Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática

O MÉTODO DE NEWTON INEXATO APLICADO ÀS EQUAÇÕES DE NAVIER-STOKES

Autor: Hilbeth Parente de Deus

Orientador: Prof. Dr. Mario César Zambaldi

Florianópolis
Agosto de 2004

Universidade Federal de Santa Catarina
Curso de Pós-Graduação em Matemática e Computação Científica

O MÉTODO DE NEWTON INEXATO APLICADO ÀS EQUAÇÕES DE NAVIER-STOKES

Dissertação apresentada ao Curso de Pós-Graduação em Matemática e Computação Científica, do Centro de Ciências Físicas e Matemáticas da Universidade Federal de Santa Catarina, para a obtenção do grau de Mestre em Matemática, com Área de Concentração em Matemática Aplicada.

Hilbeth Parente de Deus

Florianópolis

Agosto de 2004

O MÉTODO DE NEWTON INEXATO APLICADO À EQUAÇÃO DE NAVIER-STOKES

por

Hilbeth Parente de Deus

Esta Dissertação foi julgada para a obtenção do Título de "Mestre", Área de Concentração em Matemática Aplicada, e aprovada em sua forma final pelo Curso de Pós-Graduação em Matemática.

Igor E. Mozolevski
(Coordenador)

Banca Examinadora:

Prof. Dr. Mario César Zambaldi (MTM-UFSC-Orientador)

Prof. Dr. Igor E. Mozolevski (MTM-UFSC)

Prof. Dr. Jauber Oliveira (MTM-UFSC)

Prof. Dr. Marcelo Krajnc Alves (DEM-UFSC)

Florianópolis, Agosto de 2004.

Agradecimentos

... à D'us por ter-me permitido nascer de duas pessoas tão maravilhosas como meus pais, à estes por "tudo ", ao meu irmão e amigo Helilton, e ao Club de Regatas Vasco da Gama pela sua existência, companhia e constante presença em minha vida ao longo de todos anos.
... ao professor, orientador e grande amigo Mário César Zambaldi por todo seu incentivo e apoio.
... aos meus amigos que fiz ao longo desta jornada, e em especial a Anderson Borba, Anderson Luiz Maciel , André Krindges, Claudia Borba (pelo carinho, incentivo, apoio e companheirismo), Cleverson da Luz , Danilo Royer, Fermin Bazán, Fernando Deschamps, Franco Yukio Kagoiki , Lindaura Steffens, Maicon Marques Alves, Melissa Weber, Ronie Dário, Vanderlei Martins.

... à minha família.

Resumo

O trabalho aqui presente destina-se a fazer uma análise comparativa, no contexo do método de Newton inexato, os desempenhos das metodologias iterativas baseadas em subespaços de Krylov: GMRES ("Generalized Minimum Residual Method") e Bi-CGStab ("Biconjugate Gradient Stabilized") e um método direto (LU esparso). As características das desempenhos (número de iterações e tempo computacional) das metodologias investigadas são acessadas com o uso de alguns testes padrão largamente utilizados como "benchmark" em mecânica dos fluidos computacional.

O método de Newton inexato baseado em GMRES e Bi-CGStab é aplicado no sistema não linear gerado pelo método de elementos finitos (MEF) sobre o problema de valor de contorno composto pelas equações de Navier-Stokes. Uma importante observação diz respeito a condição necessária e suficiente de Brezzi-Babuška (ou condições inf-sup), a qual é satisfeita com o uso de parâmetros de estabilização.

Palavras-chave: [GMRES],[Bi-CGStab],[MEF],[Navier-Stokes].

Abstract

The present work intent to make a comparative analysis, inside of Inexact Newton method, between the performances of the Krylov subspace methodologies, GMRES (Generalized Minimum Residual Method) and Bi-CGStab (Biconjugate Gradient Stabilized), and a direct method (sparse LU). The performances characteristics (number of iterations and computational time) of the investigated methodologies are assessed using results of some standard tests widely used as a benchmark in computational fluid mechanics.

The Inexact Newton method based on GMRES and Bi-CGStab is applied on nolinear system, that was generated by finite element method (FEM) over the boundary value problem compoused by Navier-Stokes equations. An important remark, that must be done, is related to the necessary suficient condition of Brezzi-Babuška (or inf-sup condition), that is satisfied by the stability parameters.

Key-words: [GMRES],[Bi-CGStab],[MEF],[Navier-Stokes].

Simbologia

∇ ..... Gradiente
$\triangle$ ..... Operador Laplaciano
$\otimes$ ..... Produto Tensorial
$\vee$ ..... Conectivo Lógico "ou"
$\wedge$ ..... Conectivo Lógico "e"
Re ..... Número de Reynolds
$|.|_{(.)}$ ..... Norma Vetorial
$\cdot$ ..... Produto Escalar
$\|.\|_{(.)}$ ..... Norma Matricial Compatível com $|.|_{(.)}$
$\langle . \rangle_{(.)}$ ..... Produto Interno

Lista de Figuras

Lista de Tabelas

Lista de Algorítmos

Conteúdo

Capítulo 1

INTRODUÇÃO

1.1 Motivação

Este trabalho foi motivado com o intuito de se comparar o desempenho do método de Newton inexato associado a métodos baseados em aproximações em subespaços de Krylov, mais especificamente GMRES ("Generalized Minimum Residual Method") e Bi-CGStab ("Biconjugate Gradient Stabilized").

Os métodos baseados em subespaços de Krylov são aplicados em larga escala frente a grandes sistemas lineares esparsos. A principal explicação deste fato, se deve ao elevado custo computacional dos métodos diretos, em termos de tempo computacional e/ou armazenamento. Tópicos estes que são de grande importância em diversos tipos de aplicações práticas, onde geralmente exige-se uma resposta precisa, rápida e que não onere maiores custos computacionais.

Os métodos GMRES e Bi-CGStab, tem sido extensivamente explorados recentemente, principalmente associados a problemas nas áreas de mecânica, elétrica e controle e automação, tornando-se um campo muito fértil em termos de estudo e produção científica. O trabalho aqui desenvolvido tem o intuito de aplicar tais teorias e conceitos a problemas de escoamento bidimensional e incompressível de fluidos Newtonianos em regime permanente.

1.2 Revisão Bibliográfica

Os métodos apresentadas nas referências podem ser divididas didaticamente em dois grupos principais: as que fazem referência à abordagem da equação de Navier-Stokes pelo Método de Elementos Finitos e as que se referem aos métodos iterativos para sistemas lineares.

A literatura especializada mostra que os resultados de simulações numéricas da equação de Navier-Stokes para escoamentos incompressíveis, com a utilização do método clássico de Galerkin, podem sofrer oscilações espúrias (instabilidades numéricas). Isto se deve a

duas fontes principais: a primeira é o caráter advectivo-difusivo das equações, que pode permitir a contaminação do campo de pressão, assim como a conseqüente contaminação do campo de velocidade (quando o escoamento é submetido a elevados números de Reynolds), a segunda fonte destas oscilações numéricas seria à formulação de caráter "misto" (envolvendo campos de pressão e velocidade) das equações, o que limita fortemente a escolha das combinações das funções de interpolação elementares usadas para aproximar os campos de velocidade e pressão.

O contorno destes problemas, i.e. a estabilização da solução, tem sido estudada de forma incessante com o passar dos anos. Como exemplo das metodologias com esta finalidade pode-se citar: o método de Galerkin descontínuo, que é conservativo e em que as formas bilineares associadas produzem matrizes positivas definidas e bem condicionadas. Tal método surgiu no início dos anos setenta (Reed e Hill (1973)) e tem se desenvolvido bastante até os dias atuais. Dentre os principais trabalhor relacionados a esta metodologia podem-se destacar os de Bassi e Rebay (1997), Cockburn e Shu (1998) e Cockburn e Dawson (2000). Uma outra metodologia importante é a formulação de Petrov-Galerkin, em que a base das funções "peso" não são simplesmente iguais a uma base de funções "forma" ("shape functions"). Neste caso esta base é enriquecida pela adição de funções "perturbação" descontínuas que possuem características numéricas desejáveis. Resultando então em um esquema que foi introduzido por Hughes e Brooks (1980), Hughes e Brooks (1982), e Kondo (1994). Esta metodologia é referenciada na literatura como "SUPG – Streamline Upwind Petrov Galerkin". O "SUPG" tem-se mostrado capaz de controlar as instabilidades numéricas e com a vantagem de não produzir excessiva difusão numérica que pudesse prejudicar a solução (o que ocorre no esquema "up wind" puro). O que se percebe, porém, é que na vizinhança de regiões com elevados gradientes a precisão dos resultados se mostra vulnerável às tão indesejadas instabilidades.

Formulações adicionais com a mesma finalidade da citada acima também podem ser referenciadas, como o "Método do Gradiente Projetado" (ver Cecchi et al (1998) e Codina e Blasco (2000)), em que o gradiente de pressão é projetado no espaço do campo de vetores contínuo do elemento finito e o divergente da diferença entre estes dois vetores (gradiente de pressão e sua projeção) é incorporado na equação da continuidade. Outros exemplos de metodologias são as que utilizam as "Funções Bolha" ("Residual Free Bubbles" – ver Franca et al (1998)), onde a idéia é enriquecer o subespaço das funções peso com funções elementares pré-definidas para que agreguem precisão e estabilidade à solução. Há ainda o "Método de Mínimos Quadrados de Galerkin" ("Galerkin Least Square") baseado nas referências Achdou et al (1999), Codina (2000) e Franca e Frey (1992), que alia um termo adicional cujo objetivo é eliminar as oscilações que podem estar presentes nas regiões de gradientes elevados. Tal metodologia tem alcançado excelentes resultados. Outros bons exemplos são os esquemas "semi-implícitos" (ver Codina et al (1998) e Kjellgren (1997)), que consistem em tratar os termos difusivos de forma implícita e os termos

advectivos de forma explícita. Pode-se citar ainda os esquemas que utilizam ordem de interpolação iguais para os campos de pressão e velocidade (ver Franca e Frey (1992) e Codina e Blasco (1997)), i.e. tais formulações satisfazem a condição de Brezzi-Babuška, ou condição de inf-sup (ver Babuška (1973) e Brezzi (1974)) por meio de termos adicionais à formulação, denominados de termos de estabilização. A condição inf-sup é uma condição necessária para performance ótima do método de elementos finitos, quando este é aplicado a um conjunto bem definido de dados de entrada Λ. Sendo satisfeita a condição de inf-sup, diz-se então que tal método é robusto com relação ao conjunto Λ. Quando tal condição não é satisfeita o método apresenta uma performance sub-ótima ou pode até não convergir. Neste caso diz-se que o método não é robusto com relação ao conjunto Λ. Todavia, é possível que haja um subconjunto Λ^* de Λ, com respeito ao qual o método seja robusto (ver Babuška e Narasimhan (1997)). Existem também esquemas com captura de descontinuidades (ver Codina (1993)), em que uma parcela de "amortecimento" é inserida na formulação, de tal forma que esta só atue em regiões de elevados gradientes com o objetivo de combater as possíveis oscilações numéricas da solução. Há ainda metodologias que fazem uso de malhas adaptativas a fim de capturar melhor as regiões de elevados gradientes, e deste modo contribuir para uma melhor precisão dos resultados (ver Bugeda e Oñate (1995)).

A literatura na área de métodos iterativos fornece uma gama abrangente de possibilidades para abordagem de sistemas lineares. Tem-se os métodos iterativos de pontos fixos, dentre os quais pode-se destacar o S.O.R., Jacobi, Gauss-Seidel e variações. Há também, os métodos iterativos não estacionários dentro dos quais podem-se destacar os métodos baseados em subespaços de Krylov (ver Saad (1996)), como o GMRES, CG, BiCG, etc. Há ainda os métodos do tipo "Multigrid" (ver Greenbaum (1997) e Ghia et all (1982)) e "Domain Decomposition" (ver Greenbaum (1997)). Nesta gama de métodos iterativos há ainda a possibilidade de se trabalhar com pré-condicionadores, os quais podem ser formados de diversas maneiras e aplicados ao sistema linear, em questão, de diferentes formas. Podem-se destacar a fatoração ILU (ver Saad (1996)) e o "Multigrid" como pré-condicionador para os métodos baseados em subespaços de Krylov (ver Greenbaum (1997)).

O presente trabalho utilizará o Método de Mínimos Quadrados de Galerkin ("Galerkin Least Square"), aliado ao termo de captura de choque, para se chegar a abordagem via o método de elementos finitos do problema de valor de contorno. A metodologia para a solução do sistema não linear será o método de Newton inexato, associado aos métodos iterativos GMRES e BiCGStab, com pré-condicionamento ILU(0). Os códigos do GMRES, BiCGStab e ILU(0) pertencem a um pacote da biblioteca HSL, os quais foram impelmantados em Fortran 90, para poderem ser utilisadas numa programação orientada a objeto, que é o caso deste trabalho. O código desenvolvido foi em Fortran 90 e o compilador usado foi o "Compaq Visual Fortran 6.6". O pré e pós processamentos

dos dados ficaram a cargo do software GID 7.2..

1.3 Objetivo

O objetivo deste trabalho é desenvolver e implementar um código computacional de elementos finitos para comparar o desempenho entre o GMRES, o BiCGStab e um método direto (LU esparso com permutação), frente a problemas de escoamentos bidimensionais, incompressíveis e em regime permanente de fluidos Newtonianos.

1.4 Estrutura da Dissertação

Segue agora uma breve apresentação a respeito de cada capítulo deste trabalho:

- **Capítulo 2:** Este capítulo engloba desde a apresentação da formulação forte do escoamento até a apresentação da formulação via o método de elementos finitos.
- **Capitulo 3:** Este capítulo apresenta uma descrição desde o método de Newton até os métodos iterativos utilizados (GMRES e BiCGStab), bem como o esquema de pré-condicionamento (ILU(0)). Neste capítulo também encontram-se os algorítimos das metodologias, apresentados de uma forma simplificada e didática.
- **Capítulo 4:** Este capítulo engloba as aplicações feitas, assim como os resultados obtidos e as análises em cima destes resultados.
- **Capítulo 5:** Este capítulo apresenta as conclusões referentes aos resultados alcançados neste trabalho, bem como sugestões para o incremento e continuidade deste em um provável trabalho futuro.

Capítulo 2

A EQUAÇÃO DE NAVIER-STOKES

2.1 Introdução

Este capítulo engloba desde a apresentação da formulação forte do escoamento até a formulação via o método de elementos finitos.

O primeiro passo se realizará com a exposição do problema em sua forma forte, com todas as respectivas condições de contorno ("essenciais" e "naturais"). A equação de Navier-Stokes será aqui tratada em sua forma bi-dimensional, incompressível e em regime permanente, como se perceberá pelas condições a ela impostas. O passo seguinte é a obtenção da forma fraca do problema inicialmente apresentado. Uma importante observação que deve ser feita nesta etapa é o cuidado que se deve tomar com a "dimensão" das parcelas na forma integral ponderada do problema, as quais deverão ser dimensionalmente compatíveis. Terminada a obtenção da forma fraca do problema, prossegue-se com a incorporação, a esta, dos parâmetros de estabilização aliados ao parâmetro de captura de "descontinuidade" (ou "choque"). Estes parâmetros se tornam necessários ao se trabalhar com números de Reynolds elevados (Re próximo de 1000) e também devido a condição de Brezzi-Babuška ("inf-sup condition"). Feito isto, segue-se então com a aplicação da formulação via elementos finitos.

A formulação de elementos finitos do problema é sem dúvida a etapa mais importante deste processo inicial, muito embora esta seja fortemente influenciada pelas anteriores (formulações forte e fraca). Nesta etapa serão obtidas todas as matrizes e vetores "elementares" que contribuirão para a formação da matriz de "rigidez" global.

Uma vez definida a estrutura das matrizes de elementos finitos, apresentam-se, em capítulos subsequentes, o esquema de solução do sistema não linear de equações gerado e por fim são feitas algumas aplicações.

2.2 Formulação Forte

O problema do escoamento a ser resolvido pode ser descrito da forma como se segue.

Seja $\Omega \subset \mathbb{R}^2$ um domínio limitado e seja Γ o seu contorno. O problema consiste em determinar $\vec{u}(\vec{x})$ e $p(\vec{x})$, $\forall\ \vec{x} \in \Omega \cup \Gamma$ tal que:

$$
\begin{cases}
(\nabla \vec{u}(\vec{x}))\vec{u}(\vec{x}) - 2\nu\ \nabla \cdot \epsilon(\vec{u}(\vec{x})) + \frac{1}{\rho}\nabla p(\vec{x}) = \vec{b}(\vec{x}) \ \text{ em } \ \Omega; \\
\nabla \cdot \vec{u}(\vec{x}) = 0 \ \text{ em } \ \Omega; \\
\vec{u}(\vec{x}) = \vec{g}(\vec{x}) \ \text{ em } \ \Gamma_u; \\
2\nu\ \epsilon(\vec{u}(\vec{x})).\widehat{n} - \frac{1}{\rho}p(\vec{x})\widehat{n} = \vec{h}(\vec{x}) \ \text{ em } \ \Gamma_t, \\
\\
\text{onde} \\
\\
\epsilon(\vec{u}(\vec{x})) = \dfrac{\nabla \vec{u}(\vec{x}) + \nabla^T \vec{u}(\vec{x})}{2}; \\
\nu \text{ - viscosidade dinâmica,} \\
\rho \text{ - densidade,} \\
\vec{b} \in L^2(\Omega) \times L^2(\Omega),
\end{cases}
\tag{2.1}
$$

sendo

$\Gamma = \Gamma_u \cup \Gamma_t$;

Γ_u = Região do contorno com o campo de velocidade prescrito;

Γ_t = Região do contorno com a tensão prescrita,

como mostra a figura abaixo:

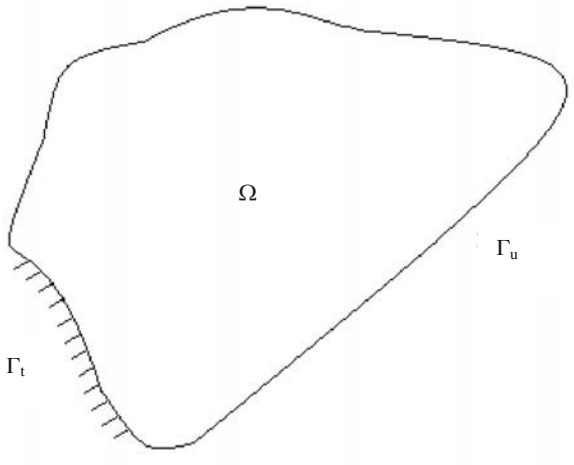

Fig 1: Dominio

Definindo os conjuntos de trabalho:

Conjunto do campo de velocidades admissíveis:
$Kinu(\Omega) = \{\vec{u}(\vec{x}) \in H^1(\Omega) \times H^1(\Omega) \mid \vec{u}(\vec{x}) = \vec{g}(\vec{x}) \text{ em } \Gamma_u\}$;

Conjunto das variações do campo de velocidades admissíveis:
$Varu(\Omega) = \{\vec{v}(\vec{x}) \in H^1(\Omega) \times H^1(\Omega) \mid \vec{u}(\vec{x}) = \vec{0} \text{ em } \Gamma_u\}$;

Conjunto do campo de pressões admissíveis:
$Kinp(\Omega) = \{p(\vec{x}) \in L^2(\Omega)\}$;

Conjunto das variações do campo de pressões admissíveis:
$Varp(\Omega) = \{\hat{p}(\vec{x}) \in L^2(\Omega)\}$,

onde $L^2(\Omega)$ é o espaço normado das funções Lebesgue quadrado integráveis em Ω e $H^1(\Omega)$ é o espaço Sobolev das funções Lebesgue quadrado integráveis, deriváveis em Ω, e cujas derivadas também são Lebesgue quadrado integráveis e normado ($\|.\|_1$).

2.3 Formulação Fraca

Visando a obtenção da forma fraca, apresenta-se agora o problema em sua formulação integral ponderada.

Determinar $(\vec{u}, p) \in Kinu \times Kinp$, tal que

$$\int_\Omega \{(\nabla\vec{u})\vec{u} - 2\nu\ \nabla\cdot\epsilon(\vec{u}) + \frac{1}{\rho}\nabla p - \vec{b}\}\cdot\vec{v}\ d\Omega - \int_\Omega \{\frac{1}{\rho}\nabla\cdot\vec{u}\}\hat{p}\ d\Omega = 0, \forall(\vec{v},\hat{p}) \in Varu(\Omega)\times Varp(\Omega), \tag{2.2}$$

onde convenciona-se que $\vec{b} = \vec{b}(\vec{x})$. Alternativamente pode-se escrever:

$$\int_\Omega \{(\nabla\vec{u})\vec{u}\}\cdot\vec{v}\ d\Omega - \int_\Omega 2\nu\{\nabla\cdot\epsilon(\vec{u})\}\cdot\vec{v}\ d\Omega + \int_\Omega \frac{1}{\rho}\{\nabla p\cdot\vec{v}\}\ d\Omega - \int_\Omega \{\vec{b}\cdot\vec{v}\}\ d\Omega - \int_\Omega \{\frac{1}{\rho}\nabla\cdot\vec{u}\}\hat{p}\ d\Omega = 0, \forall(\vec{v},\hat{p}) \in Varu(\Omega)\times Varp(\Omega). \tag{2.3}$$

Agora

$$\begin{aligned} div(\epsilon(\vec{u})^T\vec{v}) &= \vec{v}\cdot div(\epsilon(\vec{u})) + \nabla(\vec{v})\cdot\epsilon(\vec{u}); \\ &= \vec{v}\cdot div(\epsilon(\vec{u})) + \epsilon(\vec{v})\cdot\epsilon(\vec{u}), \end{aligned} \tag{2.4}$$

e

$$div(p\vec{v}) = p\cdot div(\vec{v}) + \vec{v}\cdot\nabla p. \tag{2.5}$$

Assim substituindo as relações de 2.4 e 2.5, respectivamente, na segunda e terceira integral de 2.3, obtém-se:

$$\int_\Omega \{(\nabla\vec{u})\vec{u}\}\cdot\vec{v}\ d\Omega - \int_\Omega 2\nu\ \{div(\epsilon(\vec{u})^T\vec{v}) - \epsilon(\vec{v})\cdot\epsilon(\vec{u})\}d\Omega + \int_\Omega \frac{1}{\rho}\{div(p\vec{v}) - p\cdot div(\vec{v})\}d\Omega - \int_\Omega \{\vec{b}\cdot\vec{v}\}d\Omega - \int_\Omega \{\frac{1}{\rho}div(\vec{u})\}\widehat{p}\ d\Omega = 0, \quad (2.6)$$

isto é

$$\int_\Omega \{(\nabla\vec{u})\vec{u}\}\cdot\vec{v}\ d\Omega - \int_\Omega 2\nu\{div(\epsilon(\vec{u})^T\vec{v})\}d\Omega + \int_\Omega 2\nu\{\epsilon(\vec{v})\cdot\epsilon(\vec{u})\}d\Omega + \int_\Omega \frac{1}{\rho}\{div(p\vec{v})\}d\Omega - \int_\Omega \{p\cdot div(\vec{v})\}d\Omega - \int_\Omega \{\vec{b}\cdot\vec{v}\}d\Omega - \int_\Omega \{\frac{1}{\rho}div(\vec{u})\}\widehat{p}\ d\Omega = 0, \quad (2.7)$$

porém tem-se

$$\begin{aligned}\int_\Omega \{div\ (\epsilon(\vec{u})^T\vec{v})\}d\Omega &= \int_\Gamma \{\epsilon(\vec{u})^T\vec{v}\cdot\vec{n}\}\ d\Gamma;\\ &= \int_{\Gamma_\mathbf{u}} \{\vec{v}\cdot(\epsilon(\vec{u})\vec{n}\}\ d\Gamma + \int_{\Gamma_\mathbf{t}} \{\vec{v}\cdot(\epsilon(\vec{u})\vec{n}\}\ d\Gamma. \quad (2.8)\end{aligned}$$

Contudo, $\vec{v}\in Varu$, logo $\vec{v}=\vec{0}$ em $\Gamma_\mathbf{u}$, o que implica em

$$\int_\Omega \{div\ (\epsilon(\vec{u})^T\vec{v})\}d\Omega = \int_{\Gamma_\mathbf{t}} \{\vec{v}\cdot\epsilon(\vec{u})\vec{n}\}\ d\Gamma, \quad (2.9)$$

e também

$$\begin{aligned}\int_\Omega \{div\ (p\vec{v})\}d\Omega &= \int_\Gamma \{p\vec{v}\cdot\vec{n}\}\ d\Gamma;\\ &= \int_{\Gamma_\mathbf{u}} \{\vec{v}\cdot p\vec{n}\}\ d\Gamma + \int_{\Gamma_\mathbf{t}} \{\vec{v}\cdot p\vec{n}\}\ d\Gamma;\\ &= \int_{\Gamma_\mathbf{t}} \{\vec{v}\cdot p\vec{n}\}\ d\Gamma. \quad (2.10)\end{aligned}$$

Agora substituindo os resultados de 2.9 e 2.10 em 2.7, obtém-se

$$\int_\Omega \{(\nabla\vec{u})\vec{u}\}\cdot\vec{v}\ d\Omega - 2\nu\int_{\Gamma_\mathbf{t}} \{\vec{v}\cdot\epsilon(\vec{u})\vec{n}\}\ d\Gamma + \int_\Omega 2\nu\{\epsilon(\vec{v})\cdot\epsilon(\vec{u})\}d\Omega + \int_{\Gamma_\mathbf{t}} \frac{1}{\rho}\{\vec{v}\cdot p\vec{n}\}\ d\Gamma - \int_\Omega \{p\cdot div(\vec{v})\}d\Omega - \int_\Omega \{\vec{b}\cdot\vec{v}\}d\Omega - \int_\Omega \{\frac{1}{\rho}div(\vec{u})\}\widehat{p}\ d\Omega = 0, \quad (2.11)$$

o que implica em

$$\int_\Omega \{(\nabla \vec{u})\vec{u}\} \cdot \vec{v}\ d\Omega + \int_\Omega 2\nu\ \{\epsilon(\vec{u}) \cdot \epsilon(\vec{v})\} d\Omega - \int_\Omega \frac{1}{\rho}\{p \cdot div(\vec{v})\} d\Omega - \int_\Omega \{\vec{b}\} \cdot \vec{v} d\Omega - \int_\Omega \frac{1}{\rho}\{div(\vec{u})\}\widehat{p}\ d\Omega - \int_{\Gamma_t} \{(2\nu\ \epsilon(\vec{u}).\widehat{n} - \frac{1}{\rho}p\widehat{n}) \cdot \vec{v}\} d\Gamma = 0. \quad (2.12)$$

Aplicando, agora, a condição de contorno natural $2\nu\ \epsilon(\vec{u}).\widehat{n} - \frac{1}{\rho}p\widehat{n} = \vec{h}(\vec{x})$ em Γ_t, obtém-se

$$\int_\Omega \{(\nabla \vec{u})\vec{u}\} \cdot \vec{v}\ d\Omega + \int_\Omega 2\nu\ \{\epsilon(\vec{u}) \cdot \epsilon(\vec{v})\} d\Omega - \int_\Omega \frac{1}{\rho}\{p \cdot div(\vec{v})\} d\Omega - \int_\Omega \{\vec{b}\} \cdot \vec{v} d\Omega - \int_\Omega \frac{1}{\rho}\{div(\vec{u})\}\widehat{p}\ d\Omega - \int_{\Gamma_t} \{\vec{h} \cdot \vec{v}\} d\Gamma = 0. \quad (2.13)$$

Neste ponto, simboliza-se por

$$\langle f, g\rangle_\Omega = \int_\Omega \{f(\vec{x}) \cdot g(\vec{x})\} d\Omega, \quad (2.14)$$

como o produto interno de quaisquer funções reais arbitrárias $f, g \in L^2(\Omega)$.

Desta forma obtém-se a seguinte formulação fraca:

Determinar $(\vec{u}, p) \in Kinu \times Kinp$, tal que

$$\begin{aligned} &\langle(\nabla \vec{u})\vec{u}, \vec{v}\rangle_\Omega + 2\nu\ \langle\epsilon(\vec{u}) \cdot \epsilon(\vec{v})\rangle_\Omega - \tfrac{1}{\rho}\langle p, div(\vec{v})\rangle_\Omega \\ &-\langle\vec{b}, \vec{v}\rangle_\Omega - \tfrac{1}{\rho}\langle div(\vec{u}), \hat{p}\rangle_\Omega - \langle\vec{h}, \vec{v}\rangle_{\Gamma_t} = 0, \forall(\vec{v}, \widehat{p}) \in Varu(\Omega) \times Varp(\Omega). \end{aligned} \quad (2.15)$$

Definição. 2.3.1 *: Uma subdivisão/partição/malha $\wp_h(\Omega)$ de um domínio Ω é uma coleção finita $N(\wp_h(\Omega))$ de conjuntos abertos, ou subdomínios (elementos), $\{K_i\}$, tal que*

$$\begin{aligned} &i)\ K_i \cap K_j = \varnothing,\ \text{se } i \neq j; \\ &ii)\ \bigcup_i \overline{K_i} = \bar{\Omega}. \end{aligned} \quad (2.16)$$

Observação. 2.3.1 *: Uma partição $\wp_h(\Omega)$ é dita ser regular se $\exists C \in \mathbb{R}^*_+$, tal que*

$$\frac{h_k}{h^*} \leq C,\ \forall K \in \wp_h(\Omega),$$

onde

$$h_k = diam(K);$$
$$h^* = \sup\{diam(B) \mid B \text{ é uma bola contida em } K\}.$$

Supõe-se agora que cada elemento $K \in \wp_h(\Omega)$ é a imagem de um elemento mestre $\hat{K}$ por uma aplicação afim F_K, isto é, $K = F_K(\hat{K})$, $\forall K \in \wp_h(\Omega)$, onde $\hat{K}$ é um cubo canônico ($\hat{K} = (-1,1)^n$, no caso aqui estudado $n = 2$). Sobre o elemento mestre, em $\mathbb{R}^n$, considera-se agora o espaço dos polinômios de grau $m \geq 0$, conforme segue:

$$\begin{array}{l} P_m(\hat{K}) = span\{\hat{x}^\alpha \mid 0 \leq \alpha_i \leq m,\ 0 \leq i \leq n\}; \\ Q_m(\hat{K}) = span\{\hat{x}^\alpha \mid 0 \leq |\alpha| \leq m\}. \end{array} \tag{2.17}$$

Observação. 2.3.2 *Note que $P_m(.)$ é o conjunto de todos os produtos tensoriais de polinômios de grau menor ou igual a p, definidos sobre o elemento mestre em cada direção coordenada.*

Agora define-se

$$R_m(K) = \left\{ \begin{array}{l} \vec{v} \in L^2(\Omega) \times L^2(\Omega); \\ \left[\begin{array}{l} \vec{v} = \vec{v}|_K \circ F_K \in P_m(K) \times P_m(K), \text{ se } K \in \wp_h(\Omega) \text{ é quadrilátero}; \\ \vec{v} = \vec{v}|_K \circ F_K \in Q_m(K) \times Q_m(K), \text{ se } K \in \wp_h(\Omega) \text{é trialgular}. \end{array} \right. \end{array} \right\} \tag{2.18}$$

Neste ponto faz-se necessária a introdução dos seguintes subespaços de dimensão finita:

$$\begin{array}{l} Kinu_h(\wp_h) = \{\vec{u} \in Kinu(\wp_h) \mid \vec{u}|_K \in R_m(K)^2, \forall K \in \wp_h(\Omega)\}; \\ \\ Varu_h(\wp_h) = \{\vec{v} \in Varu(\wp_h) \mid \vec{v}|_K \in R_m(K)^2, \forall K \in \wp_h(\Omega)\}; \\ \\ Kinp_h(\wp_h) = \{p \in Kinp(\wp_h) \mid p|_K \in R_l(K), \forall K \in \wp_h(\Omega)\}; \\ \\ Varp_h(\wp_h) = \{\hat{p} \in Varp(\wp_h) \mid \hat{p}|_K \in R_l(K), \forall K \in \wp_h(\Omega)\}. \end{array} \tag{2.19}$$

Assim, objetivando possibilitar o trabalho com as mais diversas condições (trabalhando ou não com elevados números de Reynolds, satisfazendo ou não as restrições de interpolação para velocidade e pressão), aplica-se agora o Método de Mínimos Quadrados de Galerkin proposto por Franca e Frey (1992). Este método contém um termo adicional, cujo objetivo é eliminar as oscilações que podem se apresentar em regiões de elevados gradientes. Tal termo é denominado de termo de captura de descontinuidade ou captura de choque. Os parâmetros de estabilização são obtidos por meio da solução de um problema de mínimos quadrados residual, aplicando-se a condição de estacionaridade (derivada de Gateaux nula) ao funcional em questão. Acrescentando estes parâmetros propostos por Franca e Frey (1992), o problema pode ser reformulado como se segue.

Determinar $(\vec{u}, p) \in Kinu_h(\wp_h) \times Kinp_h(\wp_h)$, tais que

$$B(\vec{u}, p, \vec{v}, \hat{p}) = F(\vec{v}, \hat{p}),\ \forall (\vec{v}, \widehat{p}) \in Varu_h(\wp_h) \times Varp_h(\wp_h), \tag{2.20}$$

com

$$\begin{aligned} B(\vec{u},p,\vec{v},\hat{p}) &= \langle(\nabla\vec{u})\vec{u},\vec{v}\rangle_{\wp_h(\Omega)} + 2\nu\ \langle\epsilon(\vec{u})\cdot\epsilon(\vec{v})\rangle_{\wp_h(\Omega)} \\ &- \frac{1}{\rho}\langle p, div(\vec{v})\rangle_{\wp_h(\Omega)} - \frac{1}{\rho}\langle div(\vec{u}),\hat{p}\rangle_{\wp_h(\Omega)} \\ &+ \langle div(\vec{u}), \delta div(\vec{v})\rangle_{\wp_h(\Omega)} \\ &+ \sum_{K\in\wp_h(\Omega)} \langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u} + \frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v} - \frac{1}{\rho}\nabla\hat{p}\right)\rangle_K, \end{aligned} \tag{2.21}$$

e

$$F(\vec{v},\hat{p}) = \langle\vec{b},\vec{v}\rangle_{\wp_h(\Omega)} + \langle\vec{h},\vec{v}\rangle_{\Gamma_{\mathfrak{t}}\cap\aleph_h(\Gamma)} + \sum_{K\in\wp_h(\Omega)} \langle\vec{b}, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v} - \frac{1}{\rho}\nabla\hat{p}\right)\rangle_K, \tag{2.22}$$

onde $\aleph_h(\Gamma)$ denota o conjunto dos bordos (ou de parte dos bordos) da partição $\wp_h(\Omega)$ na fronteira Γ. Os demais parâmetros de estabilidade são apresentados como se segue

$$\delta = \lambda\|\vec{u}(\vec{x})\|_p h_k \xi(\mathrm{Re}_k(\vec{x})); \tag{2.23}$$

$$\tau = \frac{h_k}{\mathbf{2}\|\vec{u}(\vec{x})\|_p}\xi(\mathrm{Re}_k(\vec{x})); \tag{2.24}$$

$$\mathrm{Re}_k(\vec{x}) = \frac{m_k\|\vec{u}(\vec{x})\|_p h_k}{4\nu}; \tag{2.25}$$

$$\xi(\mathrm{Re}_k(\vec{x})) = \begin{cases} \mathrm{Re}_k(\vec{x}),\ 0 \le \mathrm{Re}_k(\vec{x}) < 1; \\ 1,\ \mathrm{Re}_k(\vec{x}) \ge 1; \end{cases} \tag{2.26}$$

$$\|\vec{u}(\vec{x})\|_p = \left(\sum_{i=1}^{n}|u_i(\vec{x})|^p\right)^{\frac{1}{p}},\ 1 \le p < \infty; \tag{2.27}$$

$$m_k = \min\left\{\frac{1}{3}, 2C_k\right\}; \tag{2.28}$$

$$C_k \sum_{K\in\wp_h(\Omega)} h_k^2 \| \bigtriangledown\cdot\epsilon(\vec{v})\|_{0,K}^2 \le \|\epsilon(\vec{v})\|_0^2,\ \forall\vec{v}\in Varu; \tag{2.29}$$

$$\lambda > 0, \tag{2.30}$$

onde o subescrito k rafere-se ao elemento K. Considera-se agora a seguinte decomposição

$$\vec{u} = \vec{u}^* + \vec{u}^{\bigtriangledown},$$

onde $\vec{u}^{\bigtriangledown}$ é um campo de velocidade conhecido que satisfaz a condição de contorno essencial e não homogênea do campo prescrito na fronteira (i.e. $\vec{u}^{\bigtriangledown} \in Kinu_h(\wp_h)$) e $\vec{u}^*$ é um campo de velocidades desconhecido, tal que $\vec{u}^* \in Varu_h(\wp_h)$. Então o problema em questão pode ser reformulado da seguinte maneira:

Dado $\vec{u}^{\triangledown} \in Kinu_h(\wp_h)$, determine $(\vec{u}^*, p) \in Varu_h(\wp_h) \times Varp_h(\wp_h)$, tal que

$$B(\vec{u}^*, p, \vec{v}, \hat{p}) = F(\vec{v}, \hat{p}),\ \forall (\vec{v}, \hat{p}) \in Varu_h(\wp_h) \times Varp_h(\wp_h), \tag{2.31}$$

com

$$\begin{aligned} B(\vec{u}^*, p, \vec{v}, \hat{p}) = & \langle (\nabla(\vec{u}^* + \vec{u}^{\triangledown}))(\vec{u}^* + \vec{u}^{\triangledown}), \vec{v} \rangle_{\wp_h(\Omega)} + 2\nu\ \langle \epsilon(\vec{u}^* + \vec{u}^{\triangledown}) \cdot \epsilon(\vec{v}) \rangle_{\wp_h(\Omega)} \\ & - \frac{1}{\rho} \langle p, div(\vec{v}) \rangle_{\wp_h(\Omega)} - \frac{1}{\rho} \langle div(\vec{u}^* + \vec{u}^{\triangledown}), \hat{p} \rangle_{\wp_h(\Omega)} \\ & + \langle div(\vec{u}^* + \vec{u}^{\triangledown}), \delta div(\vec{v}) \rangle_{\wp_h(\Omega)} \\ & + \sum_{K \in \wp_h(\Omega)} \langle (\nabla(\vec{u}^* + \vec{u}^{\triangledown}))(\vec{u}^* + \vec{u}^{\triangledown}) - \nu \triangle(\vec{u}^* + \vec{u}^{\triangledown}) \\ & + \frac{1}{\rho} \nabla p, \tau \left((\nabla \vec{v})(\vec{u}^* + \vec{u}^{\triangledown}) \pm \nu \triangle \vec{v} - \frac{1}{\rho} \nabla \hat{p} \right) \rangle_K, \end{aligned} \tag{2.32}$$

e

$$\begin{aligned} F(\vec{v}, \hat{p}) = & \langle \vec{b}, \vec{v} \rangle_{\wp_h(\Omega)} + \langle \vec{h}, \vec{v} \rangle_{\Gamma_{\mathbf{t}} \cap \aleph_h(\Gamma)} \\ & + \sum_{K \in \wp_h(\Omega)} \langle \vec{b}, \tau \left((\nabla \vec{v})(\vec{u}^* + \vec{u}^{\triangledown}) \pm \nu \triangle \vec{v} - \frac{1}{\rho} \nabla \hat{p} \right) \rangle_K. \end{aligned} \tag{2.33}$$

2.4 Formulação do Problema pelo Método de Elementos Finitos

Tomando agora $\vec{u} = u\hat{e}_1 + v\hat{e}_2$, onde $\hat{e}_1$e $\hat{e}_2$ são os vetores da base canônica, o domínio Ω é particionado em elementos K, nos quais os campos de velocidade e pressão serão interpolados. Seguindo então os procedimentos clássicos utilizados no método de elementos finitos, pode-se escrever os campos de interpolação, para um elemento K, na seguinte forma matricial:

$$\begin{aligned} u^* &= [\mathbb{N}_i]\ \vec{q}_k^{u^*} \qquad u^{\triangledown} = [\mathbb{N}_i]\ \vec{q}_k^{u^{\triangledown}} \qquad \therefore \quad u = [\mathbb{N}_i]\ \vec{q}_k^{u} \\ v^* &= [\mathbb{N}_i]\ \vec{q}_k^{v^*} \qquad v^{\triangledown} = [\mathbb{N}_i]\ \vec{q}_k^{v^{\triangledown}} \qquad \therefore \quad v = [\mathbb{N}_i]\ \vec{q}_k^{v} \\ p &= [\mathbb{N}_i]\ \vec{q}_k^{p} \end{aligned} \tag{2.34}$$

onde $[\mathbb{N}_i]_{i=\vec{u},p}$ é o vetor que contém as funções de interpolação elementares, e $\vec{q}_k^{u} = \vec{q}_k^{u^*} + \vec{q}_k^{u^{\triangledown}}$ e $\vec{q}_k^{v} = \vec{q}_k^{v^*} + \vec{q}_k^{v^{\triangledown}}$, onde convenciona-se que o subescrito k refere-se ao elemento K.

No caso particular de um elemento quadrilateral ("quad-four"), os vetores incógnitas ficam representados da seguinte forma:

$$\begin{aligned} \vec{q}_k^{\,u^*} &= \left(u_1^*, u_2^*, u_3^*, u_4^*\right); \\ \vec{q}_k^{\,v^*} &= \left(v_1^*, v_2^*, v_3^*, v_4^*\right); \\ \vec{q}_k^{\,p} &= \left(p_1, p_2, p_3, p_4\right). \end{aligned} \tag{2.35}$$

Seguindo com o cálculo elementar (para cada elemento K) de cada uma das parcelas da formulação apresentada em (2.32) e (2.33), tem-se:

- **A)** Determinação de $\langle(\nabla\vec{u})\vec{u},\vec{v}\rangle_K = \langle[\nabla(\vec{u}^* + \vec{u}^{\triangledown})](\vec{u}^* + \vec{u}^{\triangledown}),\vec{v}\rangle_K$. Então:

$$(\bigtriangledown\vec{u})\,\vec{u} = \left\{ \begin{array}{c} u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} \\ u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} \end{array} \right\},$$

todavia,

$$\begin{aligned} \frac{\partial u}{\partial x} &= \left[\frac{\partial \mathbb{N}_u}{\partial x}\right]\vec{q}_k^{\,u}\,; \qquad \frac{\partial u}{\partial y} = \left[\frac{\partial \mathbb{N}_u}{\partial y}\right]\vec{q}_k^{\,u}; \\ \frac{\partial v}{\partial x} &= \left[\frac{\partial \mathbb{N}_u}{\partial x}\right]\vec{q}_k^{\,v}\,; \qquad \frac{\partial v}{\partial y} = \left[\frac{\partial \mathbb{N}_u}{\partial y}\right]\vec{q}_k^{\,v}, \end{aligned}$$

logo

$$(\bigtriangledown\vec{u})\,\vec{u} = \left\{ \begin{array}{cc} u\left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & \vec{q}_k^{\,u} + v\left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \quad \vec{q}_k^{\,u} \\ u\left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & \vec{q}_k^{\,v} + v\left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \quad \vec{q}_k^{\,v} \end{array} \right\} = \left[\mathbb{N}^{\nabla u\ u}\right]\vec{q}_k,$$

onde

$$\left[\mathbb{N}^{\nabla u\ u}\right] = \left[\begin{array}{ccc} u\left[\frac{\partial \mathbb{N}_u}{\partial x}\right] + v\left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & [0] & [0] \\ [0] & u\left[\frac{\partial \mathbb{N}_u}{\partial x}\right] + v\left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & [0] \end{array} \right] \quad \text{e} \quad \vec{q}_k = \left(\vec{q}_k^{\,u}, \vec{q}_k^{\,v}, \vec{q}_k^{\,p}\right).$$

Note ainda que

$$\vec{u} = \left\{ \begin{array}{c} u \\ v \end{array} \right\} = \left[\begin{array}{ccc} [\mathbb{N}_u] & [0] & [0] \\ [0] & [\mathbb{N}_u] & [0] \end{array} \right]\vec{q}_k = \left[\mathbb{N}^{disp}\right]\vec{q}_k,$$

e analogamente,

$$\vec{v} = \left\{ \begin{array}{c} \hat{u} \\ \hat{v} \end{array} \right\} = \left[\mathbb{N}^{disp}\right]\hat{q}_k \quad \text{onde} \quad \hat{q}_k = \left(\hat{q}_k^u, \hat{q}_k^v, \hat{q}_k^p\right),$$

o que permite escrever

$$\begin{aligned}\langle(\nabla\vec{u})\vec{u},\vec{v}\rangle_K &= \left\langle\left[\mathbb{N}^{\nabla u\,u}\right]\vec{q}_k,\left[\mathbb{N}^{disp}\right]\hat{q}_k\right\rangle_K;\\ &= \left\langle\left[\mathbb{N}^{disp}\right]^T\left[\mathbb{N}^{\nabla u\,u}\right]\vec{q}_k,\hat{q}_k\right\rangle_K;\\ &= \left[\int_K\left[\mathbb{N}^{disp}\right]^T\left[\mathbb{N}^{\nabla u\,u}\right]dK\right]\vec{q}_k\cdot\hat{q}_k;\\ &= \left[\mathbb{K}_k^A\right]\vec{q}_k\cdot\hat{q}_k,\end{aligned}$$

onde

$$\left[\mathbb{K}_k^A\right] = \int_K\left[\mathbb{N}^{disp}\right]^T\left[\mathbb{N}^{\nabla u\,u}\right]dK.$$

Observando melhor a matriz $\left[\mathbb{K}_k^A\right]$, tem-se,

$$\left[\mathbb{N}^{disp}\right]^T\left[\mathbb{N}^{\nabla u\,u}\right] = \begin{bmatrix}[\mathbb{N}_u]^T & [0]\\ [0] & [\mathbb{N}_u]^T\\ [0] & [0]\end{bmatrix}\begin{bmatrix}u\left[\frac{\partial\mathbb{N}_u}{\partial x}\right]+v\left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0] & [0]\\ [0] & u\left[\frac{\partial\mathbb{N}_u}{\partial x}\right]+v\left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0]\end{bmatrix};$$

$$= \begin{bmatrix}u[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial x}\right]+v[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0] & [0]\\ [0] & u[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial x}\right]+v[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0]\\ [0] & [0] & [0]\end{bmatrix}.$$

Então,

$$\left[\mathbb{K}_k^A\right] = \begin{bmatrix}[\mathbb{K}_k^{u\,u}] & [0] & [0]\\ [0] & [\mathbb{K}_k^{v\,v}] & [0]\\ [0] & [0] & [0]\end{bmatrix},$$

onde

$$[\mathbb{K}_k^{u\,u}] = [\mathbb{K}_k^{v\,v}] = \int_K\left\{u[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial x}\right]+v[\mathbb{N}_u]^T\left[\frac{\partial\mathbb{N}_u}{\partial y}\right]\right\}dK.$$

- **B)** Determinação de $2\nu\langle\epsilon(\vec{u}),\epsilon(\vec{v})\rangle_K = 2\nu\langle\epsilon(\vec{u}^*+\vec{u}^\nabla),\epsilon(\vec{v})\rangle_K$. Agora:

$$\epsilon(\vec{v}) = \frac{\nabla\vec{v}+\nabla^T\vec{v}}{2}\ \wedge\ \epsilon(\vec{u}) = \frac{\nabla\vec{u}+\nabla^T\vec{u}}{2},$$

mas como,

$$\vec{u} = \vec{u}^* + \vec{u}^\nabla\ \ \therefore\ \ \epsilon(\vec{u}) = \epsilon(\vec{u}^*) + \epsilon(\vec{u}^\nabla),$$

todavia,

$$\epsilon(\vec{u})\cdot\epsilon(\vec{v}) = tr(\epsilon^T(\vec{u})\epsilon(\vec{v})) = \epsilon(\vec{u})_{11}\epsilon(\vec{v})_{11}+\epsilon(\vec{u})_{12}\epsilon(\vec{v})_{12}+\epsilon(\vec{u})_{22}\epsilon(\vec{v})_{22}+\epsilon(\vec{u})_{21}\epsilon(\vec{v})_{21}.$$

Note porém que com $\epsilon(\cdot)$ é a parte simétrica do gradiente de velocidade, logo,

$$\epsilon(\vec{u})\cdot\epsilon(\vec{v}) = \epsilon(\vec{u})_{11}\epsilon(\vec{v})_{11}+2\epsilon(\vec{u})_{12}\epsilon(\vec{v})_{12}+\epsilon(\vec{u})_{22}\epsilon(\vec{v})_{22}.$$

Definindo agora

$$\vec{\varepsilon}(\vec{u}) = \begin{bmatrix} \epsilon(\vec{u})_{11} \\ \epsilon(\vec{u})_{22} \\ 2\epsilon(\vec{u})_{12} \end{bmatrix},$$

ou

$$\vec{\varepsilon}(\vec{u}) = \begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \end{bmatrix}.$$

Seguindo com raciocínio análogo ao do item anterior, tem-se que

$$\frac{\partial u}{\partial x} = \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] \vec{q}_k^{\,u}; \quad \frac{\partial v}{\partial y} = \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \vec{q}_k^{\,v}; \quad \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \vec{q}_k^{\,u} + \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] \vec{q}_k^{\,v},$$

então

$$\vec{\varepsilon}(\vec{u}) = \begin{bmatrix} \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] \vec{q}_k^{\,u} \\ \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \vec{q}_k^{\,v} \\ \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] \vec{q}_k^{\,u} + \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] \vec{q}_k^{\,v} \end{bmatrix},$$

logo

$$\vec{\varepsilon}(\vec{u}) = \begin{bmatrix} \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & [0] & [0] \\ [0] & \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & [0] \\ \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & [0] \end{bmatrix} \vec{q}_k = [B^{\mathbf{v}}]\, \vec{q}_k,$$

onde

$$[B^{\mathbf{v}}] = \begin{bmatrix} \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & [0] & [0] \\ [0] & \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & [0] \\ \left[\frac{\partial \mathbb{N}_u}{\partial y}\right] & \left[\frac{\partial \mathbb{N}_u}{\partial x}\right] & [0] \end{bmatrix} \quad \text{e} \quad \vec{q}_k = (\vec{q}_k^{\,u}, \vec{q}_k^{\,v}, \vec{q}_k^{\,p}).$$

Portanto, pode-se escrever

$$2\nu\epsilon(\vec{u}) \cdot \epsilon(\vec{v}) = \begin{bmatrix} 2\nu & 0 & 0 \\ 0 & 2\nu & 0 \\ 0 & 0 & \nu \end{bmatrix} \vec{\varepsilon}(\vec{u}) \cdot \vec{\varepsilon}(\vec{v}) = [H^{\mathbf{v}}]\, \vec{\varepsilon}(\vec{u}) \cdot \vec{\varepsilon}(\vec{v});$$
$$= [H^{\mathbf{v}}]\, [B^{\mathbf{v}}]\, \vec{q}_k \cdot [B^{\mathbf{v}}]\, \hat{q}_k,$$

o que permite obter

$$\begin{aligned}
2\nu\langle\epsilon(\vec{u}),\epsilon(\vec{v})\rangle_K &= \langle[H^{\mathbf{v}}]\,\vec{\varepsilon}(\vec{u}),\vec{\varepsilon}(\vec{v})\rangle_K\,;\\
&= \langle[H^{\mathbf{v}}]\,[B^{\mathbf{v}}]\,\vec{q}_k,[B^{\mathbf{v}}]\,\hat{q}_k\rangle_K\,;\\
&= \left\langle[B^{\mathbf{v}}]^T\,[H^{\mathbf{v}}]\,[B^{\mathbf{v}}]\,\vec{q}_k,\hat{q}_k\right\rangle_K\,;\\
&= \left[\int_K[B^{\mathbf{v}}]^T\,[H^{\mathbf{v}}]\,[B^{\mathbf{v}}]\,dK\right]\vec{q}_k\cdot\hat{q}_k;\\
&= \left[\mathbb{K}_k^B\right]\vec{q}_k\cdot\hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{K}_k^B\right] = \int_K[B^{\mathbf{v}}]^T\,[H^{\mathbf{v}}]\,[B^{\mathbf{v}}]\,dK.$$

- **C)** Determinação de $\frac{1}{\rho}\langle p, div(\vec{v})\rangle_K$. Agora

$$p = [\mathbb{N}_p]\,\vec{q}_k^{\,p};\quad \hat{u} = [\mathbb{N}_u]\,\hat{q}_k^u;\quad \hat{v} = [\mathbb{N}_u]\,\hat{q}_k^v,$$

daí

$$\begin{aligned}
div(\vec{v}) &= \frac{\partial\hat{u}}{\partial x}+\frac{\partial\hat{v}}{\partial y};\\
&= \left[\frac{\partial\mathbb{N}_u}{\partial x}\right]\hat{q}_k^u+\left[\frac{\partial\mathbb{N}_u}{\partial y}\right]\hat{q}_k^v;\\
&= \left[\begin{array}{ccc}\left[\frac{\partial\mathbb{N}_u}{\partial x}\right] & \left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0]\end{array}\right]\cdot\hat{q}_k = \left[B^{div}\right]\cdot\hat{q}_k,
\end{aligned}$$

onde

$$\left[B^{div}\right] = \left[\begin{array}{ccc}\left[\frac{\partial\mathbb{N}_u}{\partial x}\right] & \left[\frac{\partial\mathbb{N}_u}{\partial y}\right] & [0]\end{array}\right]\quad\text{e}\quad\hat{q}_k = (\hat{q}_k^u,\hat{q}_k^v,\hat{q}_k^p)\,.$$

Denotando,

$$[\mathbb{N}^{press}] = \left[\begin{array}{ccc}[0] & [0] & [\mathbb{N}_p]\end{array}\right],$$

então

$$p = [\mathbb{N}^{press}]\cdot\vec{q}_k,\ \text{onde}\ \vec{q}_k = (\vec{q}_k^{\,u},\vec{q}_k^{\,v},\vec{q}_k^{\,p})\,,$$

logo tem-se que,

$$\begin{aligned}
\frac{1}{\rho}\langle p, div(\vec{v})\rangle_K &= \frac{1}{\rho}\langle[\mathbb{N}^{press}]\cdot\vec{q}_k,\left[B^{div}\right]\cdot\hat{q}_k\rangle_K;\\
&= \frac{1}{\rho}\int_K([\mathbb{N}^{press}]\cdot\vec{q}_k)\cdot\left(\left[B^{div}\right]\cdot\hat{q}_k\right)dK;\\
&= \frac{1}{\rho}\left[\int_K\left[\left[B^{div}\right]\otimes[\mathbb{N}^{press}]\right]dK\right]\vec{q}_k\cdot\hat{q}_k;\\
&= \left[\mathbb{K}_k^C\right]\vec{q}_k\cdot\hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{K}_k^C\right] = \frac{1}{\rho}\left[\int_K \left[\left[B^{div}\right] \otimes \left[\mathbb{N}^{press}\right]\right] dK\right].$$

- **D)** Determinação de $\frac{1}{\rho}\langle div(\vec{u}), \hat{p}\rangle_K = \frac{1}{\rho}\langle div(\vec{u}^* + \vec{u}^\nabla), \hat{p}\rangle_K$. Tem-se por analogia ao item anterior,

$$div(\vec{u}^* + \vec{u}^\nabla) = \left[B^{div}\right] \cdot \vec{q}_k \quad \wedge \quad \hat{p} = \left[\mathbb{N}^{press}\right] \cdot \hat{q}_k,$$

então,

$$\begin{aligned}
\frac{1}{\rho}\langle div(\vec{u}), \hat{p}\rangle_K &= \frac{1}{\rho}\langle \left[B^{div}\right] \cdot \vec{q}_k, \left[\mathbb{N}^{press}\right] \cdot \hat{q}_k\rangle_K; \\
&= \frac{1}{\rho}\int_K \left(\left[B^{div}\right] \cdot \vec{q}_k\right) \cdot \left(\left[\mathbb{N}^{press}\right] \cdot \hat{q}_k\right) dK; \\
&= \frac{1}{\rho}\left[\int_K \left[\left[\mathbb{N}^{press}\right] \otimes \left[B^{div}\right]\right] dK\right] \vec{q}_k \cdot \hat{q}_k; \\
&= \left[\mathbb{K}_k^D\right] \vec{q}_k \cdot \hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{K}_k^D\right] = \frac{1}{\rho}\left[\int_{\Omega_e} \left[\left[\mathbb{N}^{press}\right] \otimes \left[B^{div}\right]\right] d\Omega\right].$$

Note ainda que

$$\left[\mathbb{K}_k^D\right] = \left[\mathbb{K}_k^C\right]^T.$$

- **E)** Determinação de $\langle div(\vec{u}), \delta div(\vec{v})\rangle_K = \langle div(\vec{u}^* + \vec{u}^\nabla), \delta div(\vec{v})\rangle_K$. Por analogia, novamente,

$$div(\vec{u}^* + \vec{u}^\nabla) = \left[B^{div}\right] \cdot \vec{q}_k \quad \wedge \quad div(\vec{v}) = \left[B^{div}\right] \cdot \hat{q}_k,$$

então

$$\begin{aligned}
\langle div(\vec{u}), \delta div(\vec{v})\rangle_K &= \langle \left[B^{div}\right] \cdot \vec{q}_k, \delta\left[B^{div}\right] \cdot \hat{q}_k\rangle_K; \\
&= \int_K \delta\left(\left[B^{div}\right] \cdot \vec{q}_k\right) \cdot \left(\left[B^{div}\right] \cdot \hat{q}_k\right) dK; \\
&= \left[\int_K \delta\left[\left[B^{div}\right] \otimes \left[B^{div}\right]\right] dK\right] \vec{q}_k \cdot \hat{q}_k; \\
&= \left[\mathbb{K}_k^E\right] \vec{q}_k \cdot \hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{K}_k^E\right] = \int_K \delta\left[\left[B^{div}\right] \otimes \left[B^{div}\right]\right] dK.$$

Este é o termo de captura de descontinuidade (ou captura de choque). Note ainda que sua atuação é dirigida a regiões de elevados gradientes, i.e. regiões nas quais a

equação da continuidade é violada.

- **F)** Determinação de $\langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u} + \frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v} - \frac{1}{\rho}\nabla\hat{p}\right)\rangle_K$. Tem-se então

$$\begin{aligned}\langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u} + \frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v} - \frac{1}{\rho}\nabla\hat{p}\right)\rangle_K &= \langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u}, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v}\right)\rangle_K \\ &- \langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u}, \tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\rangle_K \\ &+ \langle(\frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v}\right)\rangle_K \\ &- \langle\frac{1}{\rho}\nabla p, \tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\rangle_K.\end{aligned}$$

Este é o termo de estabilização, obtido pela aplicação da condição de estacionaridade ao funcional do quadrado da norma do resíduo da Equação de Navier-Stokes. A atuação desta parcela é dirigida a combater instabilidades numéricas que podem ser geradas, por exemplo, pela não satisfação das restrições de interpolações para os campos pressão e velocidade. Procedendo então ao cálculo de cada uma das parcelas isoladamente, segue:

- **F.1)** Determinação de $\langle(\nabla\vec{u})\vec{u} - \nu\triangle\vec{u}, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v}\right)\rangle_K$, tem-se então por analogia,

$$(\bigtriangledown\vec{u})\,\vec{u} = \left[\mathbb{N}^{\nabla u\,u}\right]\vec{q}_k \quad \wedge \quad (\bigtriangledown\vec{v})\,\vec{u} = \left[\mathbb{N}^{\nabla u\,u}\right]\hat{q}_k,$$

todavia,

$$\nu\triangle\vec{u} = \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)\hat{e}_1 + \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right)\hat{e}_2,$$

então,

$$\begin{aligned}\triangle\vec{u} &= \begin{bmatrix} \nu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \\ \nu\left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right)\end{bmatrix}; \\ &= \begin{bmatrix} \nu\left(\frac{\partial^2 \mathbb{N}_u}{\partial x^2} + \frac{\partial^2 \mathbb{N}_u}{\partial y^2}\right)\vec{q}_k^{\,u} \\ \nu\left(\frac{\partial^2 \mathbb{N}_u}{\partial x^2} + \frac{\partial^2 \mathbb{N}_u}{\partial y^2}\right)\vec{q}_k^{\,v}\end{bmatrix}; \\ &= \left[\mathbb{N}^{\triangle u}\right]\vec{q}_k \;\;\therefore\;\; \triangle\vec{v} = \left[\mathbb{N}^{\triangle u}\right]\hat{q}_k,\end{aligned}$$

onde,

$$\left[\mathbb{N}^{\triangle u}\right] = \begin{bmatrix} \frac{\partial^2 \mathbb{N}_u}{\partial x^2} + \frac{\partial^2 \mathbb{N}_u}{\partial y^2} & [0] & [0] \\ [0] & \frac{\partial^2 \mathbb{N}_u}{\partial x^2} + \frac{\partial^2 \mathbb{N}_u}{\partial y^2} & [0]\end{bmatrix},$$

por consequência pode-se escrever

$$\begin{aligned}(\nabla\vec{u})\vec{u}-\nu\triangle\vec{u} &= \left[\mathbb{N}^{\nabla u\ u}\right]\vec{q}_k-\left[\mathbb{N}^{\triangle u}\right]\vec{q}_k;\\ &= \left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]\vec{q}_k;\end{aligned}$$

$$\begin{aligned}(\nabla\vec{v})\vec{u}\pm\nu\triangle\vec{v} &= \left[\mathbb{N}^{\nabla u\ u}\right]\hat{q}_k\pm\left[\mathbb{N}^{\triangle u}\right]\hat{q}_k;\\ &= \left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]\hat{q}_k.\end{aligned}$$

Daí tem-se finalmente,

$$\begin{aligned}\left\langle(\nabla\vec{u})\vec{u}-\nu\triangle\vec{u},\tau\left((\nabla\vec{v})\vec{u}\pm\nu\triangle\vec{v}\right)\right\rangle_K &= \left\langle\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]\vec{q}_k,\tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]\hat{q}_k\right\rangle_K;\\ &= \left\langle\tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]\vec{q}_k,\hat{q}_k\right\rangle_K;\\ &= \left[\int_K\tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]dK\right]\vec{q}_k\cdot\hat{q}_k;\\ &= \left[\mathbb{K}_k^{F.1}\right]\vec{q}_k\cdot\hat{q}_k,\end{aligned}$$

onde

$$\left[\mathbb{K}_k^{F.1}\right]=\int_K\tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]dK.$$

- **F.2)** Determinação de $\left\langle(\nabla\vec{u})\vec{u}-\nu\triangle\vec{u},\tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\right\rangle_K$, tem-se então por analogia

$$\begin{aligned}\left\langle(\nabla\vec{u})\vec{u}-\nu\triangle\vec{u},\tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\right\rangle_K &= \frac{1}{\rho}\left\langle\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]\vec{q}_k,\tau\left[\mathbb{N}^{\nabla p}\right]\hat{q}_k\right\rangle_K;\\ &= \frac{1}{\rho}\left\langle\tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]\vec{q}_k,\hat{q}_k\right\rangle_K;\\ &= \frac{1}{\rho}\left[\int_K\tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]dK\right]\vec{q}_k\cdot\hat{q}_k;\\ &= \left[\mathbb{K}_k^{F.2}\right]\vec{q}_k\cdot\hat{q}_k,\end{aligned}$$

onde

$$\left[\mathbb{K}_k^{F.2}\right]=\frac{1}{\rho}\int_K\tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla u\ u-\triangle u}\right]dK,$$

e

$$\begin{aligned}\nabla\hat{p} &= \begin{bmatrix}\frac{\partial\hat{p}}{\partial y}\\ \frac{\partial\hat{p}}{\partial y}\end{bmatrix};\\ &= \begin{bmatrix}\frac{\partial\mathbb{N}_p}{\partial y}\hat{q}_k^p\\ \frac{\partial\mathbb{N}_p}{\partial y}\hat{q}_k^p\end{bmatrix};\\ &= \left[\mathbb{N}^{\nabla p}\right]\hat{q}_k,\end{aligned}$$

com

$$\left[\mathbb{N}^{\nabla p}\right] = \begin{bmatrix} [0] & [0] & \frac{\partial \mathbb{N}_p}{\partial x} \\ [0] & [0] & \frac{\partial \mathbb{N}_p}{\partial y} \end{bmatrix}.$$

- **F.3)** Determinação de $\langle(\frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v}\right)\rangle_K$. Tem-se então por analogia,

$$\begin{aligned}
\langle(\frac{1}{\rho}\nabla p, \tau\left((\nabla\vec{v})\vec{u} \pm \nu\triangle\vec{v}\right)\rangle_K &= \frac{1}{\rho}\langle\left[\mathbb{N}^{\nabla p}\right]\vec{q}_k, \tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]\hat{q}_k\rangle_K; \\
&= \frac{1}{\rho}\langle\tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla p}\right]\vec{q}_k, \hat{q}_k\rangle_K; \\
&= \frac{1}{\rho}\left[\int_K \tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla p}\right]dK\right]\vec{q}_k\cdot\hat{q}_k; \\
&= \left[\mathbb{K}_k^{F.3}\right]\vec{q}_k\cdot\hat{q}_k,
\end{aligned}$$

onde,

$$\left[\mathbb{K}_k^{F.3}\right] = \frac{1}{\rho}\int_K \tau\left[\mathbb{N}^{\nabla u\ u\pm\triangle u}\right]^T\left[\mathbb{N}^{\nabla p}\right]dK.$$

Note ainda que:

$$\left[\mathbb{K}_k^{F.3}\right] = \left[\mathbb{K}_k^{F.2}\right]^T.$$

- **F.4)** Determinação de $\langle\frac{1}{\rho}\nabla p, \tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\rangle_K$. Por analogia

$$\begin{aligned}
\langle\frac{1}{\rho}\nabla p, \tau\left(\frac{1}{\rho}\nabla\hat{p}\right)\rangle_K &= \frac{1}{\rho}\langle\left[\mathbb{N}^{\nabla p}\right]\vec{q}_k, \tau\left[\mathbb{N}^{\nabla p}\right]\hat{q}_k\rangle_K; \\
&= \frac{1}{\rho}\langle\tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla p}\right]\vec{q}_k, \hat{q}_k\rangle_K; \\
&= \frac{1}{\rho}\left[\int_K \tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla p}\right]dK\right]\vec{q}_k\cdot\hat{q}_k; \\
&= \left[\mathbb{K}_k^{F.4}\right]\vec{q}_k\cdot\hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{K}_k^{F.4}\right] = \frac{1}{\rho}\int_K \tau\left[\mathbb{N}^{\nabla p}\right]^T\left[\mathbb{N}^{\nabla p}\right]dK.$$

- **G)** Determinação de $\langle\vec{b}, \vec{v}\rangle_K$. Por analogia

$$\vec{v} = \begin{Bmatrix} \hat{u} \\ \hat{v} \end{Bmatrix} = \left[\mathbb{N}^{disp}\right]\hat{q}_k, \text{ onde } \hat{q}_k = \left(\hat{q}_k^u, \hat{q}_k^v, \hat{q}_k^p\right),$$

e

$$\vec{b} = \begin{bmatrix} b_x \\ b_y \end{bmatrix},$$

então

$$\begin{aligned}\langle \vec{b}, \vec{v} \rangle_K &= \langle \vec{b}, \left[\mathbb{N}^{disp}\right] \hat{q}_k \rangle_K; \\ &= \int_K \left\{\vec{b}\right\} \cdot \left[\mathbb{N}^{disp}\right] \hat{q}_k dK; \\ &= \left[\int_K \left[\mathbb{N}^{disp}\right]^T \left\{\vec{b}\right\} dK\right] \cdot \hat{q}_k; \\ &= \left[\mathbb{K}_k^G\right] \cdot \hat{q}_k,\end{aligned}$$

onde

$$\left[\mathbb{F}_k^G\right] = \int_{\Omega_\mathbf{e}} \left[\mathbb{N}^{disp}\right]^T \left\{\vec{b}\right\} d\Omega.$$

- **H)** Determinação de $\langle \vec{h}, \vec{v} \rangle_{\Gamma_\mathbf{t} \cap \partial K}$. Por analogia

$$\begin{aligned}\langle \vec{h}, \vec{v} \rangle_{\Gamma_\mathbf{t} \cap \partial K} &= \langle \vec{h}, \left[\mathbb{N}^{disp}\right] \hat{q}_k \rangle_{\Gamma_\mathbf{t} \cap \partial K}; \\ &= \int_{\Gamma_\mathbf{t} \cap \partial K} \left\{\vec{h}\right\} \cdot \left[\mathbb{N}^{disp}\right] \hat{q}_k d\Gamma; \\ &= \left[\int_{\Gamma_\mathbf{t}} \left[\mathbb{N}^{disp}\right]^T \left\{\vec{h}\right\} d\Gamma\right] \cdot \hat{q}_k; \\ &= \left[\mathbb{F}_k^H\right] \cdot \hat{q}_k,\end{aligned}$$

onde

$$\left[\mathbb{F}_k^H\right] = \int_{\Gamma_\mathbf{t}} \left[\mathbb{N}^{disp}\right]^T \left\{\vec{h}\right\} d\Gamma,$$

sendo

$$\vec{h} = \begin{bmatrix} h_x \\ h_y \end{bmatrix}.$$

- **I)** Determinação de $\langle \vec{b}, \tau \left((\nabla \vec{v}) \vec{u} \pm \nu \triangle \vec{v} - \frac{1}{\rho} \nabla \hat{p} \right) \rangle_K$. Então

$$\langle \vec{b}, \tau \left((\nabla \vec{v}) \vec{u} \pm \nu \triangle \vec{v} - \frac{1}{\rho} \nabla \hat{p} \right) \rangle_K = \langle \vec{b}, \tau \left((\nabla \vec{v}) \vec{u} \pm \nu \triangle \vec{v} - \frac{1}{\rho} \nabla \hat{p} \right) \rangle_K - \langle \vec{b}, \tau \left(\frac{1}{\rho} \nabla \hat{p} \right) \rangle_K,$$

novamente procedendo ao cálculo de cada parcela isoladamente, tem-se

- **I.1)** Determinação de $\langle \vec{b}, \tau\left((\nabla \vec{v})\vec{u} \pm \nu \triangle \vec{v}\right)\rangle_K$. Por analogia

$$\begin{aligned}
\langle \vec{b}, \tau\left((\nabla \vec{v})\vec{u} \pm \nu \triangle \vec{v}\right)\rangle_K &= \langle \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \tau \left[\mathbb{N}^{\nabla u \; u \pm \triangle u}\right] \hat{q}_k \rangle_K; \\
&= \langle \tau \left[\mathbb{N}^{\nabla u \; u \pm \triangle u}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \hat{q}_k \rangle_K; \\
&= \left[\int_K \tau \left[\mathbb{N}^{\nabla u \; u \pm \triangle u}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix} dK\right] \cdot \hat{q}_k; \\
&= \left[\mathbb{F}_k^{I.1}\right] \cdot \hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{F}_k^{I.1}\right] = \int_K \tau \left[\mathbb{N}^{\nabla u \; u \pm \triangle u}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix} dK.$$

- **I.2)** Determinação de $\langle \vec{b}, \tau\left(\frac{1}{\rho}\nabla \hat{p}\right)\rangle_K$. Por analogia

$$\begin{aligned}
\langle \vec{b}, \tau\left(\frac{1}{\rho}\nabla \hat{p}\right)\rangle_K &= \frac{1}{\rho}\langle \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \tau \left[\mathbb{N}^{\nabla p}\right] \hat{q}_k \rangle_K; \\
&= \frac{1}{\rho}\langle \tau \left[\mathbb{N}^{\nabla p}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix}, \hat{q}_k \rangle_K; \\
&= \frac{1}{\rho}\left[\int_K \tau \left[\mathbb{N}^{\nabla p}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix} dK\right] \cdot \hat{q}_k; \\
&= \left[\mathbb{F}_k^{I.2}\right] \cdot \hat{q}_k,
\end{aligned}$$

onde

$$\left[\mathbb{F}_k^{I.2}\right] = \frac{1}{\rho}\int_K \tau \left[\mathbb{N}^{\nabla p}\right]^T \begin{bmatrix} b_x \\ b_y \end{bmatrix} dK.$$

Pode-se escrever agora

$$\begin{aligned}
B(\vec{u}, p, \vec{v}, \hat{p})|_K = &\left[\mathbb{K}_k^A\right] \vec{q}_k \cdot \hat{q}_k + \left[\mathbb{K}_k^B\right] \vec{q}_k \cdot \hat{q}_k \\
&- \left[\mathbb{K}_k^C\right] \vec{q}_k \cdot \hat{q}_k - \left[\mathbb{K}_k^D\right] \vec{q}_k \cdot \hat{q}_k \\
&+ \left[\mathbb{K}_k^E\right] \vec{q}_k \cdot \hat{q}_k + \left[\mathbb{K}_k^{F.1}\right] \vec{q}_k \cdot \hat{q}_k \\
&+ \left[\mathbb{K}_k^{F.2}\right] \vec{q}_k \cdot \hat{q}_k + \left[\mathbb{K}_k^{F.3}\right] \vec{q}_k \cdot \hat{q}_k \\
&+ \left[\mathbb{K}_k^{F.4}\right] \vec{q}_k \cdot \hat{q}_k,
\end{aligned} \tag{2.36}$$

e

$$F(\vec{v}, \hat{p})|_K = \left[\mathbb{F}_k^G\right] \cdot \hat{q}_k + \left[\mathbb{F}_k^H\right] \cdot \hat{q}_k + \left[\mathbb{F}_k^{I.1}\right] \cdot \hat{q}_k + \left[\mathbb{F}_k^{I.2}\right] \cdot \hat{q}_k. \tag{2.37}$$

É importante destacar agora os termos de cada funcional associado à formulação do

problema abordado

- $B(.)|_K$:
  - termos clássicos de Galerkin: $\left[\mathbb{K}_k^A\right] \vec{q}_k \cdot \hat{q}_k$, $\left[\mathbb{K}_k^B\right] \vec{q}_k \cdot \hat{q}_k$, $\left[\mathbb{K}_k^C\right] \vec{q}_k \cdot \hat{q}_k$ e $\left[\mathbb{K}_k^D\right] \vec{q}_k \cdot \hat{q}_k$;
  - termo de captura de choque: $\left[\mathbb{K}_k^E\right] \vec{q}_k \cdot \hat{q}_k$;
  - termos de estabilização: $\left[\mathbb{K}_k^{F.1}\right] \vec{q}_k \cdot \hat{q}_k$, $\left[\mathbb{K}_k^{F.2}\right] \vec{q}_k \cdot \hat{q}_k$, $\left[\mathbb{K}_k^{F.3}\right] \vec{q}_k \cdot \hat{q}_k$ e $\left[\mathbb{K}_k^{F.4}\right] \vec{q}_k \cdot \hat{q}_k$.
- $F(.)|_K$:
  - termos clássicos de Galerkin: $\left[\mathbb{F}_k^G\right] \cdot \hat{q}_k$ e $\left[\mathbb{F}_k^H\right] \cdot \hat{q}_k$;
  - termos de estabilização para força de corpo: $\left[\mathbb{F}_k^{I.1}\right] \cdot \hat{q}_k$, $\left[\mathbb{F}_k^{I.2}\right] \cdot \hat{q}_k$.

Seguindo a recomendação de Franca e Frey (1992) nos exemplos apresentados neste trabalho (capitulo 4) é utilizada a formulação "minus" (utilizando o sinal negativo) na determinação das parcelas F e I da formulação exposta anteriormente. Segundo os resultados obtidos por Franca e Frey (1992), esta formulação é mais insensível a escolha dos parâmetros de estabilidade, em que diferentes valores de "m_k" ainda produzem soluções numéricas desejáveis para esta formulação. No caso da formulação "plus" (utilizando o sinal positivo) a determinação das parcelas F e I da formulação, se mostra mais instável para valores elevados de "τ", quando são empregadas altas ordens de interpolação para o campo de velocidades, e particularmente no caso em que se utilizam ordens iguais de interpolação para os campos de velocidade e pressão. Neste caso (formulação "plus") as oscilações numéricas espúrias apresentam-se nos resultados obtidos para ambos os campos.

Capítulo 3

MÉTODOS ITERATIVOS

3.1 Introdução

O universo dos métodos iterativos para solução de sistemas de equações é constituido por diversas idéias e técnicas no que concerne a abordagem com esta finalidade. Há metodologias para os diversos tipos de sistemas que possam ser encontrados. Os métodos em destaque no caso de sistemas de equações não lineares são o método de Newton, Newton inexato e Broyden (Dennis, Schnabel (1986)). Dentre os diversos tipos de métodos para abordagem de sistemas lineares, pode-se destacar os métodos diretos, em que as soluçoes são obtidas sem a necessidade de qualquer tipo de aproximação, a excessão da precisão da máquina, e os métodos iterativos, em que a solução aproximada do sistema é encontrada sob uma certa tolerância previamente determinada, solução esta obtida a partir de uma sequência de aproximações. Dentre os métodos diretos pode-se destacar as fatorações LU, QR, Cholesky, dentre outros métodos (Duff, Erisman, Reid (1992)). Os métodos iterativos de maior destaque são os métodos de ponto fixo (ou estacionários) como Jacobi, Gauss-Seidel, SOR, etc ((Greenbaum (1997), Saad (1992)); e os não estacionários tais como GMRES, CG (Gradiente Conjugado) e Bi-CG, dentre outros (Greenbaum (1997), Saad (1992), Saad (1996)). No caso de sistemas lineares grandes e esparsos torna-se inviável a aplicação de métodos diretos devido ao elevado custo computacional. Neste caso o ambiente torna-se muito favorável aos métodos iterativos. Há ainda a possibilidade de se melhorar o desempenho dos métodos iterativos com a introdução de précondicionadores ((Greenbaum (1997), Saad (1992)) que tem por finalidade acelerar a convergência dos mesmos.

Este capítulo se destina a apresentar as idéias básicas das metodologias adotadas no trabalho. Inicialmente os métodos de Newton e Newton inexato são enfocados. Posteriormente seguem alguns comentários sobre os métodos iterativos lineares em subespaços de Krylov, categoria na qual GMRES e Bi-CGStab se encaixam, e que posteriormente serão abordados

O trabalho aqui presente fará uma abordagem via o método de Newton inexato para

o sistema não linear de equações resultante da aplicação do método de elementos finitos sobre o sistema de equações diferenciais parciais (problema de Navier-Stokes). A equacão tangente será abordada pelos métodos GMRES e Bi-CGStab.

3.2 O Método de Newton

A aplicação do método de elementos finitos à equação de Navier-Stokes gera um sistema de equações não lineares, o qual pode ser apresentado da forma que se segue

$$\vec{F}(\vec{x}) = \vec{0}, \tag{3.1}$$

onde $\vec{F} : \mathbb{R}^n \to \mathbb{R}^n$, ou seja

$$\vec{F}(\vec{x}) = \begin{bmatrix} f_1(\vec{x}) \\ \vdots \\ f_n(\vec{x}) \end{bmatrix}, \tag{3.2}$$

em que $f_i(\vec{x}) : \mathbb{R}^n \to \mathbb{R}$, $i = 1, ..., n$.

O método de Newton consiste em expandir em série de Taylor até o termo de primeira ordem. A parcela $\vec{F}(\vec{x}_k + \vec{p}_k)$, em uma dada iteração k, isto é

$$\vec{F}(\vec{x}_k + \vec{p}_k) = \vec{F}(\vec{x}_k) + J(\vec{x}_k)\vec{p}_k, \tag{3.3}$$

onde $J(\vec{x}_k)$ designa a matriz Jacobiana de $\vec{F}$ em $\vec{x}_k$:

$$J(\vec{x}_k) = \left.\frac{\partial \vec{F}}{\partial \vec{x}}\right|_{\vec{x}=\vec{x}_k}. \tag{3.4}$$

O raciocínio agora é encontrar $\vec{p}_k$ para o qual $\vec{F}(\vec{x}_k + \vec{p}_k)$ seja nulo, ou seja

$$J(\vec{x}_k)\vec{p}_k = -\vec{F}(\vec{x}_k), \tag{3.5}$$

e pela solução de tal sistema, também conhecido como equação tangente, obtem-se a atualização

$$\vec{x}_{k+1} = \vec{x}_k + \vec{p}_k. \tag{3.6}$$

A consistência do método de Newton pode ser observada teoricamente pelos resultados que se seguem.

Definição. 3.2.1 *Sejam $m, n > 0$, $G : \mathbb{R}^n \to \mathbb{R}^{m\times n}$, $\vec{x} \in \mathbb{R}^n$, $|.|$ uma norma em $\mathbb{R}^n$, e $\|.\|$ uma norma em $\mathbb{R}^{m\times n}$. G é dito ser Lipschitz contínuo em $\vec{x}$, ou $G \in Lip_\gamma(D)$, se existe um conjunto aberto $D \subset \mathbb{R}^n$, com $\vec{x} \in D$, e uma constante γ tal que para todo*

$\vec{v} \in D$, tem-se:

$$\|G(\vec{v}) - G(\vec{x})\| \leq \gamma \left|\vec{v} - \vec{x}\right|. \tag{3.7}$$

Definição. 3.2.2 *Seja uma função contínua $f : \mathbb{R}^n \to \mathbb{R}$ é dita ser continuamente diferenciável em $\vec{x} \in \mathbb{R}^n$, se $\left(\frac{\partial f}{\partial x_i}\right)(\vec{x})$ existe e é contínua, $\forall i = 1, ..., n$.*

Teorema. 3.2.1 *Sejam $|.|$ uma norma em $\mathbb{R}^n$, $\|.\|$ uma norma em $\mathbb{R}^{n\times n}$ compatível com $|.|$, e $\vec{F} : \mathbb{R}^n \to \mathbb{R}^n$, uma função continuamente diferenciável em um conjunto aberto convexo $D \subset \mathbb{R}^n$. Suponha que $\exists\ \vec{x}^* \in \mathbb{R}^n$, e que existam os números reais $r, \beta > 0$ tais que:*

$$\begin{aligned}
&i)\ B(\vec{x}^*, r) \subset D;\\
&ii)\ \vec{F}(\vec{x}^*) = \vec{0};\\
&iii)\ \exists J^{-1}(\vec{x}^*)\ t.q.\ \|J^{-1}(\vec{x}^*)\| \leq \beta;\\
&iv)\ J \in Lip_\gamma(B(\vec{x}^*, r)).
\end{aligned}$$

Então $\exists\ \varepsilon > 0$ tal que $\forall \vec{x}_0 \in B(\vec{x}^*, \varepsilon)$, a sequência $\{\vec{x}_k\}_{k=0}^{\infty}$ gerada por:

$$\vec{x}_{k+1} = \vec{x}_k + J^{-1}(\vec{x}_k)\vec{F}(\vec{x}_k),\ k = 0, 1, ...,$$

é bem definida e converge para $\vec{x}^*$, ainda obedecendo a:

$$|\vec{x}_{k+1} - \vec{x}^*| \leq \gamma\beta \left|\vec{x}_k - \vec{x}^*\right|^2,\ k = 0, 1, ...,$$

ou seja obtem-se uma taxa de convergência quadrática.

Prova. Veja Dennis e Schnabel (1996). ■

Observação. 3.2.1 *Um outro resultado que deve ser mensionado é o teorema de Kantorovich, cuja principal diferença do teorema anterior é a não exigência da existência da solução $\vec{x}^*$, ao invés disto mostra-se que se $J(\vec{x}_0)$ é não singular, J é Lipschitz contínua em uma vizinhança de $\vec{x}_0$ e o primeiro passo de Newton é suficientemente pequeno em relação a não linearidade de $\vec{F}(.)$, então deve haver nesta vizinhaça uma única raiz $\vec{x}^*$.*

O método de Newton apesar de ser uma metodologia bastante robusta, confiável e amplamente adotada no caso de sistemas de equações não lineares, também apresenta certos incovenientes. Os principais problemas que podem ser destacados são seguintes: o primeiro diz respeito quanto a obtenção da matriz Jacobiana $J(\vec{x})$, que em alguns casos se torna impraticável o seu cálculo analítico, principalmente em problemas de grande porte. Neste caso é necessário então se dispor de técnicas numéricas como diferenças finitas, dentre outras, para o sua obtenção. Uma segunda fonte de problemas também está fortemente relacionada com a matriz Jacobiana $J(\vec{x})$, a qual pode ainda não ser suficientemente bem condicionada ou até mesmo singular em uma determinada iteração, o que acarretaria sérios problemas para a obtenção solucão da equação tangente. No caso

de mau condicionamento, dentre as varias técnicas que podem ser adotadas, destacam-se as metodologias de perturbação do modelo linear e de pré-condicionamento do sistema, esta última será discutida mais detalhadamente adiante. Há ainda metodologias que perturbam a matriz Jacobiana de modo a obter o condicionamento necessário. Esta metodologia apesar de ser bastante utilizada na prática não é recomendada caso não seja avaliada sob critérios mais rigorosos.

3.2.1 O Algoritmo

Neste tópico faz-se a apresentação da idéia principal do algoritmo do método de Newton para a solução do sistema não linear de uma forma esquemática.

Formulação do Esquema Numérico

A aplicação do método de elementos finitos à equação de Navier-Stokes geram um sistema de equações não lineares com a seguinte forma

$$[K(\vec{x})]\vec{x} = \vec{F}^{ext}(\vec{x}). \tag{3.8}$$

Definindo então o vetor resíduo $\vec{R}(\vec{x})$

$$\vec{R}(\vec{x}) = \vec{F}^{ext}(\vec{x}) - [K(\vec{x})]\vec{x}, \tag{3.9}$$

pode-se agora redefinir o problema por:

Determinar $\vec{x} \in \mathbb{R}^n$ tal que

$$\vec{R}(\vec{x}) = \vec{0}. \tag{3.10}$$

Define-se então o critério de convergência como $\left|\vec{R}(\vec{x})\right| < tol$, onde a tolerância $tol > 0$ é suficientemente pequena. Para o método de Newton o vetor das variáveis $\vec{x}$ é atualizado, como já foi mencionado anteriormente por

$$\vec{x}_{k+1} = \vec{x}_k + \vec{p}_k, \tag{3.11}$$

onde

k : iteração atual;

$\vec{x}_{k+1}$:estimativa $(k+1)$-ésima para solução de $\vec{R}(\vec{x}) = \vec{0}$;

$\vec{x}_k$: estimativa (k)-ésima para solução de $\vec{R}(\vec{x}) = \vec{0}$;

$\vec{p}_k$: correção (ou passo) para (k)-ésima estimativa de solução de $\vec{R}(\vec{x}) = \vec{0}$.

Desta forma tem-se então

$$[J_k]\vec{p}_k = -\vec{R}(\vec{x}_k), \tag{3.12}$$

onde $[J_k]$ designa a matriz Jacobiana (ou a matriz de rigidez tangente) de $\vec{R}(.)$ em $\vec{x}_k$:

$$[J_k] = \left.\frac{\partial \vec{R}}{\partial \vec{x}}\right|_{\vec{x}=\vec{x}_k}, \tag{3.13}$$

e

$$\vec{R}(\vec{x}_k) = \vec{F}^{ext}(\vec{x}_k) - [K(\vec{x}_k)]\vec{x}_k, \tag{3.14}$$

Determinação de $[J_k]$

A matriz de rigidez tangente avaliada em $\vec{x}_k$, pode ser apresentada da seguinte forma:

$$[J_k]_{ij} = \left.\frac{\partial R_i}{\partial x_j}\right|_{\vec{x}=\vec{x}_k}. \tag{3.15}$$

Com a finalidade de se obter uma aproximação para esta matriz tangente, pode-se considerar

$$[J_k] \cong [\bar{J}_k(\vec{x}_k, \vec{x}_{k-1})], \tag{3.16}$$

na qual

$$[\bar{J}_k(\vec{x}_k, \vec{x}_{k-1})]_{ij} = \left.\frac{\partial R_i(\vec{x}_k, \vec{x}_{k-1})}{\partial x_j}\right|_{\vec{x}=\vec{x}_k}, \tag{3.17}$$

onde

$$\vec{R}(\vec{x}_k, \vec{x}_{k-1}) = \vec{F}^{ext}(\vec{x}_{k-1}) - [K(\vec{x}_{k-1})]\vec{x}_k. \tag{3.18}$$

Definindo-se então $[J_k^{eff}]$ a partir do exposto anteriormente, tem-se

$$[J_k^{eff}(\vec{x}_{k-1})] = [\bar{J}_k(\vec{x}_k, \vec{x}_{k-1})] \Rightarrow [J_k^{eff}(\vec{x}_{k-1})] = -[K(\vec{x}_{k-1})], \tag{3.19}$$

e desta forma segue que

$$[J_k^{eff}(\vec{x}_{k-1})]\vec{p}_k = -\vec{R}(\vec{x}_k). \tag{3.20}$$

Com a finalidade de resolver o problema representado pelo sistema $\vec{R}(\vec{x}) = \vec{0}$ com o método de Newton, tem-se então o seguinte algorítmo para uma aproximação inicial $\vec{x}_0$ e uma tolerância *tol* previamente definidos:

Algoritmo 1

1. [Inicialize para $k = 1 : \vec{x}_{k-1}, \vec{x}_k \Leftarrow \vec{x}_0$ e $erro = 1,0$;

2. $\left[\begin{array}{l} \text{Enquanto: } erro > tol, \text{ faça:} \\ \text{2.1. Calcule } \vec{R}(\vec{x}_k) \text{ por: } \vec{R}(\vec{x}_k) = \vec{F}^{ext}(\vec{x}_k) - [K(\vec{x}_k)]\vec{x}_k; \\ \text{2.2. Calcule } [J_k^{eff}(\vec{x}_{k-1})] \text{ por: } [J_k^{eff}(\vec{x}_{k-1})] = -[K(\vec{x}_{k-1})]; \\ \text{2.3. Resolva para } \vec{p}_k \text{ o sistema: } [J_k^{eff}(\vec{x}_{k-1})]\vec{p}_k = -\vec{R}(\vec{x}_k), \text{ por um método direto;} \\ \text{2.4. Calcule } \vec{x}_{k+1} \text{ por: } \vec{x}_{k+1} = \vec{x}_k + \vec{p}_k; \\ \text{2.5. Calcule a medida do novo erro: } erro = \left|\vec{R}(\vec{x}_{k+1})\right|, \\ \text{onde } \vec{R}(\vec{x}_{k+1}) = \vec{F}^{ext}(\vec{x}_{k+1}) - [K(\vec{x}_{k+1})]\vec{x}_{k+1}; \\ \text{2.6. Se } \left\{\begin{array}{l} erro > tol : \text{atualizar } \vec{x}_{k-1} \Leftarrow \vec{x}_k \wedge \vec{x}_k \Leftarrow \vec{x}_{k+1} \text{ e voltar a (2.1);} \\ erro \leq tol : \text{ Fim Enquanto.} \end{array}\right. \end{array}\right.$

3.3 O Método de Newton Inexato

O método de Newton inexato difere do método de Newton basicamente pelo passo (2.3). O método de Newton inexato faz uso de um método iterativo linear para resolver o sistema linear, na k-ésima iteração de Newton inexato, ou seja

$$[J(\vec{x}_k)]\vec{p}_k^{\,I} = -\vec{R}(\vec{x}_k), \tag{3.21}$$

em que $\vec{p}_k^{\,I}$ é o passo de Newton inexato e $J(\vec{x}_k)$ é o Jacobiano de $\vec{R}(\vec{x}_k)$. Definindo o vetor resíduo por

$$\vec{r}(\vec{x}_k) = [J(\vec{x}_k)]\vec{p}_k^{\,I} + \vec{R}(\vec{x}_k), \tag{3.22}$$

o critério de parada do método iterativo será

$$|\vec{r}(\vec{x}_k)|_2 \leq \eta_k \left|\vec{R}(\vec{x}_k)\right|_2, \tag{3.23}$$

em que a sequência $\{\eta_k\}_k$ é conhecida como sequência de termos forçantes. A consistência do método de Newton inexato e a influência da sequência de termos forçantes na taxa de convergência pode ser teoricamente melhor observada pelos resultados que se seguem.

Definição. 3.3.1 *Seja $\{\vec{x}_k\}_{k=0}^{\infty}$ uma sequência tal que $\vec{x}_k \overset{k\to\infty}{\longrightarrow} \vec{x}^*$, então diz-se que:*

i) Se $\exists c \in [0,1)$ e um inteiro $\exists k_0 \geq 0$ tal que $|\vec{x}_{k+1} - \vec{x}^*| \leq c\,|\vec{x}_k - \vec{x}^*|$,
$\forall k > k_0$, então $\{\vec{x}_k\}_{k=0}^{\infty}$ converge para $\vec{x}^*$ linearmente;
ii) Se $\exists c \in [0,1)$ e um inteiro $\exists k_0 \geq 0$ tal que $|\vec{x}_{k+1} - \vec{x}^*| \leq c\,|\vec{x}_k - \vec{x}^*|^2$,
$\forall k > k_0$, então $\{\vec{x}_k\}_{k=0}^{\infty}$ converge para $\vec{x}^*$ quadraticamente;
iii) Se $\exists\{c_k\}_{k=0}^{\infty}$, com $c_k \overset{k\to\infty}{\longrightarrow} 0$, e um inteiro $\exists k_0 \geq 0$ tal que $|\vec{x}_{k+1} - \vec{x}^*| \leq c_k\,|\vec{x}_k - \vec{x}^*|$,
$\forall k > k_0$, então $\{\vec{x}_k\}_{k=0}^{\infty}$ converge para $\vec{x}^*$ superlinearmente.

Teorema. 3.3.1 *Seja $\vec{F} : \mathbb{R}^n \to \mathbb{R}^n$, e $\vec{x}^* \in \mathbb{R}^n$ tal que $\vec{F}(\vec{x}^*) = \vec{0}$, com $\vec{F}$ contínuamente diferenciável em uma vizinhança de $\vec{x}^*$ e $J(\vec{x}^*)$ positiva definida. Considere a iteração $\vec{x}_{k+1} = \vec{x}_k + \vec{p}_k^{\,I}$, com $\vec{x}_k \overset{k\to\infty}{\longrightarrow} \vec{x}^*$, em que $\vec{p}_k^{\,I}$ satisfaz:*

$$|\vec{r}(\vec{x}_k)|_2 \leq \eta_k \left|\vec{F}(\vec{x}_k)\right|_2. \tag{3.24}$$

Então se $\vec{x}_0$ suficientemente próximo de $\vec{x}^*$

i) Se $\eta_k \leq \eta$ onde $\eta \in [0,1) \Rightarrow \left|\vec{F}(\vec{x}_k)\right|_2 \overset{k\to\infty}{\longrightarrow} 0$ linearmente;
ii) Se $\eta_k \overset{k\to\infty}{\longrightarrow} 0 \Rightarrow \left|\vec{F}(\vec{x}_k)\right|_2 \overset{k\to\infty}{\longrightarrow} 0$ superlinearmente;
iii) Se $\eta_k \leq K \left|\vec{F}(\vec{x}_k)\right|_2$, para alguma constante $K \Rightarrow \left|\vec{F}(\vec{x}_k)\right|_2 \overset{k\to\infty}{\longrightarrow} 0$ quadraticamente.

Prova. Como $J(\vec{x}^*)$ é positiva definida, então existe $\delta > 0$ e $M > 0$ tais que $\forall\, \vec{x} \in B(\vec{x}^*, \delta)$ tem-se $\left\|[J(\vec{x}^*)]^{-1}\right\|_2 \leq M$. Além disso $J(\vec{x}_k) \overset{k\to\infty}{\longrightarrow} J(\vec{x}^*)$, logo existe um inteiro $k_0 > 0$ tal que $\left\|[J(\vec{x}_k)]^{-1}\right\|_2 < M, \forall k > k_0$, o que permite escrever:

$$\left|\vec{p}_k^{\,I}\right|_2 \leq M\left(|\vec{r}(\vec{x}_k)|_2 + \left|\vec{F}(\vec{x}_k)\right|_2\right) \leq \hat{M}\left|\vec{F}(\vec{x}_k)\right|_2.$$

Agora pelo teorema de Taylor, tem-se

$$\begin{aligned}\vec{F}(\vec{x}_{k+1}) &= \vec{F}(\vec{x}_k) + J(\vec{x}_k)\vec{p}_k^I + O(\left|\vec{p}_k^I\right|_2^2);\\ &= \vec{r}(\vec{x}_k) + O(\left|\vec{p}_k^I\right|_2^2);\\ &= \vec{r}(\vec{x}_k) + O\left(\hat{M}^2\left|\vec{F}(\vec{x}_k)\right|_2^2\right);\\ &= \vec{r}(\vec{x}_k) + O\left(\left|\vec{F}(\vec{x}_k)\right|_2^2\right).\end{aligned}$$

O implica em

$$\left|\vec{F}(\vec{x}_{k+1})\right|_2 \leq \eta_k\left|\vec{F}(\vec{x}_k)\right|_2 + O\left(\left|\vec{F}(\vec{x}_k)\right|_2^2\right).$$

Deste ponto em diante os ítens $(i), (ii)$ e (iii) são facilmente verificados. ■

Este resultado evidencia a importância da sequência de termos forçante $\{\eta_k\}$ na taxa de convergência do método de Newton inexato. Esta sequência determina a escolha da precisão ζ requerida para a resolução do sistema linear. Sendo assim no k-ésimo passo de Newton inexato a precisão requerida para a solução do sistema

$$[J(\vec{x}_k)]\vec{p}_k^I = -\vec{F}(\vec{x}_k),$$

é $\zeta = \eta_k\left|\vec{F}(\vec{x}_k)\right|_2$.

3.3.1 Métodos Iterativos em Subespaços de Krylov

Há vários tipos métodos iterativos lineares que podem ser usados na solução do sistema descriminado, por exemplo, os métodos de ponto fixo como (ou estacionários) SOR, Jacobi e Gauss-Seidel, assim como suas variações. Porém, dentre as metodologias mais bem sucedidas e robustas para se abordar este tipo de problema, encontram-se os métodos baseados em subespaços de Krylov.

Dado o sistema linear

$$A\vec{x} = \vec{b}, \tag{3.25}$$

em que $A \in \mathbb{R}^{n\times n}$ e $\vec{x}, \vec{b} \in \mathbb{R}^n$, o subespaço de Krylov m-dimensional associado a matriz $A \in \mathbb{R}^{n\times n}$ pode ser apresentado da forma como segue:

$$K_m(A, \vec{v}) = \{\vec{v}, A\vec{v}, ..., A^{m-1}\vec{v}\}, \tag{3.26}$$

para algum vetor $\vec{v} \in \mathbb{R}^n$. Os métodos iterativos baseados em subespaços de Krylov consistem em obter, dada uma aproximação inicial $\vec{x}_0 \in \mathbb{R}^n$ ($\Rightarrow \vec{r}_0 = \vec{b} - A\vec{x}_0$), numa iteração arbitrária m, a solução $\vec{x}_m \in \mathbb{R}^n$ no subespaço afim $\vec{x}_0 + K_m(A, \vec{r}_0)$ de dimensão

m, em adição com as condições de Petrov-Galerkin

$$\vec{r}_m = \vec{b} - A\vec{x}_m \perp L_m, \tag{3.27}$$

em que L_m um subespaço de dimensão m, convenientemente escolhido. O subespaço de Krylov $K_m(A, \vec{r}_0) = \{\vec{r}_0, A\vec{r}_0, ..., A^{m-1}\vec{r}_0\}$, aqui será simplemente designado por K_m.

Os vários tipos de métodos iterativos em subespaços de Krylov fazem referência as diferentes maneiras de se escolher o subespaço L_m e de que forma é feito o precondicionamento do sistema linear. De fato, as soluções são obtidas por uma aproximação polinomial da forma seguinte:

$$A^{-1}\vec{b} \simeq \vec{x}_m = \vec{x}_0 + p_{m-1}(A)\vec{r}_0, \tag{3.28}$$

em que p_{m-1} um polinômio de grau $m-1$, chamado de polinômio residual. Note que quando $\vec{x}_0 = \vec{0}$, tem-se

$$\vec{x}_m \approx p_{m-1}(A)\vec{b}, \tag{3.29}$$

ou seja, a solução $\vec{x}_m$ é definida por $p_{m-1}(A)\vec{b}$.

Uma expectativa que se torna evidente neste momento é que a escolha do subespaço L_m deverá ter uma grande influência no desempenho método iterativo. Dentre os vários tipos de escolha para L_m a literatura destaca duas classes de escolhas. A primeira que consiste em se tomar $L_m = K_m$ e para as variações de resíduo mínimo $L_m = AK_m$. A segunda classe consiste em definir L_m um subespaço de Krylov associado com a matriz A^T, $L_m = K_m(A^T, \vec{r}_0)$. As diferenças básicas entre os métodos baseiam-se no tipo de projeção utilizada para encontrar a solução aproximada $\vec{x}_m \in \vec{x}_0 + K_m$:

$$\begin{aligned}
&i)\ \vec{r}_m = \vec{b} - A\vec{x}_m \perp K_m;\\
&ii)\ \vec{x}_m = \arg\left(\min_{\vec{x}\in\vec{x}_0+K_m} \left|\vec{b} - A\vec{x}\right|_2\right);\\
&iii)\ \vec{r}_m = \vec{b} - A\vec{x}_m \text{ ortogonal a um subespaço } m\text{-dimendional adequado;}\\
&iv)\ |\ \vec{x}^* - \vec{x}_m|_2 \text{ ser mínimo sobre } A^T K_m(A^T, \vec{r}_0).
\end{aligned} \tag{3.30}$$

O ítem (i) resulta no método de ortogonalização completa (FOM) e suas variações; do ítem (ii) resulta o método dos resíduos mínimos generalizados (GMRES); no caso (iii) escolhido o suberspaço m-dimensional $K_m(A^T, \vec{r}_0)$, resultam os métodos do bi-gradiente conjugado (Bi-CG) e do resíduo quase-mínimo (QMR), ambos baseados na biortogonalização de Lanczos; já do caso (iv) resultam os métodos que não necessitam da operações produto matriz vetor por A e por A^T simultaneamente, diferentemente do caso anterior, daí tem-se o método do gradiente conjugado quadrado (CGS) e o método do bi-gradiente conjugado estabilizado (Bi-CGStab).

A idéia principal dos métodos iterativos é fazer uso de técnicas que usam aproximações sucessivas, para obter soluções cada vez mais precisas. Atualmente, os métodos mais elaborados pertencem à classe dos métodos não estacionários, fato este que constituiu a

grande motivação deste trabalho.

O presente trabalho fará a abordagem da equação tangente via GMRES e Bi-CGStab, com o objetivo de comparar o desempenho destes métodos frente a aplicações aqui abordadas. Seguindo esta linha, uma apresentação mais detalhada destes dois métodos se faz necessária, bem como as suas nuances computacionais.

Método dos Resíduos Mínimos Generalizado (GMRES)

O GMRES é um método de projeção proposto por Y. Saad e M. Schultz, em 1986. Os princípios básicos deste método podem ser apresentados considerando, novamente, sistema linear

$$A\vec{x} = \vec{b}, \tag{3.31}$$

onde $A \in \mathbb{R}^{n\times n}$ e $\vec{x}, \vec{b} \in \mathbb{R}^n$. Considerando uma aproximação inicial $\vec{x}_0 \in \mathbb{R}^n$, a cada iteração, o método busca a solução em um dado subespaço afim ($\vec{x}_0 + K_m = \vec{x}_0 + K_m(A, \vec{v}_1)$), de tal modo que o resíduo resultante seja perpendicular a um outro subespaço dado ($AK_m = AK_m(A, v_1)$), ou seja, para a m-ésima iteração, a aproximação $\vec{x}_m$ é tal que

$$\vec{x}_m \in \vec{x}_0 + K_m(A, \vec{v}_1), \tag{3.32}$$

e o resíduo,

$$\vec{r}_m = \vec{b} - A\vec{x}_m \perp AK_m(A, \vec{v}_1), \tag{3.33}$$

em que $\vec{v}_1 = \vec{r}_0 / |\vec{r}_0|_2$. Então na m-ésima iteração do GMRES, encontrar a aproximação $\vec{x}_m$ implica em resolver o problema de quadrados mínimos

$$\min_{\vec{x}\in\vec{x}_0+K_m} \left|\vec{b} - A\vec{x}\right|_2. \tag{3.34}$$

Assim seja $\vec{x} \in \vec{x}_0 + K_m$, então

$$\vec{x} = \vec{x}_0 + V_m\vec{y}, \text{ para algum } \vec{y} \in \mathbb{R}^m, \tag{3.35}$$

em que V_m é a matriz cujos vetores coluna compõem uma base ortonormal para o subespaço de Krylov K_m associado.

Nesse contexto a meneira de obter uma base ortonormal para o subespaço K_m tem uma importância fundamental. Dentre as principais metodologias com esta finalidade pode-se destacar as que se baseiam no algorítmo de Arnoldi, em que o método de Gram-Schmidt (GSM) é bastante evidenciado na literatura especializada. Os algorítmos destes procedimentos são descritos a seguir

Algorítmo 2 (Arnoldi)

1. [Escolha um vetor $\vec{v}_1$ t.q. $|\vec{v}_1|_2 = 1$;
2. [
   - 2.1. Para $j = 1, 2, \ldots, m$;
     - 2.1.1. Para $i = 1, 2, \ldots, j$;
       - 2.1.1.1. $h_{ij} =< A\vec{v}_i, \vec{v}_j >$;
       - 2.1.1.2. $\vec{w}_j = A\vec{v}_j - \sum_{i=1}^{j} h_{ij}\vec{v}_i$;
       - 2.1.1.3. $h_{j+1,j} = |\vec{w}_j|_2$, se $h_{j+1,j} = 0 \Rightarrow$ Pare !!!;
       - 2.1.1.4. $\vec{v}_{j+1} = \vec{w}_j / h_{j+1,j}$;
     - 2.1.2. Fim para;
   - 2.2. Fim para.

Algorítmo 3 (GSM)

1. [Escolha um vetor $\vec{v}_1$ t.q. $|\vec{v}_1|_2 = 1$;
2. [
   - 2.1. Para $j = 1, 2, \ldots, m$;
   - 2.2. $\vec{w}_j := A\vec{v}_j$;
     - 2.2.1. Para $i = 1, 2, \ldots, j$;
       - 2.2.1.1. $h_{ij} =< \vec{w}_j, \vec{v}_i >$;
       - 2.2.1.2. $\vec{w}_j = \vec{w}_j - h_{ij}\vec{v}_i$;
     - 2.1.2. Fim para;
   - 2.3. $h_{j+1,j} = |\vec{w}_j|_2$, se $h_{j+1,j} = 0 \Rightarrow$ Pare !!!;
   - 2.4. $\vec{v}_{j+1} = \vec{w}_j / h_{j+1,j}$;
   - 2.5. Fim para.

O algorítmo de Arnoldi, a cada passo, multiplica os vetores $\vec{v}_j$ anteriores por A e, então, ortogonaliza o vetor resultante $\vec{w}_j$ com relação todos os $\vec{v}_j$´s anteriores por um procedimento Gram-Schmidt padrão. Na prática pode-se melhorar numericamente o processo usando o processo Gram-Schmidt modificado (GSM), ao invés do procedimento Gram-Schmidt padrão. Vale ressaltar que há métodos mais robustos como os que utilizam técnicas de

reortogonalização e o método de de Householder-Arnoldi (Saad (1996)), porém estes são computacionalmente mais caros.

Retornando, agora, ao problema anteriormente apresentado, define-se:

$$J(\vec{y}) := \left|\vec{b} - A\vec{x}\right|_2 = \left|\vec{b} - A(\vec{x}_0 + V_m\vec{y})\right|_2, \tag{3.36}$$

note ainda que, como V_m é uma matriz ortonormal, pode-se escrever

$$AV_m = V_m H_m + h_{m+1,m}\ \vec{v}_{m+1}\vec{e}_m^T; \tag{3.37}$$

$$= V_{m+1}\bar{H}_m, \tag{3.38}$$

em que H_m é uma matriz de Hessenberg superior $m \times m$, ou seja, $h_{i,j} = 0, \forall(i,j)$ t.q. $i > j+1$. A matriz $\bar{H}_m$ é de ordem $(m+1) \times m$, cujo bloco superior $m \times m$ corresponde a matriz H_m com a última linha composta por zeros, a excessão do elemento $(m+1, m)$, $h_{m+1,m}$. V_{m+1} é uma matriz $n \times (m+1)$, cujo bloco esquerdo $n \times m$ corresponde a matriz V_{m+1} e a última coluna corresponde ao vetor $\vec{v}_{m+1}$, e $\vec{e}_m$ é o m-ésimo vetor da base canônica em $\mathbb{R}^m$. Assim:

$$\bar{H}_m = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & \cdots & h_{1,m} \\ h_{2,1} & h_{2,2} & \cdots & \cdots & h_{2,m} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & h_{m,m-1} & h_{m,m} \\ 0 & \cdots & \cdots & 0 & h_{m+1,m} \end{bmatrix} \in \mathbb{R}^{(m+1)\times m}; \tag{3.39}$$

$$V_{m+1} = \begin{bmatrix} [V_m]_{n\times m} & \vec{v}_{m+1} \end{bmatrix} \in \mathbb{R}^{n\times(m+1)}; \tag{3.40}$$

$$\vec{e}_m^T = (0, 0, ..., 1)^T \in \mathbb{R}^m. \tag{3.41}$$

Desenvolvendo a expressão para o resíduo tem-se

$$\begin{aligned} \vec{b} - A\vec{x} &= \vec{r}_0 - AV_m\vec{y}; \\ &= \rho_0\vec{v}_1 - V_{m+1}\bar{H}_m\vec{y}; \\ &= V_{m+1}(\rho_0\vec{e}_1 - \bar{H}_m\vec{y}), \end{aligned} \tag{3.42}$$

em que $\rho_0 = |\vec{r}_0|_2$. Pode-se escrever

$$J(\vec{y}) = \left|V_{m+1}(\rho_0\vec{e}_1 - \bar{H}_m\vec{y})\right|_2, \tag{3.43}$$

mas como a matriz V_{m+1} é ortonormal tem-se

$$J(\vec{y}) = \left|\rho_0\vec{e}_1 - \bar{H}_m\vec{y}\right|_2. \tag{3.44}$$

A aproximação GMRES busca, a cada passo, o único vetor de $\vec{x}_0 + K_m$ que minimiza $J(\vec{y})$. Esta aproximação é obtida por

$$\vec{x}_m = \vec{x}_0 + V_m \vec{y}_m, \tag{3.45}$$

em que

$$\vec{y}_m = \arg\left(\min_{\vec{y}\in\mathbb{R}^m} \left| \rho_0 \vec{e}_1 - \bar{H}_m \vec{y} \right|_2 \right). \tag{3.46}$$

O algorítmo GMRES pode ser então apresentado como se segue:

Algorítmo 4 (GMRES)

1. 1.1. Defina a tolerância tol e $\vec{x}_0$;
   1.2. Calcule $\vec{r}_0 = \vec{b} - A\vec{x}_0$, $\rho_0 = |\vec{r}_0|_2$ e $\vec{v}_1 = \vec{r}_0/\rho_0$;
2. 2.1. Para $m = 1, 2, \ldots$;
   - 2.1.1. Calcule $V_m, H_m, \vec{v}_{m+1}$ e $h_{m+1,m}$ pelo algorítmo 2 ou 3;
   - 2.1.2. Resolva o problema de mínimos quadrados:
     $$\vec{y}_m = \arg\left(\min_{\vec{y}\in\mathbb{R}^m} \left| \rho_0 \vec{e}_1 - \bar{H}_m \vec{y} \right|_2 \right);$$
   - 2.1.3. Calcule $\vec{x}_m = \vec{x}_0 + V_m \vec{y}_m$, para $m > 1 \Rightarrow res = \left| \vec{b} - A\vec{x}_m \right|_2$;
   - 2.1.4. Se $\begin{cases} res > tol: & \text{voltar a (2.1.);} \\ res \leq tol: & \text{Fim para.} \end{cases}$

   2.2. Fim para.

No algoritmo acima (GMRES), vale ainda ressaltar que neste trabalho o passo 2.1.1. é efetuado com o uso do algorítmo 3, ou seja, com o processo de ortogonalização de Gram-Schmidt modificado (GSM). Uma alternativa seria usar o processo de ortogonalização de Householder que é numericamente mais robusto, porém tem um custo computacional mais elevado que o GSM, como já havia sido mencionado.

O Problema de Mínimos Quadrados. Observando o algoritmo 4 acima, percebe-se que os passos 2.1.2. e 2.1.3. estão intimamente ligados, com a finalidade de se obter a solução aproximada $\vec{x}_m$ explicitamente a cada passo. Uma maneira computacionalmente eficiente de se obter o resíduo está relacionada com a maneira pela qual o problema de mínimos quadrados (PMQ) é resolvido.

Uma forma de se resolver o PMQ:

$$\min_{\vec{y}\in\mathbb{R}^m} \left|\rho_0\vec{e}_1 - \bar{H}_m\vec{y}\right|_2, \tag{3.47}$$

para o m-ésimo passo de GMRES, é efetuar a transformação da matriz Hessenberg $\bar{H}_m$ em uma matriz triangular superior, com o uso de matrizes de rotação ($\Omega_i \in \mathbb{R}^{(m+1)\times(m+1)}$), as quais podem ser apresentadas por:

$$\Omega_i = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c_i & s_i & & & \\ & & -s_i & c_i & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix} \begin{matrix} \\ \\ \leftarrow\ i\text{-ésima linha} \\ \leftarrow\ (i+1)\text{-ésima linha} \\ \\ \\ \\ \end{matrix} \tag{3.48}$$

em que c_i e s_i representam respectivamente o cosseno e o seno de um dado ângulo θ_i. Objetivando-se transformar $\bar{H}_m$ em uma matriz $\bar{H}_m^{(i)}$ triangular superior, basta escolher:

$$s_i = \frac{h_{i+1,i}}{\sqrt{(h_{ii}^{(i-1)})^2 + h_{i+1,i}^2}} \quad \text{e} \quad c_i = \frac{h_{ii}^{(i-1)}}{\sqrt{(h_{ii}^{(i-1)})^2 + h_{i+1,i}^2}}, \tag{3.49}$$

em que $h_{ii}^{(i-1)}$ é o elemento (i,i) da matriz $\bar{H}_m^{(i)}$, com $h_{11}^{(0)} = h_{11}$. Definindo agora a matriz $Q_m \in \mathbb{R}^{(m+1)\times(m+1)}$ como o seguinte produtório:

$$Q_m := \prod_{k=0}^{m-1} \Omega_{(m-k)}, \tag{3.50}$$

e

$$\bar{R}_m := \bar{H}_m^{(i)} = Q_m\bar{H}_m; \tag{3.51}$$

$$\overrightarrow{\bar{g}}_m := Q_m(\rho_0\vec{e}_1) = (\gamma_1, \ldots, \gamma_{m+1}), \tag{3.52}$$

em que o vetor da base canônica $\vec{e}_1 \in \mathbb{R}^{m+1}$. Procedendo-se então a exclusão das $(m+1)$-ésimas linhas de $\bar{R}_m$ e $\overrightarrow{\bar{g}}_m$, obtém-se uma matriz quadrada triangular superior, a qual será designada por $R_m \in \mathbb{R}^{m\times m}$ e um vetor $\vec{g}_m \in \mathbb{R}^m$, respectivamente. Como Q_m é uma matriz ortogonal, pode-se escrever,

$$\min_{\vec{y}\in\mathbb{R}^m} \left|\rho_0\vec{e}_1 - \bar{H}_m\vec{y}\right|_2 = \min_{\vec{y}\in\mathbb{R}^m} \left|\overrightarrow{\bar{g}}_m - \bar{R}_m\vec{y}\right|_2, \tag{3.53}$$

daí tem-se a seguinte proposição:

Proposição. 3.3.2 *Sejam Ω_i, $i = 1, ..., m$, as matrizes de rotação usadas para transformar $\bar{H}_m$ em uma matriz triangular superior $\bar{R}_m$, e sejam Q_m, R_m, $\overrightarrow{\bar{g}}_m$ e $\vec{g}_m$ como designadas anteriormente. Então se R_m é não singular, tem-se*

i) $\vec{y}_m = R_m^{-1}\vec{g}_m = \arg\left(\min_{\vec{y}\in\mathbb{R}^m} \left|\rho_0\vec{e}_1 - \bar{H}_m\vec{y}\right|_2\right)$;

ii) $\vec{r}_m = \vec{b} - A\vec{x}_m = \gamma_{m+1}V_{m+1}Q_{m+1}\vec{e}_{m+1}$, *e consequentemente* $|\vec{r}_m|_2 = |\gamma_{m+1}|$.

Prova. Tem-se que:

$$\begin{aligned}\left|\rho_0 e_1 - \bar{H}_m\vec{y}\right|_2^2 &= \left|Q_m(\overrightarrow{\bar{g}}_m - \bar{R}_m\vec{y})\right|_2^2;\\ &= \left|\overrightarrow{\bar{g}}_m - \bar{R}_m\vec{y}\right|_2^2;\\ &= |\gamma_{m+1}|^2 + |\vec{g}_m - R_m\vec{y}|_2^2.\end{aligned}$$

O mínimo é atingido quando

$$|\vec{g}_m - R_m\vec{y}|_2^2 = 0 \Rightarrow \vec{y}_m = R_m^{-1}\vec{g}_m,$$

o que prova o ítem (i). Agora para o ítem (ii) pode-se escrever

$$\begin{aligned}\vec{b} - A\vec{x}_m &= \vec{b} - A(\vec{x}_0 + V_m\vec{y}_m);\\ &= \vec{r}_0 - V_{m+1}\bar{H}_m\vec{y}_m;\\ &= V_{m+1}(\rho_0\vec{e}_1 - \bar{H}_m\vec{y}_m);\\ &= V_{m+1}(\rho_0\vec{e}_1 - Q_m^T\bar{R}_m\vec{y}_m);\\ &= V_{m+1}Q_m^T(\overrightarrow{\bar{g}}_m - \bar{R}_m\vec{y}_m).\end{aligned}$$

Daí como $\vec{y}_m$ é a solução do problema de mínimos quadrados, ou seja:

$$\vec{y}_m = R_m^{-1}g_m,$$

então

$$\begin{aligned}\vec{b} - A\vec{x}_m &= V_{m+1}Q_m^T[(\vec{g}_m, \gamma_{m+1}) - (\vec{g}_m, 0)];\\ &= V_{m+1}Q_m^T[(\vec{0}, \gamma_{m+1})];\\ &= V_{m+1}Q_m^T(\gamma_{m+1}\vec{e}_{m+1});\\ &= \gamma_{m+1}V_{m+1}Q_m^T\vec{e}_{m+1}.\end{aligned}$$

Então note que $V_{m+1}Q_m^T$ é uma matriz unitária, logo tem-se que $|\vec{r}_m|_2 = |\gamma_{m+1}|$. ■

Ainda

$$\gamma_{j+1} = -s_j\gamma_j, \tag{3.54}$$

e desta maneira tem-se

$$\left|\vec{b} - A\vec{x}_m\right|_2 = \rho_0|s_1 s_2 \dots s_m|. \tag{3.55}$$

Então se $s_j = 0$, a solução exata é atingida no passo j.

Percebe-se agora que, de acordo com a proposição anterior, pode-se em cada m-ésimo passo de GMRES, obter o resíduo do sistema de uma forma bem eficiente, o que torna possível interromper o processo num "loop" intermediário, caso o critério de parada seja satisfeito.

Observação. 3.3.1 *Pode-se ainda ocorrer situações onde o vetor $\vec{v}_{j+1} = \vec{0}$ no algorítmo 2 ou 3, o que implica em $h_{j+1,j} = 0$, para um certo passo j. Esta situação é denominada de " break down ", e neste caso o algorítmo para, pois o próximo vetor da base não pode ser calculado. Mas como*

$$s_j = \frac{h_{j+1,j}}{\sqrt{(h_{jj}^{(j-1)})^2 + h_{j+1,j}^2}}, \tag{3.56}$$

resulta que a solução exata foi obtida.

3.3.2 GMRES com Reinício

O GMRES torna-se impraticável quando a dimensão do subespaço K_m cresce demasiadamente, e consequentemente o custo computacional. A reinicialização do processo de ortogonalização é uma maneira eficaz de se contornar este incoveniente. Consiste em reiniciar o algoritmo após um certo número de iterações. O algorítmo que se segue incorpora a reinicialização.

Algorítmo 5 (RGMRES)

1. 1.1. Defina a tolerância tol e $\vec{x}_0$;
   1.2. Calcule $\vec{r}_0 = \vec{b} - A\vec{x}_0$, $\rho_0 = |\vec{r}_0|_2$ e $\vec{v}_1 = \vec{r}_0/\rho_0$;
   1.3. Defina $M \in \mathbb{N}$;
2. 2.1. Para $m = 1, 2, \dots, M$;
   - 2.1.1. Calcule $V_m, H_m, \vec{v}_{m+1}$ e $h_{m+1,m}$ pelo algorítmo 2 ou 3;
   - 2.1.2. Resolva o problema de mínimos quadrados:
     $$\vec{y}_m = \arg\left(\min_{\vec{y}\in\mathbb{R}^m} \left|\rho_0\vec{e}_1 - \bar{H}_m\vec{y}\right|_2\right);$$
   - 2.1.3. Calcule $\vec{x}_m = \vec{x}_0 + V_m\vec{y}_m$, para $m > 1 \Rightarrow res = \left|\vec{b} - A\vec{x}_m\right|_2$;
   - 2.1.4. Se
     - $(res > tol) \wedge (m = M)$: faça $\vec{x}_0 = \vec{x}_m$ e volte a (1.2.);
     - $(res > tol) \wedge (m < M)$: voltar a (2.1.);
     - $res \leq tol$: Fim para.

   2.2. Fim para.

Uma observação muito importante que ainda deve ser feita é a respeito de problemas com a convergência de que o GMRES com reinício pode enfrentar quando a matriz não é positiva definida. Este problema é denominado de "estagnação". Neste caso, um precondicionador deve ser implementado, o que será discutido mais adiante.

3.4 Análise de convergência do GMRES

O tópico aqui apresentado destina-se a analisar a convergência do método GMRES, análise esta na qual os polinômios de Chebyshev desempenham um um papel fundamental. Os polinômios de Chebyshev, além de serem usados no estudo da convergência de alguns métodos iterativos, ainda podem ser adotados com a finalidade de acelerar as iterações e consequentemente o processo como um todo.

3.4.1 Polinômios de Chebyshev

Caso real:

Os polinômios de Chebyshev reais (variável e coeficientes) de primeiro tipo de grau k são da forma

$$C_k(t) := cos(kcos^{-1}(t)),\ t \in [-1, 1], \tag{3.57}$$

fazendo uso da relação seguinte

$$cos[(k+1)\theta] + cos[(k-1)\theta] = 2cos\theta cosk\theta, \tag{3.58}$$

e do princípio de indução finita sobre k, chega-se a seguinte relação de recorrência

$$C_{k+1}(t) = 2tC_k(t) - C_{k-1}(t), \tag{3.59}$$

em que $C_0 = 1$ e $C_1 = t$. Um fato importante que deve ser comentado é que pode-se estender a definição anterior dos polinômios de Chebyshev para os casos em que $|t| \geq 1$. Tem-se então

$$C_k(t) := cosh(kcosh^{-1}(t)),\ |t| \geq 1, \tag{3.60}$$

donde deriva-se a expressão,

$$C_k(t) = \frac{1}{2}[(t + \sqrt{t^2 - 1})^k + (t + \sqrt{t^2 - 1})^{-k}], |t| \geq 1. \tag{3.61}$$

A última expressão também pode ser estendida para o caso $|t| < 1$. Observe ainda que quando k é grande o segundo termo:

$$(t + \sqrt{t^2 - 1})^{-k}, \tag{3.62}$$

torna-se pequeno, daí obtem-se a aproximação,

$$C_k(t) \simeq \frac{1}{2}(t + \sqrt{t^2 - 1})^k, \ |t| \geq 1. \tag{3.63}$$

Caso complexo

O caso anterior fornece,

$$C_k(t) := cosh(kcosh^{-1}(t)), \ |t| \geq 1, \tag{3.64}$$

dessa definição, pode-se escrever para o caso complexo (variável complexa e coeficientes reais):

$$C_k(z) = cosh(k\xi), \text{ onde } cosh(\xi) = z. \tag{3.65}$$

Definindo agora a variável $w \equiv e^{\xi}$, a fórmula acima pode ser escrita como

$$C_k(z) = \frac{1}{2}[w^k + w^{-k}], \text{ onde } z = \frac{1}{2}[w + w^{-1}]. \tag{3.66}$$

A definição acima é a dada aos polinômios de Chebyshev em $\mathbb{C}$, daí pode-se obter a seguinte relação de recorrência,

$$\begin{array}{l} C_{k+1}(z) = 2zC_k(z) - C_{k-1}(z); \\ C_0(z) = 1 \text{ e } C_1(z) = z. \end{array} \tag{3.67}$$

Observação. 3.4.1 *Os polinômios de Chebyshev estão intimamente ligados a elipses no plano complexo. Seja agora C_ρ o círculo centrado na origem de raio ρ e definindo-se o mapeamento $J : C_\rho \longrightarrow \mathbb{C}$ por*

$$J(w) = \frac{1}{2}[w + w^{-1}], \tag{3.68}$$

chamada aplicação de Joukowski, que transforma o círculo C_ρ em uma elipse centrada na origem com focos em -1 e 1, e semi-eixos principais $\frac{1}{2}[\rho + \rho^{-1}]$, $\frac{1}{2}[\rho - \rho^{-1}]$.

Os polinômios de Chebyshev são assintoticamente ótimos, sendo ótimos em apenas alguns casos. Com a finalidade de se verificar esse fato tem-se o Lema de a Zarantonello.

Lema. 3.4.1 *(Zarantonello): Sejam P_k o conjunto de todos polinômios de grau k, $C(0, \rho)$ um círculo centrado na origem com raio ρ, e seja $\gamma \in \mathbb{C} \setminus \{C(0, \rho) \cup \ int(C(0, \rho))\}$. Então:*

$$\min_{p \in P_k, p(\gamma)=1} \ \max_{z \in C(0,\rho)} |p(z)| = \left(\frac{\rho}{|\gamma|}\right)^k, \tag{3.69}$$

em que o mínimo é alcançado pelo polinômio

$$p(z) = (z/\gamma)^k. \tag{3.70}$$

Prova. Veja Rivlin. ■

Observação. 3.4.2 *O resultado anterior pode ser estendido para qualquer círculo centrado em c e raio ρ e para qualquer γ tal que $\gamma > \rho$, para tanto basta apenas efetuar uma mudança de variáveis no polinômio.*

Observação. 3.4.3 *Vale ressaltar ainda que existe um lema análogo ao anterior para o caso real.*

Retornando agora ao caso de uma elipse centrada na origem, com focos em 1, -1 e semi-eixo principal a; esta pode ser considerada como a imagem $J(C(0,\rho))$, em que $C(0,\rho)$ é um cículo de raio ρ e centro na origem.

Proposição. 3.4.1 *Seja $E_\rho = J(C(0,\rho))$ a elipse como acima, e seja $\gamma \in \mathbb{C}$ como no Lema de Zarantonello. Então:*

$$\frac{\rho^k}{|w_\gamma|^k} \leq \min_{p\in P_k, p(\gamma)=1} \max_{z\in E_\rho} |p(z)| \leq \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|}, \tag{3.71}$$

onde w_γ é a raiz dominante da equação $J(w) = \gamma$.

Prova. $\forall p \in P_k$, com $p(\gamma) = 1$, pode ser escrito da forma que segue

$$p(z) = \frac{< \vec{\alpha}, \vec{v}(z) >}{< \vec{\alpha}, \vec{v}(\gamma) >},$$

em que

$$\vec{\alpha} = (\alpha_1, ..., \alpha_k)^T;$$
$$\vec{v}(z) = (z, ..., z^k)^T.$$

Um ponto $z \in E_\rho$, pode ser observado como a transformação J de um dado ponto $w \in C(0,\rho)$, daí pode-se escrever

$$p(z) = \frac{\sum_{j=0}^k \alpha_j z^j}{\sum_{j=0}^k \alpha_j \gamma^j} = \frac{\sum_{j=0}^k \alpha_j (w^j + w^{-j})}{\sum_{j=0}^k \alpha_j (w_\gamma^j + w_\gamma^{-j})}.$$

Considerando o polinômio de Chebyshev particular de primeiro tipo e de grau k

$$p^*(z) = \frac{(w^k + w^{-k})}{(w_\gamma^k + w_\gamma^{-k})},$$

segue que

$$\max_{z\in E_\rho} |p^*(z)| = \frac{(\rho^k + \rho^{-k})}{(w_\gamma^k + w_\gamma^{-k})} \Rightarrow \min_{p\in P_k, p(\gamma)=1} \max_{z\in E_\rho} |p(z)| \leq \frac{\rho^k + \rho^{-k}}{|w_\gamma^k + w_\gamma^{-k}|}.$$

Agora tem-se

$$p(z) = \left(\frac{w^{-k}}{w_\gamma^{-k}}\right) \frac{\sum_{j=0}^{k} \alpha_j (w^{k+j} + w^{k-j})}{\sum_{j=0}^{k} \alpha_j (w_\gamma^{k+j} + w_\gamma^{k-j})} \Rightarrow |p(z)| = \frac{\rho^{-k}}{|w_\gamma|^{-k}} \left| \frac{\sum_{j=0}^{k} \alpha_j (w^{k+j} + w^{k-j})}{\sum_{j=0}^{k} \alpha_j (w_\gamma^{k+j} + w_\gamma^{k-j})} \right|,$$

segue então do Lema de Zarantonello que

$$\max_{z \in E_\rho} |p(z)| \geq \frac{\rho^{-k}}{|w_\gamma|^{-k}} \frac{\rho^{2k}}{|w_\gamma|^{2k}} = \frac{\rho^k}{|w_\gamma|^k}, \forall p \in P_k, \text{ com } p(\gamma) = 1,$$

o que permite concluir que

$$\min_{p \in P_k, p(\gamma)=1} \max_{z \in E_\rho} |p(z)| \geq \frac{\rho^k}{|w_\gamma|^k}.$$

■

Observação. 3.4.4 *Vale ressaltar ainda que quando $k \to \infty$ a diferença entre os limitantes direito e esquerdo, da proposição anterior, tende a zero. Logo tem-se que para k grande, o polinômio de Chebyshev:*

$$p^*(z) = \frac{w^k + w^{-k}}{w_\gamma^k + w_\gamma^{-k}}, \text{ onde } z = \frac{w + w^{-1}}{2}, \tag{3.72}$$

está próximo do polinômio ótimo, ou seja, os polinômios de Chebyshev são assintoticamente ótimos.

Observação. 3.4.5 *Pode-se ainda estender o resultado da proposição anterior para uma dada elipse $E(c, d, a)$, centrada em c, com uma distância focal d e semi-eixo principal a. Procedendo então uma simples mudança de variável, mostra-se que o polinômio de Chebyshev mais próximo do ótimo é dado por:*

$$\widehat{C}_k(z) = \frac{C_k(\frac{c-z}{d})}{C_k(\frac{c-\gamma}{d})}, \tag{3.73}$$

observando a expressão $(w^k + w^{-k})/2$,com $w = \rho e^{i\theta}$, verifica-se que o máximo de $|\widehat{C}_k(z)|$ sobre a elipse, é alcançado no ponto $c + a$ do eixo real. Donde tem-se,

$$\max_{z \in E(c,d,a)} |\widehat{C}_k(z)| = \frac{C_k(\frac{a}{d})}{C_k(\frac{c-\gamma}{d})}. \tag{3.74}$$

Focando agora a atenção na convergência do algoritmo GMRES, um resultado de grande importância vem do conjunto de teoremas e proposição que se seguem:

Teorema. 3.4.2 *Seja $A \in \mathbb{R}^{n \times n}$ uma matriz não singular, e seja $\vec{x}_k \in \mathbb{R}^n$ a k-ésima*

iteração de GMRES. Então $\forall \bar{p}_k \in P_k^* = \{p \in P_k \mid p(0) = 1\}$*, tem-se,*

$$|\vec{r}_k|_2 = \min_{\bar{p} \in P_k^*} |\bar{p}(A)\vec{r}_0|_2 \leq |\bar{p}_k(A)\vec{r}_0|_2 \,. \tag{3.75}$$

Prova. Note que

$$\begin{aligned}
\vec{r} &= \vec{b} - A\vec{x}; \\
&= \vec{b} - A(\vec{x}_0 + p(A)\vec{r}_0), \text{onde } p \in P_{k-1}; \\
&= (\vec{b} - A\vec{x}_0) - Ap(A)\vec{r}_0; \\
&= \vec{r}_0 - Ap(A)\vec{r}_0; \\
&= \bar{p}(A)\vec{r}_0, \text{onde } \bar{p} \in P_k^*,
\end{aligned}$$

assim tem-se para k-ésima iteração de GMRES

$$\begin{aligned}
|\vec{r}_k|_2 &= \min_{\vec{x} \in \vec{x}_0 + K_k} \left|\vec{b} - A\vec{x}\right|_2 ; \\
&= \min_{p \in P_{k-1}} \left|(\vec{b} - A\vec{x}_0) - Ap(A)\vec{r}_0\right|_2 ; \\
&= \min_{p \in P_{k-1}} |\vec{r}_0 - Ap(A)\vec{r}_0|_2 \,; \\
&= \min_{\bar{p} \in P_k^*} |\bar{p}(A)\vec{r}_0|_2 \,; \\
&\leq |\bar{p}_k(A)\vec{r}_0|_2 \,, \forall \bar{p}_k \in P_k^*.
\end{aligned}$$

■

Corolário. 3.4.1 *Seja* $A \in \mathbb{R}^{n \times n}$ *uma matriz não singular, e seja* $\vec{x}_k \in \mathbb{R}^n$ *a* k*-ésima iteração de GMRES. Então* $\forall \bar{p}_k \in P_k^*$*, tem-se*

$$\frac{|\vec{r}_k|_2}{|\vec{r}_0|_2} \leq \|\bar{p}_k(A)\|_2 \,, \; com \;\; |\vec{r}_0|_2 \neq 0. \tag{3.76}$$

Prova. Do resultado do teorema anterior tem-se

$$\begin{aligned}
|\vec{r}_k|_2 &\leq |\bar{p}_k(A)\vec{r}_0|_2 \,, \forall \bar{p}_k \in P_k^*; \\
&\leq \|\bar{p}_k(A)\|_2 \, |\vec{r}_0|_2 \,,
\end{aligned}$$

donde conclui-se que:

$$\frac{|\vec{r}_k|_2}{|\vec{r}_0|_2} \leq \|\bar{p}_k(A)\|_2 \,.$$

■

Teorema. 3.4.3 *Seja* $A \in \mathbb{R}^{n \times n}$ *uma matriz não singular, então o algorítmo GMRES, encontrará a solução em* n *iterações.*

Prova. Considere agora o polinômio característico da matriz A:

$$p(x) = \det(A - xI),$$

note que $p(0) = \det(A) \neq 0$, pois A é não singular. Como o polinômio $p \in P_n$, define-se agora:

$$\bar{p}_n(x) = \frac{p(x)}{p(0)},$$

daí $\bar{p}_n \in P_n^*$, e ainda, $\bar{p}_n(A) = p(A) = 0$. Então do colorário anterior pode-se escrever,

$$|\vec{r}_n|_2 = 0 \Rightarrow \vec{b} - A\vec{x}_n = \vec{0}.$$

■

Os resultados acima comprovam a consistência da metodologia GMRES. Analogamente tem-se a seguinte proposição para o caso de GMRES(m), isto é, com reinício após m iterações,

Proposição. 3.4.4 *Seja $A \in \mathbb{R}^{n\times n}$ uma matriz positiva definida, então o algoritmo GMRES(m) converge $\forall m \geq 1$.*

Prova. Veja Saad (1996). ■

Seguem ainda dois resultados de bastante interessantes, no que diz respeito ao conhecimento de um limitante superior para a taxa de convergência do GMRES:

Proposição. 3.4.5 *Suponha que $A \in \mathbb{R}^{n\times n}$ é uma matriz diagonalizável; assim seja $A = X\Lambda X^{-1}$ em que $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ é a matriz diagonal de autovalores. Definindo:*

$$\epsilon^{(m)} = \min_{p\in P_m, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)|, \tag{3.77}$$

então, a norma do resíduo do m-ésimo passo do GMRES satisfaz,

$$|\vec{r}_m|_2 \leq k_2(X)\epsilon^{(m)} |\vec{r}_0|_2 . \tag{3.78}$$

onde $k_2(X) = \|X\|_2\|X^{-1}\|_2$.

Prova. Tem-se então por um teorema anterior que:

$$\begin{aligned}
|\vec{r}_m|_2 &= \min_{\vec{x}\in\vec{x}_0+K_m} \left|\vec{b} - A\vec{x}\right|_2 ; \\
&= \min_{\bar{p}\in P_m^*} |\bar{p}(A)\vec{r}_0|_2 ; \\
&= \min_{\bar{p}\in P_m^*} \left|X\bar{p}(\Lambda)X^{-1}\vec{r}_0\right|_2 ; \\
&\leq \min_{\bar{p}\in P_m^*} [\|X\|_2 \|\bar{p}(\Lambda)\|_2 \left\|X^{-1}\right\|_2 |\vec{r}_0|_2]; \\
&= k_2(X)[\min_{\bar{p}\in P_m^*} \|\bar{p}(\Lambda)\|_2] \, |\vec{r}_0|_2 .
\end{aligned}$$

Notando agora que

$$\|\bar{p}(\Lambda)\|_2 = \max_{\lambda\in\sigma(A)} |\bar{p}(\lambda)|_2 \text{, onde } \sigma(A) \text{ é o espectro da matriz } A,$$

o que permite escrever para m-ésima iteração de GMRES:

$$\begin{aligned}
|\vec{r}_m|_2 &\leq k_2(X)[\min_{\bar{p}\in P_m^*} \max_{\lambda\in\sigma(A)} |\bar{p}(\lambda)|_2] \, |\vec{r}_0|_2 ; \\
&= k_2(X)\epsilon^{(m)} \, |\vec{r}_0|_2 .
\end{aligned}$$

■

Um importante resultado seria obter um majorante para $\epsilon^{(m)}$. Fazendo uso dos resultados obtidos anteriormente, e supondo que o espectro da matriz A está contido na elipse $E(c, d, a)$ com centro c, distância focal d e semi-eixo principal a e que a origem está fora desta elipse. Então o corolário que se segue fornece um majorante para o resíduo do GMRES.

Corolário. 3.4.2 *Seja $A \in \mathbb{R}^{n\times n}$ uma matriz diagonalizável, isto é $A = X\Lambda X^{-1}$, em que $\Lambda = diag\{\lambda_1, \ldots, \lambda_n\}$ é a matriz diagonal de autovalores. Suponha que todos os autovalores de A estão contido em $E(c, d, a)$, na qual a origem não está contida. Então, a norma residual formada no m-ésimo passo do GMRES satisfaz a desigualdade*

$$\|\vec{r}_m\|_2 \leq k_2(X)\frac{C_k(\frac{a}{d})}{|C_k(\frac{c}{d})|}\|\vec{r}_0\|_2. \tag{3.79}$$

Prova. Sabe-se que para $\{\lambda_i\}_{i=1\ldots n} \in \mathbb{C}$, tem-se

$$\begin{aligned}
\epsilon^{(m)} &= \min_{p\in\mathbb{P}_m, p(0)=1} \max_{i=1,\ldots,n} |p(\lambda_i)|; \\
&\leq \min_{p\in\mathbb{P}_m, p(0)=1} \max_{z\in E(c,d,a)} |p(z)|,
\end{aligned}$$

por uma observação anterior tem-se que para $\gamma \notin E(c,d,a)$, com $\gamma \in \mathbb{C}$:

$$\max_{z \in E(c,d,a)} |\widehat{C}_k(z)| = \frac{C_k(\frac{a}{d})}{C_k(\frac{c-\gamma}{d})},$$

onde $\gamma = 0$, e fazendo-se uso do polinômio acima $\widehat{C}_k$ acima,

$$\begin{aligned}\epsilon^{(m)} &\leq \min_{p \in \mathbb{P}_m, p(0)=1} \max_{z \in E(c,d,a)} |p(z)|; \\ &\leq \max_{z \in E(c,d,a)} |\widehat{C}_k(z)|; \\ &= \frac{C_k(\frac{a}{d})}{|C_k(\frac{c}{d})|}.\end{aligned}$$

Daí pode-se escrever, baseado na proposição anterior,

$$\|\vec{r}_m\|_2 \leq k_2(X) \frac{C_k(\frac{a}{d})}{|C_k(\frac{c}{d})|} \|\vec{r}_0\|_2.$$

■

3.5 Método Gradiente Bi-Conjugado Estabilizado (Bi-CGStab)

O tópico aqui presente destina-se a descrição de alguns métodos iterativos em subespaços de Krylov baseados no processo de biortogonalização de Lanczos. Trata-se de uma extensão para matrizes não simétricas do algorítmo clássico de Lanczos simétrico, que pode ser observado como uma simplificação do método de Arnoldi para o caso particular quando a matriz é simétrica. Consistem em métodos de projeção intrinsicamente não ortogonais, possuindo algumas propriedades interessantes e uma teoria bem elaborada. Uma atenção especial será dada ao método do Gradiente Bi-Conjugado estabilizado (Bi-CGStab).

3.5.1 Biortogonalização de Lanczos

O processo do Lanczos não simétrico consiste em se contruir duas sequências biortogonais ao invés de uma sequência ortogonal. Considere os dois subespaços de Krylov:

$$K_m(A, \vec{v}_1) = span\{\vec{v}_1, A\vec{v}_1, \ldots, A^{m-1}\vec{v}_1\}; \tag{3.80}$$

$$K_m(A^T, \vec{w}_1) = span\{\vec{w}_1, A^T\vec{w}_1, \ldots, (A^T)^{m-1}\vec{w}_1\}, \tag{3.81}$$

para $A \in \mathbb{R}^{n \times n}$, $\vec{v}_1, \vec{w}_1 \in \mathbb{R}^n$, e tal que $< \vec{v}_1, \vec{w}_1 >= 1$. O algorítmo proposto por Lanczos para matrizes não simétricas constrói um par de bases biortogonais para estes

dois subespaços, o que é ilustrado a seguir:

Algorítmo 6 (Biortogonalização de Lanczos)

1. [Escolha dois vetores $\vec{v}_1, \vec{w}_1$ tais que $< \vec{v}_1, \vec{w}_1 >= 1$;
2. [Escolha $\beta_1 = \delta_1 = 0$ e $\vec{w}_0 = \vec{v}_0 = \vec{0}$;
3. [3.1. Para $j = 1, 2, \ldots, m$;
   - [3.1.1. $\alpha_j =< A\vec{v}_j, \vec{w}_j >$;
   - 3.1.2. $\hat{v}_{j+1} = A\vec{v}_j - \alpha_j\vec{v}_j - \beta_j\vec{v}_{j-1}$;
   - 3.1.3. $\hat{w}_{j+1} = A^T\vec{w}_j - \alpha_j\vec{w}_j - \delta_j\vec{w}_{j-1}$;
   - 3.1.4. $\delta_{j+1} = | < \hat{v}_{j+1}, \hat{w}_{j+1} > |^{1/2}$. Se $\delta_{j+1} = 0 \Rightarrow$ PARE !!!;
   - 3.1.5. $\beta_{j+1} =< \hat{v}_{j+1}, \hat{w}_{j+1} > /\delta_{j+1}$;
   - 3.1.6. $\vec{w}_{j+1} = \hat{w}_{j+1}/\beta_{j+1}$;
   - 3.1.7. $\vec{v}_{j+1} = \hat{v}_{j+1}/\delta_{j+1}$;
   -] 3.2. Fim para.]

Os escalares δ_{j+1} e β_{j+1}, definidos respectivamente em 3.1.4. e 3.1.5., garantem que $< \vec{v}_{j+1}, \vec{w}_{j+1} >= 1$. Considerando a matriz tridiagonal,

$$T_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \delta_2 & \alpha_2 & \beta_3 & & \\ & \cdot & \cdot & \cdot & \\ & & \delta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \delta_m & \alpha_m \end{bmatrix}, \tag{3.82}$$

tem-se a proposição:

Proposição. 3.5.1 *Se o algorítmo* 6 *não interrompe o seu processo até o passo* m*, então as sequências de vetores* $\{\vec{v}_i\}_{i=1,\ldots,m}$ *e* $\{\vec{w}_j\}_{j=1\ldots,m}$ *formam um sistema biortogonal, isto é,*

$$< \vec{v}_i, \vec{w}_j >= \begin{cases} 1, & se\ i = j \\ 0, & se\ i \neq j \end{cases} \quad 1 \leq i, j \leq m. \tag{3.83}$$

Além disso,$\{\vec{v}_i\}_{i=1,\ldots,m}$ *é uma base para* $K_m(A, \vec{v}_1)$ *e* $\{\vec{w}_j\}_{j=1,\ldots,m}$ *é uma base para* $K_m(A^T, \vec{w}_1)$*, sendo que as seguintes relações são válidas,*

$$AV_m = V_m T_m + \delta_{m+1}\vec{v}_{m+1}\vec{e}_m^T, \tag{3.84}$$

$$A^T W_m = W_m T_m^T + \beta_{m+1}\vec{w}_{m+1}\vec{e}_m^T, \tag{3.85}$$

$$W_m^T A V_m = T_m. \tag{3.86}$$

Prova. A idéia é fazer uso do princípio de indução finita para demonstrar que $\{\vec{v}_i\}_{i=1,\ldots,m}$ e $\{\vec{w}_j\}_{j=1\ldots,m}$ formam um sistema biortogonal. Assim tem-se:

$$
\begin{aligned}
&i)\ < \vec{v}_1, \vec{w}_1 >= 1, \text{ por construção do algorítmo;}\\
&ii)\ < \vec{v}_i, \vec{w}_j >= \begin{cases} 1, & se\ i = j \\ 0, & se\ i \neq j \end{cases} \quad 1 \leq i, j \leq m-1, \text{ hipótese de indução;}\\
&iii)\ < \vec{v}_i, \vec{w}_j >= \begin{cases} 1, & se\ i = j \\ 0, & se\ i \neq j \end{cases} \quad 1 \leq i, j \leq m, \text{ tese de indução.}
\end{aligned}
$$

Primeiramente se verificará que $< \vec{v}_{j+1}, \vec{w}_i >= 0$, para $i \leq j \leq m-1$. No caso $i = j$, tem-se:

$$
\begin{aligned}
&< \vec{v}_{j+1}, \vec{w}_j >=< \delta_{j+1}^{-1}[A\vec{v}_j - \alpha_j\vec{v}_j - \beta_j\vec{v}_{j-1}], \vec{w}_j >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_j >= \delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j > -\delta_{j+1}^{-1} < \alpha_j\vec{v}_j, \vec{w}_j > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_j >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_j >= \delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j > -\delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j >< \vec{v}_j, \vec{w}_j > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_j >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_j >= \delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j > -\delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j >= 0.
\end{aligned}
$$

Agora considerando o caso em que $i < j$, tem-se:

$$
\begin{aligned}
&< \vec{v}_{j+1}, \vec{w}_i >=< \delta_{j+1}^{-1}[A\vec{v}_j - \alpha_j\vec{v}_j - \beta_j\vec{v}_{j-1}], \vec{w}_i >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_i > -\delta_{j+1}^{-1} < \alpha_j\vec{v}_j, \vec{w}_i > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_i >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_i > -\delta_{j+1}^{-1} < A\vec{v}_j, \vec{w}_j >< \vec{v}_j, \vec{w}_i > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_i >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < \vec{v}_j, A^T\vec{w}_i > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_i >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < \vec{v}_j, [\beta_{i+1}\vec{w}_{i+1} + \alpha_i\vec{w}_i + \delta_i\vec{w}_{i-1}] > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_i >,
\end{aligned}
$$

note ainda que para $i < j-1$, tem-se:

$$
\begin{aligned}
&< \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < \vec{v}_j, \beta_{i+1}\vec{w}_{i+1} > +\delta_{j+1}^{-1} < \vec{v}_j, \alpha_i\vec{w}_i >\\
&+\delta_{j+1}^{-1} < \vec{v}_j, \delta_i\vec{w}_{i-1} > -\delta_{j+1}^{-1} < \beta_j\vec{v}_{j-1}, \vec{w}_i >;\\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= 0,
\end{aligned}
$$

e para o caso em que $i = j - 1$, tem-se:

$$
\begin{aligned}
&< \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < \vec{v}_j, \beta_{i+1}\vec{w}_{i+1} > + \delta_{j+1}^{-1} < \vec{v}_j, \alpha_i \vec{w}_i > \\
&+\delta_{j+1}^{-1} < \vec{v}_j, \delta_i \vec{w}_{i-1} > - \delta_{j+1}^{-1} < \beta_j \vec{v}_{j-1}, \vec{w}_i >; \\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1} < \vec{v}_j, \beta_{i+1}\vec{w}_{i+1} > - \delta_{j+1}^{-1} < \beta_j \vec{v}_{j-1}, \vec{w}_i >; \\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1}\beta_{i+1} < \vec{v}_j, \vec{w}_{i+1} > - \delta_{j+1}^{-1}\beta_j < \vec{v}_{j-1}, \vec{w}_i >; \\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= \delta_{j+1}^{-1}\beta_j < \vec{v}_j, \vec{w}_j > - \delta_{j+1}^{-1}\beta_j < \vec{v}_{j-1}, \vec{w}_{j-1} >; \\
&\Rightarrow < \vec{v}_{j+1}, \vec{w}_i >= 0.
\end{aligned}
$$

Analogamente mostra-se que $< \vec{v}_i, \vec{w}_{j+1} >= 0$, e também que $< \vec{v}_{j+1}, \vec{w}_{j+1} >= 1$, com $i \leq j \leq m-1$, o que mostra que as sequências de vetores $\{\vec{v}_i\}_{i=1,\ldots,m}$ e $\{\vec{w}_j\}_{j=1\ldots,m}$ formam um sistema biortogonal. Note agora que,

$$
\begin{aligned}
A\vec{v}_i &= \hat{v}_{i+1} + \alpha_i \vec{v}_i + \beta_i \vec{v}_{i-1}; \\
&= \delta_{i+1}\vec{v}_{i+1} + \alpha_i \vec{v}_i + \beta_i \vec{v}_{i-1}; \\
A^T\vec{w}_i &= \hat{w}_{i+1} + \alpha_i \vec{w}_i + \delta_i \vec{w}_{i-1}; \\
&= \beta_{i+1}\vec{w}_{i+1} + \alpha_i \vec{w}_i + \delta_i \vec{w}_{i-1},
\end{aligned}
$$

assim,

$$
\begin{aligned}
A\vec{v}_1 &= \delta_2 \vec{v}_2 + \alpha_1 \vec{v}_1; \\
A^2\vec{v}_1 &= \delta_2 A\vec{v}_2 + \alpha_1 A\vec{v}_1; \\
&= \delta_2(\delta_3 \vec{v}_3 + \alpha_2 \vec{v}_2 + \beta_2 \vec{v}_1) + \alpha_1(\delta_2 \vec{v}_2 + \alpha_1 \vec{v}_1); \\
&\vdots \\
A^T\vec{w}_1 &= \beta_2 \vec{w}_2 + \alpha_1 \vec{w}_1; \\
(A^T)^2\vec{w}_1 &= \beta_2 A^T\vec{w}_2 + \alpha_1 A^T\vec{w}_1; \\
&= \beta_2(\beta_3 \vec{w}_3 + \alpha_2 \vec{w}_2 + \delta_2 \vec{w}_1) + \alpha_1(\beta_2 \vec{w}_2 + \alpha_1 \vec{w}_1); \\
&\vdots
\end{aligned}
$$

deste modo tem-se que $span(\{\vec{v}_i\}_{i=1,\ldots,m}) = K_m(A, \vec{v}_1)$ e $span(\{\vec{w}_j\}_{j=1,\ldots,m}) = K_m(A^T, \vec{w}_1)$. Dada a definição de T_m é fácil verificar que,

$$
\begin{aligned}
AV_m &= V_m T_m + \delta_{m+1}\vec{v}_{m+1}\vec{e}_m^T; \\
A^T W_m &= W_m T_m^T + \beta_{m+1}\vec{w}_{m+1}\vec{e}_m^T,
\end{aligned}
$$

o procedimento é análogo ao feito anteriormente no método de Arnoldi. Daí tem-se que:

$$\begin{aligned} W_m^T A V_m &= W_m^T (V_m T_m + \delta_{m+1} \vec{v}_{m+1} \vec{e}_m^T); \\ &= (W_m^T V_m) T_m + \delta_{m+1} (W_m^T \vec{v}_{m+1}) \vec{e}_m^T; \\ &= T_m. \end{aligned}$$

■

Pode-se destacar ainda algumas vantagens e desvantagens do algoritmo Lanczos não simétrico. O método de Lanczos requer pouco armazenamento de vetores, por outro lado, possui maior potencialidade de ocorrer um "break down", ou seja,

$$< \hat{v}_{j+1}, \hat{w}_{j+1} >= 0. \tag{3.87}$$

Tal ocorrência se verifica quando um dos dois vetores $\hat{v}_{j+1}$, $\hat{w}_{j+1}$ se anula, ou quando ambos são não nulos e o produto interno é nulo. No primeiro caso, com $\hat{v}_{j+1} = \vec{0}$, então $span(\{\vec{v}_i\}_{i=1,\ldots,m})$ é invariante e, portanto, a solução aproximada é exata. Analogamente ao que já foi comentado anteriormente, no caso de $\hat{w}_{j+1} = \vec{0}$, então $span(\{\vec{w}_j\}_{j=1,\ldots,m})$ é invariante e nada pode ser afirmado sobre a solução aproximada para o sistema primal (A), apenas que é exata para o sistema dual (A^T). O tipo seguinte de "break down" constitue a sua forma mais grave, impossibilitando, a primeira vista, que se calcule a iteração posterior.

Observação. 3.5.1 *Existem modificações deste algoritmo que possibilitam o cáculo da iteração posterior em muitos casos. Tais modificações consistem nos algoritmos de Lanczos "look-ahead", em que a idéia é definir o par de vetores $\vec{v}_{j+2}$, $\vec{w}_{j+2}$, mesmo que o par $\vec{v}_{j+1}$, $\vec{w}_{j+1}$ não esteja definido. Se o par $\vec{v}_{j+2}$, $\vec{w}_{j+2}$ não pode ser definido, então tenta-se $\vec{v}_{j+3}$, $\vec{w}_{j+3}$ e assim por diante. Há diferentes tipos de implementações desta técnica encontradas na literatura especializada.*

3.5.2 O algoritmo Bi-CG

O algoritmo Gradiente Bi-Conjugado (Bi-CG), proposto por Fletcher em 1974, incorpora as idéias do gradiente conjugado ao algorítmmo de Lanczos não simétrico, de 1952. O método baseia-se na projeção em um subespaço afim de Krylov, de modo que o resíduo seja ortogonal a um outro subespaço de Krylov, e tal que as direções de busca satisfazem a propriedade bi-conjugada.

O procedimento para o m-ésimo passo tem por base os dois subespaço de Krylov:

$$K_m = span\{\vec{v}_1, A\vec{v}_1, \ldots, A^{m-1}\vec{v}_1\}; \tag{3.88}$$

$$\mathcal{L}_m = span\{\vec{w}_1, A^T\vec{w}_1, \ldots, (A^T)^{m-1}\vec{w}_1\}, \tag{3.89}$$

em que $\vec{v}_1 = \vec{r}_0/\left|\vec{r}_0\right|_2$ e $\vec{w}_1$ tal que $< \vec{v}_1, \vec{w}_1 > \neq 0$, geralmente escolhe-se igual a $\vec{v}_1$.

Considerado

$$T_m = L_m U_m, \tag{3.90}$$

a decomposição LU da matriz tridiagonal T_m, e

$$P_m = V_m U_m^{-1}. \tag{3.91}$$

A solução aproximada pode ser apresentada por

$$\vec{x}_m = \vec{x}_0 + V_m \vec{y}_m, \tag{3.92}$$

em que um modo muito natural de escolher o vetor $\vec{y}_m \in \mathbb{R}^m$, é tal que

$$\vec{r}_m = \vec{r}_0 - AV_m \vec{y}_m \tag{3.93}$$

seja ortogonal à

$$\{\vec{w}_j\}_{j=1,\ldots,m}, \tag{3.94}$$

que é uma base de $\mathcal{L}_m$, obtida pelo método de Lanczos não simétrico. Donde tem-se

$$\begin{aligned} W_m^T \vec{r}_m &= W_m^T \vec{r}_0 - W_m^T A V_m \vec{y}_m = 0; \\ &\Rightarrow W_m^T \vec{r}_0 - T_m \vec{y}_m = 0; \\ &\Rightarrow \rho_0 \vec{e}_1 - T_m \vec{y}_m = 0; \\ &\Rightarrow \vec{y}_m = T_m^{-1} \rho_0 \vec{e}_1, \end{aligned} \tag{3.95}$$

em que $\rho_0 = \left|\vec{r}_0\right|_2$, portanto

$$\begin{aligned} \vec{x}_m &= \vec{x}_0 + V_m T_m^{-1}(\rho_0 \vec{e}_1); \\ &= \vec{x}_0 + V_m U_m^{-1} L_m^{-1}(\rho_0 \vec{e}_1); \\ &= \vec{x}_0 + P_m L_m^{-1}(\rho_0 \vec{e}_1). \end{aligned} \tag{3.96}$$

A partir das equações acima, $\vec{x}_{m+1}$ pode ser atualizado a partir de $\vec{x}_m$ e $\vec{p}_{m+1}$ a partir de $\vec{p}_m$, como se pode perceber no algorítmo 7.

Considerando a matriz

$$P_m^* = W_m (L_m^{-1})^T, \tag{3.97}$$

observa-se que os vetores coluna p_i^* de P_m^* e os p_i de P_m são A-conjugados, pois

$$(P_m^*)^T A P_m = L_m^{-1} W_m^T A V_m U_m^{-1} = L_m^{-1} T_m U_m^{-1} = I.$$

A partir destas informações, pode-se derivar o Bi-CG do processo de Lanczos.

Algorítmo 7 (Bi-CG)

1. 1.1. Defina a tolerância tol e $\vec{x}_0$;
   1.2. Calcule $\vec{r}_0 = \vec{b} - A\vec{x}_0$ e escolha $\vec{r}_0^*$ tal que $< \vec{r}_0, \vec{r}_0^* > \neq 0$;
   1.3. Escolha $\vec{p}_0 = \vec{r}_0$ e $\vec{p}_0^* = \vec{r}_0^*$;
2. 2.1. Para $j = 0, 1, 2, \ldots,$ até $|\vec{r}_j|_2 < tol$;
   - 2.1.1. $\alpha_j = < \vec{r}_j, \vec{r}_j^* > / < A\vec{p}_j, \vec{p}_j^* >$;
   - 2.1.2. $\vec{x}_{j+1} = \vec{x}_j + \alpha_j \vec{p}_j$;
   - 2.1.3. $\vec{r}_{j+1} = \vec{r}_j - \alpha_j A\vec{p}_j$;
   - 2.1.4. $\vec{r}_{j+1}^* = \vec{r}_j^* - \alpha_j A^T \vec{p}_j^*$;
   - 2.1.5. $\beta_j = < \vec{r}_{j+1}, \vec{r}_{j+1}^* > / < \vec{r}_j, \vec{r}_j^* >$;
   - 2.1.6. $\vec{p}_{j+1} = \vec{r}_{j+1} + \beta_j \vec{p}_j$;
   - 2.1.7. $\vec{p}_{j+1}^* = \vec{r}_{j+1}^* + \beta_j \vec{p}_j^*$;

   2.2. Fim para.

Observação. 3.5.2 *: Se for de interesse resolver também o sistema dual, deve-se redefinir em* 1.2 $\vec{r}_0 = \vec{b}^* - A^T\vec{x}_0^*$*, e atualizar* $\vec{x}_{j+1}^* = \vec{x}_j^* + \alpha_j \vec{p}_j^*$ *em* 2.1.2.

3.5.3 Variações da Biortogonalização de Lanczos

O método Bi-CG, apresentado anteriormente, requer produtos matriz-vetor envolvendo A^Te A, a cada passo, o que implica em esforço computacional extra, além de ser menos conveniente operar com A^T do que com A, uma vez que nem sempre esta transposta está disponível, por exemplo, no caso em que a matriz A representa uma matriz Jacobiana. Devido a estas razões é desejável que se tenha um método iterativo que necessite de multipicações matriz vetor que envolvam somente a matriz A, e que garanta uma boa aproximação da solução. A seguir são destacados dois métodos, baseados no Bi-CG, com esta finalidade: o CGS (Gradiente Conjugado Quadrado) e o Bi-CGStab (Gradiente Bi-Conjugado Estabilizado).

O algoritmo CGS

O algoritmo Bi-CG fornece um vetor residual na iteração j da forma:

$$\vec{r}_j = \phi_j(A)\vec{r}_0, \tag{3.98}$$

em que ϕ_j é um polinômio de grau j, tal que $\phi_j(0) = 1$. Analogamente, tem-se um polinômio π_j de grau j tal que

$$\vec{p_j} = \pi_j(A)\vec{r}_0. \tag{3.99}$$

Similarmente, para $\vec{r}_j^*$ e $\vec{p}_j^*$, tem-se

$$\vec{r}_j^* = \phi_j(A^T)\vec{r}_0^*; \tag{3.100}$$
$$\vec{p}_j^* = \pi_j(A^T)\vec{r}_0^*, \tag{3.101}$$

o que resulta para o escalar α_j do algoritmo Bi-CG,

$$\begin{aligned}\alpha_j &= \frac{<\vec{r}_j, \vec{r}_j^*>}{<A\vec{p}_j, \vec{p}_j^*>}; \\ &= \frac{<\phi_j(A)\vec{r}_0, \phi_j(A^T)\vec{r}_0^*>}{<A\pi_j(A)\vec{r}_0, \pi_j(A^T)\vec{r}_0^*>}; \qquad (3.102)\\ &= \frac{<\phi_j^2(A)\vec{r}_0, \vec{r}_0^*>}{<A\pi_j^2(A)\vec{r}_0, \vec{r}_0^*>}. \qquad (3.103)\end{aligned}$$

Analogamente,

$$\begin{aligned}\beta_j &= \frac{<\vec{r}_{j+1}, \vec{r}_{j+1}^*>}{<\vec{r}_j, \vec{r}_j^*>}; \\ &= \frac{<\phi_{j+1}(A)\vec{r}_0, \phi_{j+1}(A^T)\vec{r}_0^*>}{<\phi_j(A)\vec{r}_0, \phi_j(A^T)\vec{r}_0^*>}; \\ &= \frac{<\phi_{j+1}^2(A)\vec{r}_0, \vec{r}_0^*>}{<\phi_j^2(A)\vec{r}_0, \vec{r}_0^*>}, \qquad (3.104)\end{aligned}$$

donde conclui-se que tendo uma fórmula de recorrência para os vetores $\phi_j^2(A)\vec{r}_0$ e $\pi_j^2(A)\vec{r}_0$, não haveria problemas em calcular α_j e β_j. A idéia principal do algorítmo CGS é encontrar uma sequência de iterados cujos os resíduos satisfazem

$$\vec{r}_j = \phi_j^2(A)\vec{r}_0. \tag{3.105}$$

Reescrevendo a recorrencia do Bi-CG em termos dos polinômios ϕ e π, tem-se

$$\phi_j(A)\vec{r}_0 = \phi_{j-1}(A)\vec{r}_0 - \alpha_{j-1}A\pi_{j-1}(A)\vec{r}_0; \tag{3.106}$$
$$\pi_j(A)\vec{r}_0 = \phi_j(A)\vec{r}_0 + \beta_{j-1}\pi_{j-1}(A)\vec{r}_0, \tag{3.107}$$

daí, a recorrência que define ϕ_j e π_j, é dada por:

$$\phi_{j+1}(t) = \phi_j(t) - \alpha_j t\pi_j(t); \tag{3.108}$$
$$\pi_{j+1}(t) = \phi_{j+1}(t) + \beta_j\pi_j(t), \tag{3.109}$$

e desta forma,

$$\phi_{j+1}^2(t) = \phi_j^2(t) - 2\alpha_j t\pi_j(t)\phi_j(t) + \alpha_j^2 t^2\pi_j^2(t); \tag{3.110}$$
$$\pi_{j+1}^2(t) = \phi_{j+1}^2(t) + 2\beta_j\phi_{j+1}(t)\pi_j(t) + \beta_j^2\pi_j^2(t). \tag{3.111}$$

Procedendo agora ao cálculo dos termos cruzados $\pi_j(t)\phi_j(t)$ e $\phi_{j+1}(t)\pi_j(t)$, tem-se

$$\begin{aligned}\phi_j(t)\pi_j(t) &= \phi_j(t)(\phi_j(t) + \beta_{j-1}\pi_{j-1}(t));\\ &= \phi_j^2(t) + \beta_{j-1}\phi_j(t)\pi_{j-1}(t); \qquad (3.112)\\ \phi_{j+1}(t)\pi_j(t) &= (\phi_j(t) - \alpha_j t\pi_j(t))\pi_j(t);\\ &= \phi_j(t)\pi_j(t) - \alpha_j t\pi_j^2(t);\\ &= \phi_j(t)(\phi_j(t) + \beta_{j-1}\pi_{j-1}(t)) - \alpha_j t\pi_j^2(t);\\ &= \phi_j^2(t) + \beta_{j-1}\phi_j(t)\pi_{j-1}(t) - \alpha_j t\pi_j^2(t). \qquad (3.113)\end{aligned}$$

Obtém-se então as recorrências que são a base do algorítmo

$$\phi_{j+1}^2(t) = \phi_j^2(t) - \alpha_j t(2\phi_j^2(t) + 2\beta_{j-1}\phi_j(t)\pi_{j-1}(t) - \alpha_j t\pi_j^2(t)); \tag{3.114}$$
$$\pi_{j+1}^2(t) = \phi_{j+1}^2(t) + 2\beta_j\phi_{j+1}(t)\pi_j(t) + \beta_j^2\pi_j^2(t); \tag{3.115}$$
$$\phi_{j+1}(t)\pi_j(t) = \phi_j^2(t) + \beta_{j-1}\phi_j(t)\pi_{j-1}(t) - \alpha_j t\pi_j^2(t). \tag{3.116}$$

Definindo

$$\vec{r}_j = \phi_j^2(A)\vec{r}_0; \tag{3.117}$$
$$\vec{p}_j = \pi_j^2(A)\vec{r}_0; \tag{3.118}$$
$$\vec{q}_j = \phi_{j+1}(A)\pi_j(A)\vec{r}_0, \tag{3.119}$$

tem-se

$$\vec{r}_{j+1} = \vec{r}_j - \alpha_j A(2\vec{r}_j + 2\beta_{j-1}\vec{q}_{j-1} - \alpha_j A\vec{p}_j); \tag{3.120}$$
$$\vec{q}_j = \vec{r}_j + \beta_{j-1}\vec{q}_{j-1} - \alpha_j A\vec{p}_j; \tag{3.121}$$
$$\vec{p}_{j+1} = \vec{r}_{j+1} + 2\beta_j\vec{q}_j + \beta_j^2\vec{p}_j. \tag{3.122}$$

Definindo, agora, dois vetores auxiliares com a finalidade de simplificar o algoritmo, tem-se

$$\vec{d}_j = 2\vec{r}_j + 2\beta_{j-1}\vec{q}_{j-1} - \alpha_j A\vec{p}_j; \tag{3.123}$$
$$\vec{u}_j = \vec{r}_j + \beta_{j-1}\vec{q}_{j-1}, \tag{3.124}$$

o que resulta nas relações

$$\vec{d_j} = \vec{u}_j + \vec{q_j}; \tag{3.125}$$

$$\vec{q_j} = \vec{u}_j - \alpha_j A\vec{p_j}; \tag{3.126}$$

$$\vec{p}_{j+1} = \vec{u}_{j+1} + \beta_j(\vec{q_j} + \beta_j\vec{p_j}). \tag{3.127}$$

O algorítmo CGS segue então:

Algorítmo 8 (CGS)

1. 1.1. Defina a tolerância tol e $\vec{x}_0$;
   1.2. Calcule $\vec{r}_0 = \vec{b} - A\vec{x}_0$ e escolha $\vec{r}_0^*$ arbitrário;
   1.3. Escolha $\vec{p}_0 = \vec{u}_0 = \vec{r}_0$;
2. 2.1. Para $j = 0, 1, 2, \ldots,$ até $|\vec{r}_j|_2 < tol$;
   - 2.1.1. $\alpha_j =< \vec{r}_j, \vec{r}_0^* > / < A\vec{p}_j, \vec{r}_0^* >$;
   - 2.1.2. $\vec{q_j} = \vec{u}_j - \alpha_j A\vec{p_j}$;
   - 2.1.3 $\vec{x}_{j+1} = \vec{x}_j + \alpha_j(\vec{u}_j + \vec{q_j})$;
   - 2.1.4. $\vec{r}_{j+1} = \vec{r}_j - \alpha_j A(\vec{u}_j + \vec{q_j})$;
   - 2.1.5. $\beta_j =< \vec{r}_{j+1}, \vec{r}_0^* > / < \vec{r}_j, \vec{r}_0^* >$;
   - 2.1.6. $\vec{u}_{j+1} = \vec{r}_{j+1} + \beta_j\vec{q_j}$;
   - 2.1.7. $\vec{p}_{j+1} = \vec{u}_{j+1} + \beta_j(\vec{q_j} + \beta_j\vec{p_j})$;

   2.2. Fim para.

Observação. 3.5.3 *Uma desvantagem do algoritmo CGS é que, como os polinômios estão elevados ao quadrado, os erros de arredondamento tem maior influência no algoritmo Bi-CG padrão. Esse fato pode acarretar grandes variações dos vetores residuais, refletindo imprecisões no cálculo do resíduo. Objetivando contornar os incovenientes citados anteriormente, será apresentado a seguir o algorítmo Bi-CGStab.*

O algoritmo Bi-CGStab

O algoritmo Bi-CGStab é uma variação do algorítmo CGS. A idéia é, que ao invés de procurar um vetor residual como citado acima,

$$\vec{r_j} = \phi_j^2(A)\vec{r}_0, \tag{3.128}$$

o algorítmo Bi-CGStab produz iterados cujo vetores residuais são da forma

$$\vec{r_j} = \psi_j(A)\phi_j(A)r_0, \tag{3.129}$$

em que $\phi_j(t)$ é o polinômio associado ao resíduo do algoritmo Bi-CG e $\psi_j(t)$ é um novo polinômio, definido recursivamente, tendo por objetivo suavizar o comportamento do

processo de convergência, na medida em que tenta manter uma convergência rápida alcançada pelo algorítmo. A relação de recorrência para o polinômio ψ é apresentada na forma

$$\psi_{j+1}(t) = (1 - \omega_j t)\psi_j(t), \tag{3.130}$$

em que os coeficientes ω_j´s podem ser determinados em cada passo para minimizar

$$|\vec{r}_j|_2 = |(I - \omega_j A)\psi_{j-1}(A)\phi_j(A)r_0|_2 \,. \tag{3.131}$$

O modo de obter as relações de recorrência é análoga ao do algorítmo CGS. O algorítmo Bi-CGStab necessita de recorrências para as definições,

$$\vec{r}_j = \psi_j(A)\phi_j(A)\vec{r}_0; \tag{3.132}$$

$$\vec{p}_j = \psi_j(A)\pi_j(A)\vec{r}_0, \tag{3.133}$$

o que resulta em

$$\begin{aligned}
\vec{r}_{j+1} &= \psi_{j+1}(A)\phi_{j+1}(A)\vec{r}_0; \\
&= [(I - \omega_j A)\psi_j(A)\phi_{j+1}(A)]\vec{r}_0; \\
&= [(I - \omega_j A)\psi_j(A)(\phi_j(A) - \alpha_j A\pi_j(A))]\vec{r}_0; \\
&= (I - \omega_j A)[\psi_j(A)\phi_j(A) - \alpha_j A\psi_j(A)\pi_j(A)]\vec{r}_0; \\
&= (I - \omega_j A)(\vec{r}_j - \alpha_j A\vec{p}_j); \qquad (3.134) \\
\vec{p}_{j+1} &= \psi_{j+1}(A)\pi_{j+1}(A)\vec{r}_0; \\
&= \psi_{j+1}(A)[\phi_{j+1}(A) + \beta_j\pi_j(A)]\vec{r}_0; \\
&= [\psi_{j+1}(A)\phi_{j+1}(A) + \beta_j\psi_{j+1}(A)\pi_j(A)]\vec{r}_0; \\
&= [\psi_{j+1}(A)\phi_{j+1}(A) + \beta_j(I - \omega_j A)\psi_j(A)\pi_j(A)]\vec{r}_0; \\
&= \vec{r}_{j+1} + \beta_j(I - \omega_j A)\vec{p}_j. \qquad (3.135)
\end{aligned}$$

O algoritmo Bi-CG necessita do cálculo da parcela $\beta_j = \rho_{j+1}/\rho_j$, em que:

$$\begin{aligned}
\rho_j &= < \vec{r}_j, \vec{r}_j^* >; \\
&= < \phi_j(A)r_0, \phi_j(A^T)r_0^* >; \\
&= < \phi_j^2(A)r_0, r_0^* >, \qquad (3.136)
\end{aligned}$$

já no algoritmo Bi-CGStab, define-se este parâmetro da seguinte maneira:

$$\begin{aligned}
\widetilde{\rho}_j &= < \psi_j(A)\phi_j(A)\vec{r}_0, \vec{r}_0^* >; \\
&= < \vec{r}_j, \vec{r}_0^* > \,. \qquad (3.137)
\end{aligned}$$

Analogamente tem-se a parcela $\alpha_j = \rho_j/\eta_j$, em que:

$$\begin{aligned}\eta_j &=< A\vec{p}_j, \vec{p}_j^* >;\\ &=< A\pi_j(A)\vec{r}_0, \pi_j(A^T)\vec{r}_0^* >;\\ &=< A\pi_j^2(A)\vec{r}_0, \vec{r}_0^* >,\end{aligned} \tag{3.138}$$

logo define-se $\widetilde{\eta}_j$ por:

$$\begin{aligned}\widetilde{\eta}_j &=< A\psi_j(A)\pi_j(A)\vec{r}_0, \vec{r}_0^* >;\\ &=< A\vec{p}_j, \vec{r}_0^* > .\end{aligned} \tag{3.139}$$

Pela propriedade de biortogonalidade:

$$i) < \phi_{j-1}(A)\vec{r}_0, \phi_{j-1}(A^T)\vec{r}_0^* >= \alpha^* < \phi_{j-1}(A)\vec{r}_0, (A^T)^{j-1}\vec{r}_0^* >; \tag{3.140}$$

$$ii) < A\pi_{j-1}(A)\vec{r}_0, \pi_{j-1}(A^T)\vec{r}_0^* >= \alpha^* < A\pi_{j-1}(A)\vec{r}_0, (A^T)^{j-1}\vec{r}_0^* >; \tag{3.141}$$

$$\begin{aligned}iii) &< \vec{r}_{j-1}, \vec{r}_0^* >=< \phi_{j-1}(A)\vec{r}_0, \psi_{j-1}(A^T)\vec{r}_0^* >\\ &\Rightarrow< \vec{r}_{j-1}, \vec{r}_0^* >= (-1)^{j-2}\omega_{j-2}...\omega_1 < \phi_{j-1}(A)\vec{r}_0, (A^T)^{j-1}\vec{r}_0^* >;\end{aligned} \tag{3.142}$$

$$\begin{aligned}iv) &< A\vec{p}_{j-1}, \vec{r}_0^* >=< A\pi_{j-1}(A)\vec{r}_0, \pi_{j-1}(A^T)\vec{r}_0^* >\\ &\Rightarrow< A\vec{p}_{j-1}, \vec{r}_0^* >= (-1)^{j-2}\omega_{j-2}...\omega_1 < A\pi_{j-1}(A)\vec{r}_0, (A^T)^{j-1}\vec{r}_0^* >,\end{aligned} \tag{3.143}$$

em que $\alpha^* = (-1)^{j-2}\alpha_{j-2}...\alpha_1$.

Então, relacionando os escalares ρ_j e $\widetilde{\rho}_j$, tem-se

$$\begin{aligned}\frac{\widetilde{\rho}_{j+1}}{\widetilde{\rho}_j} &= \frac{< \phi_{j+1}(A)\vec{r}_0, \psi_{j+1}(A^T)\vec{r}_0^* >}{< \phi_j(A)\vec{r}_0, \psi_j(A^T)\vec{r}_0^* >};\\ &= -\omega_j\frac{< \phi_{j+1}(A)\vec{r}_0, (A^T)^{j+1}\vec{r}_0^* >}{< \phi_j(A)\vec{r}_0, (A^T)^j\vec{r}_0^* >};\\ &= \frac{\omega_j}{\alpha_j}\left(-\alpha_j\frac{< \phi_{j+1}(A)\vec{r}_0, (A^T)^{j+1}\vec{r}_0^* >}{< \phi_j(A)\vec{r}_0, (A^T)^j\vec{r}_0^* >}\right);\\ &= \frac{\omega_j}{\alpha_j}\left(\frac{< \phi_{j+1}(A)\vec{r}_0, \phi_{j+1}(A^T)\vec{r}_0^* >}{< \phi_j(A)\vec{r}_0, \phi_j(A^T)\vec{r}_0^* >}\right);\\ &= \frac{\omega_j}{\alpha_j}\frac{\rho_{j+1}}{\rho_j} \Rightarrow \beta_j = (\frac{\widetilde{\rho}_{j+1}}{\widetilde{\rho}_j})(\frac{\alpha_j}{\omega_j}),\end{aligned} \tag{3.144}$$

de modo análogo, tem-se para α_j

$$\begin{aligned}\alpha_j &= \frac{< \phi_j(A)\vec{r}_0, \phi_j(A^T)\vec{r}_0^* >}{< A\pi_j(A)\vec{r}_0, \pi_j(A^T)\vec{r}_0^* >}; \\ &= \frac{< \phi_j(A)\vec{r}_0, (A^T)^j\vec{r}_0^* >}{< A\pi_j(A)\vec{r}_0, (A^T)^j\vec{r}_0^* >}; \qquad (3.145) \\ &= \frac{< \phi_j(A)\vec{r}_0, \psi_j(A^T)\vec{r}_0^* >}{< A\pi_j(A)\vec{r}_0, \psi_j(A^T)\vec{r}_0^* >}; \\ &= \frac{< \psi_j(A)\phi_j(A)\vec{r}_0, \vec{r}_0^* >}{< A\psi_j(A)\pi_j(A)\vec{r}_0, \vec{r}_0^* >}, \qquad (3.146)\end{aligned}$$

o que permite escrever

$$\alpha_j = \frac{\widetilde{\rho}_j}{< Ap_j, r_0^* >}. \qquad (3.147)$$

Voltando agora a atenção para os parâmetros ω_j's, tem-se,

$$\vec{r}_{j+1} = (I - \omega_j A)\vec{s}_j, \qquad (3.148)$$

em que,

$$\vec{s}_j \equiv \vec{r}_j - \alpha_j A\vec{p}_j. \qquad (3.149)$$

Então, o valor ótimo de ω_j é dado por

$$\omega_j = \frac{< A\vec{s}_j, \vec{s}_j >}{< A\vec{s}_j, A\vec{s}_j >}. \qquad (3.150)$$

Assim a equação residual acima pode ser rescrita por

$$\begin{aligned}\vec{r}_{j+1} &= \vec{s}_j - \omega_j A\vec{s}_j; \\ &= \vec{r}_j - \alpha_j A\vec{p}_j - \omega_j A\vec{s}_j, \qquad (3.151)\end{aligned}$$

donde conclui-se que a atualização da solução aproximada é fornecida pela expressão,

$$\vec{x}_{j+1} = \vec{x}_j + \alpha_j\vec{p}_j + \omega_j\vec{s}_j. \qquad (3.152)$$

Segue então o algoritmo:

Algorítmo 9 (Bi-CGStab)

1. 1.1. Defina a tolerância tol e $\vec{x}_0$;
   1.2. Calcule $\vec{r}_0 = \vec{b} - A\vec{x}_0$ e escolha $\vec{r}_0^*$ arbitrário;
   1.3. Escolha $\vec{p}_0 = \vec{r}_0$;
2. 2.1. Para $j = 0, 1, 2, \ldots,$ até $|\vec{r}_j|_2 < tol$;
   - 2.1.1. $\alpha_j = < \vec{r}_j, \vec{r}_0^* > / < A\vec{p}_j, \vec{r}_0^* >$;
   - 2.1.2. $\vec{s}_j = \vec{r}_j - \alpha_j A\vec{p}_j$;
   - 2.1.3. $\omega_j = < A\vec{s}_j, \vec{s}_j > / < A\vec{s}_j, A\vec{s}_j >$;
   - 2.1.4. $\vec{x}_{j+1} = \vec{x}_j + \alpha_j\vec{p}_j + \omega_j\vec{s}_j$;
   - 2.1.4. $\vec{r}_{j+1} = \vec{s}_j - \omega_j A\vec{s}_j$;
   - 2.1.5. $\beta_j = (< \vec{r}_{j+1}, \vec{r}_0^* > / < \vec{r}_j, \vec{r}_0^* >)(\alpha_j/\omega_j)$;
   - 2.1.6. $\vec{p}_{j+1} = \vec{r}_{j+1} + \beta_j(\vec{p}_j - \omega_j A\vec{p}_j)$;

   2.2. Fim para.

3.5.4 Técnicas de precondicionamento

A literatura especializada mostra que, em muitos problemas práticos, resolver simplesmente o sistema linear

$$A\vec{x} = \vec{b}, \tag{3.153}$$

por meio de um método iterativo, pode não resultar em boas propriedades de convergência, sendo necessário transformá-lo em um sistema linear mais favorável. Objetivando tal finalidade, faz-se necessário o conhecimento de técnicas de precondicionamento do sistema.

A eficácia e o desempenho dos métodos iterativos, de uma forma geral, está intimamente ligada às propriedades espectrais da matriz do sistema analisado. Em termos práticos a convergência dos métodos anteriormente descritos, será rápida se a matriz $A \in \mathbb{R}^{n\times n}$ for próxima da matriz identidade, no seguinte sentido,

$$\rho(I - A) << 1, \tag{3.154}$$

em que, para uma dada matriz $G \in \mathbb{R}^{n\times n}$, $\rho(G) = \max\{|\lambda| \mid \lambda \in \sigma(G)\}$ designa o raio espectral da matriz G. Infelizmente, em muitas aplicações de interesse as matrizes associadas aos sistemas não possuem tal característica, sendo necessária então a aplicação de técnicas de precondicionamento. Estas técnicas buscam transformar o sistema linear original em um outro sistema equivalente, no sentido de ter a mesma solução que o sistema original, porém tendo propriedades espectrais bem mais favoráveis.

A ecolha de um precondicionador para o sistema linear $A\vec{x} = \vec{b}$, implica em se determinar uma matriz M, precondicionadora, com as seguintes características:

- M deve ser uma boa aproximação para a matriz A, ou seja, $M^{-1}A$ ser suficiente-

mente próxima da matriz identidade;

- O custo computacional da construção de M ser baixo;
- O sistema "equivalente" $M\vec{v} = \vec{w}$ ser muito mais fácil de se resolver do que o sistema original.

Baseado nos comentários acima, tem-se que $\rho(I - M^{-1}A) << 1$, o que implica uma convergência assintótica rápida, ou com $\|I - M^{-1}A\| << 1$, o que implica em uma grande redução no erro por passo.

Há modos diferentes de efetuar o precondicionamento, dentre os quais pode-se destacar os seguintes:

- **Precondicionamento a esquerda:** consiste em aplicar o método iterativo para o sistema linear:
$$M^{-1}A\vec{x} = M^{-1}\vec{b}. \tag{3.155}$$
- **Precondicionamento a direita:** aplicar o método iterativo ao sistema:
$$AM^{-1}\vec{y} = \vec{b}, \tag{3.156}$$
e a solução $\vec{x}$ é obtida resolvendo:
$$M\vec{x} = \vec{y}. \tag{3.157}$$
- **Precondicionamento bilateral:** seja M um precondicionador com a fatoração $M = M_1M_2$. A idéia desta implementação é resolver o sistema:
$$M_1^{-1}AM_2^{-1}\vec{z} = M_1^{-1}\vec{b}, \tag{3.158}$$
e, posteriormente, encontrar a solução resolvendo:
$$M_2\vec{x} = \vec{z}. \tag{3.159}$$
Perceba ainda que caso A seja uma matriz simétrica positiva definida, pode-se proceder a fatoração de Cholesky ($M_2 = M_1^T$) da matriz M, e desta forma a matriz $M_1^{-1}AM_2^{-1}$ ainda permanece simétrica positiva definida, o que não seria possível nos casos anteriores.

A literatura fornece diferentes e variados modos de se construir um precondicionador. O presente trabalho enfocará técnicas baseadas na fatoração LU incompleta (ILU).

Fatoração LU Incompleta (ILU): Sejam a matriz $A \in \mathbb{R}^{n\times n}$ e o subconjunto de índices $P \subset \{(i,j) \mid 1 \leq i,j \leq n, i \neq j\}$, chamado de conjunto padrão zero. Um procedimento explicativo para compreenção da fatoração LU incompleta (ILU) é apresentado na demonstração do Teorema de Meijerink e van der Vorst, para o qual se fazem necessários alguns resultados preliminares:

Definição. 3.5.1 *Sejam as matrizes $A, M, N \in \mathbb{R}^{n\times n}$, tais que $A = M - N$. O par de matrizes M,N é dito ser uma partição regular da matriz A, se M é não singular e M^{-1} e N são não negativas.*

Definição. 3.5.2 *Seja a matriz $A \in \mathbb{R}^{n\times n}$, então A é chamada M-matriz se*

$$\begin{aligned} &i)\ A_{ii} > 0,\ i = 1, ..., n; \\ &ii)\ A_{ij} \leq 0,\ i, j = 1, ..., n,\ com\ i \neq j; \\ &iii)\ A\ é\ não\ singular\ e\ A^{-1} \geq [0]_{nxn}. \end{aligned} \tag{3.160}$$

Lema. 3.5.1 *Seja a M-matriz $A \in \mathbb{R}^{n\times n}$, então a matriz $A^{(1)} \in \mathbb{R}^{n\times n}$, resultante do primeiro passo da eliminação de Gauss (eliminação da primeira coluna de A), também é uma M-matriz.*

Prova. Note que os elementos fora da diagonal de $A^{(1)}$ são da forma

$$A_{ij}^{(1)} = A_{ij} - \frac{A_{i1}A_{1j}}{A_{11}}.$$

Como A_{i1}, A_{1j} e A_{ij} são não positivos, e A_{11} é positivo, donde conclui-se que $A_{ij}^{(1)} \leq 0$, para $i \neq j$. O fato de $A^{(1)}$ não ser singular é uma consequência imediata da eliminação de Gauss padrão, já que

$$A = L^{(1)}A^{(1)},$$

em que:

$$L^{(1)} = \left[\frac{A_{*1}}{A_{11}}, \vec{e}_2, \vec{e}_3, ..., \vec{e}_n\right],$$

com A_{*1} representando o vetor coluna 1 da matriz A. Agora deve-se analizar a matriz $(A^{(1)})^{-1}$. Note que, examinando-se o vetor $(A^{(1)})^{-1}\vec{e}_j$, $j = 1, ..., n$, tem-se:

$$\begin{aligned} &i)\ j = 1:\ (A^{(1)})^{-1}\vec{e}_1 = \frac{1}{A_{11}}\vec{e}_1 \geq \vec{0}; \\ &ii)\ j \neq 1:\ (A^{(1)})^{-1}\vec{e}_j = A^{-1}\ (L^{(1)})\vec{e}_j = A^{-1}\vec{e}_j \geq \vec{0}, \end{aligned}$$

donde conclui-se que $(A^{(1)})^{-1}$ é não negativa. ■

Observação. 3.5.4 *Uma conseqüência imediata desse lema é que se A é uma M-matriz, então $A^{(k)}$, proveniente do k-ésimo passo da eliminação de Gauss padrão, também o será.*

Lema. 3.5.2 *Seja a M-matriz $A \in \mathbb{R}^{n\times n}$. Se os elementos de uma matriz $B \in \mathbb{R}^{n\times n}$ satisfazem*

$$0 < A_{ii} \leq B_{ii}; \tag{3.161}$$

$$A_{ij} \leq B_{ij} \leq 0, \text{ para } i \neq j, \tag{3.162}$$

,então B também é uma M-matriz.

Prova. Ver GREENBAUM (1997). ■

Teorema. 3.5.2 *(Meijerink e van der Vorst): Seja a M-matriz $A \in \mathbb{R}^{n\times n}$, então para todo subconjunto de índices $P \subset \{(i,j) \mid 1 \leq i,j \leq n, i \neq j\}$, conjunto padrão zero, existe uma matriz triangular inferior $L = [L_{ij}]$ com diagonal unitária e uma matriz triangular superior $U = [U_{ij}]$, tal que $A = LU - R$, em que:*

$$i)\ L_{ij} = 0, \text{ se } (i,j) \in P; \tag{3.163}$$

$$ii)\ U_{ij} = 0, \text{ se } (i,j) \in P; \tag{3.164}$$

$$iii)\ R_{ij} = 0, \text{ se } (i,j) \notin P. \tag{3.165}$$

As matrizes fatores L e U são únicas, e a partição $A = LU - R$ é uma partição regular.

Prova. O processo de demonstração decorre da construção de um procedimento análogo ao processo de eliminação de Gauss, que define o a fatoração ILU. No k-ésimo passo, primeiramente substitui-se os elementos da matriz de índices $(k,j),(i,k) \in P$ por 0 (zero). Então procede-se ao passo da eliminação Gauss de maneira convencional, ou seja, eliminando os elementos na k-ésima coluna, da $(k+1)$-ésima linha até a n-ésima linha. Definindo-se então as matrizes

$$A^{(k)} := [A_{ij}^{(k)}];$$
$$\tilde{A}^{(k)} := [\tilde{A}_{ij}^{(k)}];$$
$$L^{(k)} := [L_{ij}^{(k)}];$$
$$R^{(k)} := [R_{ij}^{(k)}],$$

pelas relações

$$A^{(0)} = A;$$
$$\tilde{A}^{(k)} = A^{(k-1)} - R^{(k)};$$
$$A^{(k)} = L^{(k)}\tilde{A}^{(k)},$$

para $k = 1, ..., n-1$, onde $R^{(k)}$ tem entradas nulas exceto nas posições $(k, j) \in P$. Nas posições $(i, k) \in P$, tem-se

$$\begin{aligned} R_{kj}^{(k)} &= -A_{kj}^{(k-1)}; \\ R_{ik}^{(k)} &= -A_{ik}^{(k-1)}. \end{aligned}$$

A matriz triangular inferior $L^{(k)}$ é igual a matriz identidade exceto pela k-ésima coluna, que é dada pelo vetor

$$\left(0, ..., 0, 1, -\frac{\tilde{A}_{(k+1)k}^{(k)}}{\tilde{A}_{kk}^{(k)}}, ..., -\frac{\tilde{A}_{nk}^{(k)}}{\tilde{A}_{kk}^{(k)}}\right).$$

onde a entrada unitária pertence a k-ésima linha. Note que a matriz $A^{(k)}$ vem da eliminação dos elementos da k-ésima coluna (da $(k+1)$-ésima linha até a n-ésima linha) da matriz $\tilde{A}^{(k)}$, a qual provém da matriz $A^{(k-1)}$, pela substituição das entradas, cujos índice pertencem ao conjunto P, por 0. Agora tem-se que $A^{(0)} = A$ é uma M-matriz, então $R^{(1)} \geq [0]_{n\times n}$. Dos lemas anteriores segue que $\tilde{A}^{(1)}$ é uma M-matriz e, consequetemente, $L^{(1)} \geq [0]_{n\times n}$, além do que $A^{(1)}$ também é uma M-matriz. Seguindo com este raciocínio pede-se concluir que $\tilde{A}^{(k)}$ e $A^{(k)}$ são M-matrizes e $L^{(k)}, R^{(k)} \geq [0]_{n\times n}$ para $k = 1, ..., n-1$. As definições anteriores permitem escrever

$$\begin{aligned} L^{(k)}R^{(m)} &= R^{(m)}, \text{ se } k < m; \\ A^{(n-1)} &= L^{(n-1)}\tilde{A}^{(n-1)}; \\ &= L^{(n-1)}A^{(n-2)} + L^{(n-1)}R^{(n-1)}; \\ &\vdots \\ &= \left(\prod_{j=1}^{n-1} L^{(n-j)}\right) A^{(0)} + \sum_{i=1}^{n-1} \left(\prod_{j=1}^{n-i} L^{(n-j)}\right) R^{(i)}. \end{aligned}$$

Daí tem-se

$$A^{(n-1)} = \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)\left(A + \sum_{i=1}^{n-1} R^{(i)}\right),$$

o que leva as seguintes definições

$$\begin{aligned} U &:= A^{(n-1)}; \\ L &:= \left(\prod_{j=1}^{n-1} L^{(n-j)}\right)^{-1}; \\ R &:= \sum_{i=1}^{n-1} R^{(i)}. \end{aligned}$$

Ainda, $LU = A + R$, $(LU)^{-1} \geq [0]_{n\times n}$ e $R \geq [0]_{n\times n}$. Logo a partição da matriz A é regular. A unicidade dos fatores L e U é uma consequência direta do equacionamento dos elementos da matriz A e LU para $(i,j) \notin P$, e do fato de que L tem sua diagonal unitária. ■

Algorítmo 10 (ILU versão IKJ)

1. Para $i = 2, ..., n$;
   2 Para $k = 1, \ldots, i-1$, e se $(i,k) \notin P$;
      3. $A_{ik} := A_{ik}/A_{kk}$;
         4. Para $j = k+1, \ldots, n$, e se $(i,j) \notin P$;
         [5. $A_{ij} := A_{ij} - A_{ik}A_{kj}$;
         6. Fim para;
   7. Fim para;
8. Fim para.

Escolhendo o conjunto $P = \{(i,j) \mid A_{ij} = 0\}$, tem-se a bem conhecida *fatoração LU incompleta de nível* 0*, ou* $ILU(0)$*,* em que os fatores L e U, possuem o mesmo padrão de esparsidade das partes triangular inferior e triangular superior da matriz A respectivamente. Entretanto, o número de elementos não nulos do produto LU é maior que o da matriz A. Existem ainda problemas em que a fatoração incompleta $ILU(0)$ não produz um bom precondicionador. Objetivando melhorar a precisão da fatoração e, consequentemente, diminuir o número de iterações, introduziu-se implementações que diferem da $ILU(0)$, agregando a adição de alguns elementos na estrutura original da matriz, e desta forma, os fatores L e U terão mais elementos não nulos do que as partes triangular inferior e superior da matriz A. Tais comentários remetem ao conceito de nível de preenchimento atribuído a cada elemento processado pela eliminação Gaussiana.

Definição. 3.5.3 *Seja a matriz* $A \in \mathbb{R}^{n\times n}$*. O nível de preenchimento de um elemento* A_{ij} *é definido por*

$$niv_{ij} = \begin{cases} 0, \text{ se } A_{ij} \neq 0 \text{ ou } i = j; \\ \infty, \text{ caso contrário.} \end{cases} \tag{3.166}$$

Como a cada iteração este elemento é modificado na linha 5, do algorítmo ILU, o niv_{ij} deve ser atualizado da seguinte maneira:

$$niv_{ij} = \min\{niv_{ij}, niv_{ik} + niv_{kj} + 1\}. \tag{3.167}$$

Definindo agora o conjunto

$$P_m = \{(i,j) \mid niv_{ij} > m\}, \tag{3.168}$$

em que niv_{ij} é o nível de preenchimento depois de todas as atualizações efetuadas. Pode-se então implementar uma fatoração LU incompleta de nível m ($ILU(m)$), em que os elementos A_{ij} cujo o nível de preenchimento não excederem m são mantidos no processamento.

Algorítmo 11 (ILU(m))

1. Defina $niv_{ij} = 0$, onde $A_{ij} \neq 0$;
2. Para $i = 2, ..., n$;
   3. Para $k = 1, \ldots, i-1$, e se $niv_{ik} \leq m$;
      3. $A_{ik} := A_{ik}/A_{kk}$;
      4. $A_{i*} := A_{i*} - A_{ik}A_{k*}$;
      5. Atualize niv_{ij} para os $A_{ij} \neq 0$;
   7. Fim para;
   8. Anule os elementos i-ésima linha cujo $niv_{ik} > m$;
9. Fim para.

Observação. 3.5.5 *Caso a matriz $A \in \mathbb{R}^{n\times n}$ seja simétrica positiva definida, pode-se adaptar a fatoração Cholesky ao algoritmo acima.*

Capítulo 4

APLICAÇÕES

4.1 INTRODUÇÃO

Este capítulo apresenta alguns problemas selecionados, já bem conhecidos da literatura, para avaliar a eficiência dos códigos implementados. Alguns comentários a respeito dos resultados obtidos também serão dispostos ao longo do texto. Uma importante observação é que o sistema adotado nos exemplos é o MKS (SI), em virtude disto não serão explicitadas as dimensões das entidades físicas.

4.2 "SQUARE LID-DRIVEN CAVITY"

O problema configura-se como um escoamento incompressível em uma cavidade quadrada cuja face superior move-se com uma velocidade horizontal constante. Este problema tem servido como modelo para teste e avaliação de muitas técnicas numéricas, principalmente devido aos seus dois pontos de singularidade situados nos vértices superiores da cavidade.

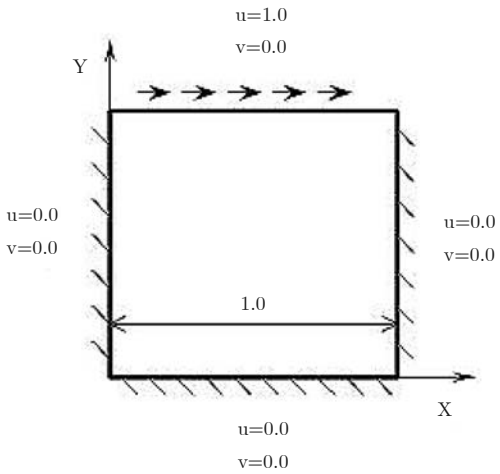

Fig. 2: Lid Driven-Cavity

Computacionalmente esses pontos são considerados como se tivessem a velocidade da face superior da cavidade Os resultados são comparados com os resultados apresentados em Ghia et al (1982). Tal trabalho apesar de ser do início da década de oitenta, ainda nos dias de hoje é largamente referenciado na literatura e constitui um importante "bentchmark" para esta aplicação. A malha utilizada é uma malha 60 X 60, estruturada e não uniforme. O domínio foi dividido em nove regiões e cada uma recebeu uma malha 20 X 20. As regiões podem ser assim destacadas: quatro regiões de dimensão 0.25 X 0.25, quatro regiões de dimensão 0.5 X 0.25 e uma de dimensão 0.5 X 0.5. O tipo de elemento utilizado, foi o "quad – four" isoparamétrico. A divisão do domínio em nove regiões está apresentada na figura que segue.

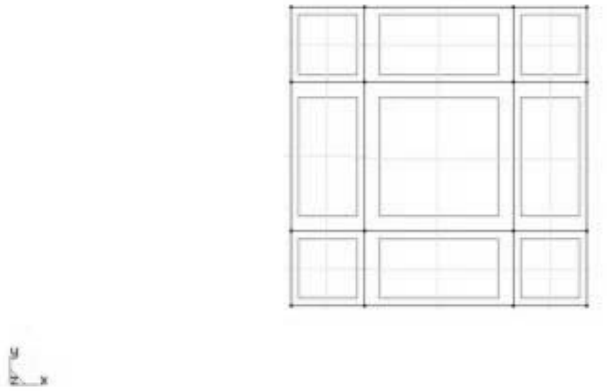

Fig. 3: Domínio dividido

A malha utilizada pode ser agora visualizada.

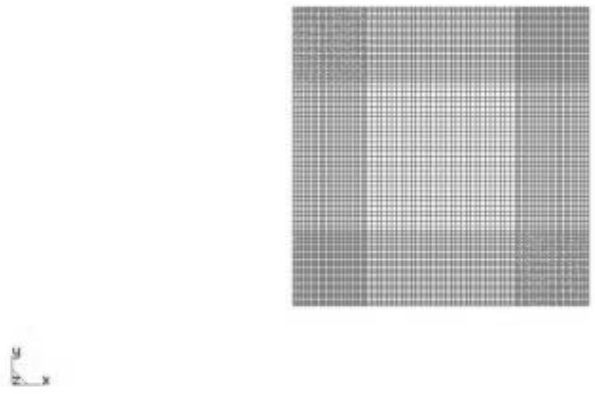

Fig. 4: Malha 60 X 60 etruturada

Vale ressaltar ainda que foram feitos testes com malhas mais refinadas e não foram detectadas disparidades significativas em relação aos resultados obtidos. Também foram feitos testes com outros tipos de elementos, inclusive triangulares; preferiu-se trabalhar com elementos de baixa ordem de interpolação devido ao custo computacional elevado dos parâmetros de estabilização, assim como devido a largura de banda da matriz associada ao problema. Novamente não se percebeu grandes disparidades entre os resultados obtidos. A integração numérica foi realizada com tablatura Gaussiana, em que utilizou-se quatro pontos de integração no elemento mestre.

A seguir são apresentados os campos de velocidade e pressão obtidos pelo método BiCGStab para o caso do número de Reynolds igual a mil (Re = 1000), uma vez que estes

resultados foram idênticos aos obtidos com o método direto e com o GMRES.

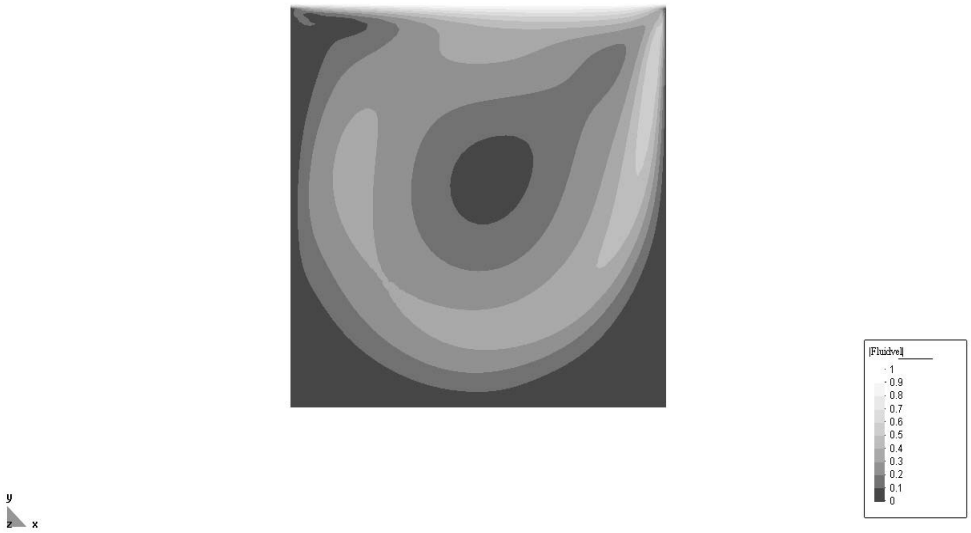

Fig. 5: Norma Euclidiana do vetor velocidade

Fig. 6: Campo de pressão

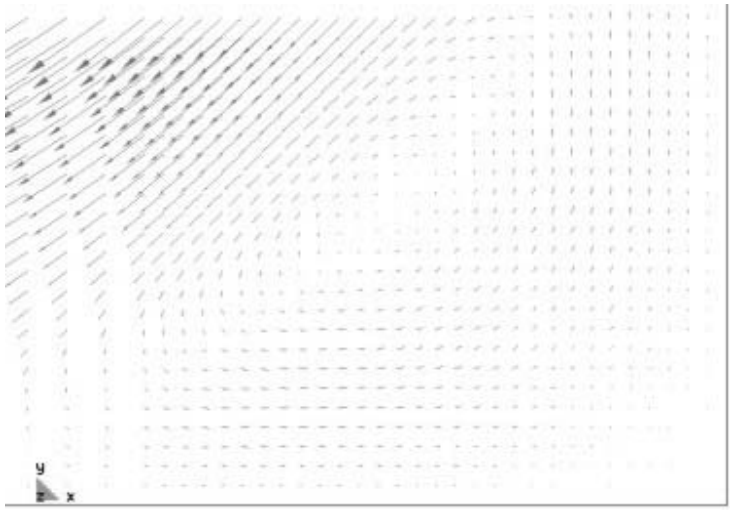

Fig. 7: Canto direito inferior da cavidade (Re = 1000)

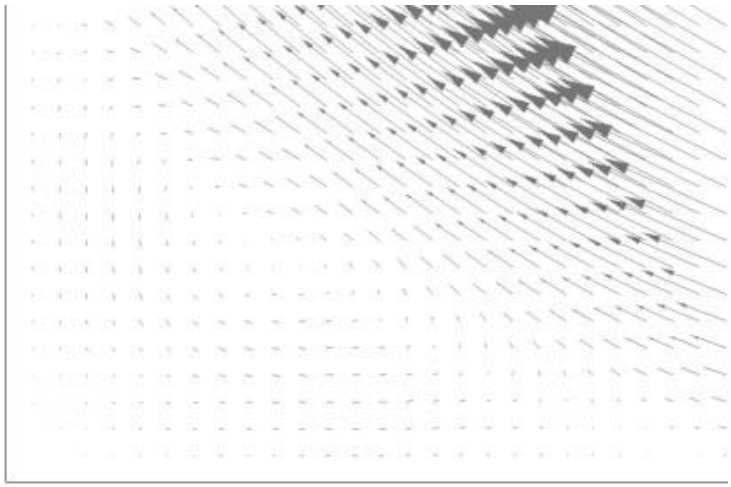

Fig. 8: Canto esquerdo inferior da cavidade (Re = 1000)

São apresentados agora os gráficos comparativos dos resultados obtidos neste trabalho com Ghia et al (1982).

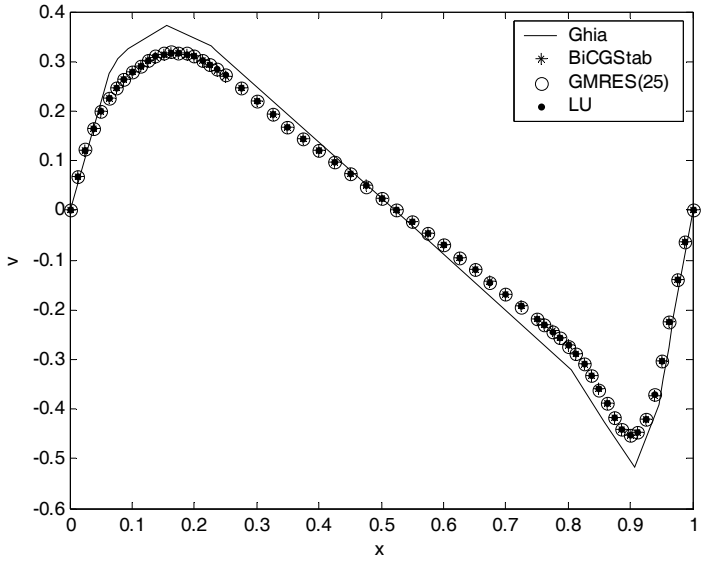

Fig. 9: Perfil da velocidade vertical ao longo da linha de centro horizontal (Re = 1000)

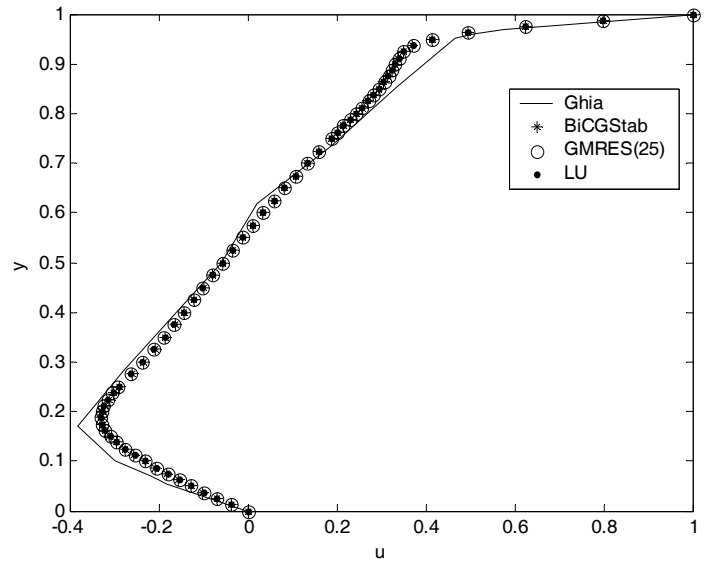

Fig. 10: Perfil da velocidade horizontal ao longo da linha de centro vertical (Re = 1000)

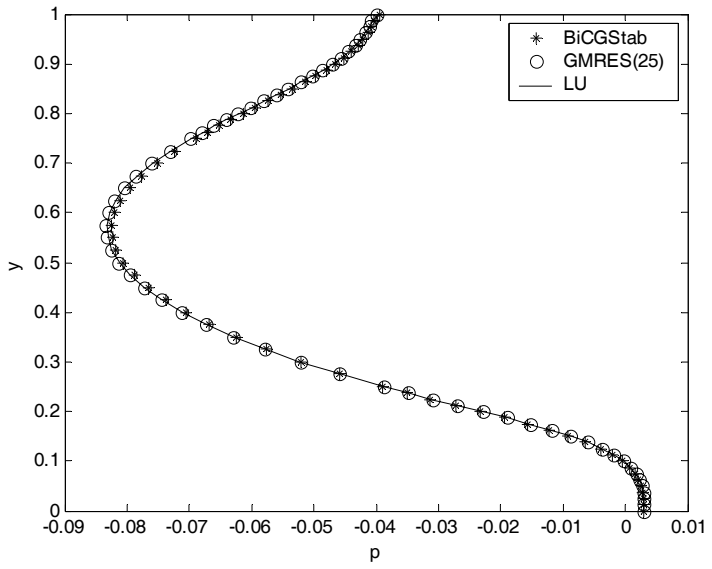

Fig. 11: Perfil da pressão ao longo da linha de centro vertical (Re = 1000)

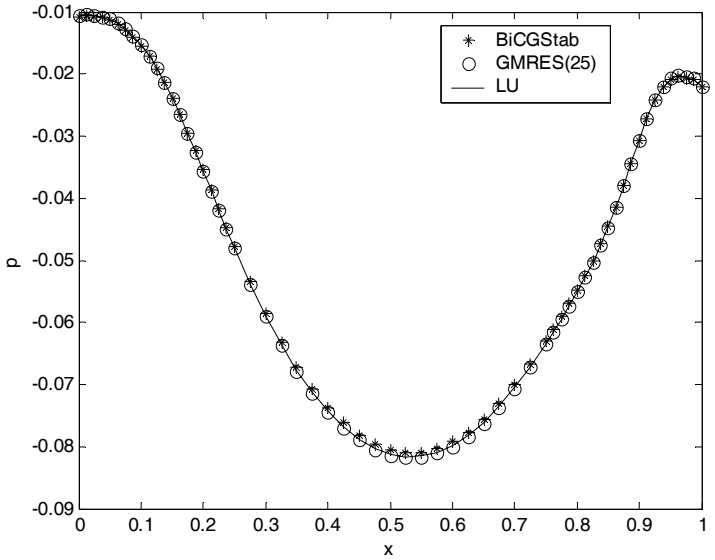

Fig. 12: Perfil da pressão ao longo da linha de centro horizontal (Re = 1000)

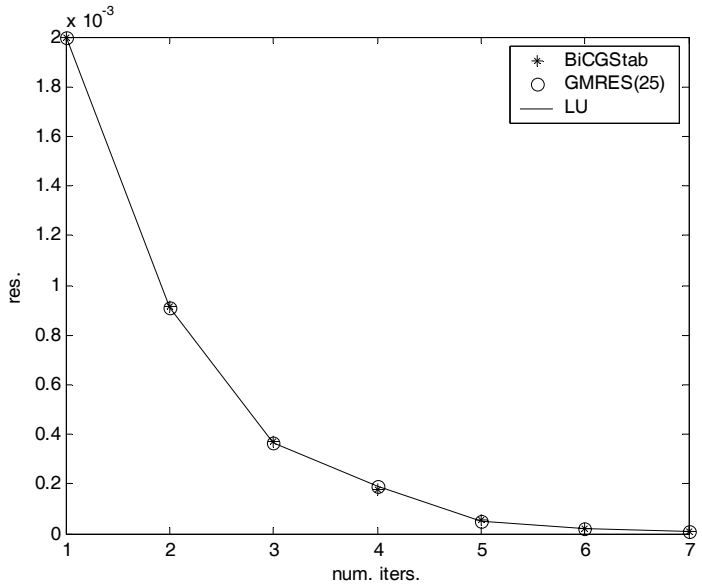

Fig. 13: Convergência dos métodos

\	BiCGStab	GMRES	LU
Tempo (min.)	11.0797667	11.8752333	12.8942000
Iterações (1^o passo)	19	31	–
Iterações (2^o passo)	46	217	–
Iterações (3^o passo)	49	279	–
Iterações (4^o passo)	63	310	–
Iterações (5^o passo)	64	744	–
Iterações (6^o passo)	66	868	–

(4.1)

As figuras 4, 5 e 6 apresentam as características de malha e do escoamento (campos de pressão e velocidade) para o caso Re=1000. Já as figuras 9 e 10 apresentam gráficos comparativos entre os resultados obtidos neste trabalho e os resultados obtidos por Ghia et al (1982). Os gráficos mostram os perfis das componentes do vetor velocidade ao longo das linhas de centro da cavidade quadrada. Nas figuras 11 e 12 são apresentados os perfis do campo de pressão ao longo das linhas de centro.

A aplicação deste caso ("Lid Driven Cavity") foi abordada com a utilização do métodos direto (LU esparso), BiCGStab e GMRES(25), cujos resultados comparativos com Ghia et al (1982) encontram-se nos gráficos apresentados nas figuras 9 e 10, como já comentado anteriormente. Observando os gráficos apresentados nestas figuras, nota-se que os resultados obtidos no trabalho aqui apresentado (método direto e métodos iterativos) tiveram uma boa concordância com os resultados de Ghia et al (1982), sendo que observou-se pela figura 10, que nas regiões do fundo da cavidade e próximo a tampa móvel houve divergências mais severas quanto ao gradiente dos resultados, o que não se observa em qualquer outra região. Já na figura 9, nota-se novamente uma boa concordância entre

os resultados, além disto observa-se ainda que os perfis obtidos são bastante similares ao perfil resultante de Ghia et al (1982).

A precisão empregada para o método de Newton foi de 10^{-5}, uma vez que testes realizados com tolerâncias menores não mostraram diferenças significativas nos resultados obtidos para os campos de pressão e velocidade, o que não ocorreu em testes com precisões maiores.

A precisão utilizada nos métodos iterativos (BiCGStab e GMRES(25)) segue a seguinte sistemática: o passo inicial é determinado com o primeiro termo da sequência forçante igual a 10^{-1}; o termo seguintes da sequência forçante, para a determinação do passo, é igual ao termo imediatamente anterior dividido por 10. Testes realizados, seguindo a mesma sistemática, com valores menores para o primeiro termo da sequência forçante, não alteraram significativamente os resultados obtidos para os campos de pressão e velocidade. Porém o que se observou foi um aumento significativo do tempo computacional, bem como do número de iterações de Newton. Um comportamento análogo também foi observado ao se empregar valores maiores para o primeiro termo da sequência forçante. Observando a figura 13, nota-se um comportamento similar entre o desempenho alcançado pelas metodologias iterativas (BiCGStab e GMRES(25)) e a metodologia direta (LU esparsa), o que não era um resultado esperado. Este fato se deve ao mau comportamento (condicionamento) da matriz tangente, sendo necessária a utilização de um esquema de précondicionamento (ILU(0)) para a abordagem do sistema linear. A tabela 4.1 mostra, contudo, uma superioridade dos métodos iterativos em relação ao método direto. O método BiCGStab levou uma vantagem sobre a GMRES(25), no que diz respeito tanto ao tempo computacional, quanto ao número de iterações por passo de Newton inexato. Vale ressaltar ainda que foram feitos testes com dimensões menores do subespaço de Krylov, para o reinício do GMRES, porém ocorreu um aumento tanto do número de iterações do método de Newton inexato, bem como do tempo computacional. No caso de dimensões inferiores a dezenove, a terceira iteração de Newton inexato não converge em um limite máximo de mil iterações. Testes feitos com dimensões maiores que 25, para o reinício, mostraram um aumento considerável no tempo computacional observado.

4.3 Difusor Divergente

O problema configura-se como um escoamento incompressível em um canal cuja geometria será apresentada à posteriori, e em que os maiores detalhamentos ficam ao encargo das condições de contorno a serem denotadas. Este problema bem como o anterior tem servido como modelo para teste e avaliação de muitas técnicas numéricas.

A malha aqui utilizada é estruturada, a qual fez uso de elementos isoparamétricos quadrilaterais ("quad – four"). O problema foi resolvido para números de Reynolds iguais a 10 e 100, que são baseados nas condições de entrada do canal, sendo a geometria do

problema dependente deste número de acordo com a seguinte expressão

$$y = \frac{1}{2}\left[\tanh(2 - \frac{30x}{\mathrm{Re}}) - \tanh(2)\right], \text{ para } 0 \leq x \leq \frac{\mathrm{Re}}{3}, \tag{4.2}$$

e as componentes cartesianas do vetor velocidade na entrada do canal são:

$$u = 3(y - \frac{y^2}{2}) \text{ e } v = 0, \text{ para } x = 0 \text{ e } 0 \leq y \leq 1. \tag{4.3}$$

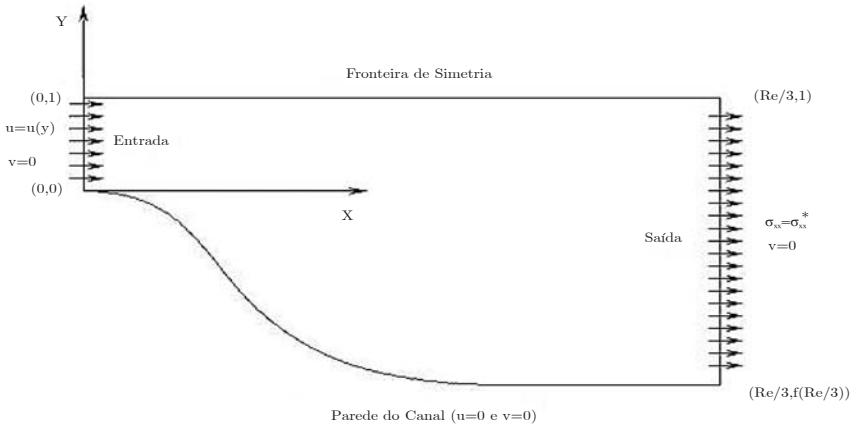

Fig. 14: Difusor divergente

4.3.1 Condições de Contorno

As condições de contorno para a equação de Navier-Stokes implicam na imposição nos contornos das variáveis principais (u – velocidade na direção x, e v – velocidade na direção y) prescritas e/ou a imposição das componentes da tensão prescrita, como se pôde observar na formulação fraca deste problema. As componentes do tensor tensão são apresentadas

da seguinte forma

$$\sigma_{ij} = -p\delta_{ij} + \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{4.4}$$

em que $u_1 = u$, $u_2 = v$, $x_1 = x$, $x_2 = y$, δ_{ij} é o delta de Kronecker, e μ é o coeficiente de viscosidade cinemática. Conseqüentemente escreve-se:

i) Condição de Saída:

$$\begin{aligned}\overline{\overline{\sigma}} \cdot \hat{n} &= [\sigma_{ij}(\hat{e}_i \otimes \hat{e}_j)] \cdot \hat{e}_k; \\ &= \sigma_{ij}\delta_{jk}\hat{e}_i; \\ &= \sigma_{ik}\hat{e}_i,\end{aligned} \tag{4.5}$$

onde $i = 1, 2$ e $k = 1$, assim:

$$\begin{aligned}\overline{\overline{\sigma}} \cdot \hat{n} &= \begin{pmatrix} -p + 2\mu\frac{\partial u}{\partial x} & \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) \\ \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & -p + \mu\frac{\partial v}{\partial y} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \\ &= \begin{pmatrix} -p + 2\mu\frac{\partial u}{\partial x} \\ \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \end{pmatrix}.\end{aligned} \tag{4.6}$$

Impõe-se então as seguintes condições nesta face:

i.1) Essencial:

$$v = 0, \tag{4.7}$$

o que implica

$$\frac{\partial v}{\partial y} + \frac{\partial u}{\partial x} = 0 \Rightarrow \frac{\partial u}{\partial x} = 0. \tag{4.8}$$

i.2) Natural:

$$\left(\overline{\overline{\sigma}} \cdot \hat{n}\right)_x = \sigma^*_{xx}. \tag{4.9}$$

ii) Condição de Simetria:

$$\begin{aligned}\overline{\overline{\sigma}} \cdot \hat{n} &= [\sigma_{ij}(\hat{e}_i \otimes \hat{e}_j)] \cdot \hat{e}_k; \\ &= \sigma_{ij}\delta_{jk}\hat{e}_i; \\ &= \sigma_{ik}\hat{e}_i,\end{aligned} \tag{4.10}$$

onde $i = 1, 2$ e $k = 2$, assim:

$$\begin{aligned}\overline{\overline{\sigma}} \cdot \hat{n} &= \begin{pmatrix} -p + 2\mu\frac{\partial u}{\partial x} & \mu\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right) \\ \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & -p + \mu\frac{\partial v}{\partial y} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \\ &= \begin{pmatrix} \mu\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ -p + \mu\frac{\partial v}{\partial y} \end{pmatrix}.\end{aligned} \tag{4.11}$$

Impõe-se então as seguintes condições nesta face:

ii.1) Essencial:

$$v = 0. \tag{4.12}$$

ii.2) Natural:

$$\left(\overline{\overline{\sigma}} \cdot \hat{n}\right)_x = 0, \tag{4.13}$$

o que implica

$$\mu \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) = 0 \Rightarrow \frac{\partial u}{\partial y} = 0. \tag{4.14}$$

4.3.2 Caso Re = 10

Visando comparar os resultados aqui obtidos com os de Perez (1987), assume-se que $\sigma^*_{xx} = 0,046$. A malha aqui utilizada pode ser visualizada na figura abaixo.

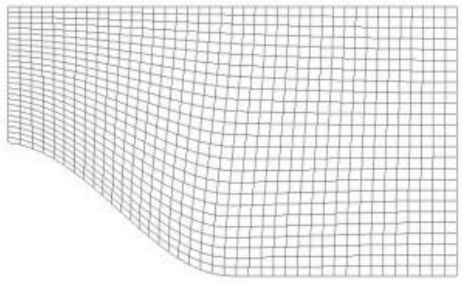

Fig. 15: Malha 25 X 40

A seguir encontram-se apresentados os campos de velocidade e pressão obtidos pelo método GMRES, uma vez que estes resultados foram idênticos aos resultados obtidos

com o método direto e com o BiCGStab.

Fig. 16: Norma Euclidiana do vetor velocidade (Re =10)

Fig. 17: Campo de pressão (Re =10)

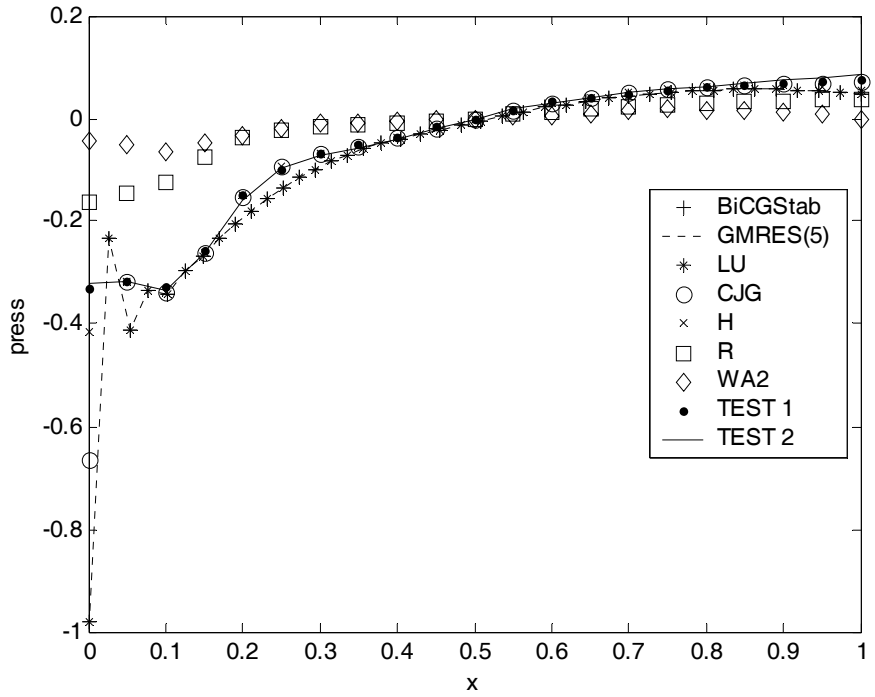

Fig. 18: Perfil do campo de pressão na parede do canal (Re = 10)

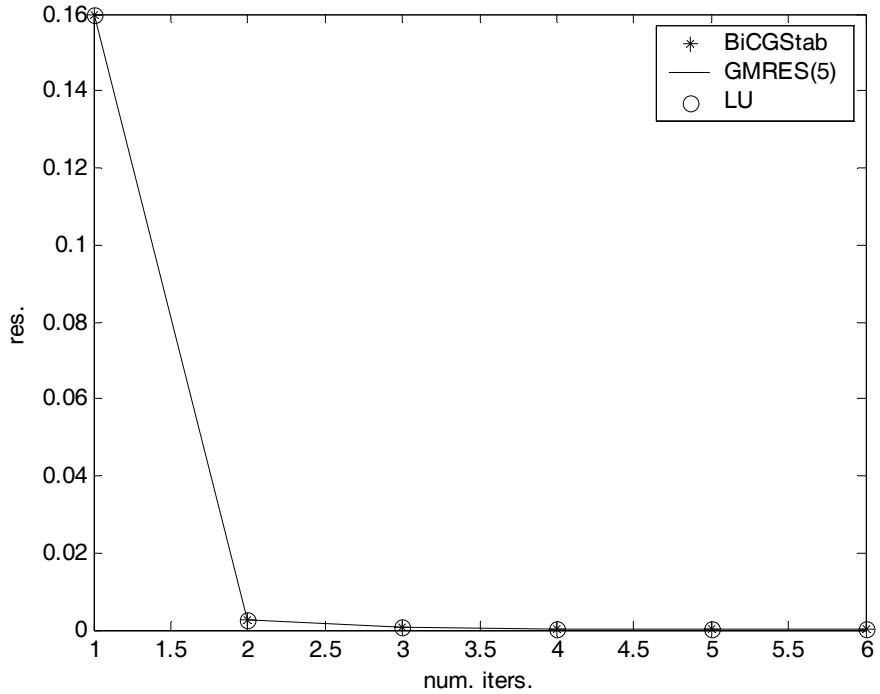

Fig. 19: Convêrgencia dos métodos

\	BiCGStab	GMRES	LU
Tempo (min.)	1.1154333	1.1017500	1.3873333
Iterações (1^o passo)	15	35	–
Iterações (2^o passo)	19	50	–
Iterações (3^o passo)	22	55	–
Iterações (4^o passo)	22	65	–
Iterações (5^o passo)	24	70	–

(4.15)

4.3.3 Caso Re = 100

Visando comparar os resultados aqui obtidos com os de Perez (1987), assume-se que $\sigma^*_{xx} = 0,019$. A malha aqui utilizada pode ser visualizada na figura abaixo.

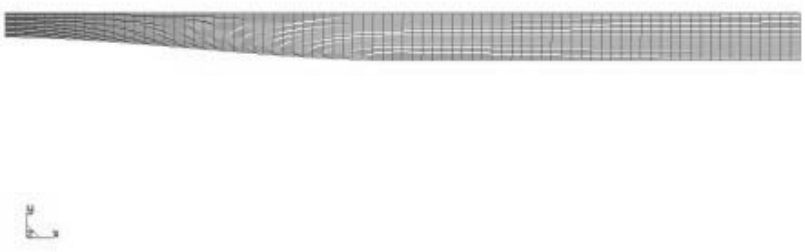

Fig. 20: Malha 25 X 80

A seguir são apresentados os campos de velocidade e pressão obtidos pelo método GMRES, uma vez que estes resultados foram idênticos aos resultados obtidos com o método direto

e com o BiCGStab.

Fig. 21: Norma Euclidiana do vetor velocidade (Re = 100)

Fig. 22: Campo de pressão (Re = 100)

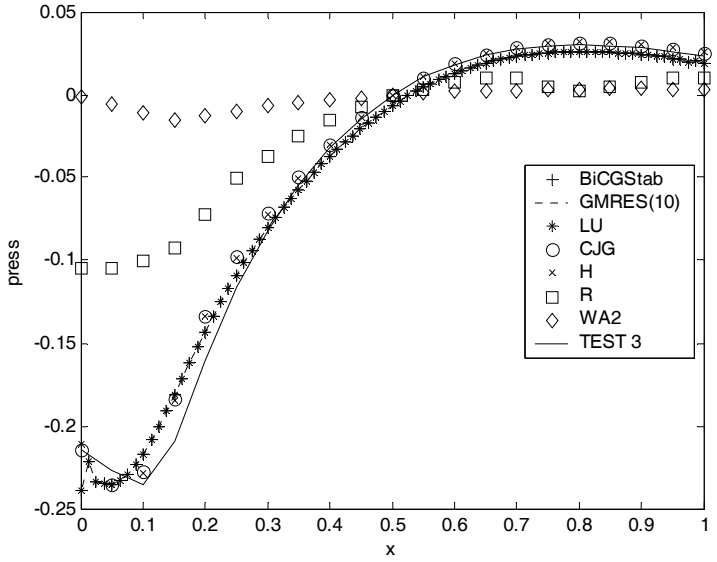

Fig. 23: Perfil do campo de pressão na parede do canal (Re = 100)

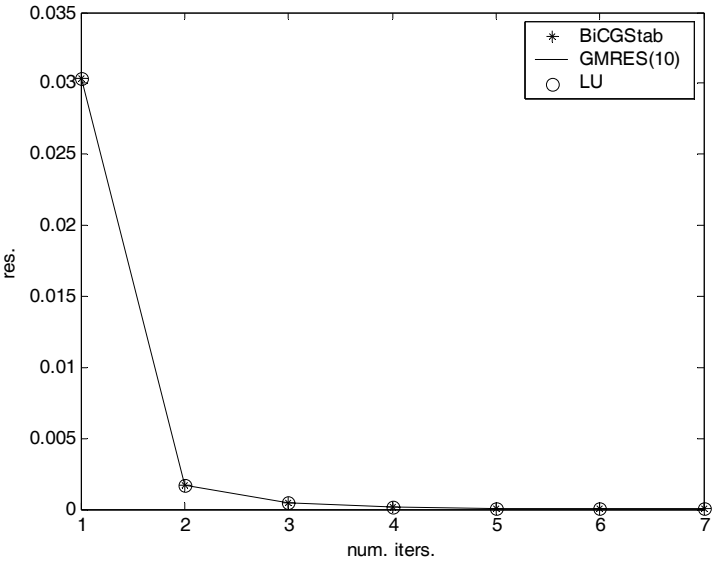

Fig. 24: Convergência dos métodos

\	BiCGStab	GMRES	LU
Tempo (min.)	4.4751000	4.3861500	5.2369167
Iterações (1^o passo)	50	370	–
Iterações (2^o passo)	31	150	–
Iterações (3^o passo)	61	200	–
Iterações (4^o passo)	58	230	–
Iterações (5^o passo)	56	270	–
Iterações (6^o passo)	56	250	–

(4.16)

As características dos métodos cujos resultados foram apresentados juntamente com os obtidos por este trabalho, são as seguintes:

- CJG, Cliffe et al. (Perez (1987)) – Método de Elementos Finitos com elementos isoparamétricos de nove pontos (quadrilateral quadrático), utilizando como variáveis primitivas u, v e P, respectivamente, as componentes do vetor velocidade nas direções x e y, e a pressão.

- H, A. G. Hutton (Perez (1987)) – Método de Elementos Finitos associado a formulação de Galerkin, com a equação da continuidade incorporada através do uso dos Multiplicadores de Lagrange. Também utiliza como variáveis primitivas u, v e P.

- R, A. K. Rasgoti (Perez (1987)) – As equações de Navier-Stokes são modeladas em coordenadas curvilíneas ortogonais por diferenças finitas. Emprega-se a integração elementar para a solução das equações de continuidade e momento, além de aproximações híbridas para os termos difusivos e convectivos. Também utiliza como variáveis primitivas u, v e P.

- WA2, A. Wada & K. Adachi (Perez (1987)) – As equações de Navier-Stokes são modeladas por diferenças finitas e em coordenadas cartesianas . Também utiliza como variáveis primitivas u, v e P.

- Test 1, Test 2, Test 3, J. O. Perez (Perez (1987)) – As equações governantes são transformadas de um sistema de coordenadas cartesianas para um sistema de coordenadas não ortogonais, gerado este, a partir da solução de um sistema elíptico de equações. O Método dos Volumes Finitos é então aplicado as equações, que são discretizadas no plano transformado. Também utiliza como variáveis primitivas u, v e P.

As figuras 15, 16, 17, 21 e 22 apresentam as características de malha e do escoamento (campos de pressão e velocidade) para ambos os casos (Re=10 e Re=100). Já as figuras 18 e 23 apresentam gráficos comparativos, para os mesmos casos, entre os resultados obtidos neste trabalho e os resultados obtidos pelas metodologias descritas anteriormente. Os gráficos mostram o perfil da pressão ao longo da parede do difusor.

Esta aplicação (Difusor Divergente) foi avaliada com a utilização do métodos direto (LU esparço), BiCGStab e GMRES, em que para o caso Re=10 utilizou-se o GMRES(5), e para o caso Re=100 utilizou-se GMRES(10), e cujos resultados comparativos encontram-se nos gráficos apresentados nas figuras 18 e 23. Observando os gráficos apresentados na figura 18, caso Re=10, nota-se que os resultados obtidos (método direto e métodos iterativos) tiveram uma melhor concordância com as metodologias CJG, H, Test 1 e Test 2. No primeiro nó o método que obteve maior proximidade foi a CJG. O caso Re=100 (fig. 23) mostra novamente um comportamento bem similar na comparação dos resultados. As metodologias que melhor concordaram com os resultados deste trabalho foram: CJG, H e Test 3.

A precisão usada para o método de Newton foi novamente de 10^{-5}, tal valor foi obtido mediante a testes com precisões menores, em que não se observou alterações relevantes nos resultados obtidos para os campos de pressão e velocidade. Esse fato não foi observado em testes com precisões maiores.

A precisão adotada nos métodos iterativos segue a mesma sistemática apresentada na apliacação anterior ("Lid Driven Cavity"), sendo que o passo inicial é determinado com o primeiro termo da sequência forçante igual a 10^{-2}, e para o caso Re=10, e igual a 10^{-1}, para o caso Re=100. Testes realizados com precisões iniciais menores não alteraram significativamente os resultados obtidos para os campos de pressão e velocidade, para ambos os casos. Porém o que se observou, em ambos, foi um aumento significativo do tempo computacional, bem como do número de iterações do método de Newton inexato. Um comportamento similar foi observado em testes com precisões iniciais maiores. Observando as figuras 19 e 24, nota-se um comportamento similar entre o desempenho alcançado pelas metodologias iterativas (BiCGStab e GMRES(5), para o caso Re=10, e BiCGStab e GMRES(10), para o caso Re=100) e a metodologia direta (LU esparsa). Esse fato é atribuido ao mau comportamento (condicionamento) da matriz tangente, sendo novamente necessária a aplicação de um esquema de precondicionamento (ILU(0)) para o sistema linear. As tabelas 4.15 e 4.16 mostram novamente uma superioridade dos métodos iterativos em relação a metodologia direta, no que diz respeito ao tempo computacional. Neste sentido o método GMRES levou uma vantagem sobre o BiCGStab, em ambos os casos abordados, embora o número de iterações por passo de Newton inexato seja bem superior. Vale ressaltar ainda que foram feitos testes, tanto para GMRES(5) como para GMRES(10), com dimensões do subespaço de Krylov menores, para o reinício, porém isso resultou num aumento tanto do número de iterações do método de Newton inexato, bem como do tempo computacional. Dimensões inferiores a quatro, no caso Re=10, fazem com que a segunda iteração de Newton inexato não convirja em um limite máximo de mil iterações. Para o caso Re=100, dimensões inferiores a oito, a terceira iteração de Newton inexato não converge, no mesmo limite máximo de iterações. Testes também foram feitos com dimensões do subespaco de Krylov maiores para o reinício, em ambos

os casos, mostraram um aumento considerável no tempo computacional observado.

Convém ainda observar que as malhas usadas nas aplicações deste ítem foram as seguintes: para o caso no qual Re=10 utilizou-se uma malha 25 x 40 de elementos quadrilaterais, enquanto que para o caso no qual Re=100 utilizou-se uma malha 25 x 80 também de elementos quadrilaterais. As malhas para ambos os casos resultaram de testes feitos com malhas mais refinadas e não foram detectadas disparidades significativas no que diz respeito aos resultados obtidos. Também foram feitos testes com outros tipos de elementos, inclusive triangulares. Preferiu-se trabalhar com elementos de baixa ordem de interpolação devido ao custo computacional elevado dos parâmetros de estabilização e também devido a largura de banda da matriz associada ao problema; aqui também não se percebeu disparidades consideráveis entre os resultados obtidos. A integração numérica foi feita com uma tablatura Gaussiana, em que utilizou-se quatro pontos de integração, no elemento mestre.

Capítulo 5

CONCLUSÃO

5.1 Conclusões

Os métodos iterativos utilizados neste trabalho (GMRES e BiCGStab) mostraram-se eficientes no tratamento do tipo de problema abordado. Pode-se observar ainda que as metodologias iterativas foram superiores a metodologia direta, no que diz respeito computacional, gerando quedas de quase um minuto.

Outro ponto relevante foi que as disparidades entre as metodologias iterativas e a metodologia direta ficam mais discrepantes quanto mais refinada é a malha. Embora as malhas finais tenham se mostrado boas frente aos testes feitos, uma análise de estimativa de erro (a posteriori) acoplado com um procedimento de refino adaptativo seria de grande utilidade para um futuro trabalho. Assim, certamente as disparidades entre as metodologias se mostrariam mais evidenciadas, tanto em tempo computacional como em taxa de convergência

5.2 Sugestões

A continuidade deste trabalho seguirá uma linha incremental, com a aplicação das ferramentas aqui estudadas, para refino adaptativo e otimização de forma em escoamentos compressíveis, turbulentos, em escoamentos de fluidos não newtonianos, assim como na área de mecânica dos sólidos.

Outras idéias a serem exploradas em trabalhos futuros são:

- A implementação/incorporação de uma metodologia sem malha *free mesh method* para a abordagem do problema,
- A incorporação do software MSC/PATRAN, nas fases de pré e pós processamento do problema,
- A implementação/incorporação de uma metodologia de estimativa de erro à porteriori, acoplada a uma estratégia de refino adaptativo.

Referências Bibliográficas

ACHDOU, Y. ; PIRONEAU, O. ; VALENTIN, F. – A Stabilazed Finite Element Method for Incompressible Navier-Stokes Equestions Satisfying Wall Laws – XX CIL-AMCE - Computational Methods in Engineering'99.

BABUŠKA, I. – The Finite Element Method with Lagrangian Multipliers - Numer. Math. 20 (1973) 179-192.

BABUŠKA, I. ; NARASIMHAN, R. – The Babuška-Brezzi Condition and the Patch Test: An Example – Comput. Methods Appl. Mech. Engrg. 140 (1997) 183-199.

BRENNER, S. C. ; SCOTT, L. R. – The Mathematical Theory of Finite Element Methods – Springer-Verlag, New York, 1996.

BASSI, F. ; REBAY, S. – A High-order Accurate Discontinuous Finite Element Method for the Numerical Solution of the Compressible Navier-Stokes Equations – J. Comput. Phy. 131 (1997), 267-279.

BREZZI, F. – On The Existence, Uniqueness and Aproximation of Saddle-Point Problems Arising from Lagrangian Multipliers – RAIRO Ser. Rouge 8 (1974) 129-151.

BREZZI, F. ; FORTIN, M. – Mixed and Hybrid Finite Element Methods – Springer-Verlag, New York, 1991.

BROOKS, A. N. ; HUGHES, T. J. R. – Streamline Up Wind / Petrov Galerkin Methods for Advection Dominated Flows – Third International Conference on Finite Element Method in Fluid Flows, Bnaff, Canadá, 1980.

BROOKS, A. N. ; HUGHES, T. J. R. – Streamline Up Wind / Petrov Galerkin Formulations for Convective Dominated Flows with Particular Enphasis on the Incompressible Navier-Stokes Equations – Comp. Methods Appl. Mech. Engrg. 32 (1982) 199-259.

BROWN, P. N. – A Theoretical Comparison of the Arnoldi and GMRES Algorithms – SIAM Journal on Scientific and Statistics Computing, vol. 12, 58-78, 1991.

BUGEDA, G. ; OÑATE, E. – Optimum Aerodynamics Shape Design Including Mesh Adapativity – International Jurnal for Numer. Meth. in Fluids 20 (1995) 915-934.

CECCHI, M. M. ; PICA, A. ; SECCO, E. – A Projection Method for Shallow Water Equations – International Jurnal for Numer. Meth. in Fluids 27 (1998) 81-95.

CHENEY, C. C. – Introduction to Approximation Theory – McGraw Hill, New York,

1996.

CODINA, R. – A Discontinuity-Capturing Crosswind-Dissipation for the Finite Element Solution of the Convection Diffusion Equation – Comp. Methods Appl. Mech. Engrg. 110 (1993) 325-342.

COCKBURN, B. ; DOWSON, C. – Some Extensions of the Local Discontinuous Galerkin Method for the Convection-Diffusion Equations in Multidimentions – The Proceedings of the Conference on the Mathematics of the Finite Elementes and Applications: MAFELAP X, Elsevier, 2000, 225-238.

COCKBURN, B. ; SHU, C. W. – The Local Discontinuous Galerkin Finite Element Method for Convection-Diffusion Systems – SIAM, J. Numer. Anal. 35 (1998), 2440-2463.

CODINA, R. – On Stabilized Finite Element Methods for Linear Sistems of Convection-Diffusion-Reaction Equations – Comp. Methods Appl. Mech. Engrg. 188 (2000) 61-82.

CODINA, R. ; BLASCO, J. – A Finite Element Formulation for the Stokes Problem Allowing Equal Velocity-Pressure Interpolation – Comp. Methods Appl. Mech. Engrg. 143 (1997) 373-391.

CODINA, R. ; BLASCO, J. – Stabilized Finite Element Method for Transient Navier-Stokes Equations Based on a Pressure Gradiente Projection – Comp. Methods Appl. Mech. Engrg. 182 (2000) 277-300.

CODINA, R. ; SOTO, O. – Finite Element Solution of the Stokes Problem with Dominating Coriolis Force – Comp. Methods Appl. Mech. Engrg. 142 (1997) 215-234.

CODINA, R. ; VÁZQUES, M. ; ZIECKIEWICZ, O. C. – A General Algorithm for Compressible and Incompressible Flows. Part III: The Semi-Implicit Form – International Jurnal for Numer. Meth. in Fluids 27 (1998) 13-32.

CODINA, R. ; VÁZQUES, M. ; ZIECKIEWICZ, O. C. ; MORGAN, K ; SATYA SAI, B. V. K. – A General Algorithm for Compressible and Incompressible Flows. Part II: Tests on Explicit Form – International Jurnal for Numer. Meth. in Fluids 20 (1995) 887-913.

CODINA, R. ZIECKIEWICZ, O. C. – A General Algorithm for Compressible and Incompressible Flows. Part I: The Split, Characteristic-Based Scheme – International Jurnal for Numer. Meth. in Fluids 20 (1995) 869-885.

DEMMEL, J. W. – Applied Numerical Linear Algebra – SIAM, Philadelphia, 1997.

DENNIS, J. E. Jr. ; SCHNABEL, R. B. – Numerical Methods for Unconstrained Optimization and Nonlinear Equations – SIAM, Philadelphia, 1996.

DHATT, G. ; TOUZOT, G. – The Finite Element Method Displayed – John Wiley & Sons, New York, 1984.

DONGARRA, J. J. ; DUFF, I. S. ; SORENSEN, D. C. ; VAN DER VORST, H. A. – Numerical Linear Algebra for High-Performance Computers – SIAM, Philadelphia, 1998.

DONGARRA, J. J. ; DUFF, I. S. ; SORENSEN, D. C. ; VAN DER VORST, H. A. – Solving Linear Systems on Vector and Shared Memory Computers. SIAM, Philadelphia,

USA,1991.

DUFF, I. S. ; ERISMAN, A. M. ; REID, J. K. – Direct Methods for Sparse Matrices – Oxford University Press Inc., New York, 1992.

FRANCA, L. P. ; FAHAT, C. ; LESOINNE, M. ; RUSSO, A. – Unusual Stabilazed Finite Element Methods and Residual Free Bubbles Form – International Jurnal for Numer. Meth. in Fluids 27 (1998) 159-168.

FRANCA, L. P. ; FREY, S. L. – Stabilazed Finite Element Methods: II. The Incompressible Navier-Stokes Equestions Interpolation – Comp. Methods Appl. Mech. Engrg. 99 (1992) 209-233.

FRANCA, L. P. ; FREY, S. L. ; HUGHES, T. J. R. – Stabilazed Finite Element Methods: I. The Advective-Difusive Model – Comp. Methods Appl. Mech. Engrg. 95 (1992) 253-276.

GHIA, U. ; GHIA, K. N. ; SHIN, C. T. – High-Re Solutions for Incompressible Flow Using the Navier-Stokes Equations and a Multigrid Method – Journal of Computational Physics 48 (1982) 387-411.

GOLUB, G. H. ; LOAN, C. F. V. – Matrix Computations – The Johns Hopkins University Press, London, 1996.

GREENBAUM, A. – Iterative Methods for Solving Linear Systems – SIAM, Philadelphia, 1997.

GURTIN, M. E. – An Introduction to Continuum Mechanics – Academic Press, New York, 1981.

HASLINGER, J. ; NEITTAANMÄKI, P. – Finite Element Approximation for Optimal Shape Design: Theory and Applications – John Wiley & Sons, New York, 1988.

HEINRICH, J. C. ; VAIONNET, C. A. – The Penalty Method for the Navier-Stokes Equations – Archieves of Computational Methods in Engineering, vol 2, 2 (1995) 51-65.

HUGHES, T. J. R. – The Finite Element Method: Linear Static and Dynamic Finite Element Analysis – Prentice-Hall Inc., New Jersey, 1987.

IDELSOHN, S. ; STORTI, M. ; NIGRO, N. – Stability Analysis of Mixed Finite Element Formulations with Special Mention of Equal-Order Interpolations – International Jurnal for Numer. Meth. in Fluids 20 (1995) 1003-1022.

KJELLGREN, P. – A Semi-Implicit Fractional Step Finite Element Method for Viscous Incompressible Flows – Computational Mechanics 20 (1997) 541-550.

KONDO, N. – Third Order Up Wind Finite Element Solutions of High Reynolds Number Flows – Comp. Methods Appl. Mech. Engrg. 122 (1994) 227-251.

KOSMA, Z. – Computing Laminar Incompressible Flows Over a Backward-Facing Step Using Newton Interations – Mechanics Research Communications, vol 27, No 2 (2000) 235-240.

LEWIS, R. W. ; RAVINDRAN, K. ; USMANI, A. S. – Finite Element Solution of Imcompressible Flows Using Segmented Approach – Archieves of Computational Methods

in Engineering, vol 2, 4 (1995) 69-93.

LIONS, P. L. – Mathematical Topics in Fluid Mechanics, vol.1: Incompressible Models – Oxford Science Publications, New York, 1996.

LIONS, P. L. – Mathematical Topics in Fluid Mechanics, vol.2: Compressible Models – Oxford Science Publications, New York, 1998.

LIU, C. H. ; LEUNG, D. Y. C. – Development of Finite Element Solution for Unsteady Navier-Stokes Equations Using Projection Method and Fractional-θ-Scheme – Comput. Methods Appl. Engrg 190 (2001) 4301-4317.

LUND, E. – Finite Element Based Design Sensitivity Analysis and Optimization – Tese de Dotorado, Universidade de Aalborg, Dinamarca, 1994.

MALVERN, L. E. – Introduction to the Mechanics of a Continuous Medium – Prentice-Hall Inc., New Jersey, 1969.

PEREZ, J. O. – Simulação Numérica de Descargas Térmicas em Corpos D'Água Rasos de Geometria e Profundidade Variáveis – Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis, Brasil, 1987.

PIRONNEAU, O. – Optimal Shape Design for Elipptic Systems – Springer-Verlag, New York, 1984.

REED, W. H. ; HILL, T. R. – Triangular Mesh Methods for the Neutron Transport Equation – Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

RIVLIN, T. J. – The Chebyshev Polynomials from Approximation Theory to Algebra and Number Theory – John Wiley and Sons, New York, 1990.

SAAD, Y. – Iterative Methods for Sparse Linear Systems – PWS Publishing Company, Boston, USA, 1996.

SAAD, Y. – Numerical Methods For Large Eigenvalue Problems – John Wiley and Sons, New York, 1992.

SAAD, Y ; SCHULTZ, M. H. –. GMRES: A Generalized Minimal Residual Method for Solving Nonsymmetric Linear

Systems. SIAM J. Sci. Stat. Comput., 7, 856-869,1986.

SONEVELD, P. – CGS, A Fast Lanczos-Type Solver for Nonsymmetric Linear Systems – SIAM Journal on Scientific and Statistics Computing, vol. 10, 36-52, 1989.

TEMAM, R. – Navier-Stokes Equations – North-Holland, Amsterdam, 1984.

VAN der VORST, H. – Bi-CGSTAB: A Fast and Smoothly Converging Variant of Bi-CG for The Solution of Nonsymmetric Linear Systems. SIAM J. Sci. Stat. Comput., 13, 631-644,1992.

WARSI, Z. U. A. – Fluid Dynamics: Theorical and Computational Approaches – CRC Press Inc., Florida, 1996.

YANG, R. J. – Finite Element Computation of Structural Design Sensitivity Analysis – Tese de Dotorado, Universidade de Iowa, Estados Unidos da América, 1984.