Gamma \tag{25}$$

sea $p^0 = r^0$ y $u^0 = 0$.

Usando el método de Gradiente Conjugado, entonces para $n = 0, 1, 2, ...$

4.- Construir $\psi^n \in D$ tal que

$$[[\psi^n]] = 0 \text{ y } \dot{\widehat{\psi^n}} = \dot{\widehat{p^n}} \text{ en } \Gamma \tag{26}$$

5.-

$$\alpha^n = \frac{p^n \cdot p^n}{p^n \cdot p^n + \psi^n \cdot \psi^n} \tag{27}$$

6.-

$$u^{n+1} = u^n + \alpha^n p^n \tag{28}$$

7.- Además, construir $q^n \in D$ tal que

$$\left[\left[\frac{\partial q^n}{\partial n}\right]\right] = 0 \text{ y } \dot{\widehat{\frac{\partial q^n}{\partial n}}} = \dot{\widehat{\frac{\partial \psi^n}{\partial n}}} \text{ en } \Gamma \tag{29}$$

8.-

$$r^{n+1} = r^n - \alpha^n q^n \tag{30}$$

9.-

$$\beta^n = \frac{r^{n+1} \cdot r^{n+1}}{r^n \cdot r^n} \tag{31}$$

10.-

$$p^{n+1} = r^{n+1} + \beta^n p^n \tag{32}$$

11.-

$$n = n + 1, \tag{33}$$

y regresar a 4.

Observación 1 *Notemos que en estas formulaciones son tales que de manera directa se obtienen transformaciones positivo definidas sin recurrir a los multiplicadores de Lagrange.*

1.2. Discretización Axiomática

Sea $\overline{D}$ un espacio de Hilbert de funciones de dimensión finita definido en Ω de dimensión $\overline{N}$, sea $\Pi = \{\Omega_1, ..., \Omega_E\}$ una partición, definiendo, para cada $\alpha = 1, ..., E$,

$$D(\Omega_\alpha) = \left\{v \mid v = u_{|_{\Omega_\alpha}} \text{ y } u \in \overline{D}\right\} \tag{34}$$

entonces, escribimos

$$\widetilde{D} = D(\Omega_1) \oplus ... \oplus D(\Omega_E) \tag{35}$$

por lo tanto, $\widetilde{D}$ es un espacio de funciones definidas por pedazos y bajo la inmersión natural de $\overline{D}$ dentro de $\widetilde{D}$, tenemos que $\overline{D} \subset \widetilde{D}$. Una función $\widetilde{w} \in \widetilde{D}(\Omega)$ se dice que tiene soporte local cuando existe un $\alpha \in \{1, ..., E\}$ tal que el soporte de $\widetilde{w}$ esta contenido en la clausura de Ω_α.

Dada cualquier función $\overline{w} \in \overline{D}$, decimos que una función $\widetilde{w} \in \widetilde{D}$ es hija de $\overline{w}$, cuando $\widetilde{w}$ es la restricción de $\overline{w}$ a un subdominio de la partición, claramente, todas las hijas de una función $\overline{w} \in \overline{D}$ tienen soporte local. En cuanto al producto interior en estos espacios, asumiremos que ellos satisfacen

$$u \cdot w = \sum_{\alpha=1}^{E} u_\alpha \cdot w_\alpha \tag{36}$$

donde, $u = \left\{u^1, ..., u^E\right\}$ y $w = \left\{w^1, ..., w^E\right\}$.

Sea $\overline{\mathcal{B}} \subset \overline{D}$ una base de $\overline{D}$

$$\overline{\mathcal{B}} = \left\{\overline{w}^1, ..., \overline{w}^{\overline{N}}\right\} \tag{37}$$

entonces para cada $i = 1, ..., \overline{N}$, $\mathcal{B}^i \subset \overline{D}$ es una colección de hijas de $\overline{w}^i$. Además escribimos

$$\mathcal{B} = \bigcup_{i=1}^{\overline{N}} \mathcal{B}^i \subset \widetilde{D} \tag{38}$$

claramente, lo elementos de $\mathcal{B}$ tienen soporte local, y también asumimos que $\mathcal{B}$, como fue definida es una base linealmente independiente de $\widetilde{D}$.

La colección de conjuntos $\left\{\mathcal{B}^1, ..., \mathcal{B}^{\overline{N}}\right\}$ es clasificado en dos subfamilias $\left\{\mathcal{B}_I^1, ..., \mathcal{B}_I^{N_I}\right\} \subset \left\{\mathcal{B}^1, ..., \mathcal{B}^{\overline{N}}\right\}$ y $\left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{N_\Gamma}\right\} \subset \left\{\mathcal{B}^1, ..., \mathcal{B}^{\overline{N}}\right\}$; ellos han sido definidos por las siguientes condiciones $\mathcal{B}^i \in \left\{\mathcal{B}_I^1, ..., \mathcal{B}_I^{N_I}\right\}$ si y sólo si la cardinalidad $\mathcal{B}^i$ es uno, $\mathcal{B}^i \in \left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{N_\Gamma}\right\}$ si y sólo si la cardinalidad de $\mathcal{B}^i$ es más grande que uno. Entonces

$$\left\{\mathcal{B}^1, ..., \mathcal{B}^{\overline{N}}\right\} = \left\{\mathcal{B}_I^1, ..., \mathcal{B}_I^{N_I}\right\} \cup \left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{N_\Gamma}\right\} \tag{39}$$

definiendo

$$\mathcal{B}_I = \bigcup_{i=1}^{N_I} \mathcal{B}_I^i \qquad \text{y} \qquad \mathcal{B}_\Gamma = \bigcup_{i=1}^{N_\Gamma} \mathcal{B}_\Gamma^i \tag{40}$$

tal que

$$\mathcal{B} = \mathcal{B}_I \cup \mathcal{B}_\Gamma. \tag{41}$$

Ahora definimos una familia de conjuntos $\mathcal{B}_\Gamma^i$, $i = 1, ..., \overline{N}_\Gamma$, donde cada conjunto $\overline{\mathcal{B}}_\Gamma^i$ es definido al remplazar el conjunto $\mathcal{B}_\Gamma^i$ por un conjunto equivalente linealmente independiente (en el sentido que cada uno de $\mathcal{B}_\Gamma^i$ y $\overline{\mathcal{B}}_\Gamma^i$ generan el mismo espacio lineal). La notación siguiente es adoptada

$$\mathcal{B}_\Gamma^i = \left\{w_M^i, w_{J_1}^i, ..., w_{J_{m(i)}}^i\right\} \tag{42}$$

donde w_M^i es definida como la función madre de la función $\overline{w}_M^i \in \overline{\mathcal{B}} \subset \overline{D}$, i.e.

$$w_M^i \equiv \overline{w}^i \in \overline{\mathcal{B}} \subset \overline{D} \tag{43}$$

además, el conjunto

$$\mathcal{B}_J^i = \left\{w_{J_1}^i, ..., w_{J_{m(i)}}^i\right\} \tag{44}$$

es un complemento algebraico del conjunto $\left\{\overline{w}^i\right\}$, con la propiedad que $\overline{\mathcal{B}}_\Gamma^i$ cuando es definida por la Ec. (42), genera el mismo espacio lineal que $\overline{\mathcal{B}}_\Gamma^i$, otras definiciones son

$$\begin{aligned} \mathcal{B}_{\Gamma M} &= \left\{w_M^1, ..., w_M^{\overline{N}_\Gamma}\right\} \\ \mathcal{B}_{\Gamma J} &= \bigcup_{i=1}^{\overline{N}_\Gamma} \mathcal{B}_J^i \\ \mathcal{B}_\Gamma &= \mathcal{B}_{\Gamma M} \cup \mathcal{B}_{\Gamma J} \end{aligned} \tag{45}$$

notemos que $\mathcal{B}_{\Gamma M} \subset \overline{\mathcal{B}} \subset \overline{D}$. Con esta definición $\overline{\mathcal{B}}_\Gamma$ y $\mathcal{B}_\Gamma$ generan el mismo espacio lineal, sin embargo la diferencia significativa entre $\overline{\mathcal{B}}_\Gamma$ y $\mathcal{B}_\Gamma$ es que todos

los elementos de $\mathcal{B}_\Gamma$ tienen soporte local, lo cual no es cierto para $\overline{\mathcal{B}}_\Gamma$. Además una propiedad adicional es

$$\overline{\mathcal{B}} = \mathcal{B}_{\Gamma M} + \mathcal{B}_I. \tag{46}$$

Los subespacios generados por las funciones $\mathcal{B}_I, \mathcal{B}_\Gamma, \mathcal{B}_{\Gamma J}$ y $\mathcal{B}_{\Gamma M}$ serán denotados por $\widetilde{D}_I, \widetilde{D}_\Gamma, \widetilde{D}_{\Gamma_1}$, y $\widetilde{D}_{\Gamma_2}$ respectivamente y cuyas dimensiones de los espacios $\widetilde{D}_I, \widetilde{D}_\Gamma$ y $\widetilde{D}$ son N_I, N_Γ y $\widetilde{N}$ respectivamente y satisfacen $\widetilde{N} = N_I + N_\Gamma$.

Además

$$\left.\begin{array}{rl} \widetilde{D} = \widetilde{D}_I + \widetilde{D}_\Gamma & \text{con } \widetilde{D}_I \cap \widetilde{D}_\Gamma = \{0\} \\ \widetilde{D}_\Gamma = \widetilde{D}_{\Gamma_1} + \widetilde{D}_{\Gamma_2} & \text{con } \widetilde{D}_{\Gamma_1} \cap \widetilde{D}_{\Gamma_2} = \{0\} \end{array}\right\} \tag{47}$$

y

$$\overline{D} = \widetilde{D}_I + \widetilde{D}_{\Gamma_2}, \tag{48}$$

la Ec. (47) implica que cualquier función $\widetilde{v} \in \widetilde{D}$ y cualquier función $\widetilde{v}_\Gamma \in \widetilde{D}_\Gamma$ puede ser escrita de forma única de las siguientes maneras

$$\widetilde{v} = \widetilde{v}_\Gamma + \widetilde{v}_I, \qquad \text{con } \widetilde{v}_\Gamma \in \widetilde{D}_\Gamma \text{ y } \widetilde{v}_I \in \widetilde{D}_I \tag{49}$$

$$\widetilde{v} = \widetilde{v}_J + \widetilde{v}_M, \qquad \text{con } \widetilde{v}_J \in \widetilde{D}_{\Gamma_1} \text{ y } \widetilde{v}_M \in \widetilde{D}_{\Gamma_2} \tag{50}$$

$$\widetilde{v} = \widetilde{v}_J + \widetilde{v}_M + \widetilde{v}_I, \qquad \text{con } \widetilde{v}_J \in \widetilde{D}_{\Gamma_1}, \widetilde{v}_M \in \widetilde{D}_{\Gamma_2} \text{ y } \widetilde{v}_I \in \widetilde{D}_I. \tag{51}$$

Por otro lado, si el espacio $D \subset \widetilde{D}$ es definido como el complemento ortogonal, con respecto a $\widetilde{D}$, de $\widetilde{D}_I \subset \widetilde{D}$, i.e. $D = \left\{ v \in \widetilde{D} \mid v \cdot w = 0, \forall w \in \widetilde{D}_I \right\}$, entonces

$$\widetilde{D} = D + \widetilde{D}_I \text{ y } D \cap \widetilde{D}_I = \{0\} \tag{52}$$

introduciendo la notación $Proy_D : \widetilde{D} \to D$ es introducida por el operador proyección de vectores de $\widetilde{D}$ sobre D, recordando que $\widetilde{D} = \widetilde{D}_\Gamma + \widetilde{D}_I$ y $\widetilde{D}_\Gamma \cap \widetilde{D}_I = \{0\}$, entonces la

$$Proy_D \widetilde{D}_\Gamma = D \tag{53}$$

además la función $Proy_D : \widetilde{D} \to D$ es una biyección.

En lo que sigue de este capitulo, el complemento ortogonal de los subespacios de D serán tomados con respecto a D. Usando tal notación, adicionalmente definimos

$$D_{11} \equiv Proy_D \widetilde{D}_{\Gamma_1} \qquad \text{y} \qquad D_{12} \equiv Proy_D \widetilde{D}_{\Gamma_2} \tag{54}$$

junto con

$$D_{21} \equiv (D_{11})^\perp \qquad \text{y} \qquad D_{22} \equiv (D_{12})^\perp. \tag{55}$$

Entonces

$$D = D_{11} + D_{12} \qquad \text{y} \qquad D_{11} \cap D_{12} = \{0\} \tag{56}$$

ya que

$$D = Proy_D \widetilde{D}_{\Gamma_1} + Proy_D \widetilde{D}_{\Gamma_2} \tag{57}$$

por la Ec. (47).

1.3. Esquema General

En esta sección, se establecerá el esquema general en términos de los distintos métodos de subestructuración pueden ser formulados.

Axiom 2 *La única suposición de este esquema es que existe un espacio de Hilbert D y un par de subespacios cerrados de D, $\{D_{11}, D_{12}\}$ con la propiedad de que*

$$D = D_{11} + D_{12} \qquad y \qquad D_{11} \cap D_{12} = \{0\}. \tag{58}$$

Definición 3 *Sea*

$$D_{21} \equiv (D_{11})^{\perp} \qquad y \qquad D_{22} \equiv (D_{12})^{\perp}. \tag{59}$$

Teorema 4 *Asumiendo el axioma y definición anteriores se tiene*

$$\left.\begin{array}{lll} D = D_{11} + D_{21} & y & D_{11} \cap D_{21} = \{0\} \\ D = D_{12} + D_{22} & y & D_{12} \cap D_{22} = \{0\} \\ D = D_{21} + D_{22} & y & D_{21} \cap D_{22} = \{0\} \end{array}\right\} \tag{60}$$

De las formulaciones dadas por las Ecs(58 y 60) implican que cualquier función $u \in D$ puede ser escrita de forma única como

$$u = u_{11} + u_{12} = u_{21} + u_{22} \tag{61}$$

con

$$u_{\alpha\beta} \in D_{\alpha\beta} \qquad \text{con } \alpha, \beta = 1, 2. \tag{62}$$

Múltiples métodos iterativos de subestructuración pueden ser formulados en términos de los siguientes dos problemas abstractos que se formulan a continuación:

Problema 1: En este problema $u_{21} \in D_{21}$ es un dato: Entonces, dado $u_{21} \in D_{21}$, encontrar $u \in D_{12}$ tal que $u = u_{21} + u_{22}$, para alguna $u_{22} \in D_{22}$.

Problema 2: En este problema $u_{11} \in D_{11}$ es un dato: Entonces, dado $u_{11} \in D_{11}$, encontrar $u \in D_{22}$ tal que $u = u_{11} + u_{12}$, para alguna $u_{12} \in D_{12}$.

Dependiendo de la forma en que los subespacios de D sean escogidos, entonces estos problemas llevan a una generalización de versiones de aproximaciones tipo Neumann-Neumann y Dirichlet-Dirichlet.

De la Ec.(61), se deduce lo siguiente

$$u_{2\alpha} = \sum_{\beta=1}^{2} (u_{1\beta})_{2\alpha} \qquad \text{y} \qquad u_{1\alpha} = \sum_{\beta=1}^{2} (u_{2\beta})_{1\alpha} \tag{63}$$

con $\alpha = 1, 2$. Definiendo para cada $\alpha, \beta = 1, 2$ y para cada $u \in D$, la función $\tau_{\alpha\beta} : D_{1\alpha} \to D_{2\beta}$ y $\mu_{\alpha\beta} : D_{2\alpha} \to D_{1\beta}$ por

$$\tau_{\alpha\beta} u \equiv (u_{1\alpha})_{2\beta} \qquad \text{y} \qquad \mu_{\alpha\beta} u \equiv (u_{2\alpha})_{1\beta}. \tag{64}$$

Lema 5 *Cuando $u \in D_{12}$ y $w \in D_{22}$ se tiene que*

$$w \cdot \tau_{22} u = -u \cdot \mu_{22} w \tag{65}$$

Corolario 6 *Definiendo la transformación $T_D : D_{12} \to D_{12}$, para toda $u \in D_{12}$ por*

$$T_D u \equiv -\mu_{22} \tau_{22} u \tag{66}$$

y la transformación $T_N : D_{22} \to D_{22}$, para toda $u \in D_{22}$ por

$$T_N u \equiv -\tau_{22} \mu_{22} u \tag{67}$$

entonces, cada una de estas transformaciones es no-negativa definida.

Teorema 7 *Dadas las formulaciones de los problemas 1 y 2, sea I la transformación identidad, entonces:*

A) Una función $u \in D_{12}$ es la solución del Problema 1, si y sólo si

$$(I + T_D)\, u = \mu_{12} u_{21} \tag{68}$$

B) Una función $u \in D_{22}$ es la solución del Problema 2, si y sólo si

$$(I + T_N)\, u = \tau_{12} u_{11}. \tag{69}$$

Aquí las transformaciones $(I + T_D) : D_{12} \to D_{12}$ y $(I + T_N) : D_{22} \to D_{22}$ son positivas definida. Ya que ambas $I + T_D$ y $I + T_N$ son transformaciones positivas definida, el método de Gradiente Conjugado (5.1.2) es aplicable a este tipo de problemas. Para obtener los nuevos algoritmos se usa la siguiente secuencia de pasos:

Algoritmo 1

Sea u^0 dado, se calcula $r^0 = b - Au^0$, $p^0 = r^0$.

Para $n = 0, 1, ...$

1.-

$$\alpha^n = \frac{p^n \cdot p^n}{p^n \cdot Ap^n} \tag{70}$$

2.-

$$u^{n+1} = u^n + \alpha^n p^n \tag{71}$$

3.-

$$r^{n+1} = r^n - \alpha^n A p^n \tag{72}$$

4.-

$$\beta^n = \frac{r^{n+1} \cdot r^{n+1}}{r^n \cdot r^n} \tag{73}$$

5.-

$$p^{n+1} = r^{n+1} + \beta^n p^n \tag{74}$$

6.-

$$n = n + 1 \tag{75}$$

y regresar a 1.

Cuando se aplica la Ec.(68) y se usa

$$(I - \mu_{22}\tau_{22})\, u = \mu_{12}u_{21}$$

tal que $A = I - \mu_{12}u_{21} = I + T_D$ y $b = \mu_{12}u_{21}$, entonces el esquema general toma la forma:

Algoritmo 2

Sea $p^0 = r^0 = b = \mu_{12}u_{21}$ y $u^0 = 0$.

Para $n = 0, 1, ...$

1.-

$$\psi^n = \tau_{22}p^n \tag{76}$$

2.-

$$\alpha = \frac{p^n \cdot p^n}{p^n \cdot p^n + \psi^n \cdot \psi^n} \tag{77}$$

3.-

$$u^{n+1} = u^n + \alpha^n p^n \tag{78}$$

4.-

$$q^n = \mu_{22}\psi^n \tag{79}$$

5.-

$$r^{n+1} = r^n - \alpha^n q^n \tag{80}$$

6.-

$$\beta^n = \frac{r^{n+1} \cdot r^{n+1}}{r^n \cdot r^n} \tag{81}$$

7.-

$$p^{n+1} = r^{n+1} + \beta^n p^n \tag{82}$$

8.-

$$n = n + 1 \tag{83}$$

y regresar a 1.

El algoritmo para la ecuación

$$(I - \tau_{22}\mu_{22})\, u = \tau_{12} u_{11} \tag{84}$$

se obtiene de forma similar.

Ahora, Consideremos el operador diferencial de segundo orden en un dominio Ω y sin perdida de generalidad consideramos $\Pi = \{\Omega_1, \Omega_2\}$ una descomposición en subdominios del dominio Ω y asumiendo condiciones de frontera tipo Dirichlet igual a cero, i.e.

$$\begin{aligned} \mathcal{L}u &= -\Delta u + u, \qquad \text{en } \Omega_1 \text{ y } \Omega_2 \\ u &= 0, \qquad \text{en } \partial\Omega. \end{aligned} \tag{85}$$

Entonces tenemos la siguiente formula de Green-Herrera

$$\begin{aligned} G(u,w) &= \int_{\Omega} w\mathcal{L}u d\underline{x} + \int_{\Gamma} \left\{ [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} - \dot{w} \left[\left[\frac{\partial u}{\partial n}\right]\right] \right\} d\underline{x} = \\ &= \int_{\Omega} u\mathcal{L}w d\underline{x} + \int_{\Gamma} \left\{ [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} - \dot{u} \left[\left[\frac{\partial w}{\partial n}\right]\right] \right\} d\underline{x} = G(w,u) \end{aligned} \tag{86}$$

con las siguientes propiedades

$$G(u,w) = \int_{\Omega} \{\nabla u \cdot \nabla w + uw\}\, d\underline{x} + \int_{\Gamma} \left\{ [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} + [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} \right\} d\underline{x} \tag{87}$$

y

$$\begin{aligned} \int_{\Omega} \{\nabla u \cdot \nabla w + uw\}\, d\underline{x} &= \int_{\Omega} w\mathcal{L}u d\underline{x} - \int_{\Gamma} \left\{ \dot{w} \left[\left[\frac{\partial u}{\partial n}\right]\right] + [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} \right\} d\underline{x} = \\ &= \int_{\Omega} u\mathcal{L}w d\underline{x} - \int_{\Gamma} \left\{ \dot{u} \left[\left[\frac{\partial w}{\partial n}\right]\right] + [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} \right\} d\underline{x} \end{aligned} \tag{88}$$

reescribiendo estas en términos de funciones discontinuas, se tiene que

$$\begin{aligned} -\Delta u + u &= f_{\Omega}, \qquad \text{en } \Omega_1 \text{ y } \Omega_2 \\ u &= 0, \qquad \text{en } \partial\Omega. \\ &\left.\begin{array}{c} [[u]] = 0 \\ \left[\left[\frac{\partial u}{\partial n}\right]\right] = 0 \end{array}\right\}, \qquad \text{en } \Gamma \end{aligned} \tag{89}$$

cuya formulación débil es: u es solución si y sólo si

$$G(u,w) = \int_{\Omega} w f_{\Omega} d\underline{x}, \qquad \text{para toda } w \in \hat{H}(\Omega). \tag{90}$$

considerando ahora a las funciones armónicas definidas como

$$\mathcal{L}u = 0, \qquad \text{en } \Omega_1 \text{ y } \Omega_2. \tag{91}$$

Entonces

$$G(u,w) = \int_{\Gamma} \left\{ [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} - \dot{w} \left[\left[\frac{\partial u}{\partial n} \right] \right] \right\} d\underline{x} = \int_{\Gamma} \left\{ [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} - \dot{u} \left[\left[\frac{\partial w}{\partial n} \right] \right] \right\} d\underline{x} \tag{92}$$

con las siguientes propiedades

$$G(u,w) = \int_{\Omega} \{\nabla u \cdot \nabla w + uw\} d\underline{x} + \int_{\Gamma} \left\{ [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} + [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} \right\} d\underline{x} \tag{93}$$

y

$$\begin{aligned} \int_{\Omega} \{\nabla u \cdot \nabla w + uw\} d\underline{x} &= -\int_{\Gamma} \left\{ \dot{w} \left[\left[\frac{\partial u}{\partial n} \right] \right] + [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} \right\} d\underline{x} = \\ &= -\int_{\Gamma} \left\{ \dot{u} \left[\left[\frac{\partial w}{\partial n} \right] \right] + [[u]] \dot{\widehat{\frac{\partial w}{\partial n}}} \right\} d\underline{x} \end{aligned} \tag{94}$$

expresando estas últimas como una formulación harmónica usando funciones discontinuas tenemos

$$\begin{aligned} -\Delta u + u &= f_{\Omega}, \qquad \text{en } \Omega_1 \text{ y } \Omega_2 \\ u &= 0, \qquad \text{en } \partial\Omega. \\ & \left. \begin{array}{c} [[u]] = j_{\Gamma}^0 \\ \left[\left[\frac{\partial u}{\partial n} \right] \right] = j_{\Gamma}^1 \end{array} \right\}, \qquad \text{en } \Gamma \end{aligned} \tag{95}$$

donde j_{Γ}^0 y j_{Γ}^1 son dadas, cuya formulación débil es: u es solución si y sólo si

$$G(u,w) = \int_{\Gamma} \left\{ j_{\Gamma}^0 \dot{\widehat{\frac{\partial w}{\partial n}}} + j_{\Gamma}^1 \dot{w} \right\} d\underline{x}, \qquad \text{para toda } w \in \hat{H}(\Omega). \tag{96}$$

Definición 8 *Sea $\overline{D}$ el espacio donde la solución u es buscada y se define como*

$$\overline{D} = \left\{ v \in H^2(\Omega) \mid \textit{traza } v = 0, \qquad \textit{sobre } \partial\Omega \right\} \tag{97}$$

así también definimos al espacio $\tilde{D}$ *como*

$$\tilde{D} = \left\{ v \in \hat{H}^2(\Omega, \Pi) \mid traza\ v = 0, \qquad sobre\ \partial\Omega \right\}. \tag{98}$$

Definición 9 *Definimos al espacio de las funciones armónicas como*

$$D \equiv \left\{ u \in \tilde{D} \mid -\Delta u = 0, \qquad en\ \Omega_\alpha; \alpha = 1, 2, ..., E \right\}. \tag{99}$$

Definición 10 *Usando el espacio de las funciones armónicas definimos ahora los siguientes subespacios*

$$\begin{aligned} D_{11} &\equiv \{ w \in D \mid \dot{w} = 0, \qquad en\ \Gamma \} \\ D_{12} &\equiv \{ w \in D \mid [[w]] = 0, \qquad en\ \Gamma \} \\ D_{21} &\equiv \left\{ w \in D \mid \dot{\widehat{\frac{\partial w}{\partial n}}} = 0, \qquad en\ \Gamma \right\} \\ D_{22} &\equiv \left\{ w \in D \mid \left[\left[\frac{\partial w}{\partial n} \right] \right] = 0, \qquad en\ \Gamma \right\} \end{aligned} \tag{100}$$

donde $D_{11} \perp D_{12}$ *y* $D_{21} \perp D_{22}$ *en el producto interior euclidiano y* $D_{22} \perp D_{12}$ *y* $D_{11} \perp D_{21}$ *en el producto interior de energía.*

Entonces, cada función $w \in D$ puede ser escrita de una única forma como

$$w = w_{11} + w_{12} \text{ con } w_{11} \in D_{11} \text{ y } w_{12} \in D_{12} \tag{101}$$

además, D_{21} y D_{11} como también D_{22} y D_{12} son ortogonales con respecto al producto interior

$$u \cdot w \equiv \sum_{\alpha=1}^{2} \int_{\Omega_\alpha} \nabla w \cdot \nabla u dx \tag{102}$$

esto puede verse usando la relación

$$\sum_{\alpha=1}^{2} \int_{\Omega_\alpha} \nabla w \cdot \nabla u dx = -\int_\Gamma \left\{ \dot{w} \left[\left[\frac{\partial u}{\partial n} \right] \right] + [[w]] \dot{\widehat{\frac{\partial u}{\partial n}}} \right\} dx \tag{103}$$

la cual se demuestra usando la Ec.(**??**). La transformación $\tau_{22} : D_{12} \to D_{22}$ es caracterizada por: Dada una función $w \in D$ tal que

$$[[w]] = 0 \text{ sobre } \Gamma \tag{104}$$

entonces $(\tau_{22} w) \in D$ es tal que

$$\left[\left[\frac{\partial (\tau_{22} w)}{\partial n} \right] \right] = 0 \text{ y } \frac{\partial (\tau_{22} w)}{\partial n} = \dot{\widehat{\frac{\partial w}{\partial n}}} \text{ sobre } \Gamma \tag{105}$$

similarmente, la transformación $\mu_{22} : D_{22} \to D_{12}$ es caracterizada por: Dada una función $w \in D$ tal que

$$\left[\left[\frac{\partial w}{\partial n}\right]\right] = 0 \text{ sobre } \Gamma \tag{106}$$

entonces $(\mu_{22} w) \in D$ es tal que

$$[[\mu_{22} w]] = 0 \text{ y } \mu_{22} w = \dot{w} \text{ sobre } \Gamma. \tag{107}$$

Nótese que la evaluación de $\tau_{22} w$ requiere resolver un problema con condiciones Neumann sobre Γ en cada una de las particiones del subdominio Ω_1 y Ω_2, mientras que la evaluación de $\mu_{22} w$ requiere resolver un problema con condiciones Dirichlet sobre Γ en cada una de las particiones del subdominio Ω_1 y Ω_2. Entonces la evaluación de $T_D = -\mu_{22}\tau_{22}$ involucra la resolución de un problema Neumann seguido de uno Dirichlet, mientras que la evaluación de $T_N = -\tau_{22}\mu_{22}$ involucra la resolución de un problema Dirichlet seguido de un problema Neumann.

1.4. Procedimiento para Evaluar la Transformación de Componentes

Para aplicar el esquema general de la sección anterior, necesitamos un procedimiento efectivo de evaluación de las transformaciones

$$\tau_{\alpha\beta} : D_{1\alpha} \to D_{2\beta} \qquad \text{y} \qquad \mu_{\alpha\beta} : D_{2\alpha} \to D_{1\beta} \tag{108}$$

requeridas. Esto queda totalmente establecido si $\mu_{\alpha\beta} \in D_{\alpha\beta}, \alpha, \beta = 1, 2$ tal que

$$u = u_{11} + u_{12} = u_{21} + u_{22} \tag{109}$$

y puede ser evaluado efectivamente cuando $u \in D$ es dado.

Considerando las suposiciones de la sección anterior, entonces existen subconjuntos linealmente independientes

$$\left.\begin{array}{ccc} \mathcal{B} \subset \widetilde{D}, & \mathcal{B}_I \subset \widetilde{D}_I, & \mathcal{B}_\Gamma \subset \widetilde{D}_\Gamma \\ \overline{\mathcal{B}}_\Gamma \subset \widetilde{D}_\Gamma, & \mathcal{B}_{\Gamma M} \subset \widetilde{D}_{\Gamma_2}, & \mathcal{B}_{\Gamma_J} \subset \widetilde{D}_{\Gamma_1} \end{array}\right\} \tag{110}$$

los cuales satisfacen

$$\mathcal{B} = \mathcal{B}_I \cup \mathcal{B}_\Gamma \qquad \text{y} \qquad \overline{\mathcal{B}}_\Gamma = \mathcal{B}_{\Gamma M} \cup \mathcal{B}_{\Gamma_J} \tag{111}$$

el espacio generado por cada uno de los subconjuntos $\mathcal{B}_\Gamma$ y $\overline{\mathcal{B}}_\Gamma$, es $\widetilde{D}_\Gamma$, sin embargo una propiedad distintiva de $\mathcal{B}_\Gamma$ es que sus miembros tienen soporte local, a continuación se detalla un procedimiento para calcular $u_{11} \in D_{11}$ y $u_{12} \in D_{12}$.

De acuerdo a la Ec.(51), podemos escribir

$$u = \tilde{u}_j + \tilde{u}_M + \tilde{u}_I \tag{112}$$

donde $\tilde{u}_j \in \widetilde{D}_{\Gamma_1}, \tilde{u}_M \in \widetilde{D}_{\Gamma_2}$ y $\tilde{u}_I \in \widetilde{D}_I$, además

$$\begin{aligned} u_{11} &= (\tilde{u}_{11})_J + (\tilde{u}_{11})_I, \text{ donde } (\tilde{u}_{11})_J \in \widetilde{D}_{\Gamma_1} \text{ y } (\tilde{u}_{11})_I \in \widetilde{D}_I \\ u_{12} &= (\tilde{u}_{12})_M + (\tilde{u}_{12})_I, \text{ donde } (\tilde{u}_{12})_M \in \widetilde{D}_{\Gamma_2} \text{ y } (\tilde{u}_{12})_I \in \widetilde{D}_I \end{aligned} \tag{113}$$

las Ecs.(109) y (113) juntas implican que

$$(\tilde{u}_{11})_J = \tilde{u}_J \qquad \text{y} \qquad (\tilde{u}_{12})_M = \tilde{u}_M \tag{114}$$

por lo tanto

$$\begin{aligned} u_{11} &= \tilde{u}_J + (\tilde{u}_{11})_I \text{ donde } (\tilde{u}_{11})_I \in \widetilde{D}_I \\ u_{12} &= \tilde{u}_M + (\tilde{u}_{12})_I, \text{ donde } (\tilde{u}_{12})_I \in \widetilde{D}_I \end{aligned} \tag{115}$$

entonces $(\tilde{u}_{11})_I \in \widetilde{D}_I$ puede ser determinada por el sistema de ecuaciones

$$(\tilde{u}_{11})_I \cdot \tilde{w} = -\tilde{u}_J \cdot \tilde{w}, \ \forall \tilde{w} \in \mathcal{B}_I \tag{116}$$

mientras $(\tilde{u}_{12})_I \in \widetilde{D}_I$ se determina por

$$(\tilde{u}_{12})_I \cdot \tilde{w} = -\tilde{u}_M \cdot \tilde{w}, \ \forall \tilde{w} \in \mathcal{B}_I \tag{117}$$

ya que $u_{11} \in D$ y $u_{12} \in D$ son ortogonales a $\widetilde{D}_I$. Cada una de las Ecs(116) y (117) constituyen una sistema de "E" problemas locales independientes.

por otro lado, cuando $u \in D$ es dado, la función $u_{21} \in D_{21}$ es caracterizada por

$$\begin{aligned} u_{21} \cdot \tilde{w} &= 0, \ \forall \tilde{w} \in \widetilde{D}_I \\ (u_{21} - u) \cdot \tilde{w} &= 0, \ \forall \tilde{w} \in D_{12} \\ u_{21} \cdot \tilde{w} &= 0, \ \forall \tilde{w} \in D_{11} \end{aligned} \tag{118}$$

usando el hecho de que cualquier $w \in D_{12}$ es

$$w = \tilde{w}_M + \tilde{w}_I, \text{ donde } \tilde{w}_M \in \widetilde{D}_{\Gamma_2} \text{ y } \tilde{w}_I \in \widetilde{D}_I \tag{119}$$

y que cada $w \in D_{11}$ es

$$w = \tilde{w}_J + \tilde{w}_I, \text{ donde } \tilde{w}_J \in \widetilde{D}_{\Gamma_1} \text{ y } \tilde{w}_I \in \widetilde{D}_I \tag{120}$$

y se ve en el sistema de Ecs.(118) es equivalente a

$$\begin{aligned} u_{21} \cdot \tilde{w}_I &= 0, \ \forall \tilde{w}_I \in \widetilde{D}_I \\ u_{21} \cdot \tilde{w}_M &= u \cdot \tilde{w}_M, \ \forall \tilde{w}_M \in \widetilde{D}_{\Gamma_2} \\ u_{21} \cdot \tilde{w}_J &= 0, \ \forall \tilde{w}_J \in \widetilde{D}_{\Gamma_1} \end{aligned} \tag{121}$$

además, $\widetilde{D}_\Gamma = \widetilde{D}_{\Gamma_1} + \widetilde{D}_{\Gamma_2}$ y por lo tanto las Ecs.(121) se satisface si y sólo si

$$\begin{aligned} u_{21} \cdot \tilde{w}_I &= 0, \ \forall \tilde{w}_I \in \widetilde{D}_I \\ u_{21} \cdot \tilde{w}_\Gamma &= u \cdot (\tilde{w}_\Gamma)_M, \ \forall \tilde{w} \in \widetilde{D}_\Gamma \end{aligned} \tag{122}$$

aquí, se sobreentiende que cada $\tilde{w}_\Gamma \in \widetilde{D}_\Gamma$ se puede escribir como

$$\tilde{w}_\Gamma = (\tilde{w}_\Gamma)_J + (\tilde{w}_\Gamma)_M\,,\ \text{con}\ (\tilde{w}_\Gamma)_J \in \widetilde{D}_{\Gamma_1}\ \text{y}\ (\tilde{w}_\Gamma)_M \in \widetilde{D}_{\Gamma_2} \tag{123}$$

finalmente, introduciendo las bases $\mathcal{B}_I$ y $\mathcal{B}_\Gamma$ de $\widetilde{D}_I$ y $\widetilde{D}_\Gamma$ respectivamente, las Ecs.(122) pueden ser remplazadas por

$$\begin{aligned} u_{21}\cdot\tilde{w}_I &= 0,\ \forall\tilde{w}_I \in \mathcal{B}_I \\ u_{21}\cdot\tilde{w}_\Gamma &= u\cdot(\tilde{w}_\Gamma)_M\,,\ \forall\tilde{w}_\Gamma \in \mathcal{B}_\Gamma \end{aligned} \tag{124}$$

usando el hecho de que todas las funciones de $\tilde{w}_\Gamma \in \mathcal{B}_\Gamma$ tienen soporte local, se ve que las Ecs.(124) constituyen un sistema de "E" ecuaciones locales independientes. De forma similar, se muestra que $u_{22} \in D_{22}$ satisface

$$\begin{aligned} u_{22}\cdot\tilde{w}_I &= 0,\ \forall\tilde{w}_I \in \mathcal{B}_I \\ u_{22}\cdot\tilde{w}_\Gamma &= u\cdot(\tilde{w}_\Gamma)\,,\ \forall\tilde{w}_\Gamma \in \mathcal{B}_\Gamma \end{aligned} \tag{125}$$

aquí, nuevamente las Ecs.(125) constituyen un sistema de "E" ecuaciones locales independientes. Finalmente, observamos que generalmente solo una de los dos sistemas (124) y (125) necesitan ser resueltos ya que $u = u_{21} + u_{22}$. Un comentario similar aplica en el caso del par de sistemas (116) y (117).

Aplicando el algoritmo de Gradiente Conjugado de la sección (1.3), también requieren el cálculo de u_{11} o de u_{21}. La versión discreta estándar del problema original, usando funciones continuas exclusivamente es: Encontrar $\overline{u} \in \overline{D}$ tal que

$$\overline{u}\cdot\overline{w} = \int_\Omega \overline{w} f_\Omega dx, \forall\overline{w} \in \overline{D} \tag{126}$$

es equivalente a: Encontrar $\overline{u} \in \overline{D}$ tal que

$$\begin{aligned} \overline{u}\cdot\tilde{w}_I &= \int_\Omega \tilde{w}_I f_\Omega dx, \forall\tilde{w}_I \in \widetilde{D}_I \\ \overline{u}\cdot\tilde{w}_M &= \int_\Omega \tilde{w}_M f_\Omega dx, \forall\tilde{w}_M \in \widetilde{D}_{\Gamma_2} \\ (\overline{u})_J &= 0 \end{aligned} \tag{127}$$

sea $\tilde{u}_p \in \widetilde{D}$ cualquier función que satisface

$$\begin{cases} \tilde{u}_p\cdot\tilde{w}_I = \int_\Omega \tilde{w}_I f_\Omega dx, \forall\tilde{w}_I \in \widetilde{D}_I \\ (\tilde{u}_p)_J = 0 \end{cases} \tag{128}$$

y definiendo $u = \overline{u} - \tilde{u}_p$, entonces tenemos que

$$\begin{cases} u \in D \\ u\cdot\tilde{w}_M = \int_\Omega \tilde{w}_M f_\Omega dx - \tilde{u}_p\cdot\tilde{w}_M, \forall\tilde{w}_M \in \widetilde{D}_{\Gamma_2} \\ \tilde{u}_J = -(\tilde{u}_p)_J = 0 \end{cases} \tag{129}$$

entonces, $u \in D_{12}$ y aplicando Ecs.(124) para obtener u_{21}. Una segunda opción es definir $\tilde{u}_p \in \widetilde{D}$ como una función que satisface

$$\begin{cases} \tilde{u}_p \cdot \tilde{w}_I = \int_\Omega \tilde{w}_I f_\Omega dx, \forall \tilde{w}_I \in \widetilde{D}_I \\ \tilde{u}_p \cdot \tilde{w}_\Gamma = \int_\Omega \tilde{w}_\Gamma f_\Omega dx, \forall \tilde{w}_\Gamma \in \widetilde{D}_\Gamma \end{cases} \tag{130}$$

en este caso

$$u \cdot \tilde{w}_M = 0, \forall \tilde{w}_M \in \widetilde{D}_{\Gamma_2} \text{ y } \tilde{u}_J = -(\tilde{u}_p)_J \tag{131}$$

tal que $u \in D_{22}$, en virtud de que las Ecs.(122), mientras que las Ecs.(115) y (116) pueden ser usadas para obtener u_{11}.

1.5. Métodos Dual-Primal

Los métodos Dual-Primal son procedimientos que permiten tratar con particiones de vértices. La idea básica de tales métodos consiste en mantener sin dividir las funciones asociadas con los vértices y tratarlas como nodos internos. Un efecto de tal procedimiento es, sin embargo, un acoplamiento de los sistemas de ecuaciones correspondientes a la partición de los subdominios que comparten un vértice, lo cual puede ser un inconveniente en algunas circunstancias. Por completes, en esta sección se incorpora los métodos Dual-Primal en nuestro esquema.

La colección de conjuntos $\left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{\overline{N}_\Gamma}\right\} \subset \left\{\mathcal{B}^1, ..., \mathcal{B}^{\overline{N}}\right\}$ de la sección (1.2) es dividido en dos subfamilias $\left\{\mathcal{B}_\Delta^1, ..., \mathcal{B}_\Delta^{N_\Delta}\right\} \subset \left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{\overline{N}_\Gamma}\right\}$ y $\left\{\mathcal{B}_\Pi^1, ..., \mathcal{B}_\Pi^{N_\pi}\right\} \subset \left\{\mathcal{B}_\Gamma^1, ..., \mathcal{B}_\Gamma^{\overline{N}_\Gamma}\right\}$, ellas son definidas por las siguientes condiciones:$\mathcal{B}_\Gamma^i \in \left\{\mathcal{B}_\Delta^1, ..., \mathcal{B}_\Delta^{N_\Delta}\right\}$ si y sólo si la cardinalidad de $\mathcal{B}_\Gamma^i$ es dos, y $\mathcal{B}_\Gamma^i \in \left\{\mathcal{B}_\Pi^1, ..., \mathcal{B}_\Pi^{N_\pi}\right\}$ si y sólo si la cardinalidad de $\mathcal{B}_\Gamma^i$ es mayor de dos. Uno de estos conjuntos $\left\{\mathcal{B}_\Delta^1, ..., \mathcal{B}_\Delta^{N_\Delta}\right\}$ es llamado en conjunto "Dualτ el otro conjunto $\left\{\mathcal{B}_\Pi^1, ..., \mathcal{B}_\Pi^{N_\Pi}\right\}$ es llamado el conjunto "Primal". Claramente las funciones del conjunto primal corresponden a las funciones asociadas con los vértices.

Sea $\mathcal{B}_\Pi^i$ el conjunto primal, entonces definimos $\overline{\mathcal{B}}_\Pi^i = \{\tilde{w}_M\}$ donde $\tilde{w}_M^i \in \overline{D}$ es una función madre del conjunto primal $\mathcal{B}_\Pi^i$. Además

$$\mathcal{B}_\Pi = \bigcup_{i=1}^{\overline{N}_\Pi} \overline{\mathcal{B}}_\Pi^i = \left\{\tilde{w}_M^1, ..., \tilde{w}_M^{\overline{N}_\Pi}\right\} \tag{132}$$

y

$$\mathcal{B}_\mathcal{J} = \mathcal{B}_\Pi \cup \mathcal{B}_I \tag{133}$$

por otro lado, cuando la cardinalidad de $\mathcal{B}_\Gamma^i$ es dos, definimos

$$\mathcal{B}_\Delta = \bigcup_{i=1}^{N_\Delta} \mathcal{B}_\Delta^i, \overline{\mathcal{B}}_\Pi^i = \left\{w_M^i, w_J^i\right\}, \overline{\mathcal{B}}_\Delta = \bigcup_{i=1}^{N_\Delta} \overline{\mathcal{B}}_\Delta^i \tag{134}$$

$$\mathcal{B} = \mathcal{B}_\Delta \cup \mathcal{B}_\mathcal{J} \tag{135}$$

como también

$$\mathcal{B}_{\Delta M} = \left\{w_M^1, ..., w_M^N\right\} \text{ y } \mathcal{B}_{\Delta J} = \left\{w_J^1, ..., w_J^N\right\} \tag{136}$$

entonces $\mathcal{B}_{\Delta M} \subset \overline{\mathcal{B}} \subset \overline{D}$, cada conjunto $\overline{\mathcal{B}}_\Delta$ y $\mathcal{B}_\Delta$ generan el mismo subespacio lineal y tienen una propiedad evidente y es que todos lo elementos de $\mathcal{B}_\Delta$ tienen soporte local, lo cual no es cierto en $\overline{\mathcal{B}}_\Delta$. Además

$$\overline{\mathcal{B}} = \mathcal{B}_{\Delta M} \cup \mathcal{B}_{\mathcal{J}}. \tag{137}$$

El subespacio generado por el conjunto de funciones $\mathcal{B}$,$\mathcal{B}_{\mathcal{J}}, \mathcal{B}_\Delta, \mathcal{B}_{\Delta J}$ y $\mathcal{B}_{\Delta M}$ será denotado por $\widetilde{D}, \widetilde{D}_{\mathcal{J}}, \widetilde{D}_\Sigma, \widetilde{D}_{\Sigma_1}$ y $\widetilde{D}_{\Sigma_2}$ respectivamente. Haciendo las siguientes sustituciones

$$\left.\begin{array}{c} \widetilde{D}_{\mathcal{J}} \to \widetilde{D}_I, \qquad \widetilde{D}_\Sigma \to \widetilde{D}_\Gamma, \qquad \widetilde{D}_{\mathcal{J}} \to \widetilde{D}_{\mathcal{J}} \\ \widetilde{D}_{\Sigma_1} \to \widetilde{D}_{\Gamma_1}, \qquad \widetilde{D}_{\Sigma_2} \to \widetilde{D}_{\Gamma_2} \end{array}\right\} \tag{138}$$

ya que corresponden con las suposiciones hechas anteriormente en este capitulo

$$\left.\begin{array}{c} \widetilde{D} = \widetilde{D}_{\mathcal{J}} + \widetilde{D}_\Sigma \qquad \text{y} \qquad \widetilde{D}_{\mathcal{J}} \cap \widetilde{D}_\Sigma = \{0\} \\ \widetilde{D}_\Sigma = \widetilde{D}_{\Sigma_1} + \widetilde{D}_{\Sigma_2} \qquad \text{y} \qquad \widetilde{D}_{\Sigma_1} \cap \widetilde{D}_{\Sigma_2} = \{0\} \\ \overline{D} = \widetilde{D}_{\Sigma_2} + \widetilde{D}_{\mathcal{J}} \end{array}\right\} \tag{139}$$

por lo tanto definimos

$$D = \left(\widetilde{D}_{\mathcal{J}}\right)^\perp \tag{140}$$

como se había hecho anteriormente, una vez $D \subset \widetilde{D}$ ha sido definido, el complemento ortogonal puede ser tomado con respecto a D. Y las siguientes definiciones son adoptadas

$$D_{11} = Proy_D \widetilde{D}_{\Sigma_1} \qquad \text{y} \qquad D_{12} = Proy_D \widetilde{D}_{\Sigma_2} \tag{141}$$

$$D_{21} = (D_{11})^\perp \qquad \text{y} \qquad D_{22} = (D_{12})^\perp \tag{142}$$

entonces

$$D = D_{11} + D_{12} \qquad \text{y} \qquad D_{11} \cap D_{12} = \{0\}. \tag{143}$$

La diferencia con respecto a lo antes desarrollado es que todas las funciones de el espacio $\widetilde{D}_I$ tienen soporte local, el cual no es el caso del espacio $\widetilde{D}_{\mathcal{J}}$ como aquí se definió. Debido a este hecho, tenemos algunos acoplamientos entre los sistemas de ecuaciones correspondientes a diferentes subdominios que comparten vértices.

2. Unificación y Simplificación de los Métodos Dual-Primal de Descomposición de Dominio

Entre los métodos más conocidos de descomposición de dominio esta el método FETI Dual-Primal, este destaca por permitir resolver un gran número de problemas y su solución se basa en separar los nodos del dominio en tres tipos de nodos: Nodos Interiores, Nodos Primales (nodos vértices de la frontera interior del subdominio) y nodos Duales (nodos distintos a los vértices de la frontera interior del subdominio).

La teoría propuesta por el Dr. Herrera [18], [19] y [20] permite hacer una unificación entre diversos métodos, entre los que destacan los de Subestructuración, FETI, FETI-DP, entre otros.

2.1. Espacio Dual-Primal

Sea Ω un conjunto finito de cardinalidad d los cuales por definición son tomados como $\{1, ..., d\}$ y a cuyos miembros nos referiremos como los *grados de libertad originales*. Además, sea el conjunto $\{\Omega_1, ..., \Omega_E\}$ una cubierta de Ω, i.e. $\{\Omega_1, ..., \Omega_E\}$ es una familia de subconjuntos de Ω tal que

$$\Omega = \bigcup_{\alpha=1}^{E} \Omega_\alpha. \tag{144}$$

Definición 11 *Un nodo derivado es el par* $\underline{p} \equiv (p, \alpha)$ *tal que* $p \in \Omega$ *y* $\alpha \in \{1, ..., E\}$. *Y el conjunto total de nodos derivados* $\overline{\Omega}^T$ *es definido por*

$$\overline{\Omega}^T \equiv \left\{ \underline{p} = (p, \alpha) \mid p \in \Omega_\alpha \right\}. \tag{145}$$

Definición 12 *Cuando* $p \in \Omega$, *escribimos*

$$Z^T(p) \equiv \left\{ \alpha \in \{1, ..., E\} \mid (p, \alpha) \in \overline{\Omega}^T \right\}. \tag{146}$$

Definición 13 *Para todo* $p \in \Omega$, *la multiplicidad total de* p, $m^T(p)$, *es definido por la cardinalidad de* $Z^T(p)$.

Distinguiremos dos clases de nodos originales:

Definición 14 $p \in \Omega$ *será llamado nodo interior si* $m^T(p) = 1$, *y lo llamaremos nodo de la frontera interior cuando* $m^T(p) > 1$.

El conjunto de nodos interiores y nodos de frontera son disjuntos, los cuales los denotaremos por Ω^I y Ω^Γ respectivamente, claramente el par $\left\{\Omega^I, \Omega^\Gamma\right\}$ constituyen una partición de Ω, ya que

$$\Omega = \Omega^I \cup \Omega \qquad \text{y} \qquad \varnothing = \Omega^I \cap \Omega^\Gamma. \tag{147}$$

Escogeremos a el conjunto $\Omega_0 \subset \Omega^\Gamma$ y definimos los conjuntos

$$\begin{cases} I \equiv \left\{\underline{p} = (p,\alpha) \in \Omega^T \mid p \in \Omega^I\right\} \\ \pi \equiv \left\{\underline{p} = (p,0) \mid p \in \Omega_0\right\} \\ \Delta \equiv \left\{\underline{p} = (p,\alpha) \in \Omega^T \mid p \in \Omega^\Gamma - \Omega_0\right\} \end{cases} \tag{148}$$

junto con

$$\overline{\Omega} = I \cup \pi \cup \Delta \qquad \text{y} \qquad \Pi \equiv I \cup \pi. \tag{149}$$

En particular, si $\overline{\Omega} = \overline{\Omega}^T$ entonces $\Omega_0 = \emptyset$. Para los algoritmos que se discutirán en este capítulo, el conjunto $\overline{\Omega}$ jugara un papel preponderante.

Definición 15 *A los elementos $p \in \overline{\Omega}$, les llamaremos nodos derivados, los cuales pueden ser nodos interiores, primales o duales dependiendo de que si pertenecen a I, π o a Δ respectivamente.*

Definición 16 *Para todo $p \in \Omega$, tenemos el conjunto*

$$Z(p) \equiv \left\{\alpha \in \{0,1,...,E\} \mid (p,\alpha) \in \overline{\Omega}\right\} \tag{150}$$

entonces definiremos la multiplicidad de $m(p)$ de cualquier $p \in \Omega$, como la cardinalidad de $Z(p)$, en particular, $m(p) = 1$ si y sólo si $p \in \Pi$.

2.1.1. Espacio de Vectores

Notemos que cada función real-valuada definida en Ω o en $\overline{\Omega}$ es un vector (y en lo que resta de este trabajo no haremos distinción entre una función y vector).

Definición 17 *Los espacios lineales $\tilde{D}(\Omega)$ y $\tilde{D}\left(\overline{\Omega}\right)$ están constituidos por funciones (vectores) definidos en Ω y $\overline{\Omega}$ respectivamente. Similarmente, $\tilde{D}(\Pi) \subset \tilde{D}\left(\overline{\Omega}\right)$ y $\tilde{D}(\Delta) \subset \tilde{D}\left(\overline{\Omega}\right)$ serán subespacios lineales de $\tilde{D}\left(\overline{\Omega}\right)$ cuyos elementos se nulifican fuera de Π y Δ respectivamente.*

Entonces

$$\tilde{D}\left(\overline{\Omega}\right) = \tilde{D}(\Pi) \oplus \tilde{D}(\Delta) \tag{151}$$

donde $\oplus$ es la suma directa de dos espacios lineales. Esta última ecuación se satisface si y sólo si

$$\begin{cases} \tilde{D}\left(\overline{\Omega}\right) = \tilde{D}(\Pi) + \tilde{D}(\Delta) \\ \{0\} = \tilde{D}(\Pi) \cap \tilde{D}(\Delta) \end{cases} \tag{152}$$

por lo tanto, los vectores de $\tilde{D}\left(\overline{\Omega}\right)$ pueden representarse de una única manera como

$$\underline{u} = (u_\Pi, u_\Delta) = u_\Pi + u_\Delta \qquad \text{con } u_\Pi \in \tilde{D}(\Pi) \text{ y } u_\Delta \in \tilde{D}(\Delta). \tag{153}$$

Definición 18 *Un vector $\underline{u} \in \tilde{D}\left(\overline{\Omega}\right)$ es llamado continuo cuando para todo $p \in \Omega$, se tiene que $u(p,\alpha)$ es independiente de α.*

Observemos que cuando $(p,\alpha)\in\Pi$, entonces la cardinalidad de $Z(p)$ es uno; por lo tanto

$$\tilde{D}(\Pi)\subset\bar{D}\left(\overline{\Omega}\right). \tag{154}$$

Definición 19 *El producto interior Euclideano es definido por*

$$\begin{cases}\underline{\hat{u}}\cdot\underline{\hat{w}}\equiv\sum\limits_{p\in\Omega}\underline{\hat{u}}(p)\,\underline{\hat{w}}(p)\,,\ \forall\underline{\hat{u}},\underline{\hat{w}}\in\tilde{D}(\Omega)\\ \underline{u}\cdot\underline{w}\equiv\sum\limits_{\underline{p}\in\overline{\Omega}}\underline{u}\left(\underline{p}\right)\underline{w}\left(\underline{p}\right)=\sum\limits_{q\in\Omega}\sum\limits_{\underline{p}\in Z(q)}\underline{u}\left(\underline{p}\right)\underline{w}\left(\underline{p}\right),\ \forall\underline{u},\underline{w}\in\tilde{D}\left(\overline{\Omega}\right).\end{cases} \tag{155}$$

Definición 20 *En sistemas de ecuaciones diferenciales parciales, se requieren usar matrices cuyo tratamiento adecuado en nuestro esquema requiere la introducción de funciones vector valuadas. En cuyo caso $\underline{\hat{u}}(p)$ y $\underline{u}(p)$ deben de ser vectores, así que las Ecs(155) que definen un producto interior deben de ser remplazadas por*

$$\begin{cases}\underline{\hat{u}}\cdot\underline{\hat{w}}\equiv\sum\limits_{p\in\Omega}\underline{\hat{u}}(p)\odot\underline{\hat{w}}(p)\,,\ \forall\underline{\hat{u}},\underline{\hat{w}}\in\tilde{D}(\Omega)\\ \underline{u}\cdot\underline{w}\equiv\sum\limits_{\underline{p}\in\overline{\Omega}}\underline{u}\left(\underline{p}\right)\odot\underline{w}\left(\underline{p}\right)=\sum\limits_{q\in\Omega}\sum\limits_{\underline{p}\in Z(q)}\underline{u}(p)\odot\underline{w}(p)\,,\ \forall\underline{u},\underline{w}\in\tilde{D}\left(\overline{\Omega}\right).\end{cases} \tag{156}$$

en este caso el símbolo $\odot$ denota al producto interior del espacio de vectores al que pertenezca $\underline{\hat{u}}(p)$ y $\underline{u}(p)$.

Definición 21 *Sea el operador promedio $\underline{\underline{a}}:\tilde{D}\left(\overline{\Omega}\right)\to\tilde{D}\left(\overline{\Omega}\right)$ y el operador salto $\underline{\underline{j}}:\tilde{D}\left(\overline{\Omega}\right)\to\tilde{D}\left(\overline{\Omega}\right)$ dos matrices definidas por*

$$\underline{\underline{a}}\underline{u}=Proy_{\bar{D}}\underline{u}\qquad y\qquad\underline{\underline{j}}=\underline{\underline{I}}-\underline{\underline{a}} \tag{157}$$

aquí, $\underline{\underline{I}}$ es la matriz identidad y la proyección sobre $\bar{D}$ es tomada con respecto al producto interior Euclideano.

Notemos que $\underline{\underline{j}}\underline{u}$ y $\underline{\underline{a}}\underline{u}$ son ortogonales, ya que $\underline{\underline{j}}\underline{u}=\underline{u}-\underline{\underline{a}}\underline{u}$, ya que el vector $\underline{u}-\underline{\underline{a}}\underline{u}$ es ortogonal a $\underline{\underline{a}}\underline{u}$. En vista de esta definición tenemos

$$\bar{D}\left(\overline{\Omega}\right)\equiv\underline{\underline{a}}\tilde{D}\left(\overline{\Omega}\right) \tag{158}$$

una obvia e importante propiedad es que

$$\underline{\underline{I}}=\underline{\underline{a}}+\underline{\underline{j}} \tag{159}$$

además, $\underline{\underline{j}}$ es también una proyección, verdaderamente, esta es la proyección sobre el complemento ortogonal de $\bar{D}$. Por lo tanto, $\underline{\underline{a}}$ y $\underline{\underline{j}}$ son ambas simétricas, no negativas e idempotentes. También notemos que

$$\underline{\underline{a}}\underline{\underline{j}}=\underline{\underline{j}}\underline{\underline{a}}=\underline{\underline{0}} \tag{160}$$

ya que $\underline{\underline{wajw}} = \left(\underline{\underline{wa}}\right) \cdot \left(\underline{\underline{jw}}\right) = 0$ pues $\underline{\underline{a}}$ y $\underline{\underline{j}}$ son ortogonales, en particular

$$\underline{\underline{j}}\bar{D}\left(\overline{\Omega}\right) = \{0\}. \tag{161}$$

La construcción de la matriz $\underline{\underline{a}}$ es sencilla "dado un vector $\underline{w} \in \tilde{D}\left(\overline{\Omega}\right)$, en cada uno de los grados de libertad pertenecientes a un *nodo* el valor de $\underline{\underline{a}}\underline{w}$ igual al promedio de $\underline{w}$ sobre este *nodo*", i.e,

$$\underline{\underline{a}}\underline{w}\left(\underline{p}\right) = \frac{1}{\hat{m}\left(p\right)} \sum_{\underline{q} \in Z(p)} w\left(\underline{q}\right), \qquad \text{siempre que } \underline{p} \in Z\left(p\right) \tag{162}$$

aquí, $\hat{m}\left(p\right)$ es la cardinalidad de $Z\left(p\right)$.

Además $\underline{\underline{a}}$ puede definirse también como

$$\underline{\underline{a}} \equiv \left(a_{(i,\alpha)(j,\beta)}\right) \tag{163}$$

donde

$$a_{(i,\alpha)(j,\beta)} = \frac{1}{m(i)}\delta_{ij}, \qquad \forall \alpha \in Z\left(i\right) \text{ y } \forall \beta \in Z\left(j\right) \tag{164}$$

para la matriz $\underline{\underline{j}}$, no es necesario calcularla, ya que

$$\underline{\underline{j}}\underline{w} = \left(\underline{\underline{I}} - \underline{\underline{a}}\right)\underline{w} = \underline{w} - \underline{\underline{a}}\underline{w}, \qquad \forall \underline{w} \in \tilde{D}\left(\overline{\Omega}\right). \tag{165}$$

Por ejemplo, la estructura de $\underline{\underline{a}}^{(q)}$ y $\underline{\underline{j}}^{(q)}$, donde q es un nodo de multiplicidad 2 queda como

$$\underline{\underline{a}}^{(q)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \qquad y \qquad \underline{\underline{j}}^{(q)} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{166}$$

y la estructura de $\underline{\underline{a}}^{(q)}$ y $\underline{\underline{j}}^{(q)}$, donde q es un nodo de multiplicidad 4 queda como

$$\underline{\underline{a}}^{(q)} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \qquad y \qquad \underline{\underline{j}}^{(q)} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{bmatrix}. \tag{167}$$

Las siguientes relaciones son usadas en el resto del trabajo

$$\underline{\underline{a}}\underline{w} = \underline{w} \qquad \text{y} \qquad \underline{\underline{j}}\underline{w} = 0, \qquad \forall \underline{w} \in \tilde{D}\left(\Pi\right), \tag{168}$$

$$\begin{cases} \underline{\underline{a}}\tilde{D}\left(\Pi\right) \subset \tilde{D}\left(\Pi\right) \text{ y } \underline{\underline{j}}\tilde{D}\left(\Pi\right) = \{0\} \\ \underline{\underline{a}}\tilde{D}\left(\Delta\right) \subset \tilde{D}\left(\Delta\right) \text{ y } \underline{\underline{j}}\tilde{D}\left(\Delta\right) \subset \tilde{D}\left(\Delta\right) \end{cases} \tag{169}$$

y

$$\underline{\underline{j}}\tilde{D}(\overline{\Omega}) = \underline{\underline{j}}\tilde{D}(\Delta) \subset \tilde{D}(\Delta).$$

Definición 22 *Definimos los espacios* $\tilde{D}_{11}(\Delta)$ *y* $\tilde{D}_{12}(\Delta)$ *por*

$$\begin{cases} \tilde{D}_{11}(\Delta) \equiv \underline{\underline{j}}\tilde{D}(\Delta) = \underline{\underline{j}}\tilde{D}(\overline{\Omega}) \subset \tilde{D}(\Delta) \\ \tilde{D}_{12}(\Delta) \equiv \underline{\underline{a}}\tilde{D}(\Delta) \end{cases} \tag{170}$$

Usando estas definiciones, tenemos las siguientes propiedades de estos subespacios de $\tilde{D}(\Delta)$:

- $\tilde{D}_{11}(\Delta)$ es el complemento ortogonal, con respecto al producto interior Euclideano de $\bar{D}\left(\overline{\Omega}\right) = \underline{\underline{a}}\tilde{D}(\overline{\Omega}) = \tilde{D}_{12}(\Delta)$.
- $\tilde{D}(\overline{\Omega}) = \tilde{D}_{11}(\Delta) \oplus \bar{D}(\overline{\Omega})$.
- $\tilde{D}(\Delta) = \tilde{D}_{11}(\Delta) \oplus \tilde{D}_{12}(\Delta)$.
- $\tilde{D}_{11}(\Delta)$ y $\tilde{D}_{12}(\Delta)$ son complementos ortogonales relativos a $\tilde{D}(\Delta)$.
- $\bar{D}(\overline{\Omega}) = \tilde{D}_{12}(\Delta) \oplus \tilde{D}(\Pi)$.
-

$$\begin{aligned} \tilde{D}\left(\overline{\Omega}\right) &= \underline{\underline{a}}\tilde{D}\left(\overline{\Omega}\right) \oplus \underline{\underline{j}}\tilde{D}\left(\overline{\Omega}\right) = \bar{D}\left(\overline{\Omega}\right) \oplus \underline{\underline{j}}\tilde{D}\left(\overline{\Omega}\right) \\ &= \tilde{D}\left(\Pi\right) \oplus \tilde{D}_{11}\left(\Delta\right) \oplus \tilde{D}_{12}\left(\Delta\right). \end{aligned} \tag{171}$$

Otras propiedades implicadas por los resultados anteriores son:

$$\begin{aligned} \bar{D}\left(\overline{\Omega}\right) &= \left\{\underline{u} \in \tilde{D}\left(\overline{\Omega}\right) \mid \underline{\underline{j}}\underline{u} = 0\right\} = \left\{\underline{u} \in \tilde{D}\left(\overline{\Omega}\right) \mid \underline{\underline{a}}\underline{u} = \underline{u}\right\} \\ \tilde{D}_{11}\left(\Delta\right) &= \left\{\underline{u} \in \tilde{D}\left(\overline{\Omega}\right) \mid \underline{\underline{a}}\underline{u} = 0\right\} = \left\{\underline{u} \in \tilde{D}\left(\overline{\Omega}\right) \mid \underline{\underline{j}}\underline{u} = \underline{u}\right\} \\ \tilde{D}_{12}\left(\Delta\right) &= \left\{\underline{u} \in \tilde{D}\left(\Delta\right) \mid \underline{\underline{j}}\underline{u} = 0\right\} = \left\{\underline{u} \in \tilde{D}\left(\Delta\right) \mid \underline{\underline{a}}\underline{u} = \underline{u}\right\} \end{aligned} \tag{172}$$

estas relaciones serán también usadas en lo que sigue.

Otra notación, la cual también será usada, es que para cada una de las funciones $\underline{u} \in \tilde{D}\left(\overline{\Omega}\right)$, escribiremos

$$\dot{\underline{\widehat{u}}} \equiv \underline{\underline{a}}\underline{u} \qquad y \qquad [[\underline{u}]] \equiv \underline{\underline{j}}\underline{u} \tag{173}$$

Entonces $\dot{\underline{\widehat{u}}} \in \bar{D}\left(\overline{\Omega}\right)$, mientras $[[\underline{u}]]$ pertenecen a $\tilde{D}_{11}\left(\Delta\right) \subset \tilde{D}\left(\Delta\right)$. Nótese que, en vista de la Ec.(171) cualquier $\underline{u} \in \tilde{D}\left(\overline{\Omega}\right)$ puede ser escrita de forma única como

$$\underline{u} = \underline{u}_{\Pi} + \underline{u}_{\Delta} = \underline{u}_{\Pi} + \underline{u}_{\Delta 1} + \underline{u}_{\Delta 2} \tag{174}$$

donde $\underline{u}_{\Pi} \in \tilde{D}\left(\Pi\right)$, $\underline{u}_{\Delta 1} \in \tilde{D}_{11}\left(\Delta\right)$ y $\underline{u}_{\Delta 2} \in \tilde{D}_{12}\left(\Delta\right)$ con

$$\underline{u}_{\Delta 1} = [[\underline{u}_{\Delta}]] = [[\underline{u}]], \underline{u}_{\Delta 2} = \dot{\underline{\widehat{u}_{\Delta}}} \qquad \text{y} \qquad \underline{u}_{\Delta} = \underline{u}_{\Delta 1} + \underline{u}_{\Delta 2}. \tag{175}$$

2.1.2. La Inmersión Natural

En esta subsección detallaremos la inmersión natural del espacio original dentro del espacio Dual-Primal.

Definición 23 *La inmersión natural de* $\tilde{D}(\Omega)$ *dentro de* $\tilde{D}\left(\overline{\Omega}\right)$, *denotado por* $\tau : \tilde{D}(\Omega) \to \tilde{D}\left(\overline{\Omega}\right)$ *es definida por cada* $\underline{\hat{u}} \in \tilde{D}(\Omega)$ *por*

$$(\tau\underline{\hat{u}})\left(\underline{q}\right) = \underline{\hat{u}}(p), \qquad \forall \underline{q} \in Z(p) \subset \overline{\Omega}. \tag{176}$$

Más explícitamente, se satisface

$$(\tau\underline{\hat{u}})_{(p,\alpha)} = \underline{\hat{u}}(p), \qquad \forall (p,\alpha) \in \overline{\Omega}. \tag{177}$$

La imagen $\tau\tilde{D}(\Omega)$ de $\tilde{D}(\Omega)$ bajo $\tau : \tilde{D}(\Omega) \to \overline{D}(\Omega)$ cubre a $\overline{D}(\Omega)$, además, la función τ es una biyección. Esto permite definir la función $\tau^{-1} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}(\Omega)$.

Dos matrices auxiliares introduciremos, ellas son $\underline{\underline{\hat{m}}} : \tilde{D}(\Omega) \to \tilde{D}(\Omega)$ y $\underline{\underline{m}} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}\left(\overline{\Omega}\right)$ las cuales definimos como:

Definición 24 *Para cada* $\underline{\hat{u}} \in \tilde{D}(\Omega)$ *y cada* $\underline{u} \in \tilde{D}\left(\overline{\Omega}\right)$, *las matrices de multiplicidades* $\underline{\underline{\hat{m}}} : \tilde{D}(\Omega) \to \tilde{D}(\Omega)$ *y* $\underline{\underline{m}} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}\left(\overline{\Omega}\right)$ *son definidas como*

$$\begin{aligned} \underline{\underline{\hat{m}}}\underline{\hat{u}}(p) &= \hat{m}(p)\,\underline{\hat{u}}(p), \qquad \forall p \in \Omega \\ \underline{\underline{m}}\underline{u}\left(\underline{p}\right) &= m(p)\,\underline{u}\left(\underline{p}\right), \qquad \forall \underline{p} = (p,\alpha) \in \overline{\Omega}. \end{aligned} \tag{178}$$

Ambas matrices $\underline{\underline{\hat{m}}}$ y $\underline{\underline{m}}$ son matrices diagonales. El valor en la diagonal principal de las matrices $\underline{\underline{\hat{m}}}$ y $\underline{\underline{m}}$ son las multiplicidades de $\hat{m}(p)$ y $m\left(\underline{p}\right)$ respectivamente. Algunos resultados derivados de este hecho son

$$\tau\underline{\underline{\hat{m}}}\underline{\hat{u}} = \underline{\underline{m}}\tau\underline{\hat{u}} \qquad \text{y} \qquad \tau\underline{\underline{\hat{m}}}^{-1}\underline{\hat{u}} = \underline{\underline{m}}^{-1}\tau\underline{\hat{u}}, \qquad \forall \underline{u} \in \tilde{D}(\Omega) \tag{179}$$

junto con

$$\underline{\underline{m}}\overline{D}(\Omega) = \overline{D}\left(\overline{\Omega}\right) \qquad \text{y} \qquad \underline{\underline{m}}^{-1}\overline{D}(\Omega) = \overline{D}(\Omega). \tag{180}$$

Lema 25 *cuando* $\underline{\hat{u}}, \underline{\hat{w}} \in \tilde{D}(\Omega)$, *entonces, los siguientes enunciados se satisfacen*

$$\left.\begin{array}{l} \underline{\hat{u}} \cdot \underline{\underline{\hat{m}}}\underline{\hat{w}} = \tau\left(\underline{\hat{u}}\right) \cdot \tau\left(\underline{\hat{w}}\right) \\ \underline{\hat{u}} \cdot \underline{\hat{w}} = \tau\left(\underline{\hat{u}}\right) \cdot \underline{\underline{m}}^{-1}\tau\left(\underline{\hat{w}}\right) \end{array}\right\}, \qquad \forall \underline{\hat{u}}, \underline{\hat{w}} \in \tilde{D}(\Omega). \tag{181}$$

Demostración. Escribimos $\underline{u} = \tau\underline{\hat{u}}$ y $\underline{w} = \tau\underline{\hat{w}}$ y aplicando la Ec.(155) obtenemos

$$\underline{u} \cdot \underline{w} = \sum_{p\in\Omega} \sum_{\underline{q}\in Z(p)} \underline{u}\left(\underline{q}\right) \underline{w}\left(\underline{q}\right) = \sum_{p\in\Omega} m(p)\,\underline{\hat{u}}(p)\,\underline{\hat{w}}(p) = \underline{\hat{u}} \cdot \underline{\underline{\hat{m}}}\underline{\hat{w}} \tag{182}$$

entonces usando la Ec.(176), tenemos que

$$\underline{u}\cdot\underline{w}=\sum_{p\in\Omega}\hat{m}\left(p\right)\underline{u}\left(\underline{q}\right)\underline{w}\left(\underline{q}\right)=\sum_{p\in\Omega}\hat{m}\left(p\right)\underline{\hat{u}}\left(p\right)\underline{\hat{w}}\left(p\right)=\underline{\hat{u}}\cdot\underline{\underline{\hat{m}}}\underline{\hat{w}} \tag{183}$$

esto muestra la primera relación del lema, entonces aplicando $\underline{\underline{\hat{m}}}^{-1}\underline{\hat{w}}$ en lugar de $\underline{\hat{w}}$ y usando la Ec.(179) la segunda relación del lema es obtenida. ■

Corolario 26 *Sea* $\underline{u}\in\tau\tilde{D}\left(\Omega\right)=\overline{D}\left(\overline{\Omega}\right)$ *tal que para alguna* $\underline{\hat{u}}\in D\left(\Omega\right)$ *se satisface*

$$\underline{\hat{u}}\cdot\underline{\hat{w}}=\underline{u}\cdot\tau\left(\underline{\hat{w}}\right),\qquad\forall\underline{\hat{w}}\in D\left(\Omega\right) \tag{184}$$

entonces

$$\underline{u}=\underline{\underline{m}}^{-1}\tau\left(\underline{\hat{u}}\right)=\tau\left(\underline{\underline{\hat{m}}}^{-1}\underline{\hat{u}}\right). \tag{185}$$

Demostración. Como se vio en el lema, tenemos que

$$\underline{\hat{w}}\cdot\underline{\hat{u}}=\tau\left(\underline{\hat{w}}\right)\cdot\underline{\underline{m}}^{-1}\tau\left(\underline{\hat{u}}\right),\qquad\forall\underline{\hat{w}}\in\tilde{D}\left(\Omega\right)$$

lo cual implica, que cuando la Ec.(184) se satisface, tenemos

$$\left(\underline{u}-\underline{\underline{m}}^{-1}\tau\left(\underline{\hat{u}}\right)\right)\cdot\underline{w}=0,\forall\underline{w}\in\tau\tilde{D}\left(\Omega\right)=\overline{D}\left(\Omega\right)$$

el corolario se sigue de esta última ecuación, ya que ambas $\underline{u}$ y $\underline{\underline{m}}^{-1}\tau\left(\underline{\hat{u}}\right)$ pertenecen a $\overline{D}\left(\Omega\right)$. ■

2.1.3. La Formulación Matricial Discontinua Libre de Multiplicadores de Lagrange

En los que sigue consideremos dos matrices simétricas

$$\underline{\underline{\hat{A}}}:\tilde{D}\left(\Omega\right)\rightarrow\tilde{D}\left(\Omega\right)\qquad\text{y}\qquad\underline{\underline{A}}:\tilde{D}\left(\overline{\Omega}\right)\rightarrow\tilde{D}\left(\overline{\Omega}\right) \tag{186}$$

la matriz $\underline{\underline{\hat{A}}}$ es generada por alguna discretización y referida como la matriz original, y escribimos

$$\underline{\underline{\hat{A}}}\equiv\left(\underline{\underline{\hat{A}}}_{pq}\right),\text{ donde }p.q\in\Omega \tag{187}$$

y también asumiremos en lo que sigue en este trabajo que:

1.- $\underline{\underline{\hat{A}}}:\tilde{D}\left(\Omega\right)\rightarrow\tilde{D}\left(\Omega\right)$ es positiva definida.

2.- Usando la Ec.(187),

$$\underline{\underline{\hat{A}}}_{pq}=0,\text{ siempre y cuando }p\in\Omega^{I}\cap\Omega_{\alpha},q\in\Omega^{I}\cap\Omega_{\beta}\text{ y }\alpha\neq\beta. \tag{188}$$

3.- La matriz $\underline{\underline{\hat{A}}}$ y $\underline{\underline{A}}$ esta relacionada por

$$\underline{\hat{w}}\cdot\underline{\underline{\hat{A}}}\underline{\hat{u}}=\tau\left(\underline{\hat{w}}\right)\cdot\underline{\underline{A}}\tau\left(\underline{\hat{u}}\right),\qquad\forall\underline{\hat{u}},\underline{\hat{w}}\in\tilde{D}\left(\Omega\right) \tag{189}$$

esta última relación no determina de manera única a la matriz $\underline{\underline{A}}$ cuando la matriz $\underline{\underline{\hat{A}}}$ es dada, pero un conveniente procedimiento para construir una de estas matrices $\underline{\underline{A}}$ esta dada en la sección (2.1.4). Entonces la matriz $\underline{\underline{A}}$ es positiva definida sobre las funciones continuas.

Definición 27 *Sea* $\hat{\underline{f}} \in \tilde{D}(\Omega)$. *Entonces el problema original consiste en buscar una función* $\hat{\underline{u}} \in \tilde{D}(\Omega)$ *que satisface*

$$\hat{\underline{\underline{A}}}\hat{\underline{u}} = \hat{\underline{f}}. \tag{190}$$

Así, una $\hat{\underline{u}} \in \tilde{D}(\Omega)$ satisface la última ecuación si y sólo si

$$\hat{\underline{w}} \cdot \hat{\underline{\underline{A}}}\hat{\underline{u}} = \hat{\underline{w}} \cdot \hat{\underline{f}}, \qquad \forall \hat{\underline{w}} \in \tilde{D}(\Omega) \tag{191}$$

lo cual es equivalente a

$$\tau(\hat{\underline{w}}) \cdot \underline{\underline{\hat{A}}}\tau(\hat{\underline{u}}) = \tau(\hat{\underline{w}}) \cdot \underline{\underline{m}}^{-1}\tau\left(\hat{\underline{f}}\right), \qquad \forall \hat{\underline{w}} \in \tilde{D}(\Omega) \tag{192}$$

o

$$\tau(\hat{\underline{w}}) \cdot \underline{\underline{a\hat{A}}}\tau(\hat{\underline{u}}) = \tau(\hat{\underline{w}}) \cdot \underline{\underline{m}}^{-1}\tau\left(\hat{\underline{f}}\right), \qquad \forall \hat{\underline{w}} \in \tilde{D}(\Omega) \tag{193}$$

tomando en cuenta que $\tau\tilde{D}(\Omega) = \overline{D}\left(\overline{\Omega}\right)$, esta última ecuación se puede remplazar por

$$\underline{\underline{aA}}\tau(\hat{\underline{u}}) = \underline{\underline{m}}^{-1}\tau\left(\hat{\underline{f}}\right), \qquad \forall \hat{\underline{w}} \in \tilde{D}(\Omega). \tag{194}$$

Teorema 28 *Una función* $\tilde{\underline{u}} \in \tilde{D}\left(\overline{\Omega}\right)$ *es solución de las ecuaciones*

$$\begin{cases} \underline{\underline{aA}}\tilde{\underline{u}} = \underline{\underline{m}}^{-1}\tau\left(\hat{\underline{f}}\right) \\ \underline{\underline{j}}\tilde{\underline{u}} = 0 \end{cases} \tag{195}$$

si y sólo si

$$\hat{\underline{u}} \equiv \tau^{-1}(\tilde{\underline{u}}) \tag{196}$$

es solución del problema original.

Demostración. $\Rightarrow$] Asumiendo que $\tilde{\underline{u}} \in \tilde{D}\left(\overline{\Omega}\right)$ esta relacionada con $\hat{\underline{u}}$ por $\hat{\underline{u}} \equiv \tau^{-1}(\underline{u})$, entonces tenemos que $\tilde{\underline{u}} = \tau(\hat{\underline{u}})$ y satisface las Ecs.(195) y (196).

[$\Leftarrow$ La Ec.(195) implica que $\tilde{\underline{u}} \in \tilde{D}\left(\overline{\Omega}\right)$, tal que τ^{-1} esta bien definida. Tomando $\tilde{\underline{u}} \in \tilde{D}\left(\overline{\Omega}\right)$ que satisfaga la Ec.(196), entonces $\hat{\underline{u}}$ satisface la Ec.(190). ■

En vista del teorema (28), podemos definir $\overline{\underline{f}} \in \overline{D}\left(\overline{\Omega}\right)$ como

$$\overline{\underline{f}} \equiv \begin{pmatrix} \overline{\underline{f}}_{\Pi} \\ \underline{f}_{\Delta} \end{pmatrix} \equiv \underline{\underline{m}}^{-1}\tau\left(\hat{\underline{f}}\right) \tag{197}$$

aquí, $\overline{\underline{f}}_{\Delta} = \overline{\underline{f}}_{\Delta 2}$ ya que $\overline{\underline{f}}_{\Delta 1} = 0$ ya que $\overline{\underline{f}} \in \overline{D}\left(\overline{\Omega}\right)$.

Definición 29 *Dados*

$$\underline{\overline{f}} = \underline{\overline{f}}_{\Pi} + \underline{\overline{f}}_{\Delta 2} = \begin{pmatrix} \underline{\overline{f}}_{\Pi} \\ \underline{\underline{\overline{f}}}_{\Delta} \end{pmatrix} \in \overline{D}\left(\overline{\Omega}\right) \tag{198}$$

el problema transformado consiste en buscar una función $\underline{\tilde{u}} \in \tilde{D}\left(\overline{\Omega}\right)$ *que satisface*

$$\begin{cases} \underline{\underline{\underline{aA\tilde{u}}}} = \underline{\overline{f}} \\ \underline{\underline{\underline{j\tilde{u}}}} = 0. \end{cases} \tag{199}$$

La existencia de la solución esta garantizada. Una manera de ver esto es recordando que la matriz $\underline{\underline{A}}$ es positiva definida sobre funciones continuas, i.e. $\underline{\underline{A}} : \tilde{D}_{12}\left(\overline{\Omega}\right) \to \tilde{D}_{12}\left(\overline{\Omega}\right)$ es positiva definida.

2.1.4. Construcción de la Matriz $\underline{\underline{A}}$

En esta sección se desarrollará el procedimiento para construir una matriz $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}\left(\overline{\Omega}\right)$ que satisfaga la condición dada por la Ec.(189) y empezaremos por desarrollar esta condición de manera más explicita.

Recordando que d es la cardinalidad de Ω y asumiendo que $\overline{\Omega} = \overline{\Omega}^T$, tenemos que $Z(p) = Z^T(p)$ y $\tilde{D}\left(\overline{\Omega}\right) = \tilde{D}\left(\overline{\Omega}^T\right)$. Entonces, obsérvese que el conjunto de vectores $\{\underline{e}_1, ..., \underline{e}_d\} \subset \tilde{D}(\Omega)$ es una base para $\tilde{D}(\Omega)$, donde para cada $i \in \Omega$, $\underline{e}_i$ es definida por

$$\underline{e}_i \equiv (\delta_{i1}, ..., \delta_{id}) \tag{200}$$

aquí y en lo que sigue el símbolo δ_{ij} es la delta de Kronecker.

La inmersión natural de este conjunto es $\{\tau(\underline{e}_1), ..., \tau(\underline{e}_d)\} \subset \tilde{D}\left(\overline{\Omega}\right)$, donde

$$\tau(\underline{e}_i)_{(j,\alpha)} = \delta_{ij}, \qquad \forall (j,\alpha) \in \overline{\Omega}^T. \tag{201}$$

Cuando la Ec.(189) es aplicada a ese conjunto de vectores, una condición equivalente es obtenida. Esta es

$$\sum_{\beta \in Z(j)} \sum_{\alpha \in Z(i)} \underline{\underline{A}}_{(i,\alpha)(j,\beta)} = \hat{\underline{\underline{A}}}_{ij}, \qquad \forall i,j \in \Omega \tag{202}$$

Aquí, para las matrices como $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}\left(\overline{\Omega}\right)$, usaremos la siguiente notación

$$\underline{\underline{A}} \equiv \left(\underline{\underline{A}}_{(i,\alpha)(j,\beta)}\right), \text{ con } (i,\alpha)(j,\beta) \in \overline{\Omega} \tag{203}$$

Definición 30 *Para cada* $\alpha \in \{1, ..., E\}$ *y para cada par* $i, j \in \Omega$, *definimos el símbolo* δ^{α}_{ij} *por*

$$\begin{cases} \delta^{\alpha}_{ij} = 1, \ si \ i,j \in \Omega_{\alpha} \\ \delta^{\alpha}_{ij} = 0, si \ i \ \acute{o} \ j \notin \Omega_{\alpha} \end{cases} \tag{204}$$

además la multiplicidad $m(i,j)$ *del par* (i,j) *es definida por*

$$m(i,j) \equiv \sum_{\alpha=1}^{E} \delta_{ij}^{\alpha}. \tag{205}$$

Usando la notación de la condición dad por la Ec.(188), tenemos

$$m(i,j) = 0 \Rightarrow \hat{\underline{\underline{A}}}_{ij} = 0. \tag{206}$$

Definición 31 *La matriz total* $\underline{\underline{A}}^{t} : \tilde{D}\left(\overline{\Omega}^{T}\right) \rightarrow \tilde{D}\left(\overline{\Omega}^{T}\right)$ *es ahora definida por*

$$\left.\begin{array}{l} \underline{\underline{A}}^{t}_{(i,\alpha)(j,\beta)} \equiv 0, \ \textit{si } m(i,j) = 0 \\ \underline{\underline{A}}^{t}_{(i,\alpha)(j,\beta)} \equiv \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \delta_{ij}^{\alpha} \delta_{\alpha\beta}, \ \textit{si } m(i,j) \neq 0 \end{array}\right\} \forall (i,\alpha)(j,\beta) \in \overline{\Omega}^{T}. \tag{207}$$

Y es fácil verificar que $\underline{\underline{A}}^{t}$ como se definió, satisface la condición de la Ec.(202). Y usando la Ec.(205) tenemos que

$$\begin{aligned} \sum_{\beta \in Z(j)} \sum_{\alpha \in Z(i)} \underline{\underline{A}}^{t}_{(i,\alpha)(j,\beta)} &= \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \sum_{\alpha=1} \sum_{\beta=1} \delta_{ij}^{\alpha} \delta_{\alpha\beta} = \\ &= \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \sum_{\beta=1} \delta_{ij}^{\alpha} = \hat{\underline{\underline{A}}}_{ij}. \end{aligned} \tag{208}$$

Para la actual construcción de $\underline{\underline{A}}^{t}$, cuando se implementa el algoritmo, es útil usar la siguiente notación, para cada $\gamma = 1, ..., E$, $\underline{\underline{A}}^{\gamma} : \tilde{D}\left(\overline{\Omega}\right) \rightarrow \tilde{D}\left(\overline{\Omega}\right)$ por

$$\left\{\begin{array}{l} \left(\underline{\underline{A}}^{\gamma}\right)_{(i,\alpha)(j,\beta)} \equiv \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \delta_{ij}^{\gamma} \delta_{\gamma\beta} \delta_{\gamma\alpha}, \ \text{si } m(i,j) \neq 0 \\ \left(\underline{\underline{A}}^{\gamma}\right)_{(i,\alpha)(j,\beta)} \equiv \hat{\underline{\underline{A}}}_{ij} = 0, \ \text{si } m(i,j) = 0 \end{array}\right. \tag{209}$$

entonces

$$\underline{\underline{A}}^{t} = \sum_{\gamma=1}^{E} \underline{\underline{A}}^{\gamma} \tag{210}$$

nuevamente cuando $m(i,j) \neq 0$, tenemos que

$$\begin{aligned} \sum_{\gamma=1}^{E} \left(\underline{\underline{A}}^{\gamma}\right)_{(i,\alpha)(j,\beta)} &= \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \sum_{\gamma=1}^{E} \delta_{ij}^{\gamma} \delta_{\alpha\beta} \delta_{\gamma\alpha} = \\ &= \frac{1}{m(i,j)} \hat{\underline{\underline{A}}}_{ij} \delta_{ij}^{\alpha} \delta_{\alpha\beta} = \underline{\underline{A}}^{t}_{(i,\alpha)(j,\beta)} \end{aligned} \tag{211}$$

y cuando $m(i,j) = 0$, tenemos

$$\sum_{\gamma=1}^{E} \left(\underline{\underline{A}}^{\gamma}\right)_{(i,\alpha)(j,\beta)} = \underline{\underline{A}}^{t}_{(i,\alpha)(j,\beta)} = 0. \tag{212}$$

En el caso más general en el que sólo tenemos la condición $\overline{\Omega} \subset \overline{\Omega}^T$, usando la notación de la Ec.(215), la matriz $\underline{\underline{A}}$ puede ser definida por

$$\begin{array}{ll} \left(\underline{\underline{A}}_{\Pi\Pi}\right)_{ij} = \hat{\underline{\underline{A}}}_{ij} & \left(\underline{\underline{A}}_{\Pi\Delta}\right)_{i(j,\beta)} = \frac{1}{m(i,j)}\hat{\underline{\underline{A}}}_{ij}\delta^{\beta}_{ij} \\ \left(\underline{\underline{A}}_{\Delta\Pi}\right)_{(i,\alpha)j} = \frac{1}{m(i,j)}\hat{\underline{\underline{A}}}_{ij}\delta^{\alpha}_{ij} & \left(\underline{\underline{A}}_{\Delta\Delta}\right)_{(i,\alpha)(j,\beta)} = \frac{1}{m(i,j)}\hat{\underline{\underline{A}}}_{ij}\delta^{\alpha}_{ij}\delta_{\alpha\beta} \end{array} \tag{213}$$

2.2. Fórmula de Green-Herrera para Matrices

Definición 32 *Sea $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \to \tilde{D}\left(\overline{\Omega}\right)$ una matriz simétrica y positiva definida. El 'producto interior de energía' es definido por*

$$\left(\underline{u}, \underline{w}\right) \equiv \underline{u} \cdot \underline{\underline{A}}\underline{w}, \ \forall \underline{u}, \underline{w} \in \tilde{D}\left(\overline{\Omega}\right). \tag{214}$$

El espacio lineal, $\tilde{D}\left(\overline{\Omega}\right)$, es un espacio de Hilbert (dimensionalmente finito) cuando este es dotado con el producto interior de energía. Escribimos

$$\underline{\underline{A}} \equiv \begin{pmatrix} \underline{\underline{A}}_{\Pi\Pi} & \underline{\underline{A}}_{\Pi\Delta} \\ \underline{\underline{A}}_{\Delta\Pi} & \underline{\underline{A}}_{\Delta\Delta} \end{pmatrix} \tag{215}$$

la notación aquí es tal que

$$\begin{cases} \underline{\underline{A}}_{\Pi\Pi} : \tilde{D}\left(\Pi\right) \to \tilde{D}\left(\Pi\right), & \underline{\underline{A}}_{\Pi\Delta} : \tilde{D}\left(\Delta\right) \to \tilde{D}\left(\Pi\right) \\ \underline{\underline{A}}_{\Delta\Pi} : \tilde{D}\left(\Pi\right) \to \tilde{D}\left(\Delta\right), & \underline{\underline{A}}_{\Delta\Delta} : \tilde{D}\left(\Delta\right) \to \tilde{D}\left(\Delta\right) \end{cases} \tag{216}$$

donde $\underline{\underline{A}}_{\Pi\Pi}\underline{u} = \left(\underline{\underline{A}}\underline{u}_{\Pi}\right)_{\Pi}$, $\underline{\underline{A}}_{\Pi\Delta}\underline{u} = \left(\underline{\underline{A}}\underline{u}_{\Delta}\right)_{\Pi}$, $\underline{\underline{A}}_{\Delta\Pi}\underline{u} = \left(\underline{\underline{A}}\underline{u}_{\Pi}\right)_{\Delta}$ y $\underline{\underline{A}}_{\Delta\Delta}\underline{u} = \left(\underline{\underline{A}}\underline{u}_{\Delta}\right)_{\Delta}$; notemos que tenemos la inmersión natural en $\tilde{D}\left(\overline{\Omega}\right)$, i.e.

$$\begin{aligned} \underline{\underline{A}}_{\Pi\Pi} &\equiv \begin{pmatrix} \underline{\underline{A}}_{\Pi\Pi} & 0 \\ 0 & 0 \end{pmatrix} \\ \underline{\underline{A}}_{\Pi\Delta} &\equiv \begin{pmatrix} 0 & \underline{\underline{A}}_{\Pi\Delta} \\ 0 & 0 \end{pmatrix} \\ \underline{\underline{A}}_{\Delta\Pi} &\equiv \begin{pmatrix} 0 & 0 \\ \underline{\underline{A}}_{\Delta\Pi} & 0 \end{pmatrix} \\ \underline{\underline{A}}_{\Delta\Delta} &\equiv \begin{pmatrix} 0 & 0 \\ 0 & \underline{\underline{A}}_{\Delta\Delta} \end{pmatrix}. \end{aligned} \tag{217}$$

Introducimos las siguientes definiciones

Definición 33 *Sea la matriz $\underline{\underline{L}}$ definida como*

$$\underline{\underline{L}} \equiv \begin{pmatrix} \underline{\underline{A}}_{\Pi\Pi} & \underline{\underline{A}}_{\Pi\Delta} \\ 0 & 0 \end{pmatrix} \tag{218}$$

y la matriz $\underline{\underline{R}}$ definida como

$$\underline{\underline{R}} \equiv \begin{pmatrix} 0 & 0 \\ \underline{\underline{A}}_{\Delta\Pi} & \underline{\underline{A}}_{\Delta\Delta} \end{pmatrix}. \tag{219}$$

Además, notemos la siguiente identidad $\underline{\underline{R}} = \underline{\underline{jR}} + \underline{\underline{aR}}$, implica también que $\underline{\underline{A}} = \underline{\underline{A}}^T$, así

$$\underline{\underline{L}} + \underline{\underline{aR}} - \underline{\underline{jR}} = \underline{\underline{L}}^T + \underline{\underline{R}}^T \underline{\underline{a}} - \underline{\underline{R}}^T \underline{\underline{j}}. \tag{220}$$

Definición 34 *A la identidad*

$$\underline{\underline{L}} + \underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}} = \underline{\underline{L}}^T + \underline{\underline{R}}^T \underline{\underline{a}} - \underline{\underline{jR}} \tag{221}$$

la cual se deriva de la Ec.(220), será referida como la fórmula Green-Herrera para matrices.

Notemos que los rangos de $\underline{\underline{L}}$ y $\underline{\underline{R}}$ son $\tilde{D}(\Pi)$ y $\tilde{D}(\Delta)$, respectivamente, mientras que los rangos de $\underline{\underline{aR}}$ y $\underline{\underline{jR}}$ están contenidos en $\tilde{D}_{12}(\Delta)$ y $\tilde{D}_{11}(\Delta)$ respectivamente. Inclusive más, estos últimos dos rangos son linealmente independientes.

2.3. El Operador de Steklov-Poincaré

La fórmula de Green-Herrera de la Ec.(221) es equivalente a

$$\underline{w} \cdot \underline{\underline{L}}\underline{u} + \dot{\underline{w}} \cdot \underline{\underline{aR}}\underline{u} - [[\underline{u}]]\, \underline{\underline{jR}}\underline{w} = \underline{u} \cdot \underline{\underline{L}}\underline{w} + \dot{\underline{u}} \cdot \underline{\underline{aR}}\underline{w} - [[\underline{w}]]\, \underline{\underline{jR}}\underline{u} \tag{222}$$

$\forall \underline{u}, \underline{w} \in \tilde{D}\left(\overline{\Omega}\right)$. Introduciendo la siguiente notación

$$\left[\left[\underline{\underline{R}}\right]\right] = -\underline{\underline{aR}} \qquad \text{y} \qquad \dot{\underline{\underline{\widehat{R}}}} \equiv -\underline{\underline{jR}} \tag{223}$$

usando esta notación, la Ec.(222) se reescribe como

$$\underline{w} \cdot \underline{\underline{L}}\underline{u} + [[\underline{u}]] \cdot \dot{\underline{\underline{\widehat{R}}}}\, \underline{w} - \dot{\underline{\widehat{w}}} \cdot \left[\left[\underline{\underline{R}}\right]\right] \underline{u} = \underline{u} \cdot \underline{\underline{L}}\underline{w} + [[\underline{w}]] \cdot \dot{\underline{\underline{\widehat{R}}}}\, \underline{u} - \dot{\underline{\widehat{u}}} \cdot \left[\left[\underline{\underline{R}}\right]\right] \underline{w} \tag{224}$$

$\forall \underline{u}, \underline{w} \in \tilde{D}\left(\overline{\Omega}\right)$.

Para el operador diferencial de Laplace actuando sobre funciones discontinuas definidas por tramos que satisfacen condiciones de frontera homogéneas, la fórmula Green-Herrera queda como

$$\int_\Omega w\mathcal{L}u dx + \int_\Gamma \left\{ [[u]]\, \dot{\widehat{\frac{\partial w}{\partial n}}} - \dot{\widehat{w}} \left[\left[\frac{\partial u}{\partial n}\right]\right] \right\} dx =$$

$$\int_\Omega u\mathcal{L}w dx + \int_\Gamma \left\{ [[w]]\, \dot{\widehat{\frac{\partial u}{\partial n}}} - \dot{\widehat{u}} \left[\left[\frac{\partial w}{\partial n}\right]\right] \right\} dx \tag{225}$$

comparando con las Ecs. (224) y (225) se obtienen las siguientes correspondencias

$$\begin{array}{c} \int_{\Omega} w\mathcal{L}u dx \leftrightarrow \underline{w} \cdot \underline{\underline{L}}\underline{u} \\ \int_{\Gamma} [[u]] \, \dot{\widehat{\frac{\partial w}{\partial n}}} dx \leftrightarrow [[\underline{u}]] \cdot \dot{\underline{\underline{\widehat{R}}}} \, \underline{w} \\ \int_{\Gamma} \dot{\widehat{w}} \left[\left[\frac{\partial u}{\partial n} \right] \right] dx \leftrightarrow \dot{\underline{\widehat{w}}} \cdot \left[\left[\underline{\underline{R}} \right] \right] \underline{u} \end{array} \tag{226}$$

Para operadores diferenciales, en particular para el operador de Laplace, el operador de Poincaré-Steklov asociado con el salto de la derivada normal y de la funcional bilineal es

$$\int_{\Gamma} \dot{\widehat{w}} \left[\left[\frac{\partial u}{\partial n} \right] \right] dx \tag{227}$$

a nivel matricial, el operador de Steklov-Poincaré es asociado con la forma bilineal

$$\dot{\underline{w}} \cdot \left[\left[\underline{\underline{R}} \right] \right] \underline{u} = \underline{w} \cdot \underline{\underline{aR}} \underline{u}, \forall \underline{u}, \underline{w} \in \tilde{D} \left(\overline{\Omega} \right) \tag{228}$$

o más simplemente, con la matriz $\underline{\underline{aR}}$.

Definición 35 *Definimos al operador de Steklov-Poincaré como la matriz*

$$\underline{\underline{aR}}. \tag{229}$$

Otro ejemplo es la fórmula de Green-Herrera fórmula para el operador elíptico general, simétrico y de segundo orden, el cual es

$$\begin{array}{c} \int_{\Omega} w\mathcal{L}u dx + \int_{\Gamma} \left\{ [[u]] \, \dot{\widehat{\underline{a}_n \cdot \nabla}} w - \dot{\widehat{w}} \, [[\underline{a}_n \bullet \nabla u]] \right\} dx = \\ \int_{\Omega} u\mathcal{L}w dx + \int_{\Gamma} \left\{ [[w]] \, \dot{\widehat{\underline{a}_n \cdot \nabla}} u - \dot{\widehat{u}} \, [[\underline{a}_n \bullet \nabla w]] \right\} dx \end{array} \tag{230}$$

La correspondencia de la ecuación (226) todavía permanece para este caso, excepto que

$$\left\{ \begin{array}{l} [[\underline{a}_n \bullet \nabla u]] \leftrightarrow - \left[\left[\underline{\underline{R}} \right] \right] \underline{u} \\ \dot{\widehat{\underline{a}_n \cdot \nabla}} w \leftrightarrow - \dot{\underline{\underline{\widehat{R}}}} \underline{u} \end{array} \right. . \tag{231}$$

Correspondencias similares a las dadas por las Ecs. (226) y (231) puede ser establecida en general; aplicaciones incluyen los sistemas gobernantes de ecuaciones de elasticidad lineal y muchos problemas. Más notemos que las Ecs. (226) y (231) implican una nueva fórmula para el operador *Steklov-Poincaré* (i.e., el salto de la derivada normal) en el nivel discreto, el cual es diferente a interpretaciones estándar que han sido presentadas por muchos autores. Nuestra fórmula para el operador *Steklov-Poincaré* es

$$- \left[\left[\underline{\underline{R}} \right] \right] \underline{u} \equiv - \underline{\underline{jR}} \tag{232}$$

en particular, esta no contiene el lado derecho de la ecuación para ser resuelta; ganando por ello, en consistencia teórica. En este respecto notemos que nuestra fórmula es aplicable para cualquier vector (función) independientemente de si está es solución del problema bajo consideración o no.

2.4. Formulación en Término de la Matriz de Complemento de Schur

La siguiente formulación del problema fue introducida por el Dr. Herrera: "Dada $\underline{\overline{f}} \in \bar{D}\left(\overline{\Omega}\right)$, encontrar $\underline{\tilde{u}} \in \bar{D}\left(\overline{\Omega}\right)$ que satisfaga

$$\left(\underline{\underline{L}} + \underline{\underline{\underline{aR}}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{\tilde{u}} = \underline{\overline{f}}." \tag{233}$$

A continuación, se mostrará que cuando $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \rightarrow \tilde{D}\left(\overline{\Omega}\right)$ es positiva definida y $\underline{\overline{f}} \in \bar{D}\left(\overline{\Omega}\right)$ es tomada igual que el vector $\underline{\overline{f}}$ de la Ec.(199), el anterior problema constituye una forma alternativa de formular el problema transformado:

Teorema 36 *Cuando* $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \rightarrow \tilde{D}\left(\overline{\Omega}\right)$ *es positiva definida, una función* $\underline{\tilde{u}} \in \bar{D}\left(\overline{\Omega}\right)$ *es solución de*

$$\left(\underline{\underline{L}} + \underline{\underline{\underline{aR}}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{\tilde{u}} = \underline{\overline{f}}$$

si y sólo si, es solución del problema transformado

$$\begin{cases} \underline{\underline{\underline{aA}}}\underline{\tilde{u}} = \underline{\overline{f}} \\ \underline{\underline{j}}\underline{\tilde{u}} = 0 \end{cases} . \tag{234}$$

Demostración. $\Rightarrow$] Si $\underline{\tilde{u}} \in \bar{D}\left(\overline{\Omega}\right)$ es solución del problema transformado, entonces $\underline{\underline{j}}\underline{\tilde{u}} = 0$. Así la Ec.(233) se reduce a

$$\left(\underline{\underline{L}} + \underline{\underline{\underline{aR}}}\right) \underline{\tilde{u}} = \underline{\overline{f}} \tag{235}$$

y notemos que

$$\underline{\underline{\underline{aA}}} = \underline{\underline{a}}\left(\underline{\underline{L}} + \underline{\underline{R}}\right) = \underline{\underline{L}} + \underline{\underline{\underline{aR}}} \tag{236}$$

esto muestra que la Ec.(235) es equivalente a

$$\underline{\underline{\underline{aA}}}\underline{\tilde{u}} = \underline{\overline{f}}. \tag{237}$$

[$\Leftarrow$ Asumiendo que $\underline{\tilde{u}} \in \bar{D}\left(\overline{\Omega}\right)$ satisface la Ec.(233), entonces multiplicando esta última ecuación por $\underline{\underline{j}}$ y usando el hecho de que $\underline{\underline{j}}\underline{\overline{f}} = 0$, obtenemos

$$\underline{\underline{j}}\underline{\underline{R}}^T \underline{\underline{j}}\underline{\tilde{u}} = 0 \tag{238}$$

esto implica que

$$\underline{\tilde{u}} \cdot \underline{\underline{j}}\underline{\underline{R}}^T \underline{\underline{j}}\underline{\tilde{u}} = \underline{\underline{j}}\underline{\tilde{u}} \cdot \underline{\underline{R}}^T \underline{\underline{j}}\underline{\tilde{u}} = \underline{\underline{j}}\underline{\tilde{u}} \cdot \underline{\underline{\underline{A}}}\underline{\underline{j}}\underline{\tilde{u}} = 0 \tag{239}$$

por lo tanto $\underline{\underline{j}}\underline{\tilde{u}} = 0$ ya que $\underline{\underline{A}}$ es positiva definida. Por lo tanto la Ec.(233) se reduce a la Ec.(235), lo cual muestra que es equivalente a la Ec.(237). ■

Generalmente es ventajoso transformar el problema que hemos considerado en otro en el cual

$$\underline{\overline{f}}_{\Pi} = 0 \tag{240}$$

esto se puede lograr al sustraer el vector auxiliar

$$\underline{u}_p = \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{\overline{f}}_{\Pi} \tag{241}$$

y notemos que esta última ecuación implica

$$\left(\underline{u}_p\right)_{\Delta} = 0 \tag{242}$$

por lo tanto $\underline{\underline{j}}\underline{u}_p = 0$. Definiendo $\underline{u} = \tilde{\underline{u}} - \underline{u}_p$, entonces la Ec.(233) se transforma en

$$\left(\underline{\underline{L}} + \underline{\underline{\underline{aR}}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{u} = \underline{\overline{f}}_{\Delta 2} - \underline{\underline{\underline{aA}}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{\overline{f}}_{\Pi} = \underline{f}_{\Delta 2}. \tag{243}$$

Esta ecuación, en vista del teorema anterior, es equivalente a

$$\left(\underline{\underline{L}} + \underline{\underline{\underline{aR}}}\right) \underline{u} = \underline{f}_{\Delta 2} \qquad \text{y} \qquad \underline{\underline{j}}\underline{u} = 0 \tag{244}$$

esta última ecuación, se satisface si y sólo si

$$\underline{\underline{L}}\underline{u} = 0, \underline{\underline{\underline{aR}}}\underline{u} = \underline{f}_{\Delta 2} \qquad \text{y} \qquad \underline{\underline{j}}\underline{u} = 0. \tag{245}$$

2.4.1. Relación con la Formulación de Multiplicadores de Lagrange

En vista de las definiciones de la sección anterior, la formulación con multiplicadores de Lagrange (FETI-DP) puede ser escrita como

$$\begin{pmatrix} \underline{\underline{A}}_{\Pi\Pi} & \underline{\underline{A}}_{\Pi\Delta} & 0 \\ \underline{\underline{A}}_{\Delta\Pi} & \underline{\underline{A}}_{\Delta\Delta} & \underline{\underline{B}}_{\Delta}^T \\ 0 & \underline{\underline{B}}_{\Delta} & 0 \end{pmatrix} \begin{pmatrix} \underline{u}_{\Pi} \\ \underline{u}_{\Delta} \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} \underline{f}_{\Pi} \\ \underline{f}_{\Delta} \\ 0 \end{pmatrix} \tag{246}$$

con nuestra notación, esto es

$$\left(\underline{\underline{L}} + \underline{\underline{R}}\right) \underline{u} + \underline{\underline{B}}_{\Delta}^T \underline{\lambda} = \underline{f}_{\Pi} + \underline{f}_{\Delta}. \tag{247}$$

por otro lado, en la formulación sin multiplicadores de Lagrange tenemos

$$\left(\underline{\underline{L}} - \left(\underline{\underline{R}}\right)^T \underline{\underline{j}} + \underline{\underline{\underline{aR}}}\right) \bar{\underline{u}} = \underline{f}_{\Pi} + \underline{f}_{\Delta} \tag{248}$$

las Ecs.(247) y (248) implican

$$\underline{\underline{B}}_{\Delta}^T \underline{\lambda} = -\left(\underline{\underline{\underline{jR}}} + \left(\underline{\underline{\underline{jR}}}\right)^T\right) \bar{\underline{u}}. \tag{249}$$

Cuando la Ec.(249) es usada en la Ec.(247) se ve que *'La formulación con Multiplicadores de Lagrange'* se reduce a *'la formulación sin Multiplicadores de Lagrange'*.

2.4.2. Espacio de Vectores Armónicos

El producto interior usado en esta sección es el producto interior de energía Ec.(214), a menos que se indique explícitamente otro producto interior.

Definición 37 *El espacio de funciones armónicas es definido por*

$$D = \left\{ \underline{u} \in \overline{D} \mid \underline{\underline{L}}\,\underline{u} = 0 \right\}. \tag{250}$$

Entonces el problema dado por la Ec.(243) puede escribirse como:
"Encontrar una función amónica $\underline{u} \in D$ que satisfaga

$$\left(\underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}} \right) \underline{u} = \underline{f}_{\Delta 2}." \tag{251}$$

Observación 38 *Dos importantes hechos de las funciones armónicas hay que destacar*

A) Las funciones armónicas son caracterizadas por sus valores duales, i.e. si $\underline{u} \in D$, *entonces*

$$\underline{u}_\Pi = -\left(\underline{\underline{A}}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \underline{u}_\Delta. \tag{252}$$

B) Cuando $\underline{u} \in D$

$$\underline{\underline{A}}\,\underline{u} = \underline{\underline{R}}\,\underline{u} = \underline{\underline{S}}\,\underline{u} \tag{253}$$

donde $\underline{\underline{S}}$ *es la matriz del complemento de Schur Dual-Primal definido por*

$$\underline{\underline{S}} = \underline{\underline{A}}_{\Delta\Delta} - \underline{\underline{A}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \tag{254}$$

además, esta matriz define una transformación $\underline{\underline{S}} : \tilde{D}(\Delta) \to \tilde{D}(\Delta)$ *de* $\tilde{D}(\Delta)$ *en si mismo, el cual es simétrico (y positivo definido, cuando* $\underline{\underline{A}} : \bar{D}\left(\overline{\Omega}\right) \to \bar{D}\left(\overline{\Omega}\right)$ *es positiva definida).*

En el siguiente teorema, se mostrará que la Ec.(251) puede ser remplazada por

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}} \right) \underline{u}_\Delta = \underline{f}_{\Delta 2} \tag{255}$$

por lo tanto, cuando se usen funciones armónicas, nuestro problema se reescribe como:

"Encontrar una función armónica $\underline{u} \in D$, cuyos valores de frontera satisfacen la ecuación

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}} \right) \underline{u}_\Delta = \underline{f}_{\Delta 2}." \tag{256}$$

Esta formulación será referida como la formulación en funciones armónicas. Además, también se mostrará que la Ec.(255) es equivalente a

$$\underline{\underline{aS}}\,\underline{u}_\Delta = \underline{f}_{\Delta 2} \qquad \text{y} \qquad \underline{\underline{j}}\,\underline{u}_\Delta = 0. \tag{257}$$

Teorema 39 *Sea* $\underline{u} \in D$ *y si* $\underline{\underline{A}} : \tilde{D}\left(\overline{\Omega}\right) \rightarrow \tilde{D}\left(\overline{\Omega}\right)$ *es positiva definida, una función* $\underline{\tilde{u}} \in \bar{D}\left(\overline{\Omega}\right)$ *es solución de*

$$\left(\underline{\underline{L}} + \underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{\tilde{u}} = \underline{\overline{f}} \tag{258}$$

las siguientes ecuaciones

$$\left(\underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{u} = \underline{f}_{\Delta 2}, \tag{259}$$

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u}_{\Delta} = \underline{f}_{\Delta 2} \; y \tag{260}$$

$$\underline{\underline{aS}} \underline{u}_{\Delta} = \underline{f}_{\Delta 2} \qquad y \qquad \underline{\underline{j}} \underline{u}_{\Delta} = 0 \tag{261}$$

son equivalentes.

Demostración. $\Rightarrow$] Notemos que la ecuación $\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u}_{\Delta} = \underline{f}_{\Delta 2}$, también puede ser escrita como $\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u} = \underline{f}_{\Delta 2}$, además recordando que las Ecs.(243) y (245) son equivalentes. Ahora, cuando $\underline{u} \in D$, la Ec.(245) se reduce a

$$\underline{\underline{aS}} \underline{u} = \underline{f}_{\Delta 2} \qquad \text{y} \qquad \underline{\underline{j}} \underline{u} = 0 \tag{262}$$

esta ecuación implica $\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u}_{\Delta} = \underline{f}_{\Delta 2}$.

$\Leftarrow$] Ahora, aplicamos la matriz $\underline{\underline{j}}$ a la ecuación $\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u} = \underline{f}_{\Delta 2}$ para obtener

$$\underline{\underline{j}} \underline{\underline{S}} \underline{\underline{j}} \underline{u}_{\Delta} = 0 \tag{263}$$

esta ecuación implica

$$\underline{\underline{j}} \underline{u} = \underline{\underline{j}} \underline{u}_{\Delta} = 0 \tag{264}$$

ya que la matriz $\underline{\underline{S}} : \tilde{D}(\Delta) \rightarrow \tilde{D}(\Delta)$ es positiva definida, la ecuación $\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u} = \underline{f}_{\Delta 2}$ se reduce a

$$\underline{\underline{aS}} \underline{u} = \underline{f}_{\Delta 2}. \tag{265}$$

■

En lo que sigue este resultado es usado para derivar una variedad de métodos de descomposición de dominio sin traslape, los cuales permiten obtener los valores en la frontera interior $\underline{u}_{\Delta} \in \tilde{D}(\Delta)$. Una vez conocidos los nodos $\underline{u}_{\Delta}$, los nodos en $\underline{u}_{\Pi} \in \tilde{D}(\Pi)$ son resueltos mediante la Ec. (252).

2.5. Métodos basados en la Formulación Matricial Discontinua

Por simplicidad de la presentación, en esta sección asumiremos que la matriz $\underline{\underline{A}}$ es positiva definida (el caso cuando no es positiva definida se discutirá en la sección (2.6)). Cuando $\underline{\underline{A}}$ es positiva definida la matriz del complemento de

Schur Dual-Primal $\underline{\underline{S}} : \tilde{D}(\Delta) \rightarrow \tilde{D}(\Delta)$ es también positiva definida. Entonces adicionalmente al producto interior Euclidiano definiremos un segundo producto interior en el espacio $\tilde{D}(\Delta)$, este será el producto interior de energía.

Definición 40 *Cuando la matriz $\underline{\underline{A}}$ es positiva definida, definimos al producto interior de energía $(\cdot,\cdot)$ por*

$$(\underline{u},\underline{w}) \equiv \underline{w}\underline{\underline{A}}\underline{u}. \tag{266}$$

Ambos productos interiores están definidos sobre $\tilde{D}\left(\overline{\Omega}\right)$, sin embargo notemos que la restricción del producto interior de energía a $\tilde{D}(\Delta)$ puede ser expresado como

$$(\underline{u},\underline{w}) \equiv \underline{w} \cdot \underline{\underline{S}}\underline{u}, \forall \underline{u},\underline{w} \in \bar{D}(\Delta). \tag{267}$$

2.5.1. Métodos Single-Trip

Partiendo de la formulación dada por la Ec.(258) obtendremos el método del complemento de Schur, usando la aproximación Dirichlet, y el método FETI de un nivel usando la aproximación Neumann.

El Enfoque Dirichlet Problema 1.- "Este problema consiste en buscar una función $\underline{u}_{\Delta} \in \bar{D}_{12}(\Delta)$ tal que satisfaga

$$\underline{\underline{aS}}\underline{u} = \underline{f}_{\Delta 2}\text{"}. \tag{268}$$

Observación 41 *Claramente esta formulación se basa en la Ec.(261), además notemos que $\underline{\underline{aS}}$ define una transformación de $\underline{\underline{aS}} : \bar{D}(\Delta) \rightarrow \bar{D}(\Delta)$, además $\underline{\underline{aS}}$ es simétrica y positiva definida sobre $\tilde{D}(\Delta)$, por lo tanto la Ec.(268) es susceptible de la aplicación del método de Gradiente Conjugado, y el método de descomposición de dominio obtenido es el conocido método de Complemento de Schur. Sin embargo en nuestra formulación también se incluye la posibilidad de incluir nodos primales algo que imposible de hacer en las formulaciones estándar.*

Así, el problema 1 se implementa como un método de Gradiente Conjugado para resolver el sistema y queda de manera esquemática como se indica a continuación:

$\underline{r}^0 = \underline{\underline{a}}\underline{f}_{\Delta_2} - \underline{\underline{aS}}\underline{u}^0$
$\underline{p} = \underline{r}^0$

$$\begin{array}{l} 1)\ \alpha^n = \frac{\underline{p}^n \cdot \underline{p}^n}{\underline{p}^n \cdot \underline{\underline{aS}}\underline{p}^n} \\ 2)\ \underline{u}^{n+1} = \underline{u}^n + \alpha^n \underline{p}^n \\ 3)\ \underline{r}^{n+1} = \underline{r}^n - \alpha^n \underline{\underline{aS}}\underline{p}^n \\ 4)\ \beta^n = \frac{\underline{r}^{n+1} \cdot \underline{r}^{n+1}}{\underline{r}^n \cdot \underline{r}^n} \\ 5)\ \underline{p}^{n+1} = \underline{r}^{n+1} + \beta^n \underline{p}^n \\ 6)\ n = n+1 \text{ y regresar a 1)} \end{array} \tag{269}$$

El Enfoque Neumann Una segunda formulación puede ser derivada de la Ec.(260) cuando $\underline{u}_{\Delta}^{FT} \in D$ es definida como

$$\underline{u}_{\Delta}^{FT} = \underline{u}_{\Delta} - \underline{\underline{S}}^{-1}\overline{\underline{f}}_{\Delta_2} \tag{270}$$

entonces

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right)\underline{u}_{\Delta}^{FT} = \underline{\underline{Sj}}\,\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2} \tag{271}$$

la cual se satisface si y sólo si

$$\underline{\underline{aS}}\,\underline{u}_{\Delta}^{FT} = 0 \qquad \text{y} \qquad \underline{\underline{j}}\,\underline{u}_{\Delta}^{FT} = -\underline{\underline{j}}\,\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2}. \tag{272}$$

Definición 42 *Definimos el subespacio* $\tilde{D}_{22}(\Delta) \subset \tilde{D}(\Delta)$ *por*

$$\tilde{D}_{22}(\Delta) \equiv \left\{\underline{w} \in \tilde{D}(\Delta) \mid \underline{\underline{aS}}\,\underline{w} = 0\right\}. \tag{273}$$

Problema 2.- "Este problema consiste en buscar una función $\underline{u}_{\Delta}^{FT} \in \bar{D}_{22}(\Delta)$ tal que satisfaga

$$\underline{\underline{j}}\,\underline{u}_{\Delta}^{FT} = -\underline{\underline{j}}\,\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2}\text{"}. \tag{274}$$

Entonces la formulación Neumann o FETI no precondicionado del problema es obtenido multiplicando la Ec.(274) por $\underline{\underline{S}}^{-1}$. Quedando definido por:

"Buscar una función $\underline{u}_{\Delta}^{FT} \in \bar{D}_{22}(\Delta)$ tal que satisfaga

$$\underline{\underline{S}}^{-1}\underline{\underline{j}}\,\underline{u}_{\Delta}^{FT} = -\underline{\underline{S}}^{-1}\underline{\underline{j}}\,\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2}\text{"}. \tag{275}$$

Observación 43 *Una importante propiedad de la formulación Neumann de la Ec.(275) es que la matriz* $\underline{\underline{S}}^{-1}\underline{\underline{j}}$ *cuando es aplicada a cualquier vector* $\underline{v} \in \tilde{D}(\Delta)$ *se obtiene un vector* $\underline{\underline{S}}^{-1}\underline{\underline{j}}\,\underline{v} \in \tilde{D}_{22}(\Delta)$, *tal que* $\underline{\underline{S}}^{-1}\underline{\underline{j}} : \tilde{D}_{22}(\Delta) \to \tilde{D}_{22}(\Delta)$. *Esta transformación es auto adjunta y positiva definida sobre* $\tilde{D}_{22}(\Delta)$ *con respecto al producto interior de energía, esto se desprende del hecho de que*

$$\underline{\underline{S}}\left(\underline{\underline{S}}^{-1}\underline{\underline{j}}\right) = \underline{\underline{j}} \tag{276}$$

además cuando $\underline{u} \in \tilde{D}_{22}(\Delta)$

$$\underline{\underline{j}}\,\underline{u} = \underline{u} - \underline{\underline{a}}\,\underline{u} \qquad y \qquad \underline{\underline{a}}\,\underline{u} \cdot \underline{\underline{S}}\,\underline{u} = 0 \tag{277}$$

por lo tanto

$$\underline{\underline{j}}\,\underline{u} \cdot \underline{\underline{S}}\,\underline{\underline{j}}\,\underline{u} = \underline{u} \cdot \underline{\underline{S}}\,\underline{u} + \underline{\underline{a}}\,\underline{u} \cdot \underline{\underline{Sa}}\,\underline{u} \tag{278}$$

entonces

$$\underline{\underline{j}}\,\underline{u} = 0 \Rightarrow \underline{u} \cdot \underline{\underline{S}}\,\underline{u} + \underline{\underline{a}}\,\underline{u} \cdot \underline{\underline{Sa}}\,\underline{u} = \underline{\underline{j}}\,\underline{u} \cdot \underline{\underline{S}}\,\underline{\underline{j}}\,\underline{u} = 0 \Rightarrow \underline{u} \cdot \underline{\underline{S}}\,\underline{u} = 0 \Rightarrow \underline{u} = 0 \tag{279}$$

esto muestra que cuando $\underline{u} \in \tilde{D}_{22}(\Delta)$, $\underline{\underline{j}}\,\underline{u} = 0 \Rightarrow \underline{\underline{a}}\,\underline{u} = 0$ *y por lo tanto*

$$\underline{u} = \underline{\underline{a}}\,\underline{u} + \underline{\underline{j}}\,\underline{u} = 0 \tag{280}$$

en conclusión, la matriz $\underline{\underline{j}}$ *es positiva definida sobre* $\tilde{D}_{22}(\Delta)$, *por lo tanto la Ec.(275) es susceptible de la aplicación del método de Gradiente Conjugado. En nuestra formulación también se incluye la posibilidad de incluir nodos primales.*

Así, el problema 2 se implementa como un método de Gradiente Conjugado para resolver el sistema y queda de manera esquemática como se indica a continuación:

$$\begin{array}{l}\underline{r}^0 = \left[-\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2}\right] - \underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{u}^0 \\ \underline{p} = \underline{r}^0 \\ \qquad 1)\ \alpha^n = \frac{\underline{p}^n \cdot \underline{\underline{S}}\underline{p}^n}{\underline{p}^n \cdot \underline{\underline{j}}\underline{p}^n} \\ \qquad 2)\ \underline{u}^{n+1} = \underline{u}^n + \alpha^n \underline{p}^n \\ \qquad 3)\ \underline{r}^{n+1} = \underline{r}^n - \alpha^n \underline{\underline{j}}\underline{p}^n \\ \qquad 4)\ \beta^n = \frac{\underline{r}^{n+1} \cdot \underline{\underline{S}}\underline{r}^{n+1}}{\underline{r}^n \cdot \underline{\underline{S}}\underline{r}^n} \\ \qquad 5)\ \underline{p}^{n+1} = \underline{r}^{n+1} + \beta^n \underline{p}^n \\ \qquad 6)\ n = n+1 \text{ y regresar a 1)}\end{array} \tag{281}$$

2.5.2. Métodos Round-Trip

En esta sección presentaremos dos algoritmos Round-Trip comúnmente nombrados como algoritmo Neumann-Neumann y el algoritmo FETI precondicionado.

El Enfoque Neumann-Neumann El procedimiento iterativo del complemento de Schur queda dado por:

Problema 1.- "Este problema consiste en buscar una función $\underline{u}_\Delta \in \tilde{D}_{12}(\Delta)$ tal que

$$\underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{u}_\Delta = \underline{\underline{aS}}^{-1}\underline{f}_{\Delta_2}\text{"} \tag{282}$$

Observación 44 *La Ec.(282) es equivalente a la Ec.(268) porque cuando esta última ecuación es multiplicada por* $\underline{\underline{aS}}^{-1}\underline{\underline{a}}$ *se obtiene la Ec.(282), además* $\underline{\underline{aS}}^{-1}\underline{\underline{a}}$ *es positivo definido sobre* $\tilde{D}_{12}(\Delta)$. *De tal forma que la Ec.(282) es susceptible de la aplicación del método de Gradiente Conjugado usando el producto interior de energía, ya que cuando* $\underline{u} \in \tilde{D}_{12}(\Delta)$, *entonces* $\underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{u} \in \tilde{D}_{12}(\Delta)$, *por lo tanto la matriz* $\underline{\underline{aS}}^{-1}\underline{\underline{aS}}$ *define una transformación de* $\tilde{D}_{12}(\Delta)$ *sobre si misma, la cual es auto adjunta y positiva definida con respecto al producto interior de energía. Esto también puede ser observado ya que la matriz* $\underline{\underline{SaS}}^{-1}\underline{\underline{aS}}$ *es simétrica y positivo definido.*

Además la Ec.(282) puede ser interpretada como un precondicionamiento de la Ec.(268), con $\underline{\underline{aS}}^{-1}$ *como precondicionador.*

La matriz $\underline{\underline{aS}}^{-1}\underline{\underline{aS}} : \tilde{D}_{12}(\Delta) \to \tilde{D}_{12}(\Delta)$ satisface la siguiente identidad

$$\underline{\underline{aS}}^{-1}\underline{\underline{aS}} = \underline{\underline{I}} - \underline{\underline{aS}}^{-1}\underline{\underline{j}}\underline{\underline{S}}$$

la cual puede ser útil cuando se aplica el método de Gradiente Conjugado, y es también necesario para llevar a cabo el análisis del número de condicionamiento.

Así, el método se implementa como un método de Gradiente Conjugado para resolver el sistema y queda de manera esquemática como se indica a continuación:

$$\underline{r}^0 = \underline{\underline{aS}}^{-1}\left[\underline{f}_{\Delta 2} - \underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_{\Pi}\right] - \underline{\underline{aS}}\underline{u}^0$$
$$\underline{p} = \underline{r}^0$$

$$\begin{aligned}
&1)\ \alpha^n = \frac{\underline{p}^n \cdot \underline{p}^n}{\underline{p}^n \cdot \underline{\underline{SaS}}^{-1}\underline{\underline{aS}}\underline{p}^n} \\
&2)\ \underline{u}^{n+1} = \underline{u}^n + \alpha^n \underline{p}^n \\
&3)\ \underline{r}^{n+1} = \underline{r}^n - \alpha^n \underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{p}^n \\
&4)\ \beta^n = \frac{\underline{r}^{n+1} \cdot \underline{\underline{S}}\underline{r}^{n+1}}{\underline{r}^n \cdot \underline{\underline{S}}\underline{r}^n} \\
&5)\ \underline{p}^{n+1} = \underline{r}^{n+1} + \beta^n \underline{p}^n \\
&6)\ n = n+1 \text{ y regresar a 1)}
\end{aligned} \tag{283}$$

El Enfoque FETI Precondicionado El procedimiento iterativo de FETI precondicionado queda dado por:

Problema 2.- "Este problema consiste en buscar una función $\underline{u}_\Delta \in \tilde{D}_{22}(\Delta)$ tal que

$$\underline{\underline{S}}^{-1}\underline{\underline{jSj}}\underline{u}^{FT} = -\underline{\underline{S}}^{-1}\underline{\underline{jSj}}\underline{\underline{S}}^{-1}\underline{f}_{\Delta_2}\text{"} \tag{284}$$

Observación 45 *La Ec.(284) es equivalente a la Ec.(274) ya que cuando la última ecuación es multiplicada por* $\underline{\underline{S}}^{-1}\underline{\underline{jSj}}$ *la Ec.(284) es obtenida, y* $\underline{\underline{S}}^{-1}\underline{\underline{jS}}$ *es positiva definida sobre* $\tilde{D}_{22}(\Delta)$. *Cuando* $\underline{u}_\Delta \in \tilde{D}_{22}(\Delta)$, *entonces* $\underline{\underline{S}}^{-1}\underline{\underline{jSj}}\underline{u} \in \tilde{D}_{22}(\Delta)$, *por lo tanto, la matriz* $\underline{\underline{S}}^{-1}\underline{\underline{jSj}}$ *define una transformación de* $\tilde{D}_{22}(\Delta)$ *en si misma. Esta transformación es auto adjunta y positiva definida con respecto al producto interior de energía, esto se sigue del hecho de que la matriz*

$$\underline{\underline{S}}\left(\underline{\underline{S}}^{-1}\underline{\underline{jSj}}\right) = \underline{\underline{jSj}} \tag{285}$$

es simétrica y positiva definida sobre $\tilde{D}_{22}(\Delta)$. *Por lo tanto, usando el producto interior de energía, es susceptible de la aplicación del método de Gradiente Conjugado.*

También la Ec.(284) puede interpretarse como una versión precondicionada de la Ec.(274), con $\underline{\underline{S}}^{-1}\underline{\underline{jS}}$ *como precondicionador*

Además, cuando se aplica el método de Gradiente Conjugado a las Ec.(284) la matriz a ser iterada es $\underline{\underline{S}}^{-1}\underline{\underline{jSj}} : \tilde{D}_{22}(\Delta) \to \tilde{D}_{22}(\Delta)$ y debe de ser usado el producto interior de energía. Cuando se usa con éxito este procedimiento, la siguiente identidad se sigue

$$\underline{\underline{S}}^{-1}\underline{\underline{jSj}} = \underline{\underline{I}} - \underline{\underline{S}}^{-1}\underline{\underline{jSa}} \tag{286}$$

la cual puede ser usada. Cuando $\underline{u} \in \tilde{D}_{22}(\Delta)$ entonces

$$\underline{\underline{S}}^{-1}\underline{\underline{jSju}} = \underline{u} - \underline{\underline{S}}^{-1}\underline{\underline{jSau}} \tag{287}$$

de la cual se sigue que

$$\underline{u} = \underline{\underline{S}}^{-1}\underline{\underline{S}}\left(\underline{\underline{a}}\underline{u} + \underline{\underline{j}}\underline{u}\right) = \underline{\underline{S}}^{-1}\underline{\underline{jS}}\left(\underline{\underline{a}}\underline{u} + \underline{\underline{j}}\underline{u}\right),\ \forall \underline{u} \in \tilde{D}_{22}(\Delta) \tag{288}$$

Así, el método se implementa como un método de Gradiente Conjugado para resolver el sistema y queda de manera esquemática como se indica a continuación:

$$\begin{aligned}
&\underline{r}^0 = -\underline{\underline{S}}^{-1}\underline{\underline{jSj}}\underline{\underline{S}}^{-1}\left[\underline{f}_{\Delta 2} - \underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_{\Pi}\right] - \underline{\underline{aS}}\underline{u}^0\\
&\underline{p} = \underline{r}^0\\
&\qquad 1)\ \alpha^n = \frac{\underline{p}^n \cdot \underline{p}^n}{\underline{p}^n \cdot \underline{\underline{jSj}}\underline{p}^n}\\
&\qquad 2)\ \underline{u}^{n+1} = \underline{u}^n + \alpha^n \underline{p}^n\\
&\qquad 3)\ \underline{r}^{n+1} = \underline{r}^n - \alpha^n \underline{\underline{S}}^{-1}\underline{\underline{jSj}}\underline{p}^n\\
&\qquad 4)\ \beta^n = \frac{\underline{r}^{n+1} \cdot \underline{\underline{S}}\underline{r}^{n+1}}{\underline{r}^n \cdot \underline{\underline{S}}\underline{r}^n}\\
&\qquad 5)\ \underline{p}^{n+1} = \underline{r}^{n+1} + \beta^n \underline{p}^n\\
&\qquad 6)\ n = n+1 \text{ y regresar a 1)}
\end{aligned} \tag{289}$$

En esta sección trataremos dos casos de la forma como se selecciono el conjunto $\overline{\Omega}$, estos dos casos son cuando todos los nodos primales son interiores y cuando no todos los nodos primales son interiores, además trataremos el caso cuando la matriz $\underline{\underline{A}}$ es no positiva definida y la correspondiente matriz del complemento de Schur también tiene esta propiedad.

2.5.3. Caso Cuando Todos los Nodos Primales son Interiores

Este caso corresponde cuando $\left\{\overline{\Omega}_1, ..., \overline{\Omega}_E\right\}$ es la partición del dominio $\overline{\Omega} \subset \mathbb{R}^n$, i.e.

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

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_i respectivamente, claramente

$$\partial\Omega \subset \bigcup_{\alpha=1}^{E} \partial\Omega_\alpha. \tag{290}$$

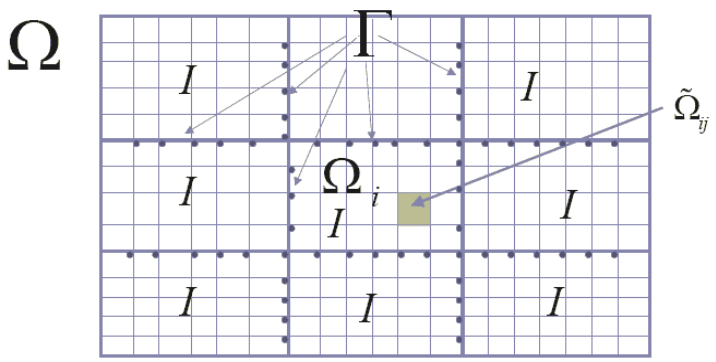

Figura 1: El dominio Ω, su frontera externa $\partial\Omega$ y la frontera interna Γ.

Adicionalmente definimos

$$\Gamma_\alpha = \Delta \cap \overline{\Omega}_\alpha \qquad \text{y} \qquad I_\alpha = \Pi \cap \overline{\Omega}_\alpha \tag{291}$$

junto con

$$\Gamma = \bigcup_{\alpha=1}^{E} \Gamma_\alpha \qquad \text{y} \qquad I = \bigcup_{\alpha=1}^{E} I_\alpha \tag{292}$$

entonces

$$\overline{\Omega} = \Gamma \cup I \text{ mientras } \Delta = \Gamma \text{ y } \Pi = I \tag{293}$$

además, para cada $\alpha = 1, ..., E$, definimos la matriz

$$\underline{\underline{A}}^\alpha : D\left(\overline{\Omega}_\alpha\right) \to D\left(\overline{\Omega}_\alpha\right) \tag{294}$$

por la condición de que la forma bilineal asociada a esta sea la restricción a $D\left(\overline{\Omega}_\alpha\right) \times D\left(\overline{\Omega}_\alpha\right)$ de la forma bilineal asociada con $\underline{\underline{A}}^\alpha : D\left(\overline{\Omega}\right) \to D\left(\overline{\Omega}\right)$, o más precisamente

$$\underline{w} \cdot \underline{\underline{A}}^\alpha \underline{v} = \underline{w}^\alpha \cdot \underline{\underline{A}} \underline{v}^\alpha, \forall \underline{w}, \underline{v} \in D\left(\overline{\Omega}_\alpha\right) \tag{295}$$

aquí

$$\underline{v} = \sum_{\alpha=1}^{E} \underline{v}^\alpha \text{ y } \underline{w} = \sum_{\alpha=1}^{E} \underline{w}^\alpha \text{ con } \underline{w}^\alpha, \underline{v}^\alpha \in D\left(\overline{\Omega}_\alpha\right) \tag{296}$$

entonces, escribimos

$$\underline{\underline{A}}^\alpha = \begin{pmatrix} \underline{\underline{A}}^\alpha_{II} & \underline{\underline{A}}^\alpha_{I\Gamma} \\ \underline{\underline{A}}^\alpha_{\Gamma I} & \underline{\underline{A}}^\alpha_{\Gamma\Gamma} \end{pmatrix} \tag{297}$$

además, para el caso discutido en esta sección, tenemos

$$\underline{\underline{A}} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^\alpha \tag{298}$$

por lo tanto, tenemos

$$\begin{cases} \underline{\underline{A}}_{II} = \sum_{\alpha=1}^{E} \underline{\underline{A}}_{II}^{\alpha}, & \underline{\underline{A}}_{I\Gamma} = \sum_{\alpha=1}^{E} \underline{\underline{A}}_{I\Gamma}^{\alpha} \\ \underline{\underline{A}}_{\Gamma I} = \sum_{\alpha=1}^{E} \underline{\underline{A}}_{\Gamma I}^{\alpha}, & \underline{\underline{A}}_{\Gamma\Gamma} = \sum_{\alpha=1}^{E} \underline{\underline{A}}_{\Gamma\Gamma}^{\alpha} \end{cases} . \tag{299}$$

Notemos las siguientes propiedades importantes

$$\left(\underline{\underline{A}}_{II}\right)^{-1} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}_{II}^{\alpha}\right)^{-1} \qquad \text{y} \qquad \left(\underline{\underline{A}}\right)^{-1} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}\right)^{-1} \tag{300}$$

esto implica que el cálculo de las inversas de las matrices $\underline{\underline{A}}_{II}$ y $\underline{\underline{A}}$ requiere calcular exclusivamente las inversas locales.

Ahora las Ecs.(**??**), (**??**) y (**??**) se convierten en

$$\left(\underline{v}_{11}\right)_I = -\sum_{\alpha=1}^{E} \left(\underline{\underline{A}}_{II}^{\alpha}\right)^{-1} \underline{\underline{A}}_{I\Gamma}^{\alpha} \underline{j\underline{v}}_{\Gamma} \tag{301}$$

junto con

$$\left(\underline{v}_{12}\right)_I = -\sum_{\alpha=1}^{E} \left(\underline{\underline{A}}_{II}^{\alpha}\right)^{-1} \underline{\underline{A}}_{I\Gamma}^{\alpha} \underline{av}_{\Gamma} \tag{302}$$

y

$$\underline{v}_{21} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}\right)^{-1} \underline{\underline{aRv}} \tag{303}$$

junto con

$$\underline{v}_{22} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}\right)^{-1} \underline{\underline{jRv}}$$

similarmente las Ecs.(**??**) y (**??**) se transforman en

$$\underline{u}_{21} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}\right)^{-1} \left(\overline{\underline{f}}_{\Delta} - \left[\left[\underline{R}\right]\right] \widetilde{\underline{u}}_{p}\right) \tag{304}$$

y

$$\left(\underline{u}_{11}\right)_{\Pi} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}_{II}^{\alpha}\right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \underline{j\widetilde{u}}_{p} \tag{305}$$

finalmente, recordemos una vez más que en vista de la Ec.(300) el cálculo de las componentes del vector harmónico requiere el cálculo de inversas locales exclusivamente.

2.5.4. Caso Cuando no Todos los Nodos Primales son Interiores

En esta sección trataremos el caso en que no todos los nodos primales son interiores. El conjunto de nodos primales que no son interiores serán denotados por $\pi \subset \Pi$ y denotaremos s a los nodos interiores por $I \subset \Pi$. Entonces

$$\Pi = \pi \cup I \text{ y } \pi \cap I = \emptyset \tag{306}$$

además, denotamos

$$\Gamma = \Delta \tag{307}$$

y usando una notación similar a la usada en las definiciones dadas por las Ecs.(215) a (219) tenemos que

$$\underline{\underline{A}} \equiv \begin{pmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\pi} & \underline{\underline{A}}_{I\Delta} \\ \underline{\underline{A}}_{\pi I} & \underline{\underline{A}}_{\pi\pi} & \underline{\underline{A}}_{\pi\Delta} \\ \underline{\underline{A}}_{\Delta I} & \underline{\underline{A}}_{\Delta\pi} & \underline{\underline{A}}_{\Delta\Delta} \end{pmatrix} \tag{308}$$

y

$$\underline{\underline{L}} \equiv \begin{pmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\pi} & \underline{\underline{A}}_{I\Delta} \\ \underline{\underline{A}}_{\pi I} & \underline{\underline{A}}_{\pi\pi} & \underline{\underline{A}}_{\pi\Delta} \\ 0 & 0 & 0 \end{pmatrix} \text{ y } \underline{\underline{R}} \equiv \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \underline{\underline{A}}_{\Delta I} & \underline{\underline{A}}_{\Delta\pi} & \underline{\underline{A}}_{\Delta\Delta} \end{pmatrix} \tag{309}$$

notemos que

$$\underline{\underline{A}}_{\Pi\Pi} \equiv \begin{pmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\pi} & 0 \\ \underline{\underline{A}}_{\pi I} & \underline{\underline{A}}_{\pi\pi} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{310}$$

Dada $\underline{v} \in D$ con componentes $\underline{v}_{11}$ y $\underline{v}_{12}$ están dadas por los dos sistemas de ecuaciones de la Ec.(**??**). Ellos son

$$\begin{pmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\pi} \\ \underline{\underline{A}}_{\pi I} & \underline{\underline{A}}_{\pi\pi} \end{pmatrix} \begin{pmatrix} (\underline{v}_{11})_I \\ (\underline{v}_{11})_\pi \end{pmatrix} = - \begin{pmatrix} \underline{\underline{A}}_{I\Delta} \underline{\underline{j}} \underline{v}_\Delta \\ \underline{\underline{A}}_{\pi\Delta} \underline{\underline{j}} \underline{v}_\Delta \end{pmatrix} \tag{311}$$

y

$$\begin{pmatrix} \underline{\underline{A}}_{II} & \underline{\underline{A}}_{I\pi} \\ \underline{\underline{A}}_{\pi I} & \underline{\underline{A}}_{\pi\pi} \end{pmatrix} \begin{pmatrix} (\underline{v}_{12})_I \\ (\underline{v}_{12})_\pi \end{pmatrix} = - \begin{pmatrix} \underline{\underline{A}}_{I\Delta} \underline{\underline{a}} \underline{v}_\Delta \\ \underline{\underline{A}}_{\pi\Delta} \underline{\underline{a}} \underline{v}_\Delta \end{pmatrix} \tag{312}$$

respectivamente. Cuando la cardinalidad de π es más pequeña que la asociada con I, una eficiente manera de resolver estos sistemas es por la aplicación del enfoque del complemento de Schur estándar a la matriz del lado derecho de estas ecuaciones. Esto obtiene $\underline{v}_{11}$ y $\underline{v}_{12}$.

Las componentes de $\underline{v}_{21}$ y $\underline{v}_{22}$, por otro lado, están dadas por la Ec.(**??**) y (**??**). Ellas son

$$\underline{\underline{A}} \underline{v}_{21} = \underline{\underline{\underline{a}}} \underline{R} \underline{v} \qquad \text{y} \qquad \underline{\underline{A}} \underline{v}_{22} = \underline{\underline{\underline{j}}} \underline{R} \underline{v}. \tag{313}$$

Definimos la restricción del Promedio y la restricción del Salto como aquellos en los que tomamos el promedio y el salto solo de los grados de libertad primales; ellos serán denotados por $\overline{\underline{\underline{a}}}$ y $\overline{\underline{\underline{j}}}$ respectivamente. Cuando el precondicionador

Dual-Primal sea usado, requerimos que $\overline{\overline{j}}\underline{u}_{21} = 0$ y $\overline{\overline{j}}\underline{u}_{22} = 0$. Pero reordenando, podemos escribir cada vector $\underline{w} \in D \subset \tilde{D}\left(\overline{\Omega}\right)$ en la forma

$$\underline{w} = \begin{pmatrix} \underline{w}_I \\ \underline{w}_\pi \\ \underline{w}_\Gamma \end{pmatrix} \rightarrow \underline{w} = \begin{pmatrix} \underline{w}_J \\ \underline{w}_\pi \end{pmatrix} \tag{314}$$

donde

$$\underline{w}_J = \begin{pmatrix} \underline{w}_I \\ \underline{w}_\Gamma \end{pmatrix} \tag{315}$$

entonces

$$\underline{\underline{A}}\underline{w} \rightarrow \underline{\underline{Q}}\underline{w} \tag{316}$$

con

$$\underline{\underline{Q}}^i = \begin{pmatrix} \underline{\underline{Q}}^i_{JJ} & \underline{\underline{Q}}^i_{J\pi} \\ \underline{\underline{Q}}^i_{\pi J} & \underline{\underline{Q}}^i_{\pi\pi} \end{pmatrix} \tag{317}$$

con $i = 1, ...E$ y

$$\underline{\underline{Q}} = \sum_{i=1}^{E} \underline{\underline{Q}}^i \tag{318}$$

entonces, definimos el complemento de Schur por medio de

$$\underline{\underline{S}}^i = \underline{\underline{Q}}^i_{\pi\pi} - \underline{\underline{Q}}^i_{\pi J}\left(\underline{\underline{Q}}^i_{JJ}\right)^{-1}\underline{\underline{Q}}^i_{J\pi} \tag{319}$$

con $i = 1, ...E$ y

$$\underline{\underline{S}} = \sum_{i=1}^{E} \underline{\underline{S}}^i \tag{320}$$

se puede ver que el complemento de Schur es positivo definido en el subespacio donde $\overline{\overline{j}}\underline{w} = 0$. Además, la Ec.(313) puede expresarse en términos del complemento de Schur, y entonces el método de Gradiente Conjugado puede aplicarse.

2.6. Problemas del tipo del Operador Laplaciano

En esta sección se extenderá la teoría de las secciones anteriores al caso cuando la matriz $\underline{\underline{A}}$ no es positiva definida y la correspondiente matriz del complemento de Schur $\underline{\underline{S}}$ tiene esta misma propiedad.

Definición 46 *El subespacio nulo de* $\underline{\underline{S}} : \tilde{D}(\Delta) \rightarrow \tilde{D}(\Delta)$ *se denotará por* N_S, *de manera que* $N_S \subset \tilde{D}(\Delta)$, *además denotaremos por* E *al rango de* $\underline{\underline{S}}$, *este también satisface* $E \subset \tilde{D}(\Delta)$.

Entonces notemos que

$$\tilde{D}(\Delta) = N_S \oplus E \tag{321}$$

esto es porque N_S es el complemento ortogonal de E con respecto al producto interior Euclidiano.

Definición 47 *Las matrices* $\underline{\underline{I}}_S^C : \tilde{D}(\Delta) \to N_S$ *y* $\underline{\underline{I}}_S^R : \tilde{D}(\Delta) \to E$ *denota las matrices proyección con respecto al producto interior Euclidiano sobre* N_S *y* E *respectivamente, i.e. para cualquier* $\underline{u} \in \tilde{D}(\Delta)$, $\underline{\underline{I}}_S^C \underline{u}$ *y* $\underline{\underline{I}}_S^R \underline{u}$ *son tales proyecciones.*

Notemos el hecho de que

$$N_S \cap \tilde{D}_{12}(\Delta) = \{0\} \tag{322}$$

esto es claro, ya que $\underline{\underline{S}}$ es positiva definida en el espacio lineal de las funciones continuas, observemos además que $\underline{\underline{j}}$ es no negativa sobre $\tilde{D}(\Delta)$ y es positiva definida sobre N_S.

Introducimos ahora las siguientes definiciones y resultados, en ellos las relaciones de ortogonalidad y de proyección, son entendidas que son con respecto al producto interior Euclidiano:

1.-

$$\tilde{D}_{11}^w(\Delta) \equiv \underline{\underline{j}} N_S \subset \underline{\underline{j}} \tilde{D}(\Delta) = \tilde{D}_{11}(\Delta). \tag{323}$$

2.- El espacio $\tilde{D}_{12}^w(\Delta) \subset \tilde{D}(\Delta)$ es definido por el complemento ortogonal de $\tilde{D}_{11}^w(\Delta)$.

3.- Las matrices $\underline{\underline{j}}^w : \tilde{D}(\Delta) \to \tilde{D}_{11}^w(\Delta)$ y $\underline{\underline{a}}^w : \tilde{D}(\Delta) \to \tilde{D}_{12}^w(\Delta)$ están definidas por las matrices proyección sobre $\tilde{D}_{11}^w(\Delta)$ y $\tilde{D}_{12}^w(\Delta)$ respectivamente, i.e.

$$\underline{\underline{j}}^w \underline{u} \equiv Proy_{\tilde{D}_{11}^w(\Delta)} \underline{u} \qquad \text{y} \qquad \underline{\underline{a}}^w \underline{u} \equiv Proy_{\tilde{D}_{12}^w(\Delta)} \underline{u}. \tag{324}$$

4.-

$$\underline{\underline{I}} = \underline{\underline{a}}^w + \underline{\underline{j}}^w. \tag{325}$$

5.- Cada $\underline{u} \in \tilde{D}(\Delta)$ puede ser escrita como

$$\underline{u} = \underline{u}_{11}^w + \underline{u}_{12}^w \tag{326}$$

donde $\underline{u}_{11}^w \equiv \underline{\underline{j}}^w \underline{u} \in \tilde{D}_{11}^w(\Delta)$ y $\underline{u}_{12}^w \equiv \underline{\underline{a}}^w \underline{u} \in \tilde{D}_{12}^w(\Delta)$.

6.- Cuando $\underline{u} \in N_S$ tenemos que

$$\underline{u} = \underline{\underline{j}} \underline{u} + \underline{\underline{a}} \underline{u} \tag{327}$$

donde $\underline{\underline{j}} \underline{u} \in \tilde{D}_{11}^w(\Delta)$ y $\underline{\underline{a}} \underline{u} \in \tilde{D}_{12}(\Delta) \subset \tilde{D}_{12}^w(\Delta)$ también

$$\underline{\underline{j}}^w \underline{u} = \underline{\underline{j}} \underline{u} \qquad \text{y} \qquad \underline{\underline{a}}^w \underline{u} = \underline{\underline{a}} \underline{u}. \tag{328}$$

7.- Por lo tanto

$$\tilde{D}_{11}^w(\Delta) + \tilde{D}_{12}^w(\Delta) \supset N_S. \tag{329}$$

8.- Cuando $\underline{u} \in N_S$ y $\underline{\underline{j}}^w \underline{u} = 0$, entonces $\underline{u} = 0$. Ya que, cuando las Ecs.(327) y (328) se satisfacen, tal que $\underline{\underline{j}}^w \underline{u} = 0$ implica que $\underline{u} \in \tilde{D}_{12}(\Delta)$ y por la Ec.(322) entonces $\underline{u} = 0$.

9.- Como corolario de la 8) tenemos que

$$\tilde{D}_{12}^{w}(\Delta) \cap N_S = \{0\}. \tag{330}$$

10.- $\underline{\underline{S}}$ es positiva definida sobre $\tilde{D}_{12}^{w}(\Delta)$ en virtud de la Ec.(330).

11.- La matriz $\underline{\underline{M}} : \tilde{D}(\Delta) \to \tilde{D}(\Delta)$ definida por

$$\underline{\underline{M}} \equiv \underline{\underline{S}} + \underline{\underline{j}}^{w} \tag{331}$$

es positiva definida sobre $\tilde{D}(\Delta)$.

12.-

$$\underline{\underline{aM}} = \underline{\underline{aS}} \tag{332}$$

ya que $\underline{\underline{j}}^{w}$ es la proyección sobre $\tilde{D}_{11}^{w}(\Delta) \subset \tilde{D}_{11}(\Delta)$. Aquí $\underline{\underline{aj}}^{w} = 0$.

13.- La ecuación

$$\left(\underline{\underline{aM}} - \underline{\underline{jM}}\right)\underline{u} = \underline{f}_{\Delta_2} \tag{333}$$

es equivalente a

$$\underline{\underline{aMu}} = \underline{f}_{\Delta_2} \qquad \text{y} \qquad \underline{\underline{j}}\underline{u} = 0 \tag{334}$$

la cual, cuando la Ec.(332) es aplicada, se reduce a

$$\underline{\underline{aSu}} = \underline{f}_{\Delta_2} \qquad \text{y} \qquad \underline{\underline{j}}\underline{u} = 0. \tag{335}$$

Nomenclatura:

a).- $\underline{\underline{j}}^{w}\underline{u}$ es el *salto débil* de la función $\underline{u} \in \tilde{D}(\Delta)$; y

b).- $\tilde{D}_{12}^{w}(\Delta)$ es el espacio de *funciones continuas débiles.*

2.6.1. Cálculo de la inversa de $\underline{\underline{M}}$

Para algunos de los algoritmos tratados en las pasadas secciones es necesario el cálculo de la inversa de la matriz $\underline{\underline{M}}$, en esta subsección derivaremos su construcción.

Teorema 48 *La matriz* $\underline{\underline{I}}_{S}^{C} : \tilde{D}_{11}^{w}(\Delta) \to N_S$ *es positiva definida*

Demostración. Claramente

$$\underline{u}\underline{\underline{I}}_{S}^{C}\underline{u} \geq 0 \tag{336}$$

ya que $\underline{\underline{I}}_{S}^{C}$ es una proyección. Así, sólo es necesario probar que cuando $\underline{u} \in \tilde{D}_{11}^{w}(\Delta)$

$$\underline{\underline{I}}_{S}^{C}\underline{u} = 0 \Rightarrow \underline{u} = 0 \tag{337}$$

ahora, cuando $\underline{u} \in \tilde{D}_{11}^{w}(\Delta)$, tenemos que

$$\underline{u} = \underline{\underline{j}}\underline{v}, \text{ para alguna } \underline{v} \in N_S \tag{338}$$

la condición $\underline{\underline{I}}_S^C \underline{u} = 0$ implica que

$$\underline{w} \cdot \underline{\underline{j}} \underline{v} = 0, \qquad \forall \underline{w} \in N_S \tag{339}$$

recordando que $\underline{\underline{j}}$ es positiva definida sobre N_S, entonces la última ecuación implica que $\underline{v} = 0$, lo cual en vista de la Ec.(338) implica que $\underline{u} = 0$. ■

El resultado anterior implica que $\underline{\underline{I}}_S^C : \tilde{D}_{11}^w(\Delta) \to N_S$ posee una inversa la cual será denotada por $\underline{\underline{l}} : N_S \to \tilde{D}_{11}^w(\Delta)$. Entonces

$$\underline{\underline{I}}_S^C \underline{\underline{l}} \underline{u} = \underline{u}, \qquad \forall \underline{u} \in N_S. \tag{340}$$

Otra matriz será usada, la matriz $\underline{\underline{k}}^w : \tilde{D}(\Delta) \to \tilde{D}_{11}^w(\Delta)$. Tal que para cada $\underline{u} \in \tilde{D}(\Delta)$, esta es definida por

$$\underline{\underline{k}}^w \underline{u} \equiv \underline{\underline{l}} \underline{\underline{I}}_S^C \underline{u} \tag{341}$$

esto define a $\underline{\underline{k}}^w \underline{u}$ de forma única, ya que $\underline{\underline{I}}_S^C \underline{u} \in N_S$. Además observemos que en vista de la Ec.(343) se siguen las siguientes propiedades: Para toda $\underline{u} \in \tilde{D}(\Delta)$ tenemos

$$\underline{\underline{I}}_S^C \underline{\underline{k}}^w \underline{u} = \underline{\underline{I}}_S^C \underline{\underline{l}} \underline{\underline{I}}_S^C \underline{u} = \underline{\underline{I}}_S^C \underline{u} \tag{342}$$

por lo tanto

$$\underline{w} \cdot \underline{\underline{k}}^w \underline{u} = \underline{w} \cdot \underline{u}, \ \forall \underline{w} \in N_S. \tag{343}$$

Para obtener $\underline{\underline{M}}^{-1}$ para cualquier $\underline{u} \in \tilde{D}(\Delta)$, asumiendo $\underline{v} = \underline{\underline{M}}^{-1} \underline{u}$, entonces

$$\left(\underline{\underline{S}} + \underline{\underline{j}}^w \right) \underline{v} = \underline{u} \tag{344}$$

aplicando $\underline{\underline{I}}_S^C$ a esta ecuación, y usando la $\underline{\underline{I}}_S^C \underline{\underline{k}}^w \underline{u} = \underline{\underline{I}}_S^C \underline{u}$, obtenemos

$$\underline{\underline{I}}_S^C \underline{\underline{j}}^w \underline{v} = \underline{\underline{I}}_S^C \underline{u} = \underline{\underline{I}}_S^C \underline{\underline{k}}^w \underline{u} \tag{345}$$

observando que $\underline{\underline{j}}^w \underline{v} \in \tilde{D}_{11}^w(\Delta)$ y $\underline{\underline{k}}^w \underline{u} \in \tilde{D}_{11}^w(\Delta)$, por lo tanto de la ecuación anterior se sigue que

$$\underline{\underline{j}}^w \underline{v} = \underline{\underline{k}}^w \underline{u}. \tag{346}$$

ya que $\underline{\underline{I}}_S^C : \tilde{D}_{11}^w(\Delta) \to N_S$ es biyectiva. Por sustitución de este resultado en la Ec.(344) obtenemos

$$\underline{\underline{S}} \underline{v} + \underline{\underline{k}}^w \underline{u} = \underline{u} \qquad \text{o} \qquad \underline{\underline{S}} \underline{v} = \underline{u} - \underline{\underline{k}}^w \underline{u}. \tag{347}$$

Definición 49 *Definimos a las siguientes matrices, espacios y vectores*

$$\begin{aligned} \underline{\underline{I}}_S^R &\equiv \underline{\underline{I}} - \underline{\underline{I}}_S^C \\ \tilde{D}^R(\Delta) &\equiv \underline{\underline{I}}_S^R \tilde{D}(\Delta) \\ \underline{v}^R &\equiv \underline{\underline{I}}_S^R \underline{v} \\ \underline{v}^C &\equiv \underline{\underline{I}}_S^C \underline{v} \end{aligned}$$

Entonces, el vector $\underline{v}$ es escrito como $\underline{v} = \underline{v}^R + \underline{v}^C$, con $\underline{v}^R \in \tilde{D}^R(\Delta)$ y $\underline{v}^C \in N_S$. Aplicando $\underline{\underline{I}}_S^R$ a la Ec.(347) tenemos

$$\underline{\underline{S}}\underline{v}^R = \underline{\underline{I}}_S^R \left(\underline{u} - \underline{\underline{k}}^w u\right) \tag{348}$$

Además, $\underline{\underline{S}} : \tilde{D}^R(\Delta) \to \tilde{D}^R(\Delta)$ es biyectiva. Efectivamente, $\underline{v}^R \in \tilde{D}^R(\Delta)$ es la única solución de

$$\left(\underline{\underline{S}} + \underline{\underline{I}}_S^C\right) \underline{v}^R = \underline{\underline{I}}_S^R \left(\underline{u} - \underline{\underline{k}}^w \underline{u}\right) \tag{349}$$

ya que la matriz $\underline{\underline{S}} + \underline{\underline{I}}_S^C$ es no singular. En resumen

$$\underline{v}^R = \left(\underline{\underline{S}} + \underline{\underline{I}}_S^C\right)^{-1} \underline{\underline{I}}_S^R \left(\underline{u} - \underline{\underline{k}}^w \underline{u}\right). \tag{350}$$

Una vez que $\underline{v}^R \in \tilde{D}^R(\Delta)$ es obtenida la Ec.(346) puede ser usada para obtener

$$\underline{v}^C = \left(\underline{\underline{j}}^w\right)^{-1} \left(\underline{\underline{k}}^w \underline{u} - \underline{\underline{j}}^w \underline{u}\right) \tag{351}$$

aquí, $\left(\underline{\underline{j}}^w\right)^{-1} : \tilde{D}_{11}^w(\Delta) \to N_S$ es la inversa de $\underline{\underline{j}}^w : N_S \to \tilde{D}_{11}^w(\Delta)$.

La última observación es que el Complemento de Schur de la matriz

$$\begin{pmatrix} \underline{\underline{A}}_{\Pi\Pi} & \underline{\underline{A}}_{\Pi\Delta} \\ \underline{\underline{A}}_{\Delta\Pi} & \underline{\underline{A}}_{\Delta\Delta} + \underline{\underline{I}}_S^C \end{pmatrix} \tag{352}$$

es

$$\underline{\underline{S}} + \underline{\underline{I}}_S^C = \underline{\underline{I}}_S^C + \underline{\underline{A}}_{\Delta\Delta} - \underline{\underline{A}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \tag{353}$$

usando esta expresión cuando se aplica $\left(\underline{\underline{S}} + \underline{\underline{I}}_S^C\right)^{-1}$.

2.6.2. Cálculo de $\underline{\underline{I}}_S^C$, $\left(\underline{\underline{j}}^w\right)^{-1}$ y $\underline{\underline{k}}^w$

Para la construcción de la matriz $\underline{\underline{I}}_S^C : \tilde{D}(\Delta) \to N_S$, es conveniente la construcción de una base ortonormal $\left\{\underline{e}^1, ..., \underline{e}^{d_N}\right\} \subset N_S$ de N_S. Aquí, d_N es la dimensión del espacio nulo de N_S. Entonces, para cada $\underline{u} \in \tilde{D}(\Delta)$ tenemos que

$$\underline{\underline{I}}_S^C \underline{u} = \sum_{\alpha=1}^{d_N} C_\alpha \underline{e}^\alpha \tag{354}$$

donde

$$C_\alpha = \underline{u} \cdot \underline{e}^\alpha, \tag{355}$$

por supuesto, otra opción es la construcción de cualquier base de N_S, la cual no necesariamente satisfaga la condición de ser una base ortonormal, en cuyo caso la determinación de los coeficientes C_α requieren resolver un sistema relativamente

pequeño de ecuaciones. Aquí asumiremos que la base $\left\{\underline{e}^1, ..., \underline{e}^{d_N}\right\}$ es tomada como ortonormal.

Sea $\underline{t} \in \tilde{D}_{11}^{w}(\Delta)$ y $\underline{w} \equiv \left(\underline{\underline{j}}^{w}\right)^{-1} \underline{t} \in N_S$. Entonces

$$\underline{w} = \sum_{\alpha=1}^{d_N} C_\alpha \underline{e}^\alpha \tag{356}$$

además,

$$\underline{\underline{j}}\underline{e}^\alpha \cdot \underline{w} = \underline{e}^\alpha \cdot \underline{\underline{j}}\underline{w} = \underline{e}^\alpha \cdot \underline{t}, \forall \alpha = 1, ..., d_N. \tag{357}$$

Por lo tanto

$$\sum_{\alpha=1}^{d_N} C_\alpha \underline{\underline{j}}\underline{e}^\alpha \cdot \underline{\underline{j}}\underline{e}^\beta = \underline{t} \cdot \underline{e}^\beta, \beta = 1, ..., d_N \tag{358}$$

obsérvese que la matriz cuadrada de $d_N \times d_N$:

$$\left(\underline{\underline{j}}\underline{e}^\alpha \cdot \underline{\underline{j}}\underline{e}^\beta\right) \tag{359}$$

es simétrica y positiva definida en el espacio R^{d_N}. Por lo tanto, usando está ultima ecuación se obtiene el conjunto de coeficientes $\{C_1, ..., C_{d_N}\}$ de forma única.

Por otro lado, dado cualquier $\underline{u} \in \tilde{D}(\Delta)$ tenemos que

$$\underline{\underline{k}}^{w}\underline{u} = \underline{\underline{j}}N_S \subset \tilde{D}_{11}(\Delta) \tag{360}$$

donde $\underline{\underline{k}}^{w}\underline{u} \in \tilde{D}_{11}^{w}(\Delta)$. Por lo tanto

$$\underline{\underline{k}}^{w}\underline{u} = \sum_{\alpha=1}^{d_N} d_\alpha \underline{\underline{j}}\underline{e}^\beta \tag{361}$$

usando las Ec.(343) y (360), obtenemos

$$\underline{\underline{j}}\underline{e}^\beta \cdot \underline{\underline{k}}^{w}\underline{u} = \underline{e}^\beta \cdot \underline{\underline{k}}^{w}\underline{u} = \underline{e}^\beta \cdot \underline{u},\ \beta = 1, ..., d_N \tag{362}$$

por lo tanto

$$\sum_{\alpha=1}^{d_N} d_\alpha \underline{\underline{j}}\underline{e}^\beta \cdot \underline{\underline{j}}\underline{e}^\beta = \underline{u} \cdot \underline{e}^\beta = C_\beta,\ \beta = 1, ..., d_N. \tag{363}$$

3. Implementación Computacional de los Métodos Round-Trip

En este capítulo veremos como implementar computacionalmente los algoritmos Single-Trip y Round-Trip partiendo de las formulaciones matriciales de cada uno de ellos, para ello primeramente es necesario hacer la discretización del espacio, para después calcular la solución particular que es necesaria para implementar los algoritmos.

Sea Ω un dominio y sean $\left\{\overline{\Omega}_1, ..., \overline{\Omega}_E\right\}$ la partición del dominio, i.e.

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.- $\overline{\Omega} = \bigcup_{\alpha=1}^{E} \overline{\Omega}_\alpha$.

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_i respectivamente, claramente

$$\partial\Omega \subset \bigcup_{\alpha=1}^{E} \partial\Omega_\alpha. \tag{364}$$

Adicionalmente definimos

$$\Gamma_\alpha = \Delta \cap \overline{\Omega}_\alpha \qquad \text{y} \qquad I_\alpha = \Pi \cap \overline{\Omega}_\alpha \tag{365}$$

junto con

$$\Gamma = \bigcup_{\alpha=1}^{E} \Gamma_\alpha \qquad \text{y} \qquad I = \bigcup_{\alpha=1}^{E} I_\alpha \tag{366}$$

entonces

$$\overline{\Omega} = \Gamma \cup I \text{ mientras } \Delta = \Gamma \text{ y } \Pi = I \tag{367}$$

3.1. Discretización del Espacio

Según el desarrollo teórico, existen dos formas de resolver el problema, la primera es partiendo de la formulación original mediante el problema original y la otra usando la discretización algebraica generada partiendo de la descomposición local de los subdominios $\overline{\Omega}_\alpha$ e integrando estos hasta recuperar la formulación original, ambas son equivalentes y las describiremos a continuación.

Discretización partiendo del problema original Dado el problema original

$$\underline{\underline{\hat{A}}}\underline{\hat{u}} = \underline{\hat{f}} \tag{368}$$

defínase la matriz virtual

$$\underline{\underline{A}} = \tau\left(\underline{\underline{\hat{A}}}\right) \tag{369}$$

y el vector virtual

$$\underline{f} = \tau\left(\underline{\hat{f}}\right) \tag{370}$$

los cuales están formados con la siguiente estructura

$$\underline{\underline{A}} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha} \tag{371}$$

y

$$\underline{f} = \sum_{\alpha=1}^{E} \underline{f}^{\alpha} \tag{372}$$

donde

$$\underline{\underline{A}}^{\alpha} = \begin{pmatrix} \underline{\underline{A}}^{\alpha}_{\Pi\Pi} & \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \\ \underline{\underline{A}}^{\alpha}_{\Delta\Pi} & \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \end{pmatrix} \tag{373}$$

$$\underline{f}^{\alpha} = \begin{pmatrix} \underline{f}^{\alpha}{}_{\Pi} \\ \underline{f}^{\alpha}{}_{\Delta} \end{pmatrix}. \tag{374}$$

donde para cada $\alpha = 1, ..., E$, son las matrices locales y están definidas de $\underline{\underline{A}}^{\alpha} : D\left(\overline{\Omega}_{\alpha}\right) \to D\left(\overline{\Omega}_{\alpha}\right)$ y $\underline{f}^{\alpha} : D\left(\overline{\Omega}_{\alpha}\right) \to D\left(\overline{\Omega}_{\alpha}\right)$ respectivamente.

Discretización partiendo de la formulación local Para cada Ω_{α} como $\alpha = 1, ..., E$, definimos las matrices locales $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}, \underline{\underline{A}}^{\alpha}_{\Pi\Delta}, \underline{\underline{A}}^{\alpha}_{\Delta\Pi}, \underline{\underline{A}}^{\alpha}_{\Delta\Delta} : D\left(\overline{\Omega}_{\alpha}\right) \to D\left(\overline{\Omega}_{\alpha}\right)$ y $\underline{f}^{\alpha}{}_{\Pi}, \underline{f}^{\alpha}{}_{\Delta} : D\left(\overline{\Omega}_{\alpha}\right) \to D\left(\overline{\Omega}_{\alpha}\right)$ tales que

$$\underline{\underline{A}}^{\alpha} = \begin{pmatrix} \underline{\underline{A}}^{\alpha}_{\Pi\Pi} & \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \\ \underline{\underline{A}}^{\alpha}_{\Delta\Pi} & \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \end{pmatrix}$$

y

$$\underline{f}^{\alpha} = \begin{pmatrix} \underline{f}^{\alpha}{}_{\Pi} \\ \underline{f}^{\alpha}{}_{\Delta} \end{pmatrix}.$$

De tal forma que podemos definir la matriz virtual

$$\underline{\underline{A}} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha} \tag{375}$$

y el vector virtual

$$\underline{f} = \sum_{\alpha=1}^{E} \underline{f}^{\alpha} \tag{376}$$

tales que

$$\underline{\underline{\hat{A}}} = \tau^{-1}\left(\underline{\underline{A}}\right) \tag{377}$$

y el vector virtual

$$\underline{\hat{f}} = \tau^{-1}\left(\underline{f}\right) \tag{378}$$

satisfacen el problema original

$$\underline{\underline{\hat{A}}}\,\underline{\hat{u}} = \underline{\hat{f}}. \tag{379}$$

En ambos casos las matrices y vectores

$$\underline{\underline{A}}^{\alpha} = \begin{pmatrix} \underline{\underline{A}}^{\alpha}_{\Pi\Pi} & \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \\ \underline{\underline{A}}^{\alpha}_{\Delta\Pi} & \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \end{pmatrix} \text{ y } \underline{f}^{\alpha} = \begin{pmatrix} \underline{f}^{\alpha}{}_{\Pi} \\ \underline{f}^{\alpha}{}_{\Delta} \end{pmatrix}$$

no coinciden, pero la matriz $\underline{\underline{A}}$ y el vector $\underline{f}$ resultantes si coinciden.

En ambos caso, notemos que

$$\begin{cases} \underline{\underline{A}}_{\Pi\Pi} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha}_{\Pi\Pi}, & \underline{\underline{A}}_{\Pi\Delta} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \\ \underline{\underline{A}}_{\Delta\Pi} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha}_{\Delta\Pi}, & \underline{\underline{A}}_{\Delta\Delta} = \sum_{\alpha=1}^{E} \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \end{cases} \tag{380}$$

$$\begin{cases} \underline{f}_{\Pi} = \sum_{\alpha=1}^{E} \underline{f}^{\alpha}{}_{\Pi} \\ \underline{f}_{\Delta} = \sum_{\alpha=1}^{E} \underline{f}^{\alpha}{}_{\Delta} \end{cases}. \tag{381}$$

así también, las siguientes propiedades importantes

$$\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1} \qquad \text{y} \qquad \left(\underline{\underline{A}}\right)^{-1} = \sum_{\alpha=1}^{E} \left(\underline{\underline{A}}^{\alpha}\right)^{-1} \tag{382}$$

esto implica que el cálculo de las inversas de la matriz $\underline{\underline{A}}_{\Pi\Pi}$ requiere calcular exclusivamente las inversas locales.

En el desarrollo de la siguiente sección usaremos al complemento de Schur el cual es definido por

$$\underline{\underline{S}} = \underline{\underline{A}}_{\Delta\Delta} - \underline{\underline{A}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}. \tag{383}$$

3.2. Construcción de la Solución Particular

En la formulación desarrollada es necesario encontrar $\underline{u} \in \tilde{D}(\Delta)$ que satisfaga

$$\left(\underline{\underline{L}} + \underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{\tilde{u}} = \underline{f} \tag{384}$$

donde el vector $\underline{u}$ y $\underline{f}$ esta formado por

$$\begin{pmatrix} \underline{\tilde{u}}_\Pi \\ \underline{\tilde{u}}_\Delta \end{pmatrix} = \underline{\tilde{u}} \text{ y } \begin{pmatrix} \underline{f}_\Pi \\ \underline{f}_\Delta \end{pmatrix} = \underline{f}. \tag{385}$$

Como es conveniente transformar el problema Ec.(384) en uno que

$$\underline{f}_\Pi = 0 \tag{386}$$

entonces introducimos un vector auxiliar

$$\underline{u}_p = \underline{\underline{A}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{f}_\Pi \tag{387}$$

en el cual $\left(\underline{u}_p\right)_\Delta = 0$, por lo tanto la Ec. (384) toma la forma

$$\left(\underline{\underline{aR}} - \underline{\underline{R}}^T \underline{\underline{j}}\right) \underline{u} = \underline{f}_\Delta - \underline{u}_p \tag{388}$$

así, la solución $\underline{u}$ al problema será $\underline{u} = \underline{\tilde{u}} - \underline{u}_p$.

Dado que necesitamos expresar el problema en términos del espacio Δ_2, entonces

$$\underline{f}_{\Delta_2} = \underline{\underline{a}} \underline{f}_\Delta, \text{ y } \underline{f}_{\Delta_1} = \underline{\underline{j}} \underline{f}_\Delta = 0 \tag{389}$$

$$\left(\underline{u}_p\right)_{\Delta 2} = \underline{\underline{aA}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{f}_\Pi \tag{390}$$

así, la expresión de la Ec.(388) se puede reescribir como

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u} = \underline{\underline{a}} \underline{f}_\Delta - \underline{\underline{aA}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1} \underline{f}_\Pi \tag{391}$$

o más compactamente como

$$\left(\underline{\underline{aS}} - \underline{\underline{Sj}}\right) \underline{u} = \underline{f}_{\Delta 2} - \left(\underline{u}_p\right)_{\Delta 2}. \tag{392}$$

3.3. Métodos Single-Trip

En los métodos SingleTrip se desarrollan dos enfoques, el Dirichlet y el Neumann, los cuales se describen a continuación.

3.3.1. El Enfoque Dirichlet

En este enfoque hay que buscar una función $\underline{u}_{\Delta} \in \tilde{D}_{12}(\Delta)$ tal que satisfaga

$$\underline{\underline{aS}}\underline{u}_{\Delta} = \underline{f}_{\Delta 2} - \left(\underline{u}_p\right)_{\Delta 2} \tag{393}$$

i.e.

$$\underline{\underline{aS}}\underline{u}_{\Delta} = \underline{\underline{af}}_{\Delta} - \underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_{\Pi}. \tag{394}$$

La implementación computacional queda como a continuación se muestra:

$\underline{r} = \left[\underline{\underline{af}}_{\Delta} - \underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_{\Pi}\right] - \underline{\underline{aS}}\underline{u}$
$\underline{w} = \underline{r}$
$\underline{v} = \underline{w}$
$\alpha = \sum_{j=1}^{n} w_j^2$
$k = 1$
Mientras que $k \leq N$

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

La salida del método será la solución aproximada $\underline{u} = (u_1, ..., u_n) = \underline{u}_{\Delta}$ y el residual $\underline{r} = (r_1, ..., r_n)$, para la solución en los nodos $\underline{u}_{\Pi}$ en cualquiera de los métodos, sólo es necesario calcular

$$\underline{u}_{\Pi} = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_{\Delta}. \tag{395}$$

3.3.2. El Enfoque Neumann

En este enfoque hay que buscar una función $\underline{u}_\Delta \in \tilde{D}_{22}(\Delta)$ tal que satisfaga

$$\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{u}_\Delta = -\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2} - \underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}^{-1}\left(\underline{u}_p\right)_{\Delta 2} \tag{396}$$

i.e.

$$\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{u}_\Delta = -\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}_\Delta^{-1}\underline{f}_\Delta - \underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi. \tag{397}$$

La implementación computacional queda como a continuación se muestra:

$\underline{r} = -\underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{\underline{S}}_\Delta^{-1}\left[\underline{f}_\Delta - \underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi\right] - \underline{\underline{S}}^{-1}\underline{\underline{j}}\underline{u}$
$\underline{w} = \underline{r}$
$\underline{v} = \underline{w}$
$\alpha = \sum_{j=1}^{n} w_j^2$
$k = 1$
Mientras que $k \leq N$

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

La salida del método será la solución aproximada $\underline{u} = (u_1, ..., u_n) = \underline{u}_\Delta$ y el residual $\underline{r} = (r_1, ..., r_n)$, para la solución en los nodos $\underline{u}_\Pi$ en cualquiera de los métodos, sólo es necesario calcular

$$\underline{u}_\Pi = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_\Delta. \tag{398}$$

3.4. Métodos Round-Trip

En los métodos Round-Trip se desarrollan dos enfoques, el Dirichlet-Dirichlet y el Neumann-Neumann, los cuales se describen a continuación.

3.4.1. El enfoque Dirichlet-Dirichlet

En este enfoque hay que buscar una función $\underline{u}_\Delta \in \tilde{D}_{22}(\Delta)$ tal que satisfaga

$$\underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{u}_\Delta = \underline{\underline{aS}}^{-1}\underline{f}_{\Delta 2} - \underline{\underline{aS}}^{-1}\left(\underline{u}_p\right)_{\Delta 2} \tag{399}$$

i.e.

$$\underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{u}_\Delta = \underline{\underline{aS}}^{-1}\underline{af}_\Delta - \underline{\underline{aS}}^{-1}\underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi \tag{400}$$

aquí, $\underline{\underline{aS}}^{-1}$ es el precondicionador.

La implementación computacional queda como a continuación se muestra:

$\underline{r} = \underline{\underline{aS}}^{-1}\underline{a}\left[\underline{f}_\Delta - \underline{\underline{A}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi\right] - \underline{\underline{aS}}^{-1}\underline{\underline{aS}}\underline{u}$
$\underline{w} = \underline{r}$
$\underline{v} = \underline{w}$
$\alpha = \sum_{j=1}^{n} w_j^2$
$k = 1$
Mientras que $k \leq N$

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

La salida del método será la solución aproximada $\underline{u} = (u_1, ..., u_n) = \underline{u}_\Delta$ y el residual $\underline{r} = (r_1, ..., r_n)$, para la solución en los nodos $\underline{u}_\Pi$ en cualquiera de los métodos, sólo es necesario calcular

$$\underline{u}_\Pi = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_\Delta. \tag{401}$$

3.4.2. El enfoque Neumann-Neumann

En este enfoque hay que buscar una función $\underline{u}_\Delta \in \tilde{D}_{22}(\Delta)$ tal que satisfaga

$$\underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{u}_\Delta = -\underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{f}_{\Delta 2} - \underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{\underline{S}}^{-1}\left(\underline{u}_p\right)_{\Delta 2} \tag{402}$$

i.e.

$$\underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{u}_\Delta = -\underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{jS}}^{-1}\underline{af}_\Delta - \underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{\underline{aA}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi \tag{403}$$

aquí, $\underline{\underline{S}}^{-1}\underline{\underline{jS}}$ es el precondicionador.

La implementación computacional queda como a continuación se muestra:

$\underline{r} = -\underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{\underline{S}}^{-1}\underline{a}\left[\underline{f}_\Delta - \underline{\underline{A}}_{\Delta\Pi}\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{f}_\Pi\right] - \underline{\underline{S}}^{-1}\underline{\underline{jS}}\underline{\underline{j}}\underline{u}$
$\underline{w} = \underline{r}$
$\underline{v} = \underline{w}$
$\alpha = \sum_{j=1}^{n} w_j^2$
$k = 1$
Mientras que $k \leq N$

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

La salida del método será la solución aproximada $\underline{u} = (u_1, ..., u_n) = \underline{u}_\Delta$ y el residual $\underline{r} = (r_1, ..., r_n)$, para la solución en los nodos $\underline{u}_\Pi$ en cualquiera de los métodos, sólo es necesario calcular

$$\underline{u}_\Pi = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_\Delta. \tag{404}$$

3.5. Consideraciones Computacionales

En esta sección describiremos como calcular la matriz $\underline{\underline{S}}, \underline{\underline{S}}^{-1}$ y los nodos interiores.

3.5.1. Cálculo de la Matriz $\underline{\underline{S}}$

La matriz $\underline{\underline{S}}$ definida por

$$\underline{\underline{S}} = \underline{\underline{A}}_{\Delta\Delta} - \underline{\underline{A}}_{\Delta\Pi} \left(\underline{\underline{A}}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}_{\Pi\Delta} \tag{405}$$

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

$$\underline{\underline{S}}^{\alpha} = \underline{\underline{A}}^{\alpha}_{\Delta\Delta} - \underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta}. \tag{406}$$

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

$$\underline{\tilde{u}_\Gamma}^{\alpha} = \left(\underline{\underline{A}}^{\alpha}_{\Delta\Delta} - \underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \right) \underline{y_\Gamma}^{\alpha}. \tag{407}$$

Para evaluar eficientemente esta expresión, realizamos las siguientes operaciones equivalentes

$$\begin{aligned} \underline{x1} &= \underline{\underline{A}}^{\alpha}_{\Delta\Delta} \underline{y_\Gamma}^{\alpha} \\ \underline{x2} &= \left(\underline{\underline{A}}^{\alpha}_{\Delta\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \right) \underline{y_\Gamma}^{\alpha} \\ \underline{\tilde{u}_\Gamma}^{\alpha} &= \underline{x1} - \underline{x2} \end{aligned} \tag{408}$$

la primera y tercera expresión no tienen ningún problema en su evaluación, para la segunda expresión tendremos que hacer

$$\underline{x3} = \underline{\underline{A}}^{\alpha}_{\Pi\Delta} \underline{y_\Gamma}^{\alpha} \tag{409}$$

con este resultado intermedio deberíamos calcular

$$\underline{x4} = \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{x3} \tag{410}$$

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

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \underline{x4} = \underline{\underline{A}}^{\alpha}_{\Pi\Pi} \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \right)^{-1} \underline{x3} \tag{411}$$

al simplificar, tenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi} \underline{x4} = \underline{x3}. \tag{412}$$

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

$$\underline{x2} = \underline{\underline{A}}^{\alpha}_{\Delta\Pi}\underline{x4} \tag{413}$$

así

$$\underline{\tilde{u}_{\Gamma}}^{\alpha} = \underline{x1} - \underline{x2} \tag{414}$$

completando la secuencia de operaciones necesaria para obtener $\underline{\underline{S}}_{\alpha}\underline{y_{\Gamma}}^{\alpha}$.

3.5.2. Cálculo de los Nodos Interiores

La evaluación de

$$\underline{u}_{\Pi} = -\left(\underline{\underline{A}}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}_{\Pi\Delta}\underline{u}_{\Delta} \tag{415}$$

involucra nuevamente cálculos locales de la expresión

$$\underline{u_I}^{\alpha} = -\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{\underline{A}}^{\alpha}_{\Pi\Delta}\underline{u_{\Gamma}}^{\alpha} \tag{416}$$

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

$$\begin{aligned} \underline{x4} &= \underline{\underline{A}}^{\alpha}_{\Pi\Delta}\underline{u_{\Gamma}}^{\alpha} \\ \underline{u_I}^{\alpha} &= \left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{x4} \end{aligned} \tag{417}$$

multiplicando por $\underline{\underline{A}}^{\alpha}_{\Pi\Pi}$ a la última expresión, obtenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{u_I}^{\alpha} = \underline{\underline{A}}^{\alpha}_{\Pi\Pi}\left(\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\right)^{-1}\underline{x4} \tag{418}$$

simplificando, tenemos

$$\underline{\underline{A}}^{\alpha}_{\Pi\Pi}\underline{u_I}^{\alpha} = \underline{x4} \tag{419}$$

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

$$\begin{aligned} \underline{\underline{L}}\underline{y} &= \underline{b} \\ \underline{\underline{U}}\underline{x} &= \underline{y} \end{aligned} \tag{420}$$

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.

3.5.3. Cálculo de la Matriz $\underline{\underline{S}}^{-1}$

En los algoritmos descritos anteriormente, interviene la evaluación 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 evaluar

$$\underline{u} = \underline{\underline{S}}^{-1}\underline{v} \tag{421}$$

se procede a multiplicar por $\underline{\underline{S}}$ a la ecuación anterior, obteniendo

$$\underline{\underline{S}}\underline{u} = \underline{\underline{S}}\underline{\underline{S}}^{-1}\underline{v} \tag{422}$$

simplificando, tenemos que

$$\underline{\underline{S}}\underline{u} = \underline{v}, \tag{423}$$

así, mediante algún procedimiento directo u iterativo (usando factorización LU o CGM) resolvemos el sistema anterior.

4. Apéndice A

4.1. 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.

4.1.1. 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).

4.1.2. 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.

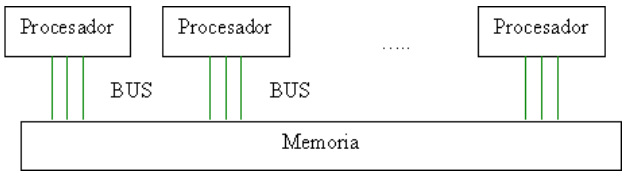

Figura 2: Arquitectura de una computadora paralela con memoria compartida

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.

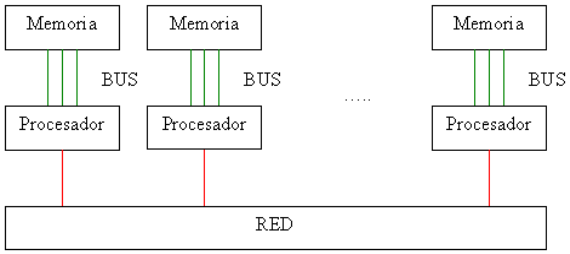

Figura 3: Arquitectura de una computadora paralela con memoria distribuida

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:

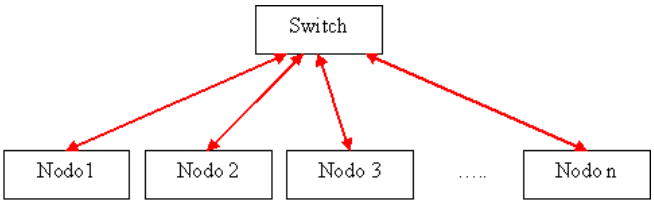

Figura 4: Arquitectura de un cluster

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.

4.2. 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

$$s = \frac{T(1)}{T(p)} \tag{424}$$

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

$$e = \frac{T(1)}{pT(p)} = \frac{s}{p}. \tag{425}$$

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

$$f = \frac{T_s}{T(1)} \tag{426}$$

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

$$f = \frac{\frac{1}{s} - \frac{1}{p}}{1 - \frac{1}{p}}. \tag{427}$$

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.

4.3. 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.

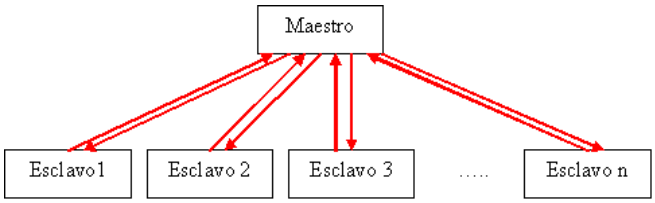

Figura 5: Esquema del maestro-esclavo

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 [33] y [12].

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.

5. Apéndice B

5.1. 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

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{428}$$

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.

5.1.1. 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

$$\sum_{j=1}^{N} a_{ij}^{(0)} x_j = a_{i,n+1}^{(0)}, \quad i = 1, 2, ..., N \tag{429}$$

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)

$$x_i + \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j = a_{i,n+1}^{(i)}, \quad i = 1, 2, ..., N \tag{430}$$

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

$$\begin{aligned} x_N &= a_{N,N+1}^{(N)}; \\ i &= N-1, N-2, ..., 1 \\ x_i &= a_{i,N+1}^{(i)} - \sum_{j=i+1}^{N} a_{ij}^{(i)} x_j. \end{aligned} \tag{431}$$

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

$$\underline{\underline{L}}\underline{y} = \underline{b} \quad \text{y} \quad \underline{\underline{U}}\underline{u} = \underline{y}. \tag{432}$$

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.

5.1.2. 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 [26] y [36].

Un método iterativo en el cual se resuelve el sistema lineal

$$\underline{\underline{A}}\underline{u} = \underline{b} \tag{433}$$

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

$$\underline{u}^k = \underline{\underline{T}}\underline{u}^{k-1} + \underline{c} \quad \forall k = 1, 2, 3, ... \tag{434}$$

La convergencia a la solución la garantiza el siguiente teorema cuya solución puede verse en [37].

Teorema 50 *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

$$cond(\underline{\underline{A}}) = \frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}} \geq 1 \tag{435}$$

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 [4].

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 (433) 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

$$\underline{u} = \underline{\underline{B}}\underline{u} + \underline{d} \tag{436}$$

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

$$x_i = \sum_{j=1}^{n} b_{ij} x_j^{(k-1)} + d_i \tag{437}$$

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. (433) se puede reescribir como

$$\left(\underline{\underline{I}} - \underline{\underline{B}}\right) \underline{u} = \underline{b}. \tag{438}$$

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

$$\underline{u}^{m+1} = \underline{\underline{B}}\underline{u}^m + \underline{k}. \tag{439}$$

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. (433) por $D^{-1} = \left[diag(\underline{\underline{A}})\right]^{-1}$ obteniendo

$$\underline{\underline{B}} = \underline{\underline{I}} - \underline{\underline{D}}^{-1}\underline{\underline{A}}, \quad \underline{k} = \underline{\underline{D}}^{-1}\underline{b}. \tag{440}$$

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. (439) puede ser escrita como

$$u_i^{m+1} = \sum_{j<i} b_{ij} u_j^{m+1} + \sum_{j>i} b_{ij} u_j^m + k_i. \tag{441}$$

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

$$\underline{u}^{m+1} = \underline{\underline{E}}\underline{u}^{m+1} + \underline{\underline{F}}\underline{u}^m + \underline{k} \tag{442}$$

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

$$\underline{u}^{m+1} - \underline{u}^{m} = \underline{b} - \underline{\underline{A}}\underline{u}^{m} \tag{443}$$

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

$$\underline{u}^{m+1} = \underline{u}^{m} + \omega\left(\underline{b} - \underline{\underline{A}}\underline{u}^{m}\right). \tag{444}$$

El método de Richardson se define como

$$\underline{u}^{m+1} = \left(\underline{\underline{I}} - \omega\underline{\underline{A}}\right)\underline{u}^{m} + \omega\underline{b} \tag{445}$$

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

$$u_i^{m+1} = (1-\omega)\,u_i^{m} + \omega\left[\sum_{j=1}^{i-1} b_{ij}u_j^{m+1} + \sum_{j=i+1}^{N} b_{ij}u_j^{m} + k_i\right] \tag{446}$$

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 [15].

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 51 *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

$$\underline{\underline{A}} \underline{u} = \underline{b} \tag{447}$$

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:

$$\begin{aligned} \beta^{k+1} &= \frac{\underline{\underline{A}} \underline{p}^k \cdot \underline{r}^k}{\underline{\underline{A}} \underline{p}^k \cdot \underline{p}^k} \\ \underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1} \underline{p}^k \\ \alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{\underline{A}} \underline{p}^{k+1} \cdot \underline{p}^{k+1}} \end{aligned} \tag{448}$$

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

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

$$\left\| \underline{u} - \underline{u}^k \right\|_{\underline{\underline{A}}} \leq \left\| \underline{u} - \underline{u}^0 \right\|_{\underline{\underline{A}}} \left[\frac{1 - \sqrt{Cond(\underline{\underline{A}})}}{1 + \sqrt{Cond(\underline{\underline{A}})}} \right]^{2k}. \tag{449}$$

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 52 *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*

$$\frac{\|e^n\|}{\|e^0\|} \leq \frac{2}{\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \right)^n + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \right)^{-n} \right]} \leq 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^n \tag{450}$$

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

$$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \simeq 1 - \frac{2}{\sqrt{\kappa}} \tag{451}$$

tal que

$$\|\underline{e}^n\|_{\underline{\underline{A}}} \simeq \left\|\underline{e}^0\right\|_{\underline{\underline{A}}} \exp\left(-2\frac{n}{\sqrt{\kappa}}\right) \tag{452}$$

de lo anterior podemos esperar un espectro de convergencia del orden de $O(\sqrt{\kappa})$ iteraciones, para mayor referencia ver [37].

Definición 53 *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.*

5.2. 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 [4] y [28].

¿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

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1}\underline{b} \tag{453}$$

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

$$Cond(\underline{\underline{C}}^{-1}\underline{\underline{A}}) < Cond(\underline{\underline{A}}), \tag{454}$$

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

$$\underline{\underline{C}}^{-1} \simeq \underline{\underline{A}}^{-1} \tag{455}$$

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.

5.2.1. Gradiente Conjugado Precondicionado

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

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{456}$$

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

$$\lambda_1 \leq \frac{\underline{u}\underline{\underline{C}}^{-1}\underline{\underline{A}} \cdot \underline{u}}{\underline{u} \cdot \underline{u}} \leq \lambda_n \tag{457}$$

entonces la Ec. (447) se pude escribir como

$$\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{u} = \underline{\underline{C}}^{-1} b \tag{458}$$

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

$$\begin{aligned} \left\langle \underline{u}, \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right\rangle &= \underline{u} \cdot \underline{\underline{C}} \left(\underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \right) \\ &= \underline{u} \cdot \underline{\underline{A}}\underline{v} \end{aligned} \tag{459}$$

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

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

$$\langle \underline{u}, \underline{v} \rangle = \underline{u} \cdot \underline{\underline{C}}^{-1}\underline{\underline{A}}\underline{v} \tag{460}$$

por ello las Ecs. (448[1]) y (448[3]), se convierten en

$$\alpha^{k+1} = \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \tag{461}$$

y

$$\beta^{k+1} = \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \tag{462}$$

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:

$$\begin{aligned} \beta^{k+1} &= \frac{\underline{p}^k \cdot \underline{\underline{C}}^{-1}\underline{r}^k}{\underline{p}^k \cdot \underline{\underline{A}}\underline{p}^k} \\ \underline{p}^{k+1} &= \underline{r}^k - \beta^{k+1}\underline{p}^k \\ \alpha^{k+1} &= \frac{\underline{r}^k \cdot \underline{r}^k}{\underline{p}^{k+1} \cdot \underline{\underline{C}}^{-1}\underline{p}^{k+1}} \\ \underline{u}^{k+1} &= \underline{u}^k + \alpha^{k+1}\underline{p}^{k+1} \\ \underline{r}^{k+1} &= \underline{\underline{C}}^{-1}\underline{r}^k - \alpha^{k+1}\underline{\underline{A}}\underline{p}^{k+1}. \end{aligned} \tag{463}$$

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 [15] y [13].

Teorema 54 *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.

5.2.2. 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 $\underline{\underline{A}}$

$$C_{ij} = \begin{cases} A_{ij} \quad si \quad i = j \\ \\ 0 \quad si \quad i \neq j. \end{cases} \tag{464}$$

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

$$\underline{\underline{A}} = \underline{\underline{D}} + \underline{\underline{L}} + \underline{\underline{L}}^T \tag{465}$$

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

$$\underline{\underline{C}}(\omega) = \frac{1}{2-w}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)\left(\frac{1}{\omega}\underline{\underline{D}}\right)^{-1}\left(\frac{1}{\omega}\underline{\underline{D}} + \underline{\underline{L}}\right)^T \tag{466}$$

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 $\underline{\underline{C}} = \underline{\underline{L}}\underline{\underline{U}}$, 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 $i, j > k$:

$$S_{ij} = \begin{cases} A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} & \text{Si} \quad (i,j) \in S \\ \\ A_{ij} & \text{Si} \quad (i,j) \notin S. \end{cases} \tag{467}$$

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

$$A_{ij} - A_{ij}A_{ij}^{-1}A_{kj} \neq 0 \tag{468}$$

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

$$\underline{\underline{C}}_0 = \arg mín_{\underline{\underline{C}} \in \mathcal{S}} \left\| \underline{\underline{AC}} - \underline{\underline{I}} \right\|. \tag{469}$$

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

$$\left\| \underline{\underline{AC}}_0 - \underline{\underline{I}} \right\| = \varepsilon < 1. \tag{470}$$

Existen dos razones para esto, primero, la ecuación (470) 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 $\underline{\underline{Au}} = \underline{b}$.

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

$$r_k = \underline{\underline{AC}}_k - \underline{e}_k \tag{471}$$

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

$$\left\| \underline{\underline{AC}}_k - \underline{e}_k \right\|_2^2 = 1 - \sum_{l=1}^{p_k} \frac{\left[\det\left(\underline{\underline{D}}_l^k\right)\right]^2}{\det\left(\underline{\underline{G}}_{l-2}^k\right) \det\left(\underline{\underline{G}}_l^k\right)} \tag{472}$$

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 (469) 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

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

$$\underline{m}_k = \sum_{l=1}^{p_k} \frac{\det\left(\underline{\underline{D}}_l^k\right)}{\det\left(\underline{\underline{G}}_{l-2}^k\right)\det\left(\underline{\underline{G}}_l^k\right)} \underline{\tilde{m}}_l \tag{473}$$

donde $\underline{\tilde{C}}_l$ 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 $\underline{\tilde{C}}_l$ debe resolverse el sistema $\underline{\underline{G}}_l^k \underline{v}_l = \underline{e}_l$, con $\underline{\tilde{C}}_{i_1^k l} = v_{hl}, (1 \leq h \leq l)$.

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

Desventaja: El proceso construcción es algo laborioso.

5.2.3. 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

$$(\underline{\underline{A}} + \underline{\underline{A}}^T)/2 \tag{474}$$

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

$$c_1 \leq \frac{\underline{x}^T \underline{\underline{A}} \underline{x}}{\underline{x}^T \underline{\underline{B}} \underline{x}} \leq c_2 \quad \forall \underline{x} \tag{475}$$

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 [22], [9], [8] y [4].

Definición 55 *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.

6. Bibliografía

Referencias

[1] A. Quarteroni y A. Valli; *Domain Decomposition Methods for Partial Differential Equations*. Clarendon Press Oxford, 1999.

[2] A. Quarteroni y A. Valli; Numerical Approximation of Partial Diferential Equations. Springer, 1994.

[3] A. Rosas Medina, Teoría General de Elementos Finitos en Funciones Discontinuas Definidas por Tramos, Tesis de Maestría, Instituto de Geofísica, UNAM, 2008.

[4] A. Toselli, O. Widlund; *Domain Decomposition Methods - Algorithms and Theory*. Springer, 2005.

[5] B. Cockburn, G. E. Karniadakis y C. W. Shu; Discontinuous Galerkin Methods: Theory, Computation and Applications. Springer, 2000.

[6] B. Dietrich, Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics, Cambridge University, 2001.

[7] B. D. Reddy; *Introductory Functional Analysis - With Applications to Boundary Value Problems and Finite Elements*. Springer 1991.

[8] B. F. Smith, P. E. Bjørstad, W. D. Gropp; *Domain Decomposition, Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, 1996.

[9] B. I. Wohlmuth; *Discretization Methods and Iterative Solvers Based on Domain Decomposition*. Springer, 2003.

[10] C. Farhat, I Harari, L. P. Franca; The Discontinuous Enrichment Method. Computer Methods in Applied mechanics and Engineering, 190 (6455-6479), 2001.

[11] E. Rubio; Métodos de Elementos Finitos con Funciones Óptimas, Tesis Doctoral, Instituto de geofísica, UNAM, 2008.

[12] I. Foster; *Designing and Building Parallel Programs*. Addison-Wesley Inc., Argonne National Laboratory, and the NSF, 2004.

[13] I. Herrera; Análisis de Alternativas al Método de Gradiente Conjugado para Matrices no Simétricas. Tesis de Licenciatura, Facultad de Ciencias, UNAM, 1989.

[14] I. Herrera, M. Díaz; *Modelación Matemática de Sistemas Terrestres* (Notas de Curso en Preparación). Instituto de Geofísica, (UNAM).

[15] I. Herrera; *Un Análisis del Método de Gradiente Conjugado.* Comunicaciones Técnicas del Instituto de Geofísica, UNAM; Serie Investigación, No. 7, 1988.

[16] I. Herrera; *Método de Subestructuración* (Notas de Curso en Preparación). Instituto de Geofísica, (UNAM).

[17] I. Herrera, R. Yates y E. Rubio. "More Efficient Procedures for Applying Collocation". Advances in Engineering Software 38 (2007) 657-667.

[18] I. Herrera. "Theory of Differential Equations in Discontinuos Piecewise-Defined Functions". Wiley InterScience, 2006.

[19] I. Herrera. "New Formulation of Iterative Substructuring Methods Without Lagrange Multipliers Neumann-Neumann and FETI". Wiley InterScience, 2007.

[20] I. Herrera y R. Yates, Ünified Multipliers-Free Theory of Dual-Primal Domain Descomposition Methods", Numerical Methods For Partial Differenctial Equations, Wiley InterScience, 2008.

[21] F. Brezzi y M. Fortin; Mixed and Hibrid Finite Element Methods, Springer, 1991.

[22] 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.

[23] J. L. Lions & E. Magenes; *Non-Homogeneous Bonduary Value Problems and Applications Vol. I*, Springer-Verlag Berlin Heidelber New York 1972.

[24] K. Hutter & K. Jöhnk; *Continuum Methods of Physical Modeling.* Springer-Verlag Berlin Heidelber New York 2004.

[25] L. F. Pavarino, A. Toselli; *Recent Developments in Domain Decomposition Methods.* Springer, 2003.

[26] M.B. Allen III, I. Herrera & G. F. Pinder; *Numerical Modeling in Science And Engineering.* John Wiley & Sons, Inc . 1988.

[27] M. Diaz; *Desarrollo del Método de Colocación Trefftz-Herrera Aplicación a Problemas de Transporte en las Geociencias.* Tesis Doctoral, Instituto de Geofísica, UNAM, 2001.

[28] 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.

[29] P.G. Ciarlet, J. L. Lions; *Handbook of Numerical Analysis, Vol. II.* North-Holland, 1991.

[30] R. L. Burden y J. D. Faires; *Análisis Numérico*. Math Learning, 7 ed. 2004.

[31] S. Friedberg, A. Insel, and L. Spence; *Linear Algebra*, 4th Edition, Prentice Hall, Inc. 2003.

[32] T. J. R. Hughes; The Finite Element Method: Linear Static and Dynamic Finite Element Analysis. Prentice Hall, 1987.

[33] W. Gropp, E. Lusk, A. Skjellem, *Using MPI, Portable Parallel Programming Whit the Message Passing Interface*. Scientific and Engineering Computation Series, 2ed, 1999.

[34] W. Rudin; *Principles of Mathematical Analysis*. McGraw-Hill International Editions, 1976.

[35] X. O. Olivella, C. A. de Sacribar; *Mecánica de Medios Continuos para Ingenieros.* Ediciones UPC, 2000.

[36] Y. Saad; *Iterative Methods for Sparse Linear Systems*. SIAM, 2 ed. 2000.

[37] Y. Skiba; *Métodos y Esquemas Numéricos, un Análisis Computacional*. UNAM, 2005.