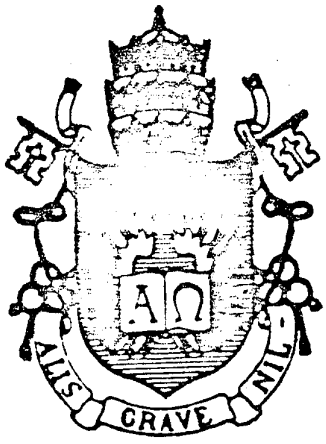

PUC

Série: Monografias em Ciência da Computação,
No. 13/89

ESTABILIDADE E CONVERGÊNCIA DO MÉTODO DA PROJEÇÃO

Fernando S. Lima

Departamento de Informática

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
RUA MARQUÊS DE SÃO VICENTE, 225 - CEP 22453
RIO DE JANEIRO - BRASIL

PUC/RJ - DEPARTAMENTO DE INFORMÁTICA

Série: Monografias em Ciência da Computação, 13/89

Editor: Paulo Augusto Silva Veloso Abril, 1989

ESTABILIDADE E CONVERGÊNCIA DO MÉTODO DA PROJEÇÃO

Fernando S. Lima

Trabalho parcialmento financiado pela FINEP

Responsável por publicações:

Rosane Teles Lins Castilho
Assessoria de Biblioteca, Documentação e Informação
PUC RIO, Departamento de Informática
Rua Marquês de São Vicente, 225 - Gávea
22453 - Rio de Janeiro, RJ
BRASIL

Tel.: (021) 529-9386 TELEX: 31078 FAX: (021) 274-4546
BITNET: userrtlc@lncc.bitnet

Abstract

The treatment of numerical stability and convergence for the projection methods at an increasing number of coordinate (test) functions is analysed. Some basic propositions on quadratic extreme problems are presented according to which the correspondent energy approach is viewed as equivalent to the Rayleigh-Ritz algorithm. Minimal systems are considered and stability conditions for the projection methods are analysed, concerning the nearly orthonormed systems, as well as relations between stability and convergence for the Finite Element Method.

Key-Words: Stability, convergence, projection methods, energy approach, Rayleigh-Ritz, minimal systems, nearly orthonormed systems, finite elements.

Resumo

O tratamento da estabilidade numérica e convergência dos métodos de projeção para um número crescente de funções coordenadas é analisado. São apresentadas algumas proposições básicas sobre problemas extremais quadráticos segundo o qual o correspondente método da energia é visto como equivalente ao algorítimo de Rayleigh-Ritz. Foram considerados os sistemas minimais e analisadas as condições de estabilidade dos métodos de projeção em relação a sistemas semiortonormados, bem como relações entre estabilidade e convergência do Método dos Elementos Finitos.

Palavras-Chave: Estabilidade, convergência, métodos de projeção, processo da energia, Rayleight-Ritz, sistemas minimais, sistemas semiortonormados, elementos finitos.

PREFÁCIO

Os problemas extremais quadráticos, baseados numa norma apropriada (norma do êrro quadrátrico, norma energética, outras) são equivalentes ao problema de resolver a tarefa da projeção. Problemas extremais quadráticos foram estudados no capítulo 1 sob o ponto de vista do método da projeção em paralelo com o método da energia na forma de Rayleigh-Ritz. A equivalência de um problema variacional com o problema da minimização da energia inerente a um sistema físico foi amplamente considerada. Como exemplo é apresentado o Método dos Elementos Finitos de uma forma simples com base no modelo de uma EDO, segundo Strang-Fix.

O principal objeto desta pesquisa é o tratamento da estabilidade numérica e convergência dos métodos de projeção para um número crescente de funções coordenadas. Outra proprieadade que foi relacionada para o estudo de estabilidade numérica é aquela que trata de sistemas minimais cuja motivação principal é o estabelecimento de condições necessárias e suficientes para que um método de projeção se torne ou seja estável.

A estabilidade do método da projeção foi estudada através do confinamento do número de condição, função da relação entre os maiores e menores autovalores da matriz do sistema considerado. Para um maior entendimento do conceito de estabilidade foi introduzida a caracterização de sistemas minimais de funções coordenadas, e as condições segundo as quais um sistema pode ser minimal, para garantir a estabilidade numérica dos métodos de projeção.

Construindo um sistema de funções coordenadas com base no método da energia, que é ortonormado em relação ao produto interno $(\ , \ )$, então ele também é semiortonormado rem relação ao produto interno energético $a(\ , \ )$, propriedades que são frequentemente usadas para demonstrar estabilidade dos métodos de projeção.

O caso ideal pode ser considerado pelos sistemas ortonormados. A existência da estabilidade numérica pode ser demonstrada pelo confinamento dos maiores e menores autovalores da matriz do sistema resultante considerado. Os sistemas submetidos a essa condição de confinamento são denominados semiortonormados em relação aos produtos internos $(\ , \ )$ e $((\ , \ ))$, fato este que pode ser verificado pelo confinamento do número de condição da matriz do sistema resultante.

Tomando por base um sistema de funções coordenadas ortonormado em relação ao produto interno $(\ , \ )$ foram analisadas as condições de estabilidade numérica dos métodos de projeção em relação ao caso do referido sistema ser semiortonormado em relação ao produto interno energético $a(\ , \ )$.

Partindo da hipótese de que existe estreita relação entre propriedades de estabilidade numérica e convergência das soluções aproximadas, propriedades de convergência (uniforme e na norma) em sistemas completos de funções coordenadas

minimais, ortonormados e também biortogonais, foram minuciosamente analisadas.

Uma aplicação especial foi tratada pela análise de estabilidade e convergência do Método dos Elementos Finitos.

Queremos aqui levar o nosso agradecimento à Dra. Therezinha Chaves[1] pela leitura do manuscrito e inúmeras sugestões. Somos gratos também ao colega Paulo Antônio Ferreira[2] pela leitura minuciosa do trabalho tendo dado valiosas sugestões e feito inúmeras correções.

Somos gratos ao CNPq pelo apoio financeiro parcial e queremos ressaltar a nossa gratidão ao Cel. Av. Dr. R. dos Santos[3] e ao Dr. A. Menezes[4], por haverem encorajado e possibilitado a publicação deste trabalho.

[1]Departamento de Informática da Pontifícia Universidade Católica do Rio de Janeiro.
[2]Divisão de Energia Nuclear do IEAv-CTA, São Joé dos Campos
[3]Diretor do IEAv - Instituto de Estudos Avançados - CTA São José dos Campos
[4]Chefe da ENU - Divisão de Energia Nuclear do IEAv - CTA

Sumário

Capítulo 1

Generalidades Sobre Alguns Métodos Computacionais

1.1 Formulações Variacionais

Seja χ um espaço linear com um subespaço W. Consideremos o mapeamento

$$P : \chi \rightarrow W. \tag{1.1}$$

Definição 1.1 *O mapeamento (1.1) chama-se projeção de χ em W se*
(i) P é linear
(ii) $P\omega = \omega$, $\forall \omega \in W$.

Um fato da análise, bastante conhecido é que sendo W e U subespaços lineares de χ, então a soma direta de W e U compreende χ, ou seja

$$\chi = W \oplus U, \tag{1.2}$$

e sendo $x \in \chi$, com $u \in U$ e $\omega \in W$, vale então

$$x = \omega + u. \tag{1.3}$$

Sejam por exemplo, dois mapeamentos P e Q tais que

$$\left.\begin{array}{l} Px = u \\ Qx = \omega \end{array}\right\} \tag{1.4}$$

onde, pela Definição 1.1, $P : \chi \to U$, $Q : \chi \to W$, são projeções. Podemos observar que P é função de Q e do operador identidade I, segundo a expressão

$$P = I - Q. \tag{1.5}$$

A expressão (1.5) é justificada pela soma de P e Q dados em (1.4), ou seja:

$$\Rightarrow (P + Q)x = u + \omega,$$

que, com (1.3) fornece

$$Ix = x,$$

de onde

$$I = P + Q \to P = I - Q.$$

1.1.1 Problemas Extremais Quadráticos

Para estudarmos e desenvolvermos alguns resultados importantes no que diz respeito à convergência e estabilidade do método da projeção, daremos a seguir alguns resultados preparatórios, sobre problemas extremais quadráticos para as tarefas lineares de valor de contorno. Seja, por exemplo, o conhecido

Teorema 1.1 (da projeção) *Seja $\chi[(,),\|.\|]$ um espaço de Hilbert real ou no mínimo um pré-Hilbert e seja W um subespaço linear fechado em relação à norma $\|.\|$. Além disso seja U o subespaço linear de todos os $u \in \chi$ com*

$$(u, \varphi) = 0, \qquad \forall \varphi \in W.$$

Então vale

$$\chi = W \oplus U,$$

isto é, cada $x \in \chi$ possui uma e somente uma representação da forma

$$x = \bar{u} + \bar{\omega}, \qquad \bar{u} \in U, \bar{\omega} \in W,$$

onde $\bar{\omega}$ é a projeção ortogonal de x sobre W, ou seja, $\bar{\omega}$ é a solução do problema extremal

$$\|\omega - x\|^2 \to min\,, \quad \omega \in W.$$

Como pela Definição 1.1, observando (1.2) e (1.3), o mapeamento projeção é linear se e só se o subconjunto W for um subespaço, resulta daí que os problemas associados com desigualdades variacionais são, em geral, não lineares. A linearidade ou não linearidade sendo reveladas pelo mapeamento

$$f \in \chi' \to \omega \in \chi$$

onde χ' é o espaço dual de χ, sendo todos os outros dados fixados de alguma forma. É óbvio que dado um espaço vetorial normado χ com uma norma $\|.\|$, uma forma bilinear $a(.,.) : \chi \times \chi \to \mathcal{R}$, uma forma linear contínua $f : \chi \to \mathcal{R}$ e um subconjunto W do espaço χ, temos reunido os ingredientes para a formulação do problema de minimização: Ache ω tal que

$$J[\omega] = \inf_{v \in W} J[v], \quad \omega \in W \tag{1.6}$$

onde o funcional

$$J : \chi \to \mathcal{R}$$

é definido pela expressão conhecida

$$J : v \in \chi \to J[v] = \frac{1}{2}a(v,v) - f(v) \tag{1.7}$$

O problema (1.6) tem solução única, desde que as seguintes hipóteses sejam satisfeitas:
(i) o espaço χ é completo,
(ii) W é fechado convexo de χ,
(iii) $a(\ ,\ )$ é simétrica e χ-elíptica no sentido de que

$$\exists \alpha > 0, \quad \forall v \in \chi', \quad \alpha\|v\|^2 \leq a(v,v).$$

O problema resultante da associação com desigualdades variacionais é linear quando minimizados sobre um subespaço. Isto se deve também a que o funcional é quadrático, ou seja, ele é da forma (1.7). A minimização de funcionais mais gerais sobre um subespaço corresponde, de maneira geral, a problemas não lineares. Suponhamos que o conjunto χ seja um espaço vetorial (não precisa ser necessariamente vetorial). Consideremos o problema (1.6) de minimização com um funcional J sob a forma

$$J[v] = F(v) - f(v),$$

com $f \in \chi'$. Há aqui dois caminhos nos quais o problema (1.6) pode tornar-se linear:

a) O funcional é quadrático

$$F(v) = \frac{1}{2}a(v,v),$$

mas o conjunto W não é um espaço vetorial, ou

b) O funcional não é quadrático.

No caso b) há tantos problemas quantos são os funcionais não quadráticos e consequentemente não dispomos de uma teoria geral. Talvez a característica mais surpreendente dos problemas não lineares, seja aquela em que suas soluções não possam ser muito suaves sobre todo o domínio considerado, até mesmo se os dados forem muito suaves. Veja o exemplo do problema da membrana onde a sua solução está em geral somente em $H^2(\Omega)$, qualquer que seja a suavidade dos dados de

$$\chi : [0,\infty] \to [0,\infty],$$

tal que

$$\lim_{t\to\infty} \chi(t) = \infty,$$

e dos dados de f e $\partial\Omega$ (Courant, Hilbert [12]).

Resolver uma tarefa de projeção é equivalente a resolver um problema extremal quadrático, sob as bases de uma norma apropriada, podendo esta ser por exemplo uma norma de erro quadrática, ou uma norma de energia. Para definir o que vem a ser problema extremal quadrático necessitamos de juntar os seguintes elementos. Sejam dados um espaço linear χ e um subspaço W; aliado a estes consideremos um funcional linear $q : \chi \to \mathcal{R}$ e também uma forma bilinear $a(,) : \chi \times \chi \to \mathcal{R}$ com as seguintes propriedades:

$$a(v,\omega) = a(\omega,v) \quad \forall v,\omega \in \chi \tag{1.8}$$

$$a(v,v) \geq 0 \qquad \forall v \in \chi \tag{1.9}$$

$$a(v,v) > 0 \qquad \forall v \in W, \quad v \neq \vec{0}. \tag{1.10}$$

Além disso muniremos χ com a seminorma definida por

$$| v |_\chi := a(v,v)^{1/2} = | v |_\chi, \tag{1.11}$$

e o subespaço W com a norma definida por

$$\| v \|_W := a(v,v)^{1/2} = | v |_W , \tag{1.11a}$$

sendo $| \, . \, |$, a norma energética em W.

Definição 1.2 *Entendemos como problema extremal quadrático aquele em que vale a seguinte descrição: Seja $v_0 \in V \subset \chi$ dado. Procuramos $v \in V \subset \chi$ como solução de*

$$J[v] = a(v,v) - 2q(v) \to \min\,, \qquad v - v_0 \in W, \tag{1.12}$$

sendo V subconjunto de χ.
Em decorrência da Definição 1.2, (Strang - Fix,[6]), podemos prosseguir objetivando compor um teorema, admitindo existência de solução para o problema (1.12). Seja por exemplo $\tilde{v}$ a solução de (1.12), ou seja, consideremos

$$J[v] \geq J[\tilde{v}] \quad , \quad para \quad v - v_0 \in W.$$

V é uma variedade linear do espaço χ, pois χ/W é o espaço quociente de χ, por $W \subset \chi$, subespaço linear de χ. O espaço χ/W é vetorial. Tomemos um $u \in W$ arbitrário com $\lambda \in \mathcal{R}$ e da mesma forma esperemos valer também

$$J[\tilde{v} + \lambda u] \geq J[\tilde{v}]. \tag{1.13}$$

Substituindo (1.11) em (1.12) obtemos

$$\begin{aligned} J[\tilde{v} + \lambda u] = a(\tilde{v} + \lambda u, \tilde{v} + \lambda u) - 2q(\tilde{v} + \lambda u) &= \\ a(\tilde{v} + \lambda u, \tilde{v} + \lambda u) - 2q(\tilde{v}) - 2\lambda q(u) &= \\ a(\tilde{v}, \tilde{v}) + \lambda a(\tilde{v}, u) + \lambda a(u, \tilde{v}) &+ \\ \lambda^2 a(u,u) - 2q(\tilde{v}) - 2\lambda q(u) &= \\ [a(\tilde{v}, \tilde{v}) - 2q(\tilde{v})] + 2\lambda[a(\tilde{v}, u) - q(u)] &+ \\ \lambda^2 a(u,u) &= \\ J[\tilde{v}] + 2\lambda[a(\tilde{v}, u) - q(u)] + \lambda^2 a(u,u) & \end{aligned} \tag{1.14}$$

Com (1.14) e (1.13) podemos escrever:

$$J[\tilde{v}] + 2\lambda[a(\tilde{v}, u) - q(u)] + \lambda^2 a(u,u) \geq J[\tilde{v}] \tag{1.15}$$

Mas (1.15) é um polinômio quadrático em λ, com $u \neq 0$ fixado. A condição de mínimo é $\frac{d(.)}{d\lambda} = 0$, e portanto temos

$$2\lambda a(u,u) + 2[a(\tilde{v}, u) - q(u)] = 0.$$

Pela hipótese (1.10) temos que

$$\lambda = 0$$

e

$$a(\tilde{v}, u) - q(u) = 0,$$

ou seja

$$a(\tilde{v}, u) = q(u), \qquad \forall u \in W. \tag{1.16}$$

Observamos aqui a equivalência do problema (1.12) com (1.16). Vejamos agora o contrário, partindo do fato de considerarmos, genericamente, a expressão variacional (1.16) como

$$a(v, \omega) = q(\omega), \qquad \forall \omega \in W \tag{1.17}$$

e de fixarmos por hipótese $v = \tilde{v}$, como sendo a solução de (1.17), tal que $v - v_0 \in W$. Assim teremos para $\omega = v$ e depois $\omega = \tilde{v}$, respectivamente :

$$a(\tilde{v}, v) = q(v) \qquad \forall v \in W \tag{1.18}$$

e

$$a(\tilde{v}, \tilde{v}) = q(\tilde{v}) \qquad \forall \tilde{v} \in W \tag{1.19}$$

e subtraindo (1.19) de (1.18) obtemos

$$a(\tilde{v}, v - \tilde{v}) = q(v - \tilde{v}) \qquad \forall (v - \tilde{v}) \in W$$

ou

$$a(\tilde{v}, \omega) = q(\omega) \quad , \quad \forall \omega \in W \Rightarrow \omega = v - \tilde{v}. \tag{1.20}$$

De (1.13) para $\lambda = 1$ temos

$$J[\tilde{v} + \omega] \geq J[\tilde{v}].$$

Mas com (1.12) resulta

$$J[\tilde{v} + \omega] = J[\tilde{v}] + a(\omega, \omega) + 2[a(\tilde{v}, \omega) - q(\omega)] \geq J[\tilde{v}]$$

e como a hipótese considerada é que $v = \tilde{v}$ é a solução de (1.17) então

$$a(\tilde{v}, \omega) - q(\omega) = 0 \quad \forall \omega \in W$$

obtemos assim

$$J[v] = J[\tilde{v} + \omega] = J[\tilde{v}] + a(\omega, \omega) \geq J[\tilde{v}], \tag{1.21}$$

isto é, $\tilde{v}$ é solução também de (1.12) (veja [5]). Da mesma forma observamos que, para $v \neq \tilde{v}$, resulta $J[v] > J[\tilde{v}]$, isto é, $\tilde{v}$ é a única solução de (1.12). Em decorrência desses resultados, apresentamos o

Teorema 1.2 *O problema extremal quadrático (1.12) possui uma única solução. Além do mais o problema extremal quadrático (1.12) é equivalente ao problema (1.17): Ache $v \in \chi$ que satisfaça a equação variacional*

$$a(v,\omega) = q(\omega) \qquad \forall \omega \in W, v - v_0 \in W.$$

A unicidade pode ser provada de outra maneira : sejam $\tilde{v}_1$ e $\tilde{v}_2$ soluções de (1.17)

$$a(\tilde{v}_1,\omega) = q(\omega) \qquad \forall \omega \in W, \tilde{v}_1 - v_0 \in W$$
$$a(\tilde{v}_2,\omega) = q(\omega) \qquad \forall \omega \in W, \tilde{v}_2 - v_0 \in W.$$

Então temos

$$a(\tilde{v}_1,\omega) = a(\tilde{v}_2,\omega) \Rightarrow \tilde{v}_1 = \tilde{v}_2.$$

As condições suficientes para que o problema (1.12), conforme os termos do teorema 1.2, tenha solução única, se apresentam efetivamente na grande maioria dos problemas variacionais equivalentes a equações diferenciais lineares, que modelam fenômenos físicos. Essas condições estão contidas no enunciado do teorema de Lax - Milgram, minuciosamente tratado por Ruas [11]. A equação (1.17) é designada na literatura como a forma variacional do problema extremal quadrático (1.12). O método do erro quadrático pode ser considerado também como método da projeção; podemos dizer que ambos os métodos se equivalem, sob certos apectos. Isso tem base no fato de que podemos munir o subespaço W com a norma do erro quadrático, quando definimos um produto interno $(\ ,\ )_L$ pela expressão

$$(v,\omega)_L = (Lv, L\omega) \tag{1.22}$$
$$\|v\|_L = (Lv, Lv)^{1/2} = \|Lv\| \tag{1.23}$$

A norma $\|.\|_L$ é chamada norma do erro quadrátrico. Observamos, entretanto, que $(\ ,\ )_L$ em W é uma forma bilinear não negativa e simétrica e por conseguinte $\|.\|_L$ pode ser no mínimo uma seminorma $|\ |_L$, conforme prescrito em (1.11) e (1.11a), para $v \in \chi$. Portanto considerando $\chi[(.,.)\ ,\|.\|]$ um espaço de Hilbert ou pré-Hilbert podemos estabelecer a equação operacional clássica

$$Lv = f, \quad v \in W. \tag{1.24}$$

Partimos da hipótese de que (1.24) possui uma solução $\tilde{v} \in W$ e a equação homogênea relativa a (1.24), ou seja,

$$Lv = 0 \qquad v \in W,$$

possui a solução trivial $v = 0$. Assim sendo a solução $\tilde{v}$ de (1.24) é unicamente determinável. A solução $\tilde{v}$ resolve, sem dúvida, o problema extremal

$$\|Lv - f\|^2 \to \min, \qquad v \in W,$$

pois sempre $\|Lv - f\| \geq 0$ e de

$$\|Lu - f\| = 0$$

resulta que $Lu = f$.
Como vimos, essas simples considerações representam ou são as bases do método do erro quadrático: Escolhamos uma base para o subespaço $W_m \subset W$, isto é tomemos $v_1, v_2, \ldots, v_n \in W$ linearmente independentes, que geram a base do subespaço W_m de dimensão m. Com isso feito, podemos em seguida considerar um $v \in W_m$ como sendo a melhor aproximação para $\tilde{v} \in W$, que resolve o problema extremal

$$\|Lv - f\|^2 \rightarrow \min . \quad , \quad v \in W_m. \tag{1.25}$$

ou seja, procuramos a melhor aproximação $v \in W_m$ que minimiza o erro quadrático. O problema extremal (1.25) é equivalente com

$$\|L(\xi_1 v_1 + \xi_2 v_2 + \ldots + \xi_n v_n) - f\|^2 \rightarrow \min, \tag{1.25a}$$

para $(\xi_1, \xi_2, \ldots \xi_n) \in \mathcal{R}^n$. A norma acima é um polinômio em $\xi_1, \xi_2, \ldots, \xi_n$ que se deixa representar na forma

$$F(x) = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j - 2\sum_{i=1}^{n} b_i\xi_i + c \tag{1.26}$$

em que

$$a_{ij} = (Lv_i, Lv_j) \quad , \quad b_i = (Lv_i, f) \quad , \quad c = (f, f).$$

A matriz a_{ij} é simétrica e a forma quadrática correspondente à mesma é positiva definida:

$$Q(x) = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j > 0, \quad x \neq 0, \tag{1.27}$$

onde $x = (\xi_1, \xi_2, \ldots, \xi_n)$ é a variável independente do polinômio quadrático F no $\mathcal{R}^n$. De (1.27) é sabido que a matriz dos elementos a_{ij} tem posto maximal e que então o sistema linear

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \quad , \quad i = 1(1)n^1 \tag{1.28}$$

[1] $i = 1(1)n$ é o mesmo que $i = 1, 2, \ldots, n$.

tem solução única. Seja por exemplo $\tilde{x} = (\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n)$ essa solução e fazendo entretanto $y = x - \tilde{x}$ obtemos

$$\begin{aligned} Q(y) = Q(x - \tilde{x}) = Q(x) - Q(\tilde{x}) = \\ \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j - \sum_{i,j=1}^{n} a_{ij}\tilde{\xi}_i\tilde{\xi}_j \\ = \sum_{i,j=1}^{n} a_{ij}\left(\xi_i\xi_j - \tilde{\xi}_i\tilde{\xi}_j\right) \end{aligned} \tag{1.29}$$

Mas para $\tilde{x} = (\tilde{\xi}_1\tilde{\xi}_2, \ldots, \tilde{\xi}_n)$ temos

$$\begin{aligned} F(\tilde{x}) = \sum_{i,j=1}^{n} a_{ij}\tilde{\xi}_i\tilde{\xi}_j - 2\sum_{i=1}^{n} b_i\tilde{\xi}_i + c \\ = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}\tilde{\xi}_i\tilde{\xi}_j - 2\sum_{i=1}^{n} b_i\tilde{\xi}_i + c \\ = \sum_{i=1}^{n}\tilde{\xi}_i\sum_{j=1}^{n} a_{ij}\tilde{\xi}_j - 2\sum_{i=1}^{n} b_i\tilde{\xi}_i + c \\ = \sum_{i=1}^{n} b_i\tilde{\xi}_i - 2\sum_{i=1}^{n} b_i\tilde{\xi}_i + c \end{aligned} \tag{1.30}$$

$$\Rightarrow F(\tilde{x}) = c - \sum_{i=1}^{n} b_i\tilde{\xi}_i, \tag{1.31}$$

tendo usado a expressão (1.28) em (1.30). Mas a expressão genérica de $F(y)$ é dada por

$$F(x) - F(\tilde{x}) = \sum_{i,j=1}^{n} a_{ij}\left(\xi_i\xi_j - \tilde{\xi}_i\tilde{\xi}_j\right) - 2\sum_{i=1}^{n} b_i\xi_i + 2\sum_{i=1}^{m} b_i\tilde{\xi}_i.$$

Por (1.28), aproveitando a simetria de Q, levando em conta que $\sum_{j=1}^{n} a_{ij}\xi_j = \sum_{j=1}^{n} a_{ij}\tilde{\xi}_j = b_i$ obtemos:

$$-2\sum_{i=1}^{n} b_i\xi_i + 2\sum_{i=1}^{n} b_i\xi_i = -2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\tilde{\xi}_j\xi_i + 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\xi_i\tilde{\xi}_j = 0,$$

o que nos leva à expressão

$$F(y) = F(x) - F(\tilde{x}) = \sum_{i,j=1}^{n} a_{ij}\left(\xi_i\xi_j - \tilde{\xi}_i\tilde{\xi}_j\right),$$

que por (1.29) fornece então

$$\begin{aligned} & F(x) - F(\tilde{x}) = Q(y) \\ \Rightarrow\ & F(x) = F(\tilde{x}) + Q(y) \\ \Rightarrow\ & F(\tilde{x}) < F(x) \quad , \quad para \quad x \neq \tilde{x}. \end{aligned} \tag{1.32}$$

E assim havemos conseguido a formulação do

Teorema 1.3 *Sob as hipóteses (1.26) e (1.28) , o polinômio $F(x)$ possui um mínimo absoluto. Isto significa que existe um e somente um $\tilde{x} \in \mathcal{R}^n$ de tal forma definido que*

$$F(\tilde{x}) < F(x) \qquad \forall x \in \mathcal{R}^n, \qquad x \neq \tilde{x}.$$

Da mesma forma

$$\tilde{x} = \left(\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_n\right) \in \mathcal{R}^n$$

é a única solução do sistema de equações lineares

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \quad , \quad i = 1(1)n.$$

Para o mínimo resulta (1.31):

$$F(\tilde{x}) = c - \sum_{i=1}^{n} b_i \tilde{\xi}_i.$$

Seja χ um espaço linear com produto interno $(,)$ e norma $\| \ \|$. E seja $W \subset \chi$ um subespaço linear de χ. O espaço χ pode ser de dimensão infinita mas W é de dimensão finita.

Consideremos em seguida o problema: Seja $v \in \chi$ elemento qualquer dado. Procuremos $\omega \in W$ que tenha distância mínima em relação a v, solução do problema extremal:

$$\|\omega - v\|^2 \rightarrow min \quad , \quad para \quad \omega \in W. \tag{1.33}$$

Estabelecendo $\omega_1, \omega_2, \ldots, \omega_n$ como uma base de W, então o problema extremal (1.33) é equivalente ao problema :

$$\|\xi_1\omega_1 + \xi_2\omega_2 + \ldots + \xi_n\omega_n - v\|^2 \rightarrow \min, \tag{1.34}$$

com $(\xi_1, \ldots, \xi_n) \in \mathcal{R}^n$, onde o membro esquerdo é um polinômio quadrático em $x = (\xi_1, \ldots, \xi_n)$ e pode ser escrito sob a forma (1.26), ou seja,

$$F(x) = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j - 2\sum_{i=1}^{n} b_i\xi_i + c,$$

onde, no caso,

$$a_{ij} = (\omega_i, \omega_j) \quad , \quad b_i = (\omega_i, v) \quad , \quad c = (v, v)$$

e sem dúvida $a_{ij} = a_{ji}$, bem como

$$Q(x) = \|\xi_1\omega_1 + \ldots + \xi_n\omega_n\|^2 > 0, \tag{1.35}$$

por causa da independência linear dos ω_i para $x \neq 0$. Podemos ver que as hipóteses do Teorema 1.3 estão aqui certamente preenchidas e assim obtivemos a formulação do

Teorema 1.4 *O problema extremal (1.34) possui exatamente uma solução, que resolve, da mesma forma, o sistema de equações lineares*

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \quad , \quad i = 1(1)n \tag{1.36}$$

com

$$a_{ij} = (\omega_i, \omega_j) \quad , \quad b_i = (\omega_i, v).$$

Se os $\omega_1, \omega_2, \ldots, \omega_n$ geram uma base ortornormal de W então

$$a_{ij} = (\omega_i, \omega_j) = \delta_{ij}$$

e assim obtemos

$$\xi_i = (\omega_i, v).$$

Nesse caso, a solução procurada $\tilde{\omega}$ será dada por

$$\tilde{\omega} = \sum_{i=1}^{n} (\omega_i, v)\omega_i.$$

Quando o resultado do Teorema 1.4 for obtido independentemente da escolha de uma base especial, então podemos formular o

Corolário 1.1 *O problema extremal*

$$\|\omega - v\|^2 \to \min \quad , \quad \omega \in W$$

possui exatamente uma solução $\tilde{\omega} \in W$. Ela é, simultaneamente, a solução única em W da equação funcional

$$f(\varphi) = B(\omega, \varphi) = (v, \varphi) \qquad \forall \varphi \in W. \tag{1.37}$$

Introduzindo agora o espaço U (geralmente de dimensão infinita) junto ao espaço W de dimensão finita, podemos munir U com a propriedade

$$(u, \varphi) = 0, \quad \forall \varphi \in W$$

para todo $u \in \chi$. Assim W e U são portanto subespaços ortogonais: $\chi = W \oplus U$, ou seja $\forall z \in \chi$ se deixa representar unicamente por

$$z = \omega + u, \qquad \omega \in W, \qquad u \in U, \tag{1.38}$$

conforme Teorema 1.1.
Isso resulta da observação de que para toda decomposição (1.38), caso ela exista, a igualdade (1.37) seja válida. Mas a equação (1.37) possui exatamente uma solução $\tilde{\omega} \in W$. Se fizermos

$$\tilde{u} = z - \tilde{\omega}$$

então temos $\tilde{u} \in U$ e $z = \tilde{u} + \tilde{\omega}$ é a decomposição procurada. A solução $\tilde{\omega} \in W$ de (1.37) e o elemento $\tilde{u} = z - \tilde{\omega} \in U$ são portanto as projeções ortogonais de z sobre os subespaços ortogonais W e U. Daqui segue como resultado, o

Corolário 1.2 *A solução $\tilde{\omega}$ do problema extremal*

$$\|\omega - z\|^2 \to min, \qquad \omega \in W$$

é a Projeção Ortogonal de z sobre o subespaço de dimensão finita W.

Podemos também formular problemas em que temos de considerar casos de ocorrência só de produto interno semidefinido, com a correspondente seminorma. Nesse caso a unicidade da solução do problema (1.33) falha. Se u é um elemento de χ com $\|u\| = 0$ e se $\omega \in \chi$ é um elemento qualquer, então sempre se verifica a igualdade

$$\|u + \omega\|^2 = \|\omega\|^2,$$

pois $\|u\|^2 = 0$, $(u,\omega) = (\omega, u) = 0.$, Portanto, podemos adicionar um u qualquer a uma solução, com $\|u\| = 0$.

A nossa indagação agora se resume no seguinte. Quais as modificações que devemos proceder no Teorema 1.4, considerando o caso de ocorrência de produto interno semidefinido e sua correlata seminorma? Ora, quando $\|.\|$ for somente uma seminorma então os elementos $u \in \chi$ compoem, com $\mid u \mid = \| u \| = 0$, um subespaço linear $\chi_0 \subset \chi$ que não somente é formado de elemento nulo. Trata-se aqui, nesse caso, de termos a evidência do espaço quociente $\hat{\chi} = \frac{\chi}{\chi_0}$, cujos elementos se compõem de classes $\hat{\omega}$ de elementos χ que se diferenciam apenas por um elemento de χ_0. Portanto dois elementos ω e ω' pertencem a uma mesma classe $\hat{\omega}$, se e somente se $\omega - \omega' \in \chi_0$. Este é exatamente o caso quando

$$\|\omega - \omega'\| = 0.$$

Podemos assim introduzir um produto interno e uma norma em $\hat{\chi}$ através de

$$(\hat{\omega}, \hat{u}) = (\omega, u), \qquad \|\hat{\omega}\| = \|\omega\|,$$

onde ω e u são representantes quaisquer tirados das classes $\hat{\omega}$ e $\hat{u}$ respectivamente. Se $\hat{z}$ é a classe a que z pertence e se $\hat{W}$ é subespaço de dimensão finita de $\hat{\chi}$, então o problema extremal

$$\|\hat{\omega} - \hat{z}\|^2 \to \min \quad , \quad \hat{\omega} \in \hat{W}$$

segundo o Corolário 1.1 possui exatamente uma solução que simultaneamente é solução do problema

$$(\hat{\omega}, \hat{\varphi}) = (\hat{z}, \hat{\varphi}) \qquad \forall \hat{\varphi} \in \hat{W}.$$

E com esse resultado temos construído o

Teorema 1.5 *Seja* $\| \, \|$ *uma seminorma em* χ *e seja* W *um subespaço com a base* $\omega_1, \omega_2, \ldots, \omega_n \in \chi$. *Então o problema extremal*

$$\| \omega - z \|^2 \to \min \quad , \quad \omega \in W,$$

é equivalente ao sistema de equações lineares

$$\sum_{j=1}^{n} (\omega_i, \omega_j) \xi_j = (\omega_i, z), \qquad i = 1(1)n,$$

que é solúvel.

O método da projeção tem também a haver com o método da energia sob a forma da versão de Rayleigh-Ritz. O método da energia se caracteriza pelos

ingredientes $\chi, (,), \| \ \|$, para formalizar o espaço de Hilbert ou pré-Hilbert no qual existe um operador simétrico e positivo com a seguinte ocupação:

$$L : W \to \chi,$$

com as propriedades

$$(L\omega, u) = (\omega, Lu) \qquad \forall u, \omega \in W \tag{1.39}$$

$$(L\omega, \omega) > 0 \qquad \forall \omega \in W, \omega \neq 0. \tag{1.40}$$

Seja o problema linear

$$L\omega = f \quad , \quad \omega \in W \tag{1.41}$$

e admitamos a solução $\tilde{\omega} \in W$. De (1.40) segue que $L\omega = 0$ em W só admite a solução $\omega = 0$ e daí $\tilde{\omega} \in W$ é a solução de (1.41) únicamente definida.
Em virtude de (1.39) e (1.40) podemos definir através de

$$((\omega, u))_L = (L\omega, u), \quad |\omega|_L = (L\omega, \omega)^{1/2} \tag{1.42}$$

produto interno energético e norma energética respectivamente, em W. Temos o seguinte conhecido teorema [5,6]:

Teorema 1.6 *A equação operacional (1.41) é equivalente ao problema extremal*

$$I[\omega] \to \min \quad , \quad \omega \in W, \tag{1.43}$$

onde

$$I[\omega] = (L\omega, \omega) - 2(f, \omega). \tag{1.44}$$

Daremos, em seguida, o prova do Teorema 1.6.
A caracterização da solução $\tilde{\omega}$ do problema (1.43) é conseguida com a aplicação do funcional energético (1.44) e o método correspondente se denomina método da energia. O processo de Rayleigh-Ritz se baseia no Teorema 1.6 para o cálculo aproximado da solução $\tilde{\omega}$: Considera a melhor aproximação como a solução do problema extremal

$$I[\omega] \to \min \quad , \quad \omega \in W_n,$$

que é equivalente ao problema aproximado

$$I[\xi_1\omega_1 + \xi_2\omega_2 + \ldots + \xi_n\omega_n] \to \min, \quad (\xi_1, \ldots, \xi_n) \in \mathcal{R}^n.$$

O membro esquerdo é um polinômio quadrático em $\xi_1, \ldots, \xi_n$ que pode ser escrito na forma (1.26), com os coeficientes

$$a_{ij} = (L\omega_i, \omega_j), \quad b_i = (\omega_i, f), \qquad c = 0.$$

Em decorrência da independência linear dos $\omega_1, \ldots, \omega_n$ a forma quadrática

$$|\xi_1\omega_1 + \ldots + \xi_n\omega_n|_L = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \tag{1.45}$$

relativa à matriz a_{ij} é positiva definida. Dai, observando o Teorema 1.3 e suas hipóteses, resulta provado o

Teorema 1.7 *Sejam $\omega_1, \ldots, \omega_n \in W$ linearmente independentes. Então o problema extremal*

$$I[\omega] \to \min \quad \omega \in W_n \tag{1.46}$$

é equivalente ao sistema de equações lineares

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i\ ,\ i = 1(1)n, \tag{1.46a}$$

com

$$a_{ij} = (Lu_i, u_j) \qquad ,\quad b_i = (u_i, f) \tag{1.46b}$$

que tem exatamente uma solução.

Isto significa que a melhor aproximação obtida pelo processo de Rayleigh–Ritz pode ser compreendida como a obtida também pelo método da Projeção. Ou seja, o método de Rayleigh-Ritz é nada mais nada menos que o método da projeção. Esse fato tem base na prova do Teorema 1.6, sob nossa versã. Detalhemos, a seguir, o andamento: consideremos $\tilde{\omega}$ solução de (1.41). Com base em (1.42) escrevemos

$$|\omega - \tilde{\omega}|_L^2 = (L\omega - L\tilde{\omega}, \omega - \tilde{\omega}) = |\omega|_L^2 - 2(L\tilde{\omega}, \omega) + |\tilde{\omega}|_L^2$$

e como $L\tilde{\omega} = f$ vem

$$|\omega - \tilde{\omega}|_L^2 = |\omega|_L^2 - 2(f, \omega) + |\omega|_L^2,$$

e obtemos assim a expressão

$$|\omega - \tilde{\omega}|_L^2 - |\tilde{\omega}|_L^2 = I[\omega]. \tag{1.47}$$

Mas ainda com (1.42) e (1.44), observando que $L\tilde{\omega} = f$, temos

$$I[\tilde{\omega}] = (L\tilde{\omega}, \tilde{\omega}) - 2(f, \tilde{\omega}) = (L\tilde{\omega}, \tilde{\omega}) - 2(L\tilde{\omega}, \omega)$$
$$\Rightarrow I[\tilde{\omega}] = -(L\tilde{\omega}, \tilde{\omega}) = -|\tilde{\omega}|_L^2.$$

E portanto (1.45) tem agora a forma:

$$|\omega - \tilde{\omega}|_L^2 + I[\tilde{\omega}] = I[\omega].$$

Assim, para $\omega \neq \tilde{\omega}$, obtemos a prova final do Teorema 1.6:

$$I[\omega] > I[\tilde{\omega}] \Rightarrow |\omega - \tilde{\omega}|_L^2 \rightarrow \min.$$

Mas (1.46) justifica também o resultado condensado no

Teorema 1.8 *O problema extremal (1.45) é equivalente ao problema*

$$|\omega - \tilde{\omega}|_L^2 \rightarrow \min \quad , \quad \omega \in W_n. \tag{1.48}$$

A melhor aproximação de $\tilde{\omega}$ no sentido do processo de Rayleigh-Ritz é a projeção ortogonal de $\tilde{\omega}$ sobre o subespaço W_n de dimensão finita n; ortogonal no sentido do produto interno energético $((,))$ em (1.42). As equações (1.46a) com coeficientes (1.46b) são conhecidas como equações de Ritz e Galerkin na literatura.

É nosso objetivo agora elaborar uma rápida descrição sobre o método de Rayleigh-Ritz para justificar o fato de que a melhor aproximação de $\tilde{\omega}$ é a projeção ortogonal de $\tilde{\omega}$ sobre a variedade linear $\omega_0 + W_n$ de dimensão n, no sentido do produto interno $a(,)$.

Fizemos análise e considerações sobre o funcional linear $q : \chi \rightarrow \mathcal{R}$ e sobre a forma bilinear $a(,) : \quad \chi \times \chi \rightarrow \mathcal{R}$ com as propriedades (1.8) , (1.9) e (1.10). Essas propriedades não são suficientes sozinhas para constituirem hipóteses para prova de existência de solução $\tilde{\omega}$ de (1.12). Aqui queremos partir da suposição de que a solução $\tilde{\omega}$ de (1.12) existe, para estudarmos o método do Rayleith-Ritz, na forma do problema extremal

$$J[\omega] = a(\omega,\omega) - 2q(\omega) \rightarrow \min \quad , \quad \omega - \omega_0 \in W. \tag{1.49}$$

Escolhamos elementos $\omega_1, \omega_2, \ldots \omega_n \in W$ linearmente independentes que gerem o subespaço $W_n \subset W$ de dimensão n. Consideremos como melhor aproximação $\tilde{\omega}$, a solução do problema extremal

$$J[\omega] \rightarrow \min \qquad \omega - \omega_0 \in W_n, \tag{1.50}$$

que é equivalente a

$$J[\omega_0 + \xi_1\omega_1 + \ldots + \xi_n\omega_n] \rightarrow \min, \quad (\xi_1, \ldots, \xi_n) \in \mathcal{R}^n. \tag{1.51}$$

Vemos que o membro esquerdo de (1.51) é normalmente um polinômio quadrático em $\xi_1, \xi_2, \ldots, \xi_n$ que pode ser expressso pela forma (1.26) com os coeficientes

$$a_{ij} = a(\omega_i, \omega_j) \quad , \quad b_i = q(\omega_i) - a(\omega_i, \omega_0)$$

e

$$c = a(\omega_0, \omega_0) - 2q(\omega_0).$$

A matriz (a_{ij}) é simétrica por causa de (1.8) e a sua forma quadrática é

$$| \xi_1\omega_1 + \ldots + \xi_n\omega_n |^2 = \sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j,$$

em vista de (1.10), positiva definida. Partindo do Teorema 1.3 resulta o

Teorema 1.9 *Sejam $\omega_1, \omega_2, \ldots \omega_n \in W$ linearmente independentes. Então o problema extremal (1.49) tem solução única e é equivalente ao sistema de equações lineares*

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \qquad i = 1(1)n \tag{1.52}$$

com

$$a_{ij} = a(\omega_i, \omega_j) \quad , \quad b_i = q(\omega_i) - a(\omega_i, \omega_0). \tag{1.53}$$

Seja $\tilde{\omega}$ a solução de (1.49) e seja ω um elemento qualquer com $\omega - \omega_0 \in W$. Então $\varphi = \omega - \tilde{\omega} \in W$, e assim obtemos com a aplicação de $a(\tilde{\omega}, \varphi) = q(\varphi)$, a equação

$$J[\omega] = | \omega - \tilde{\omega} |^2 + J[\tilde{\omega}]$$

como em (1.46), por analogia, tendo em vista que $J[\tilde{\omega}] = -| \tilde{\omega} |^2$. Portanto analogamente ao Teorema 1.8 obtemos o

Teorema 1.10 *O problema extremal (1.49) é equivalente ao problema*

$$| \omega - \tilde{\omega} |^2 \to \min \quad \omega - \omega_0 \in W_n.$$

Quer dizer, a melhor aproximação de $\tilde{\omega}$ no sentido do método de Rayleigh-Ritz é a **projeção ortogonal** de $\tilde{\omega}$ sobre a variedade linear $\omega_0 + W_n$ de dimensão n como queríamos demonstrar.

1.1.2 Método dos Elementos Finitos

Os elementos finitos se transformaram mais num método sofisticado propriamente dito, do que simplesmente técnica, porque a sua elaboração matemática tem conseguido um desenvolvimento analítico muito vasto e os elementos têm sido descobertos e aplicados em problemas os mais variados e complexos da matemática computacional ou mecânica computacional aplicada. Aqui vamos tratar do aspecto mais analítico de problemas lineares elípticos e aproximação pelos elementos finitos, sob o ponto de vista do desenvolvimento de Ritz, por questões de simplicidade.

O método dos elementos finitos é um método variacional importante e prático que leva um determinado problema linear à resolução numérica de um sistema de equações lineares. Ele é na realidade o desenvolvimento numérico do método de Ritz representado por problemas extremais

$$J[\omega] = a(\omega,\omega) - 2q(\omega) \rightarrow \min\ ,\ \omega - \omega_0 \in W, \tag{1.54}$$

em que entretanto as funções $\omega_1,\omega_2,\ldots,\omega_n$ são escolhidas de maneira especial: cada uma das ω_i é uma função que toma valores diferentes de zero, somente sobre um domínio parcial $\Omega_i \subset \Omega$ e fora deste ela se anula identicamente. Para cada dois subdomínios Ω_i e Ω_j, em que $\Omega_i \bigcap \Omega_j = \emptyset$, obtemos

$$a_{ij} = a(\omega_i,\omega_j) = 0.$$

Se u é a solução do problema (1.54) então $J[W] \rightarrow \min$ para $\omega = u$, com ω satisfazendo certas condições inerentes ao contorno de Ω. Desta forma conseguimos que a matriz dos coeficientes a_{ij} seja fracamente ocupada, ou seja, de forma esparsa mostrando muitos zeros. Este fato é desejável, pois dele depende a estabilidade do método (veja p.ex. Strang [6]).

Aproveitamo-nos também de uma outra especialidade: num problema de valor de contorno de ordem $2m$, W é um subespaço de $H^m(\Omega)$. Aqui também são especiais as funções de $E_c^m(\bar{\Omega})$, sendo este o espaço de funções em $C^{k-1}(\bar{\Omega})$ com k-ésima derivada em $E_c(\bar{\Omega})$ espaço de funções espaçadamente contínuas em $\bar{\Omega}$.

Suponhamos que o problema a ser resolvido seja dado sob forma variacional, ou seja, achar uma função u que minimize uma dada expressão de energia potencial. A propriedade minimizante leva a uma equação diferencial em u, a equação de Euler. Entretanto, normalmente é impossível conseguirmos a solução exata, e assim alguma aproximação é necessária. A idéia de Rayleigh-Ritz-(Galerkin) é escolher um número finito de funções-testes $\omega_1,\omega_2,\ldots,\omega_n$ e dentre todas as combinações lineares $\sum q_j\omega_j$ selecionar uma que seja a minimizante: essa é a aproximação de Ritz. Os pesos desconhecidos q_j são determinados não diretamente da discretização

da equação diferencial mas sim pela minimização do correspondente funcional que leva à resolução numérica de um sistema de equações lineares discretas e algébricas.

A justificativa teórica para este método é simples. O processo de minimização busca a combinação que está mais próxima de u, isto é:
$\|u_n - u\| \to 0 \quad , \quad n \to \infty$.

Por conseguinte o objetivo do método é escolher as funções ω_j que sejam convenientes o bastante para que a energia potencial seja calculada e minimizada, e ao mesmo tempo aproximar o mais possível a solução u. Mas a maior dificuldade está na tarefa de processar ω_j que sejam convenientes o bastante, para levar a energia potencial ao mínimo e alcançar computabilidade ou maleabilidade computacional. Na teoria da aproximação sabemos que sempre existe um conjunto de funções ω_j que é completo (suas combinações lineares preenchem o espaço de todas as possíveis soluções u com $n \to \infty$ e com isto as aproximações de Ritz convergem). Mas ser capaz de computar adequadamente com essas funções ω, não é tarefa tão simples. Mas a aproximação é conseguida com facilidade, usando o processo dos elementos-finitos. (Aziz [1], Strang [6], Zienkiewiez [7]).

O processo ou método dos elementos finitos tem conseguido resolver a tarefa da escolha conveniente de ω_j e da computabilidade adequada, com grande e incomparável desempenho, incluindo domínios da mais alta complexidade e com as mais variadas condições de contorno, em 2D e 3D. Uma das ferramentas mais poderosas dos elementos finitos é a utilização de malhadores automáticos e sofisticados, desenvolvidos para estudos e projetos industriais de alto nível e para pesquisa científica de ponta.

A idéia básica é simples. Primeiro subdividimos a região de interesse físico em pedaços ou subdomínios denominados elementos. Podem ser triângulos, ou quadriláteros. Dentro de cada elemento damos às funções testes as formas mais simples, como polinômios de $1^{\underline{o}}$, $2^{\underline{o}}$, ou $3^{\underline{o}}$ grau. As condições de contorno são infinitamente mais fáceis de serem impostas localmente, ao longo do lado de um triângulo ou retângulo, do que globalmente ao longo de um contorno complicado. A precisão da aproximação pode ser aumentada, se for necessário, mas não pelo método clássico de Ritz em que se incluem mais e mais complexas funções-teste. Invés disso o mesmo polinômio é mantido e refinamos a malha de subdivisões. Isso se liga ao fato de que um sistema de elementos finitos de larga escala pode usar o potencial ou capacidade de um computador para a formulação das equações de aproximação, bem como para calcular as soluções a um grau nunca alcançado antes em problemas físicos complicados. Desenvolvimento posterior de elementos mais precisos levam ao conhecimento de que, aumentando-se o grau do polinômio aproximante melhora-se o grau de precisão na aproximação, e os parâmetros desconhecidos q_j computados na aproximação discreta sempre mantém um significado físico. Nesse caso a saída

dos dados no computador é mais fácil de interpretar do que os pesos produzidos pelo método clássico.

O procedimento, na sua forma global, tornou-se matematicamente respeitável no momento em que foi descoberto ser o método dos elementos finitos o resultado da aplicação do método de Ritz, justamente quando as incógnitas q_j foram identificadas como sendo os coeficientes na aproximação de Ritz

$$u \approx \sum q_j \omega_j$$

e as equações discretas eram exatamente as condições de minimização da energia potencial. Com essa descoberta iniciou-se então o fornecimento de uma base matemática sólida para o método: Strang e Fix, Bramble, Ciarlet, Clough, Courant, Argyris, Zienkiewicz e outros.

O problema básico é descobrir como polinômios seccionais podem aproximar o máximo possível uma solução u. Significa determinar quão bem os elementos finitos, que vêm sendo desenvolvidos na base da simplicidade computacional, satisfazem às exigências, de que funções-teste possam vir a ser efetivas na aproximação. Estimar o erro de aproximação ou determinar a velocidade de diminuição desse erro acontece em geral com o refinamento da malha ou com aumento do grau do polinômio em cada elemento [6].

As idéias matemáticas chaves para o método se baseiam nos espaços de Hilbert H^k e suas normas, nas estimativas da solução em termos dos dados e no produto interno energético que está naturalmente associado ao problema específico. Com estas ferramentas a convergência do método dos elementos finitos pode ser provada mesmo para problemas de geometrias complicadas. De fato, a simplicidade dos argumentos variacionais permite uma análise que já vai muito além daquilo que as diferenças finitas podem alcançar ou oferecer.

Na teoria dos problemas de valor de contorno seja dado o problema (em uma dimensão):

$$Lu = f \quad , \quad x \in [a,b] \subset \mathcal{R}$$

onde L é um operador linear que mapeia u em f. O operador L é tal que age sobre uma determinada classe de funções que em certo sentido satisfazem às condições de contorno e podem ser diferenciadas, por exemplo, até $k = 2m$. A questão fundamental é arranjar um espaço de funções u com uma classe de termos não homogêneos f de tal maneira que a cada f corresponda uma e somente uma solução u. Uma vez que essa correspondência entre u e f esteja estabelecida, o problema $Lu = f$ está resolvido no sentido abstrato. Para fixarmos o aspecto da correspondência entre u e f na equação dada temos que estabelecer critérios de validade seguindo propriedades da análise funcional, influentes na construção dos espaços. Em que espaços, para u

e f, vale essa correspondência? A escolha específica para os dados não homogêneos é feita admitindo que os mesmos possuam energia finita, traduzida por

$$\int_\Omega f^2 dx < \infty,$$

no sentido da integração de Lebesgue. Isso vale também para uma função espaçadamente suave, mas não para o $\delta - Dirac$, por exemplo. O espaço para u é o espaço de Hilbert H^{2m} que é mapeado por L no espaço H^0, ou seja, $u \in H^{2m}(\Omega)$ e $f \in H^0(\Omega) \equiv L^2(\Omega)$.

Em relação às condições de contorno, a regra geral é a seguinte. Seja por exemplo o problema elíptico de valor de contorno:

$$\begin{aligned} Lu &= f, \quad em \quad \Omega \\ R_\mu u &= g_\mu \quad , \quad sobre \quad \partial\Omega \quad para \quad \mu = 1(1)m \end{aligned}$$

onde L é um operador linear de ordem $k = 2m$. Então o número de condições de contorno, denominadas estáveis, é dado por m.
Exemplo 1.1:

$$\begin{aligned} \Delta u &= f \quad em \quad \Omega \\ u &= 0 \quad sobre \quad \partial\Omega \end{aligned} \tag{1.55}$$

O operador Laplaciano Δ é de ordem $2m = 2$ e portanto o número de condições de contorno é igual a $m = 1$, ou seja, $R_1 u = 0$ é dado por $u = 0$. Aqui $u \in H^2(\Omega)$ e $f \in H^0(\Omega) \equiv L^2(\Omega)$.
Exemplo 1.2:

$$\begin{aligned} \Delta\Delta u &= f \quad , \quad em \quad \Omega \\ u &= 0 \quad , \quad sobre \quad \partial\Omega \\ \frac{\partial u}{\partial \nu} &= 0 \quad , \quad sobre \quad \partial\Omega. \end{aligned} \tag{1.56}$$

Observe-se que $L = \Delta\Delta$ é o bi-Laplaceano de ordem $2m = 4$. O número de condições de contorno é $m = 2$, ou seja,

$$\begin{aligned} R_1 u &= \quad u = g_1 = 0 \\ R_2 u &= \quad \tfrac{\partial u}{\partial \nu} = g_2 = 0 \end{aligned}$$

onde R_1 é o operador derivada de ordem zero e R_2 é operador de ordem 1, sendo ela definida pela derivada normal ao contorno de Ω, com ν sendo a direção da normal externa ao contorno $\partial\Omega$.

Construção da Matriz de Rigidez

Iniciaremos no capítulo 2 a análise de estabilidade do método dos elementos finitos. Mas para podermos compreender como os parâmetros de uma equação influenciam na análise da estabilidade numérica, estudaremos primeiramente o emprego dos Elementos Finitos no modelo

$$\begin{aligned} -(pu')' + qu &= f \quad em \quad [0,1] \\ u(0) = u'(1) &= 0 \end{aligned} \tag{1.57}$$

por motivos de ilustração, onde $p > 0$ é um coeficiente dependente de x, com $p \in E_c^1[0,1]$ e q, $f \in E_c[0,1]$. Como vimos em (1.54), (1.55) e (1.56) num problema de valor de contorno de ordem $2m$, W é subespaço de $H^m(\Omega)$. Nesse espaço estão também as funções seccional ou espaçadamente contínuas, com derivadas contínuas até ordem m inclusive, ou seja, $u \in E_c^m(\bar{\Omega})$. Assim para o problema (1.57) acima, a solução u é admitida em $E_c^1[0,1] \subset H^1[0,1]$.

Portanto, o método de Ritz é a formalização computacional básica para o método dos elementos finitos que, por sua vez, nada mais é que o método de Ritz onde o subespaço W agora é um espaço de dimensão finita, designado por W_n e construído segundo uma base $\omega_1, \omega_2, \ldots, \omega_n$ a ser convenientemente escolhida em função da discretização de Ω (no caso especial do intervalo $[0,1]$). Quando tratarmos de estabilidade e número de condição, usaremos o símbolo A para designar matriz global dos coeficientes a_{jk} do sistema linear resultante do emprêgo do método dos elementos finitos, ao invés de K como é frequente na literatura.

A discretização pode ser feita, nesse caso, dividindo-se $[0,1]$ em subintervalos $I_k = \left[x^{k-1}, x^k\right]$, de modo que

$$0 = x_0 < x_1 < \ldots < x_n = 1.$$

Consideremos todas as funções $u \in E_c^1[0,1]$ com $u(0) = 0$, tais que sejam, por exemplo, linear em cada I_k , $k = 1(1)n$. A dimensão de W_n é, sem dúvida n. A matriz dos coeficientes $a_{jk} = (\omega_j, \omega_k)$ do sistema linear resultante $AQ = F$ vai depender fundamentalmente da escolha da base $\omega_1, \ldots, \omega_n$ geradora de W_n. A condição da matriz A vai também depender fundamentalmente da escolha daquela base. Vejamos, por exemplo, como chegaremos à construção da matriz A. Seja escolhida, por exemplo, a função chapéu com $1 \leq j \leq n-1$:

$$\omega_j(x) = \begin{cases} (x - x_{j-1})/(x_j - x_{j-1}), & x \in I_j \\ (x_{j+1} - x)/(x_{j+1} - x_j), & x \in I_{j+1} \\ 0 \quad , & \text{caso contrário,} \end{cases} \tag{1.58}$$

definida pelo conhecido polinômio de Lagrange. Caso $j = n$, então temos

$$\omega_n(x) = \begin{cases} (x - x_{n-1})/(x_n - x_{n-1}), & x \in I_n \\ 0 & , \text{ caso contrário,} \end{cases}$$

Enfim, podemos representar $u_n \in W_n$ pela designação de S_h relativa ao tamanho do subintervalo constante h : $u^h \in S^h \equiv W_n \subset H^1[0,1]$, na forma

$$u^h = \xi_1\omega_1 + \xi_2\omega_2 + \ldots + \xi_n\omega_n,$$

onde os coeficientes ξ_k têm um significado físico:

$$\xi_k = u(x_k).$$

Os u^h são as funções-tentativa definidas por polinômios espaçadamente contínuos em I_k, por exemplo, (1.58), $u^h \in E_c^1[0,1]$.

Como já vimos o funcional (1.54) é tal que deve ser levado ao mínimo:

$$J[u] = \int_0^1 (p\, u'^2 + q\, u^2)dx - 2\int_0^1 f\, u\, dx \rightarrow min,$$

de onde temos para $u^h \in S^h$ a forma quadrátrica

$$\int_0^1 (p\, u'^2 + q\, u^2)dx = \sum_{j,k=1}^{n} a_{jk}\xi_j\xi_k, \tag{1.59}$$

cuja matriz dos coeficientes a_{jk} é, ao mesmo tempo, a matriz do sistema de equações lineares $AQ = F$ para cálculo das incógnitas $\xi_1, \ldots, \xi_n$. A integral (1.59) pode ser escrita na seguinte forma

$$\int_0^1 (p\, u'^2 + q\, u^2)dx = \sum_{r=1}^{n} \int_{I_r} (p\, u'^2 + q\, u^2)dx = \sum_{r=1}^{n} \left(\sum_{j,k=1}^{n} a_{jk}^{(r)}\xi_j\xi_k \right) \tag{1.60}$$

A procurada matriz $A = (a_{jk})$, dos coeficientes pode ser construída pelas matrizes $A_r = \left(a_{jk}^{(r)}\right)$ que são obtidas a partir da integral energética (1.60) acima:

$$a_{jk}^{(r)} = (\omega_j, \omega_k)_r = \int_{I_r} (p\omega_j'\omega_k' + q\omega_j\omega_k)dx \tag{1.61}$$

para cada elemento finito I_r. A matriz A é designada na literatura como matriz de Gram total ou matriz de rigidez total. O vetor F é o vetor de carga.

Façamos, por exemplo, $p = cte = 1$ e $q = cte = 0$ e $h_r = x_r - x_{r-1}$. Encontramos assim

$$\int_{I_r} (u')^2 dx = \begin{cases} \xi_1^2/h_1 \ , \ r = 1 \\ (\xi_r - \xi_{r-1})^2/h_r \ , \ r = 2(1)n \end{cases}$$

As matrizes A_r têm, portanto, a forma

$$A_r = \frac{1}{h_r} \begin{pmatrix} 0 & & & & & & 0 \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & 1 & -1 & & \\ & & & -1 & 1 & & \\ & & & & & 0 & \\ & & & & & & \ddots \\ 0 & & & & & & 0 \end{pmatrix} \quad , \quad \text{para } r \geq 2.$$

A matriz A é o resultado da soma

$$A = \sum_{r=1}^{n} A_r.$$

Para o caso particular em que $h_1 = \ldots = h_n = h$ obtem-se a matriz A na forma

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & & \ddots & & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 1 \end{pmatrix}$$

É claro que podemos usar h_r variável com grande vantagem. Mesmo assim, os elementos de A_r são calculados pela única integral

$$a_{jk}^{(r)} = \int_{I_r} \left(p\omega_j'\omega_k' + q\omega_j\omega_k \right) dx$$

para $j = r-1$, r e $k = r-1$, r.

O lado direito b_j do sistema linear

$$\sum_{k=1}^{n} a_{jk}\xi_k = b_j \ , \ j = 1(1)n$$

ou $AQ = F$, pode ser calculado, onde temos para b_j:

$$\int_0^1 f\ u\ dx = \sum_{r=1}^{n} \int_{I_r} fu\ dx.$$

Após alguns cálculos obtemos

$$\int_{I_r} fu\ dx = b_{r-1}^{(r)} \xi_{r-1} + b_r^{(r)} \xi_r,$$

onde $b_k^{(r)} = \int_{I_r} f\omega_k\ dx$, para todos os elementos.

Para melhorar a precisão da aproximação os elementos quadráticos ou cúbicos podem ser empregados. Mas antes devemos experimentar um refinamento da malha k_r, que poderá levar, sob certas circunstâncias, dependendo do problema dado, a uma melhor aproximação. Pelas considerações em torno dos problemas extremais quadráticos pudemos mostrar que um problema variacional é equivalente a um problema de minimização da energia inerente ao sistema físico, se a sua forma bilinear é simétrica. A solução do problema variacional é a solução (forte) de uma equação diferencial com condições de contorno dadas. O método dos elementos finitos opera diretamente no problema variacional e em consequência disso as seguintes vantagens, entre outras, podem ser apontadas:

a) Há problemas em que não se pode atribuir sentido matemático preciso à equação diferencial correspondente. Eles formam uma classe bem abrangente, que pode ser tratada pelo método dos elementos finitos.
b) Fácil ajustagem a domínios de fronteiras curvas ou irregulares. Para equações de diferenças finitas a uma dimensão é possível a prova de convergência, segundo uma teoria satisfatória. Mas para equações diferenciais parciais quase toda prova de convergência depende de um princípio de máximo. O problema é que uma simples equação diferencial admite uma enorme variedade de aproximação por diferenças finitas, sobretudo num contorno curvo. Ao contrário, os métodos variacionais são governados por regras, restritas, e isso permite uma teoria mais completa, o que possibilita garantir provas de convergência mesmo para contornos curvos complicados.
c) Em geometrias irregulares as condições naturais de contorno são, de maneira geral, as mais complexas. Essas condições naturais estão embutidas inerentemente na formulação variacional e, por isso, o método dos elementos finitos as ignoram. Elas não oferecem alteração alguma quanto ao emprego do método.

Capítulo 2

Estabilidade Numérica e Convergência

Os métodos computacionais tratados nos capítulos anteriores levam a conclusão de que a resolução de um dado problema de valor de contorno redunda em resolver, enfim, um sistema de equações lineares, com a matriz dos coeficientes simétrica e positiva definida. Nas aplicações em problemas de valor de contorno com equações diferenciais parciais aparecem os coeficientes a_{ij} e o vetor b_i do lado direito das equações lineares com integrais respectivamente de domínio ou de contorno, cujo cálculo geralmente se faz empregando-se métodos numéricos. Os coeficientes a_{ij} e b_i calculados são afetados com erros de arredondamento e erros inerentes ao próprio método numérico utilizado. Os erros relativos pequenos nos dados de saída do cálculo das referidas integrais podem, sob certas circunstâncias, gerar erros relativos muito grosseiros no cálculo final do sistema linear. - Um sistema no qual esse é o caso é denominado numericamente instável. O objetivo desse capítulo é o tratamento da estabilidade numérica e convergência, iniciando aqui oportunamente com a indagação fundamental seguinte: Numa aplicação de métodos variacionais, o que tem de ser feito ou observado quanto à escolha das funções coordenadas, para que o sistema linear resultante, também permaneça numericamente estável quando aumentamos o número de funções coordenadas $\omega_1, \ldots, \omega_n$?
Outra indagação é a de como as aproximações convergem para a solução de um dado problema, quando aumentamos o nº de funções coordenadas? A essas indagações está ligado também um estreito relacionamento entre estabilidade numérica e convergência, fato este que será tratado neste capítulo.

2.1 Estabilidade Numérica

2.1.1 A Condição de Sistema de Equações.

Seja dado o sistema de equações lineares

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \quad , \quad i = 1(1)n \tag{2.1}$$

o qual escrevemos sob forma simplificada

$$A\vec{x} = \vec{b} \tag{2.2}$$

com a matriz $A = (a_{ij})$ e os vetores-coluna $\vec{x}$ e $\vec{b}$, cujas transpostas são os vetores-linha $\vec{x}^t = (\xi_1, \ldots, \xi_n)$ e $\vec{b}^t = (b_1, b_2, \ldots, b_n)$, escritas, em seguida, respectivamente como x e b.
Se o sistema (2.1) é ou não numericamente estável, isso se deixa julgar pelo número de condição relativa à matriz A. A norma Euclideana do vetor x se escreve

$$\|x\| = \left(\sum_{i=1}^{n} \xi_i^2\right)^{1/2} . \tag{2.3}$$

e a norma da matriz A é dada pela expressão

$$\|A\| = \left(\sum_{i,j=1}^{n} a_{ij}^2\right)^{1/2} . \tag{2.4}$$

Então dada a matriz A segue, da desigualdade de Schwarz

$$\|A\,x\| \leq \|A\| \quad \|x\| \quad , \quad \forall x \in \mathcal{R}^n,$$

e definimos a menor cota superior (mcs) pela expressão

$$mcs(A) = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \rightarrow mcs(A) \geq \frac{\|Ax\|}{\|x\|},$$

de onde resulta

$$\|Ax\| \leq mcs(A)\|x\| \quad , \quad mcs(A) \leq \|A\|. \tag{2.5}$$

Para o produto AB de duas matrizes $n \times n$ achamos a desigualdade

$$mcs(AB) \leq mcs(A)mcs(B). \tag{2.6}$$

Seja agora A uma matriz regular com a inversa A^{-1}. Comparemos em seguida os vetores solução x e $\tilde{x}$ dos sistemas lineares

$$Ax = b \quad e \quad A\tilde{x} = \tilde{b}.$$

Então vale

$$A(\tilde{x} - x) = \tilde{b} - b$$

ou

$$\tilde{x} - x = A^{-1}(\tilde{b} - b).$$

De (2.5) temos

$$\|b\| \leq mcs(A)\|x\|, \quad \|\tilde{x} - x\| \leq mcs(A^{-1})\|\tilde{b} - b\|$$

e obtemos no total

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq mcs(A)mcs\left(A^{-1}\right)\frac{\|\tilde{b} - b\|}{\|b\|},$$

de onde sai o número de condição da matriz A

$$cond(A) = mcs(A)mcs\left(A^{-1}\right),$$

que é, como podemos interpretar, uma medida da sensibilidade do sistema de equações lineares diante dos erros no lado direito b_i. Se $cond(A)$ não for muito grande (em torno de 1), então o erro relativo da solução $\tilde{x}$ é no máximo da ordem do erro relativo de $\tilde{b}$, medido na norma vetorial euclideana. Sejam, entretanto, A e $\tilde{A}$ matrizes regulares. Comparemos os vetores-solução x e $\tilde{x}$ de

$$Ax = b, \qquad \tilde{A}\tilde{x} = b.$$

Vale

$$\tilde{A}(\tilde{x} - x) + (\tilde{A} - A)x = 0 \quad ou \quad (x - \tilde{x}) = \tilde{A}^{-1}(\tilde{A} - A)x.$$

De (2.5) e (2.6) segue

$$\frac{\|\tilde{x} - x\|}{\|x\|} \leq mcs\left(\tilde{A}^{-1}\right)\|\tilde{A} - A\|,$$

expressão que mostra ser $mcs\left(\tilde{A}^{-1}\right)$ a medida da sensibilidade do sistema de equações lineares diante dos erros na matriz dos coeficientes A. Em aplicação de métodos variacionais devemos ter o cuidade necessário para que os valores de $cond(A)$ e

$mcs(A^{-1})$ não atinjam grandes magnitudes, ou seja, valores em torno de 1. As matrizes A, que aparecem nos métodos variacionais, são simétricas e positivas definidas. Então, por decorrência disso, os seus autovalores são todos reais e positivos:

$$0 \leq \lambda_1, \leq \lambda_2 \leq \ldots \leq \lambda_n. \tag{2.7}$$

Além disso o quociente de Rayleigh (Isaacson, K.[13])

$$R(x) = \frac{x^t Ax}{x^t x} = \left(\sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j \right) / \sum_{i=1}^{n} \xi_i^2 \tag{2.8}$$

possui a propriedade conhecida

$$\lambda_1 \leq R(x) \leq \lambda_n. \tag{2.9}$$

Podemos justificar a afirmação acima. Da observação, de que podemos encontrar uma base $x_1, x_2, \ldots, x_n$ de autovetores no $\mathcal{R}^n$, que, em relação ao produto interno

$$x^t y = \sum_{i=1}^{n} \xi_i \eta_i,$$

pode ser suposta ortonormada, vale escrever, com $x = c_1x_1 + c_2x_2 + \ldots + c_nx_n$ o seguinte:

$$R(x) = \left(\sum_{i=1}^{n} \lambda_i c_i^2 \right) / \left(\sum_{i=1}^{n} c_i^2 \right),$$

expressão da qual resultou a expressão (2.8). Em especial temos

$$R(x_1) = \lambda_1 \quad e \quad R(x_n) = \lambda_n.$$

Além disso existe uma conexão entre autovalores e $cond(A)$, $mcs(A)$ quando trabalhamos com matrizes simétricas positivas definidas.
Tomemos a base

$$x = c_1x_1 + \ldots + c_nx_n$$

em que

$$\|x\|^2 = (x,x) = c_1^2 + c_2^2 + \ldots + c_n^2 = \sum_{i=1}^{n} c_i^2.$$

$$\Rightarrow \|Ax\|^2 = \sum_{i=1}^{n} (\lambda_i c_i)^2$$

$$\Rightarrow \|A^{-1}x\|^2 = \sum_{i=1}^{n} \left(\frac{c_i}{\lambda_i} \right)^2.$$

E assim temos realizado o

Teorema 2.1 *Vale afirmar:*

$$mcs(A) = \lambda_n \quad , \quad mcs(A^{-1}) = \frac{1}{\lambda_1} \quad , \quad cond(A) = \frac{\lambda_n}{\lambda_1}.$$

2.1.2 Estabilidade dos Métodos de Projeção

Como vimos no capítulo anterior sobre métodos computacionais pudemos observar que resolver problemas extremais quadráticos (como o baseado na norma do erro quadrático, ou o fundamentado com a norma energética) significa equivalentemente resolver o problema da projeção. Consideremos, por isso, primeiramente o espaço de Hilbert $H[(,),\|.\|]$ o mais geral, bem como um subespaço fechado $W \subset H$. Procuremos calcular, para dado $u \in H(\Omega)$, a solução de

$$\|u - \omega\|^2 \to min \qquad \forall \omega \in W.$$

Seja $\omega_1, \omega_2, \ldots$ um sistema linear de funções coordenadas em W, e designemos por $W_n \subset W$ o subespaço linear, gerado por $\omega_1, \omega_2, \ldots, \omega_n$. Daí obtemos o conjunto de problemas de aproximação

$$\|u - \omega\|^2 \to min \qquad \forall \omega \in W_n. \tag{2.10}$$

Com a base

$$\omega = \xi_1\omega_1 + \xi_2\omega_2 + \ldots + \xi_n\omega_n$$

o problema extremal (2.10) se torna equivalente a resolver o sistema de equações lineares

$$\sum_{j=1}^{n} a_{ij}\xi_j = b_i \qquad (i = 1(1)n), \tag{2.11}$$

com

$$a_{ij} = (\omega_i, \omega_j) \quad , \quad b_i = (\omega_i, u)$$

(veja p.ex. Teoremas 1.5; 1.6 ou 1.7, cap. 1)

Sob a denominação de método das projeções entendemos o seguinte:
(i) Escolha de um sistema concreto de funções coordenadas $\omega_1, \omega_2, \ldots$
(ii) O sucessivo cálculo das soluções de (2.11) para $n = 1, 2, \ldots$

A questão da convergência das aproximações obtidas tendendo para a solução procurada $\tilde{\omega} \in W$ de (2.11) será considerada e estudada oportunamente. Em seguida devemos nos ocupar com a estabilidade numérica da equação (2.11) para n crescente. Para isso vamos nominar A_n a matriz dos coeficientes (a_{ij}) de (2.11).

Além disso sejam $\lambda_1^{(n)}$ e $\lambda_n^{(n)}$ respectivamente o menor e o maior auto-valores da matriz $A_n(n \times n)$. Um método de projeção se denomina numericamente estável se existem uma constante K tal que

$$K \geq cond(A_n) \geq 1, \qquad \forall n \in \mathcal{N}.$$

Suponhamos, todavia, que $\lambda_1^{(n)}$ e $\lambda_n^{(n)}$ sejam respectivamente sequências monótonas (fracas) decrescentes ou crescentes, na forma em que uma comparação entre elas possa levar a relacioná-las com o número de condição num intervalo limitado. Partindo das expressões (2.8) e (2.9) temos:

$$\lambda_1^{(n)} \leq \frac{\sum_{i,j=1}^{n} a_{ij}\xi_i\xi_j}{\sum_{i=1}^{n} \xi_i^2} \leq \lambda_n^{(n)} \tag{2.12}$$

em que os valores $\lambda_1^{(n)}$ e $\lambda_n^{(n)}$, para valores apropriados $\xi_1, \xi_2, \ldots, \xi_n$, são aceitos. Fazendo $\xi_n = 0$, o quociente $R(x)$ reduz-se aos quocientes correspondentes a A_{n-1}, cujo domínio é porém o intervalo $[\lambda_1^{(n-1)}, \lambda_{n-1}^{(n-1)}]$, que, entretanto, está imerso no intervalo $[\lambda_1^{(n)}, \lambda_n^{(n)}]$ e daí segue o resultado de que

$$\lambda_1^{(n)} \leq \lambda_1^{(n-1)} \quad , \quad \lambda_{n-1}^{(n-1)} \leq \lambda_n^{(n)},$$

o que compõe o enunciado do seguinte

Teorema 2.2 *Os menores autovalores $\lambda_1^{(n)}$ da matriz A_n compõem uma sequência decrescente monótona e os maiores autovalores $\lambda_n^{(n)}$ da matriz A_n compõem uma sequência crescente monótona:*

$$\lambda_1^{(n)} \leq \lambda_1^{(n-1)}; \qquad \lambda_{n-1}^{(n-1)} \leq \lambda_n^{(n)}.$$

Suponhamos que existem constantes λ_0 e Λ_0 tais que $0 < \lambda_0 \leq \lambda_1^{(n)}$ e $\lambda_n^{(n)} \leq \Lambda_0 \quad \forall n \in \mathcal{N}$. Vamos relacionar essas constantes com $cond(A_n) \leq K$, para medir estabilidade:

$$\begin{aligned} \lambda_n^{(n)} &\leq \Lambda_0 \\ 1/\lambda_1^{(n)} &\leq 1/\lambda_0 \\ \Rightarrow \quad \lambda_n^{(n)}/\lambda_1^{(n)} &\leq \Lambda_0/\lambda_0, \end{aligned}$$

que pelo Teorema 2.1, fornece

$$1 \leq cond(A_n) = \frac{\lambda_n^{(n)}}{\lambda_1^{(n)}} \leq \frac{\Lambda_0}{\lambda_0},$$

o que prova o enunciado do

Teorema 2.3 *Um certo método de projeção é numericamente estável se e só se existem duas constantes λ_0 e Λ_0 tais que*

$$0 < \lambda_0 \leq \lambda_1^{(n)} \quad e \quad \lambda_n^{(n)} \leq \Lambda_0, \quad \forall n \in \mathcal{N}. \tag{2.13}$$

Como tanto a sequência dos $\lambda_n^{(n)}$ e da mesma forma a sequência dos $\frac{1}{\lambda_1^{(n)}}$, conforme Teorema 2.1, é monótona crescente, é possível que o produto só seja limitado, se as constantes λ_0 e Λ_0 existem, de tal modo que (2.13) se verifique [3].

Através do Teorema 2.3 se esclarece, o que se deve exigir da matriz A_n quando se deseja que o método da projeção seja ou deva ser numéricamente estável. A próxima indagação é, como essa estabilidade pode ser alcançada no caso concreto? Sobre essa pergunta voltaremos oportunamente nas ulteriores considerações. Para melhor entendimento da estabilidade numérica tratamos, em seguida, os

2.1.3 Sistemas Mínimais de Funções Coordenadas

Um método de projeção necessita de uma outra propriedade, também eficaz, que deve ser relacionada ao sistema linear de equações, para que a estabilidade numérica venha a ser melhor formulada ou garantida. Seja $\omega_1, \omega_2, \ldots$ um sistema infinito de funções coordenadas admitidas no espaço de Hilbert H. Com $[\omega_1, \omega_2, \ldots]$ designamos o subespaço linear de todas as combinações lineares finitas

$$\alpha_1\omega_1 + \alpha_2\omega_2 + \ldots + \alpha_r\omega_r \quad , r = 1, 2, \ldots$$

bem como $\overline{[\omega_1, \omega_2, \ldots]}$ representa o subespaço fechado em que $[\omega_1, \omega_2, \ldots]$ é denso. Um sistema $\omega_1, \omega_2, \ldots$ é denominado minimal quando nenhum dos ω_k está no subespaço fechado

$$W_k = \overline{[\omega_1, \ldots, \omega_{k-1}, \omega_{k+1}, \ldots]}$$

sendo este gerado pelas respectivas funções coordenadas restantes (Michlin [5] 1 e 2; Babŭska [3]). Então $\omega_k \notin W_k$. Assim podemos observar que um sistema minimal é sempre um sistema linearmente independente, mas a recíproca não é verdadeira, em geral; isto é, pode acontecer que um sistema linearmente independente venha a ser também um sistema minimal. Se juntarmos um elemento ω_0 ao sistema linearmente independente infinito $\omega_1, \omega_2, \ldots$ com $\omega_0 \in \overline{[\omega_1, \omega_2, \ldots]}$, $\omega_0 \notin [\omega_1, \omega_2, \ldots]$ formamos assim um sistema $\omega_0, \omega_1, \ldots$ que, em verdade, é linearmente independente, mas não é minimal.

Uma forma de caracterizarmos sistema minimal é escolhermos ω_k de tal forma que tenhamos uma medida em norma, traduzindo sensibilidade. Isso é possível se definirmos a diferença entre ω_k e $\sum \xi_j \omega_k$ excetuando a parcela para a qual $j = k$. É claro que como $r = 1, 2, \ldots\ k = 1, 2, \ldots n, \ldots$ podemos construir

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2$$

sendo esta norma da diferença, uma menor cota superior relativa a uma referência numérica que designamos por $\delta_k > 0$.
Assim temos

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2 \geq \delta_k > 0$$

para todo $n \in \mathcal{N}$ e todo $\xi_1, \ldots, \xi_n \in \mathcal{R}$. Obviamente observamos que enquanto $k \leq n$ o número de parcelas do somatório $\sum_{j=1}^{n}$ é igual a $n-1$ e quando $k > n$ então o número de parcelas será igual a n. A propriedade $\omega_k \notin W_k$ significa a mesma coisa que os ω_k não se deixam aproximar dos ω_j's restantes através de combinações lineares finitas sob uma forma qualquer arbitrária. Isso é justamente o caso quando $\delta_k > 0$ e $\delta_k \leq \|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2$. Assim acabamos de arranjar uma forma de caracterizar sistemas minimais, o que redescrevemos no enunciado do

Teorema 2.4 *Um sistema $\omega_1, \omega_2, \ldots$ é minimal, se e somente se, para todo ω_k, existe um número $\delta_k > 0$ tal que*

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2 \geq \delta_k > 0,$$

para $\forall n \in \mathcal{N}$ e $\forall \xi_1, \ldots, \xi_n \in \mathcal{R}$.

Mas o quadrado da norma da diferença é representado pela expressão

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2 = \|\omega_k\|^2 + \|\sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2 - 2\left(\omega_k, \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\right)$$

de onde obtemos, para sistemas ortogonais em que $\left(\omega_k, \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_j\right) = 0$ a expressão

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2 = \|\omega_k\|^2 + \|\sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_k\|^2.$$

Mas observamos que vale também a desigualdade

$$\|\omega_k\|^2 + \|\sum_{\substack{j=1 \\ j\neq k}}^{n} \xi_j\omega_j\|^2 \geq \|\omega_k\|^2.$$

Então pelo Teorema 2.3, podemos finalmente escrever

$$\|\omega_k - \sum_{\substack{j=1 \\ j\neq k}}^{n} \xi_j\omega_j\|^2 \geq \delta_k = \|\omega_k\|^2 > 0,$$

e assim temos provado o enunciado do

Teorema 2.5 *Para constantes* $\delta_k = \|\omega_k\|^2$ *os sistemas ortogonais* $(\omega_j,\omega_k) = 0 \quad , \quad j \neq k$, *são minimais.*

Desse ponto podemos procurar a condição necessária para que um método de projeção seja estável. Partimos da hipótese de que o sistema $\omega_1,\omega_2,\ldots$ não seja minimal e, sob essa hipótese, seja $\omega_1 \in \overline{[\omega_2,\omega_3,\ldots]}$. Então ω_1 pode ser aproximado arbitrariamente por $\omega_2,\omega_3,\ldots$. Façamos então, para um dado n,

$$d_n = \min_{\xi_i} \|\omega_1 - \sum_{i=2}^{n} \xi_i\omega_i\|^2.$$

Observamos que aumentando-se n cresce também a parcela $\sum_{j=2}^{n} \xi_j\omega_j$, diminuindo a diferença. Por isso temos $d_n \geq d_{n+1}$ para todo $n \in \mathcal{N}$ bem como $d_n \to 0$ quando $n \to \infty$. Por outro lado temos que

$$\|\sum_{j=1}^{n} \xi_j\omega_j\|^2 = \sum_{j,k=1}^{n} a_{jk}\xi_j\xi_k, \quad a_{jk} = (\omega_j,\omega_k). \tag{2.14}$$

Mas usando a expressão (2.12) chegamos, com (2.14) ao resultado

$$\|\sum_{j=1}^{n} \xi_j\omega_j\|^2 = \sum_{j,k=1}^{n} a_{jk}\xi_j\xi_k \geq \lambda_1^{(n)} \sum_{j=1}^{n} \xi_j^2. \tag{2.15}$$

Mas separando o termo para $j = 1$ obtemos de (2.15) o seguinte:

$$\|\xi_1\omega_1 + \sum_{j=2}^{n} \xi_j\omega_j\|^2 \geq \lambda_1^{(n)} \left(\xi_1^2 + \sum_{j=2}^{n} \xi_j^2\right)$$

que, particularmente, para $\xi_1 = -1$, fornece a expressão

$$\|\sum_{j=2}^{n} \xi_j \omega_j - \omega_1\|^2 \geq \lambda_1^{(n)} \left(1 + \sum_{j=2}^{n} \xi_j^2\right).$$

Agora com o fato de que, para um dado n, ω_1 pode ser aproximado por $\omega_2, \omega_3, \ldots$ (veja p.ex. Lusternik, Sobolev [9]), se escolhermos para $\xi_2, \xi_3, \ldots, \xi n$ aqueles valores para os quais vale a expressão (2.16) abaixo, obtemos então a desejada desigualdade para $\lambda_1^{(n)}$:

$$d_n = \min_{\xi_j} \|\sum_{j=2}^{n} \xi_j \omega_j - \omega_1\|^2 \geq \min_{\xi_j} \lambda_1^2 \left(1 + \sum_{j=2}^{n} \xi_j^2\right) \Rightarrow d_n \geq \lambda_1^{(n)} > 0 \qquad (2.16)$$

De (2.16) segue que $\lambda_1^{(n)} \to 0$ bem como $cond(A_n) \to \infty$ para $n \to \infty$. Portanto o resultado final mostra então que partindo da hipótese imposta o método da projeção não é numericamente estável. Assim sendo, temos provado o enunciado do

Teorema 2.6 *Uma condição necessária para que o método da projeção seja estável é que o sistema básico $\omega_1, \omega_2, \ldots$ de funções coordenadas seja minimal.*

Acontece que a recíproca desse teorema não é verdadeira. O método da projeção, mesmo com um sistema minimal de funções coordenadas, não é numericamente estável, com facilidade. Esse fato pode ser exclarecido através de
i) O sistema ortogonal $\omega_1, \omega_2, \ldots$ por exemplo. Porque $a_{jk} = (\omega_j, \omega_k) = 0$ para $j \neq k$ as matrizes A_n se reduzem a matrizes diagonais, cujos elementos são os autovalores.
i_1 - Com $a_{kk} = (\omega_k, \omega_k) = \frac{1}{k}$ encontram-se as expressões

$$\lambda_1^{(n)} = \frac{1}{n}, \quad \lambda_n^{(n)} = 1, \quad cond(A_n) = n.$$

i_2 - Com outra forma especial

$$a_{kk} = (\omega_k, \omega_k) = k$$

encontramos, de maneira análoga

$$\lambda_1^{(n)} = n, \ \lambda_n^{(n)} = \frac{1}{n} \quad , \quad cond(A_n) = n$$

Em ambos os casos i_1 e i_2 o sistema $\omega_1, \omega_2, \ldots$ ortogonal (Teorema 2.4 e 2.5) é um sistema minimal, mas os métodos de projeção são, numericamente, não-estáveis.

No caso especial de sistema ortogonal $\omega_1,\omega_2,\ldots$ a causa da instabilidade pode ser anulada, usando um produto escalar apropriado:

$$a_{jk} = (\omega_j,\omega_k) = \delta_{jk}.$$

Assim obtemos facilmente

$$\lambda_1^{(n)} = \lambda_n^{(n)} = 1,$$

bem como $cond(A_n) = 1$ para todo $n \in \mathcal{N}$.

O sistema ortogonal é portanto o caso ideal. É suficiente para a estabilidade numérica (conforme Teorema 2.3) que os menores e maiores autovalores das matrizes A_n permaneçam limitados:

$$0 < \lambda_0 \leq \lambda_1^{(n)}, \quad \lambda_n^{(n)} \leq \Lambda_0.$$

Sistemas $\omega_1,\omega_2,\ldots$ com esta propriedade são também conhecidos por semi-ortonormados. Com isto passamos então a estudar a

2.1.4 Construção de Sistemas Semiortonormados

Apoiemo-nos na norma energética $|\omega|_A^2 = ((\omega,\omega)) = (A\omega,\omega)$ e na norma natural $\|\omega\|^2 = (\omega,\omega)$ em um subespaço linear $W \subset \chi$, com χ espaço linear mais geral possível. Essas normas são aqui consideradas equivalentes:

$$c_1\|\omega\|^2 \leq |\omega|_A^2 \leq c_2\|\omega\|^2, \quad \forall\omega, \in W \quad , \quad c_1 > 0. \tag{2.17}$$

Então vale para um sistema qualquer $\omega_1,\omega_2,\ldots$ de funções coordenadas de W declarar os termos do

Teorema 2.7 *Se o sistema $\omega_1,\omega_2,\ldots$ é semiortonormado em relação ao produto interno $(,)$, então ele é também semiortonormado em relação ao produto interno energético $((,))$ e vice-versa (a recíproca é verdadeira).*

Prova:

Sejam dados $a_{jk} = (\omega_j,\omega_k)$ e $b_{jk} = ((\omega_j,\omega_k))$. Sejam A_n e B_n as matrizes dos coeficientes respectivamente a_{jk} e b_{jk}. Vamos comparar o domínio dos valores dos quocientes de Rayleigh

$$R_n(x) = \frac{x^t A_n x}{x^t x} = \left(\sum_{j,k=1}^{n} a_{jk}\xi_j\xi_k\right) / \sum_{j=1}^{n}\xi_j^2 \tag{2.18}$$

$$Q_n(x) = \frac{x^t B_n x}{x^t x} = \left(\sum_{j,k=1}^{n} b_{jk}\xi_j\xi_k \right) / \sum_{j=1}^{n} \xi_j^2. \tag{2.19}$$

De

$$\| \sum_{j=1}^{n} \xi_j \omega_j \|^2 = \sum_{j,k=1}^{n} a_{jk}\xi_j\xi_k$$

e

$$| \sum_{j=1}^{n} \xi_j \omega_j |^2 = \sum_{j,k=1}^{n} b_{jk}\xi_j\xi_k$$

obtemos com (2.17),(2.18) e (2.19):

$$c_1 R_n(x) \leq Q_n(x) \leq c_2 R_n(x),$$

baseando-nos na hipótese da equivalência das duas normas, para todo $x^t = (\xi_1, \ldots, \xi_n) \neq 0$. Agora, se $\omega_1, \omega_2, \ldots$ é um sistema semiortonormado em relação ao produto interno $(,)$, então existem números λ_0 e $\wedge_0$ com

$$0 < \lambda_0 \leq R_n(x) \leq \wedge_0, \qquad \forall n \in \mathcal{N}$$

e

$$\forall \ \ x^t = (\xi_1, \ldots, \xi_n) \neq \vec{0}.$$

Então vale também escrever

$$0 < c_1\lambda_0 \leq Q_n(x) \leq c_2 \wedge_0 .$$

Isto quer dizer que o sistema é também semiortonormado em relação ao produto interno energético $((,))$. A recíproca é verdadeira e segue analogamente, o mesmo raciocínio. Para os números de condição obtemos os confinamentos:

$$1 \leq cond(A_n) \leq \frac{\wedge_0}{\lambda_0}, \qquad 1 \leq cond(B_n) \leq \frac{c_2 \wedge_0}{c_1 \lambda_0}.$$

Corolário 2.1
Se o sistema $\omega_1, \omega_2, \ldots$ é ortonormado em relação ao produto interno $(,)$, então ele é, em relação a $((,))$ também semiortonormado. Vale escrever

$$1 \leq cond(B_n) \leq \frac{c_2}{c_1} \quad , \quad \forall n \in \mathcal{N}.$$

Prova: Consequência direta do Teorema anterior.
Exemplo

Seja dado o problema

$$(P1)\begin{cases} -\Delta\omega + p\omega &= f \quad em \quad \Omega \\ \omega &= 0 \quad sobre \quad \partial\Omega, \end{cases} \tag{2.20}$$

com $f \in L^2(\Omega)$. Seja o coeficiente p uma função de $E_c(\bar{\Omega})$ com $p(x) \geq 0$. Designemos o máximo de p por $\bar{p}$. O problema (P1) é equivalente ao problema (P2) - forma fraca: Ache $\omega \in H_0^1(\Omega)$ como solução de

$$\left.\begin{array}{ll} & a(\omega,\varphi) = l(\varphi) \qquad \forall\varphi \in H_0^1(\Omega) \\ com & \\ & a(\omega,v) = \int_\Omega (grad\,\omega\, grad\, v + p\omega v)\, dx \\ & l(\omega) = \int_\Omega f\omega dx \end{array}\right\} \quad (P2)$$

Ao lado de $a(\omega,v)$ e da norma enegética $|\,\omega\,|$ introduzimos ainda

$$(\omega,v) = \int_\Omega (grad\,\omega\, grad\, v)dx, \quad \|\omega\|^2 = (\omega,\omega).$$

É óbvio que $\|\omega\|^2 \leq |\,\omega\,|^2, \quad \forall\omega \in H_0^1(\Omega)$. Aplicando a desigualdade de Friederich [5]:

$$\int_\Omega \omega^2 dx \leq \quad c\int_\Omega |\frac{\partial\omega}{\partial x_j}|^2 dx, \qquad \forall\omega \in H_0^1(\Omega),\, c > 0,$$

onde c é constante que depende do domínio Ω, obtemos:

$$\int_\Omega p\omega^2 dx \leq \bar{p}\int_\Omega \omega^2 dx \leq \bar{p}c\int_\Omega |grad\,\omega|^2 dx.$$

E assim resulta portanto, desse exemplo:

$$\|\omega\|^2 \leq |\,\omega\,|^2 \leq (1+\bar{p}c)\|\omega\|^2 \quad , \quad \forall\omega \in H_0^1(\Omega).$$

Assim temos auferido o seguinte resultado importante, conforme versão desenvolvida acima:
Se construirmos um sistema de funções coordenadas com base no método da energia, que é ortonormado em relação à norma $(,)$, então ele é também semiortonormado em relação ao produto interno energético $a(,)$, e o método é numericamente estável. Com esses fatos até agora desenvolvidos podemos tecer algum comentário sobre

2.2 Propriedades de Convergência dos Métodos de Projeção

Consideramos, dentro desse amplo contexto, métodos de projeção levando em conta a estabilidade numérica do sistema de equações lineares resultante, para o número n de funções coordenadas $\omega_1, \ldots, \omega_n$ crescente. Mas ainda não consideramos o íntimo relacionamento entre a estabilidade numérica e certas propriedades de convergência das aproximações. Em seguida passamos a análise das

2.2.1 Sistemas Minimais-Convergência

Seja, novamente, $\omega_1, \omega_2, \ldots$ um sistema linear de funções coordenadas linearmente independentes em W e designemos por $W_n \subset W$ o subsespaço linear gerado por $\omega_1, \omega_2, \ldots, \omega_n$. Obtemos o grupo de problemas de aproximação. Para $u \in H$ dado, ache a solução do problema

$$\|u - \omega\|^2 \to min, \qquad \omega \in W_n \subset W \subset H.$$

Consideremos agora esse grupo de problemas e designemos a solução deles com $\omega^{(n)}$. Isto é, $\omega^{(n)}$ é a projeção de ω sobre o subespaço de dimensão finita $W_n \subset W$. Se o sistema $\omega_1, \omega_2, \ldots$ em W não for completo, ou seja, se $\overline{[\omega_1, \omega_2, \ldots]}$ for um subespaço próprio de W, então não obtemos em geral convergência alguma contra $\tilde{\omega} = Pu \in W$. Mas, ao contrário, vale a seguinte proposição:

Teorema 2.8 *Se o sistema $\omega_1, \omega_2, \ldots$ é completo em W, então existe a convergência na norma*

$$\|\tilde{\omega} - \omega^{(n)}\|^2 \to 0, \qquad n \to \infty.$$

Prova: Com $\tilde{v} = u - \tilde{\omega}$ obtemos para $\forall \omega \in W$

$$\|u - \omega\|^2 = \|\tilde{v}\|^2 + \|\tilde{\omega} - \omega\|^2.$$

Isto mostra que o problema extremal (2.10) é equivalente ao problema

$$\|\tilde{\omega} - \omega\|^2 \to min, \qquad \omega \in W_n.$$

Os números $d_n = \|\tilde{\omega} - \omega^{(n)}\|^2$ representam uma sequência monótona fraca decrescente. Mas como $\tilde{\omega}$ pode ser aproximada, de alguma forma, por combinações lineares finitas podemos encontrar um $n = n(\varepsilon)$, para dado $\varepsilon > 0$, tal que

$$\|\tilde{\omega} - \omega\|^2 < \varepsilon,$$

com $\omega \in W_n$ apropriado. Daí vemos que os d_n formam uma sequência nula e porisso temos convergência na norma. Fim da prova.

A aproximação pode ser escrita na forma

$$\omega^{(n)} = \alpha_1^{(n)}\omega_1 + \alpha_2^{(n)}\omega_2 + \ldots + \alpha_n^{(n)}\omega_n. \tag{2.21}$$

Os coeficientes $\alpha_1^{(n)}, \alpha_2^{(n)}, \ldots, , \alpha_n^{(n)}$ são exatamente a solução do sistema

$$\sum_{k=1}^{n} a_{jk}\alpha_k^n = b_j \qquad j = 1(1)n$$

com

$$a_{jk} = (\omega_j, \omega_k), \qquad b_j = (\omega_j, u).$$

Uma indagação que nos ressalta naturalmente é como se comportam os coeficientes $\alpha_k^{(n)}$ para k fixo e n crescente? Vemos aqui que podemos pesquisar essa indagação, com base num sistema completo e minimal e daí podemos analisar o comportamento dos $\alpha_k^{(n)}$ em relação à convergência.

Consideremos portanto o fato de que, num sistema minimal $\omega_1, \omega_2, \ldots$ não se encontra ω_k algum no espaço W_k, que é gerado pelos elementos restantes ω_j:

$$\omega_k \notin W_k = \overline{[\omega_1, \ldots, \omega_{k-1}, \omega_{k+1}, \ldots]}$$

Seja V_k o complemento ortogonal de W_k em W. Então podemos escrever

$$W = V_k \oplus W_k$$

e podemos representar ω_k de uma forma única (veja Teorema 1.1)

$$\omega_k = v_k + u_k \quad , \quad v_k \in V_k \quad , \quad u_k \in W_k.$$

Façamos $v_k \neq 0$ porque senão $\omega_k \in W_k$ (contra a hipótese). Por outro lado $\omega_j \in W_k$ para $j \neq k$. Obtemos então

$$\begin{aligned} (\omega_j, v_k) &= 0 \quad para \quad j \neq k \\ (\omega_k, v_k) &= \|v_k\|^2 \neq 0 \end{aligned}$$

É claro que podemos usar uma outra formulação usando os sistemas $\omega_1, \omega_2, \ldots$ e $v_1, v_2, \ldots$. Estes dois sistemas são biortogonais (veja [8]). Temos que passar do sistema $v_1, v_2, \ldots$ para o sistema normado $\varphi_1, \varphi_2, \ldots$, por exemplo, com

$$\varphi_k = \frac{v_k}{\|v_k\|^2} \quad , \quad \|\varphi_k\| = \frac{1}{\|v_k\|}.$$

De (2.21) obtemos primeiramente

$$\left(\omega^{(n)}, \varphi_k\right) = \alpha_k^{(n)}$$

Definimos α_k através de

$$\alpha_k := (\tilde{\omega}, \varphi_k) \qquad k = 1, 2, \ldots$$

e assim obtemos

$$\left|\alpha_k - \alpha_k^{(n)}\right| = \left|(\tilde{\omega} - \omega^{(n)}, \varphi_k)\right| \leq \|\tilde{\omega} - \omega^{(n)}\| \frac{1}{\|v_k\|}.$$

Para k fixo, o membro direito dessa última desigualdade se anula, quando $n \to \infty$. Nesse ponto a completação é utilizada, e assim temos demonstrado o

Teorema 2.9 *Se o sistema $\omega_1, \omega_2, \ldots$ é completo em W e se ele é minimal, então os $\alpha_k^{(n)}$ formam uma série convergente:*

$$\lim_{n \to \infty} \alpha_k^{(n)} = \alpha_k.$$

Corolário 2.2
Se os elementos do sistema $v_1, v_2, \ldots$ biortogonal ao sistema $\omega_1, \omega_2, \ldots$ são, na norma, uniformemente limitados inferiormente, com M sendo uma constante positiva:

$$\|v_k\| \geq \frac{1}{M} > 0,$$

então a convergência $\alpha_k^{(n)} \to \alpha_k$ é uniforme em k porque

$$|\alpha_k - \alpha_k^{(n)}| \leq M \|\tilde{\omega} - \omega^{(n)}\|. \tag{2.22}$$

Prova: Decorre diretamente do Teorema 2.9.

2.2.2 Convergência em Sistema Semiortonormal

Como foi até aqui considerado, um sistema semiortonormado de funções coordenadas leva necessariamente a um método de projeção numericamente estável. Em especial os menores autovalores $\lambda_1^{(n)}$ de todas as matrizes A_n são, para um tal sistema, limitados inferiormente por um $\lambda_0 > 0$:

$$0 < \lambda_0 \leq \lambda_1^{(n)}, \qquad \forall n \in \mathcal{N}.$$

Dessa propriedade de métodos de projeção, estáveis, podemos desenvolver ulteriores propriedades de convergência $\alpha_k^{(n)} \to \alpha_k$. Seja assim, por exemplo, W um sistema $\omega_1, \omega_2, \ldots$ completo tal que a condição (2.22) seja satisfeita. Consideremos dado k e escolhido n tal que $n \geq k$. Com relação à hipótese (2.22) λ_0 é portanto um limite inferior para os quocientes de Rayleigh construídos com A_n. Para todo $(\xi_1, \ldots, \xi_n) \in \mathcal{R}^n$ vale escrever

$$\|\sum_{j=1}^{n} \xi_j \omega_j\|^2 = \sum_{j,k=1}^{n} a_{jk} \xi_j \xi_k \geq \lambda_0 \sum_{j=1}^{n} \xi_j^2$$

e com a escolha especial $\xi_k = -1$ resulta que

$$\|\omega_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_j\|^2 \geq \lambda_0 > 0.$$

Então do Teorema 2.4 observamos que o sistema $\omega_1, \omega_2, \ldots$ é minimal. Decompondo o elemento ω_k em componentes ortogonais (Teorema 1.1)

$$\omega_k = v_k + u_k, \quad v_k \in V_k, \quad u_k \in W_k$$

obtemos, com $\omega_j \in W_k$ para $j \neq k$,

$$\|v_k\|^2 \geq \lambda_0 - \|u_k - \sum_{\substack{j=1 \\ j \neq k}}^{n} \xi_j \omega_j\|^2.$$

Todavia, $u_k \in W_k$ pode ser aproximado por uma combinação linear finita dos elementos $\omega_1, \ldots, \omega_{k-1}, \omega_{k+1}, \ldots$. Resulta daí que, para qualquer $\varepsilon > 0$ com n e $\xi_1, \ldots, \xi_n$ apropriados, vale

$$\|v_k\|^2 \geq \lambda_0 - \varepsilon$$

e também

$$\|v_k\|^2 \geq \lambda_0 \qquad \textit{para todo } k,$$

o que resulta na prova do enunciado do

Teorema 2.10 *Se para o sistema completo $\omega_1, \omega_2, \ldots$ em W é satisfeita a condição $0 < \lambda_0 \leq \lambda_1^{(n)} \quad \forall n \in \mathcal{N}$, então o sistema é também minimal. Além disso vale para os v_k do sistema biortogonal a expressão:*

$$\|v_k\|^2 \geq \lambda_0 > 0.$$

Em seguida podemos pensar em convergência uniforme em k da sequência $\alpha_k^{(n)}$ com base no Teorema 2.9 e Corolário 2.2, formulando o

Corolário 2.3. *Se para o sistema completo* $\omega_1, \omega_2, \ldots$ *em* W *a condição* $0 < \lambda_0 \neq \lambda_1^{(n)}$ *é satisfeita para* $\forall n \in N$, *então a convergência* $\alpha_k^{(n)} \to \alpha_k$ *é uniforme em* k.

Prova:
Do Teorema 2.9 e Corolário 2.2 a constatação de que $\alpha_k^{(n)} \to \alpha_k$ é uniforme em k, é imediata.

Poderíamos nos deter em continuar no exame da extensão desse tema, todavia somos da opinião de que os elementos de análise minuciosamente apresentados neste trabalho são suficientes para o entendimento global da teoria, especialmente em se tratando de estabilidade e convergência, e suficientes para o esclarecimento conceitual da teoria tratada aqui onde se colocaram indagações complexas e se consideraram aspectos intrínsicos ao desenvolvimento de sistemas minimais e semiortonormais.

Para continuidade da pesquisa sugerimos o tratamento de convergência uniforme porém também no sentido de que seja imposta a condição

$$\sum_{k=1}^{\infty} d_n^2 \to \infty \quad \text{, para} \quad n \to \infty$$

nos sistemas completos $\omega_1, \omega_2, \ldots$ em W, onde d_n seja dado, por exemplo, por

$$d_n := \left(\alpha_k - \alpha_k^{(n)}\right).$$

Além desse aspecto podemos examinar a relação ou relações que existem entre $\sum_{k=1}^{\infty} d_n^2$, $\|\tilde{\omega} - \omega^{(n)}\|$ (veja Teorema 2.8) e λ_0. Sobre convergência uniforme recomendamos Ruas [11] e Lusternik-Sobolev [8].

2.3 A Condição da Matriz de Ritz

(Elementos Finitos)

A matriz $A = (a_{jk})$ resultante da aplicação do Método dos Elementos Finitos apresenta a característica peculiar ao método que é a fraca ocupação, ou seja A é uma matriz esparsa em geral. Mas para dada região com contorno complicado

os elementos $a_{jk} \neq 0$ de A podem estar distribuídos muito irregularmente em A. Do ponto de vista de alguns casos especiais nas aplicações não é muito fácil obter-se uma boa aproximação para o número de condição da matriz A, sem que o menor e o maior autovalores de A sejam calculados numericamente.

Primeiramente, consideraremos o relacionamento existente entre os autovalores das matrizes A_r de um único elemento finito e os autovalores da matriz A. Esse relacionamento permite uma melhor inspeção no controle da estabilidade numérica, porém sua utilidade prática é muito limitada. Em seguida consideraremos o estudo da condição de A com referência ao problema (1.48) dado no capítulo 1.

Seja então considerada a matriz $A \in \mathcal{R}^{n\times n}$ e o vetor $x^t = (\xi_1, \ldots, \xi_n)$, $x \in \mathcal{R}^n$ e $\xi_\nu \in \mathcal{R}$, com $\nu = 1(1)n$. É claro que podemos estender a pertinência de A e x ao campo dos complexos, e consideraríamos, no caso, $A \in \mathcal{K}^{n\times n}$ e $x \in \mathcal{K}^n$. Além disso, seja A_n a matriz pertencente ao elemento finito ϱ_r, com $r = 1(1)N$. Então seja x_r^t o vetor que resulta de x^t através do fato de que todas as componentes ξ_j que não participam de ϱ_r são levadas a zero. Por exemplo, seja o elemento finito ϱ_r um triângulo com vértices P_j , P_{j+1} , P_{j+2}. Então temos:

$$x_r^t = (0, \ldots, 0, \xi_j, \xi_{j+1}, \xi_{j+2}, 0, \ldots, 0)\,.$$

Agora, quando um único nó $P_j (j = 1, \ldots, n)$, maximal em p, participa de vários elementos, então para um vetor qualquer arbitrário $x^t = (\xi_1, \ldots, \xi_n)$ e para os correspondentes x_r^t surgem sem dúvida as desigualdades

$$x^t x \leq \sum_{r=1}^{n} x_r^t x_r \leq p x^t x,$$

Além disso obtemos como já vimos

$$x^t A x = \sum_{r=1}^{n} x^t A_r x = \sum_{r=1}^{n} x_r^t A_r x_r,$$

em virtude de $A = A_1 + \ldots + A_n$.
Daí resulta, para $x \neq 0$:

$$\frac{x^t A_r x}{x^t x} \leq \frac{x^t A x}{x^t x} \leq p \left(\sum_{r=1}^{n} x_r^t A_r x_r \right) / \left(\sum_{r=1}^{n} x_r^t x_r \right).$$

Se Λ_{max} é o maior autovalor dentre todas as matrizes $A_1, \ldots, A_n$ e se λ_n é o maior autovalor de A, resulta portanto

$$\Lambda_{max} \leq \lambda_n \leq p\, \Lambda_{max}\,.$$

Suponhamos que Λ_{min} designe o mínimo dos valores de todos os quocientes de Rayleigh

$$R_r(x) = \frac{x_r^t A_r x_r}{x_r^t x_r} \ , \ x_r \neq 0,$$

então chega-se ao próximo seguinte resultado para o número de condição da matriz A, através de um raciocínio semelhante em relação ao menor autovalor λ_1 de A que leva à desigualdade

$$\Lambda_{min} \leq \lambda_1,$$

com a qual obtemos

$$cond(A) \leq p\frac{\Lambda_{max}}{\Lambda_{min}}.$$

Pode acontecer, entretanto, que $\Lambda_{min} = 0$ e com isso o valor de $cond(A)$ não tem sentido, ou seja, a aproximação para $cond(A)$ torna-se sem valor.

A solução de um problema variacional elíptico depende contínuamente dos dados; se a carga f e todos os deslocamentos e forças prescritas sobre o contorno $\partial\Omega$ *de* Ω forem pequenos então a energia de deformação em u é pequena, sendo u a solução do problema dado (1.47). Ou seja, o problema dado é denominado bem-posto ou bem-colocado. Além disso, independentemente da escolha do espaço de aproximação $S^h \subset H$ a energia de deformação na aproximação de Ritz $u^h \in S^h$ está automáticamente limitada pela energia de deformação na solução exata $u \in H$: ou seja, o método de Ritz projeta u sobre $S^h \subset H$ e isso só pode reduzir a referida energia. Portanto os problemas de aproximação são uniformemente bem-colocados e deve ser sempre possível construir um procedimento estável para a computação de u^h.

A dificuldade é a de conseguir-se estabilidade numérica total, pois às vezes o algorítimo usado não é definitivamente compatível, ou não é aquele que melhor se adapta às condições intrínsicas de estabilidade numérica.

Sabe-se que a chave para a estabilidade reside na independência linear uniforme das funções de base ω_j (veja Strang [6]). Embora u_h seja completamente independente da escolha das funções de base, o arredondamento que entra no seu cômputo depende sim dessa escolha (veja p.e. Michlin [5]). Michlin se refere a minimalidade forte das funções de base.

Para qualificar a independência linear das funções de base, o procedimento padrão é considerar-se a matriz de massa M cujos elementos são os produtos internos das funções de base

$$M_{jk} = (\omega_j, \omega_k) = \int_\Omega \omega_j(x)\omega_k(x)dx.$$

O método de Ritz opera sempre com a energia a(,) que é intrínsica ao problema, e com a qual a matriz de rigidez A_{jk} é construída. Ambas as matrizes são Hermitianas e positivas definidas.

Como uma primeira medida da independência da base, tem sido proposto o número de condição da matriz M dado por

$$Cond(M) := \frac{\lambda_{max}(M)}{\lambda_{min}(M)}.$$

Se a base é ortonormal, então M é a matriz identidade e daí resultando

$$Cond(M) = 1.$$

Esse não é o caso para os elementos finitos. Mas o importante para uma malha regular é que as funções de base dos elementos finitos sejam uniforme e linearmente independentes:

$$Cond(M) \leq \quad constante,$$

ou seja, os autovalores de (M) são todos da mesma ordem.

Há aplicações em que uma medida mais realística de independência é dada pelo número ótimo de condição definido por

$$Condot(M) := \min_{D}\{cond(DMD)\},$$

a matriz D aqui podendo ser qualquer matriz diagonal positiva que corresponda a um reescalonamento das funções de base. $Condot(M) = 1$ se a base original for só ortogonal ao invés de ortonormal. Com elementos irregulares algumas funções-tentativa podem ser muito menores do que outras e assim o reescalonamento poderia fazer grande diferença para o número de condição. Na realidade o reescalonamento sana, tão somente, dificuldades locais; se, por exemplo, um ω_j estiver fora de escala, como é o caso do elemento A_{jj} da diagonal ser muito pequeno, então o arredondamento tende a destruir precisões no valor computado dos Q_j, no sistema $AQ = F$.

A regra padrão do escalonamento de uma matriz esparsa positiva definida é a de manter iguais todos os elementos da diagonal. No caso dos elementos finitos, que são governados pela matriz de rigidez A, isto significa que as energias de deformação A_{jj} são iguais, segundo uma base convenientemente escolhida. Isso redunda naturalmente numa matriz diagonal D, de escalonamento, que é quase ótima (veja Van der Sluis [10]). Escalonamento significa aumentar ou diminuir, sob certos critérios, os elementos da matriz, segundo uma certa disposição, na mesma proporção.

Numa situação como no problema dado (1.48) o rescalonamento tende a ser efetivo de sucesso com a condição essencial de contorno, mas não com a condição

natural (Strang [6]). É o que acontece analogamente a uma situação física: Um sistema rígido de molas é altamente instável, mas quando ligado por um vínculo firme (condição essencial) o sistema é naturalmente estável. No primeiro caso estamos diante de um sistema físico mal-condicionado, que nem sempre pode ser alterado, exceto talvez por uma mudança de método: trocar o método da rigidez (deformação) pelo método da força (flexibilidade) onde as incógnitas são as tensões. Isso ocorre normalmente onde existe mudança brusca nas condições de rigidez do meio, ou quando o coeficiente de Poisson atingir o limite $\nu = 1/2$ de incompressibilidade (veja por exemplo Fried [9]).

Às vezes encontram-se situações em que o sistema linear resultante $AQ = F$ possui um vetor de carga F coincidente com um autovetor unitário v_{max} correspondente, em particular, ao autovalor λ_{max} de A. Esse fato pode ser importante no estabelecimento de uma relação entre número de condicão da matriz A e sua sensibilidade à perturbações. A solução, por exemplo, para o caso particular de v_{max} relacionado com λ_{max} de A é portanto

$$Q = \frac{v_{max}}{\lambda_{max}}.$$

Suponhamos a admissão de uma perturbação ε no vetor de carga produzida pelo autovetor no ponto de mínimo: $\varepsilon\, v_{min}$

$$\tilde{F} = v_{max} + \varepsilon v_{min}.$$

Então a solução $\tilde{Q}$ será dada por

$$\tilde{Q} = \frac{v_{max}}{\lambda_{max}} + \varepsilon \frac{v_{min}}{\lambda_{min}}.$$

A variação relativa na solução será então

$$\frac{|Q - \tilde{Q}|}{|Q|} \sim \varepsilon \frac{\lambda_{max}}{\lambda_{min}} = \varepsilon\, cond(M),$$

o que mostra ser a perturbação em F de ordem ε, e que a mesma é ampliada para dar uma perturbação em Q da ordem de $\varepsilon cond(M)$.
Mas por outro lado, para $\delta Q = Q - \tilde{Q}$ e $\delta F = F - \tilde{F}$, temos

$$|F| = |AQ| \leq \lambda_{max}|Q|$$
$$\lambda_{min}|\delta Q| \leq |A\delta Q| = |\delta F|\,,\ \delta F = F - \tilde{F}.$$

Multiplicando essas desigualdades membro a membro, obtemos:

$$|F|\lambda_{min}|\delta Q| \leq \lambda_{max}|Q|\,|\delta F|$$

resultando daí a desigualdade final

$$\frac{|\delta Q|}{|Q|} \leq cond(M)\frac{|\delta F|}{|F|},$$

para $Q \neq \odot$ e $F \neq \odot$, com $\odot$ designando a matriz nula. Para medida a priori de sensibilidade o fato é que o número de condição tem demonstrado ser satisfatório.

Para estimativa do número de condição da matriz A de rigidez, daremos aqui o enunciado do Teorema (2.11) (veja p.e. Strang [6])

Teorema 2.11 *Para todo problema variacional e toda escolha do elemento finito existe uma constante c tal que*

$$cond(A) \leq c.h_{min}^{-2m},$$

sendo 2m a ordem do operador diferencial do problema dado e h_{min} a menor dimensão dos elementos na malha. A constante depende inversamente do menor autovalor λ_1 de A do problema dado e cresce se a geometria dos elementos se tornar degenerada.

Conclusão: O erro de arredondamento não depende fortemente do grau do elemento polinomial. Depende principalmente de h, da ordem e do autovalor fundamental do operador que traduz o problema contínuo dado. Portanto, o meio de conseguir-se precisão numérica em face do arredondamento é aumentar o grau das funções-tentativa. O número de condição com base em elementos cúbicos é ligeiramente pior do que para os elementos lineares, de tal maneira que os erros de arredondamento são comparáveis para um dado h. O erro de discretização, contudo, é de ordem de grandeza menor para os elementos cúbicos. Portanto no ponto de cruzamento, onde o arredondamento impede qualquer melhoria oriunda da redução de h, o elemento cúbico é muito mais preciso. Isso se aplica especialmente na computação de tensões onde a diferenciação do campo de deslocamento introduz o fator extra h^{-1} no erro numérico. Mesmo nos problemas de 2a. ordem o arredondamento se torna significante, e um aumento no grau das funções-tetativa é altamente benéfico.

Strang and Fix /6/ mostram em suas observações que para malhas uniformes o número de condição da matriz de Gram é uma maneira razoavelmente aceitável de medir-se estabiblidade. Todos os polinômios seccionais usados como funções-teste fornecem uma base estável onde a estabilidade é indicada pelo fato de que o número de condição da matriz A permanece limitado, quando a malha é refinada ($h \to 0$). A condição de A resultante da aplicação do método dos elementos finitos sobre um operador uniformemente elíptico de ordem 2m é dada por $c\,h^{-2m}$ onde h é o tamanho da malha e c é uma constante que depende da escolha da base.

Isto significa que para um dado problema, enquanto usarmos uma base estável, o número de condição da matriz de Gram não se altera, a medida que aumentamos o grau dos polinômios.

A matriz $A = (a_{ij})$ é em geral, fracamente ocupada. Entretanto em regiões de contorno mais complicado os elementos não nulos da matriz A podem se apresentar distribuidos muito irregularmente. Por essa razão torna-se difícil fazer-se uma avaliação para fixarmos aproximadamente um valor para o número de condição de A , sem que seja necessário calcular numéricamente os maior e menor autovalores de A. Contudo, conseguimos estabelecer relações entre as matrizes A_r de um simples elemento com os autovalores da matriz A que fornecem uma visão mais realista, um controle mais sensitivo da estabilidade, mas que muita das vezes parecem não denotar utilidade prática, dependendo muito da situação do modelo matemático. Logicamente aqui influi muito o fato do número de condição, dependente do tamanho de h e da ordem do operador governantes do modelo, poder indicar que A seja mal condicionada. Mas no entanto as dificuldades numéricas podem estar sendo geradas não por esse mal condicionomento, mas sim pela má escolha do algorítimo. A experiência tem mostrado que esse tipo de comportamento é comum nas matrizes resultantes do uso de elementos mistos, por exemplo, especialmente quando o contorno é complicado. Atualmente podemos resolver problemas de grande porte, especialmente os de contorno complicado, usando os esquemas das matrizes eficientemente orientadas por banda ou a estratégia do multigrid, hoje amplamente empregada.

Bibliografia

[1] Aziz,A.K(1972) - The mathematical foundatiom of finite element method with applications to partial differential equations.

[2] Aubin,J.P.(1972) - Appromiation of elliptic boundary value problems. London - N.York - Sídney.

[3] Babuska,I Prager,M. Vitásek,E.(1966). Numerical Process in Differential Equations. London - N.York - Sídney.

[4] Ciarlet,P.G. - The Finite Element Method For Elliptic Problems. North Holland Publ. Co.

[5] Michlin,S.G.(1962)(1) Variationsmethoden der Mhatematischen Physik. Berlim (1965)(2) The problem of the minimum of a quadratic functional. S.Francisco, London, Amsterdam.

[6] Strang,G. Fix,G.J.(1973). An Analysis of the Finite Element Method.

[7] Zienkiewicz,).C.(1975). Methoden der finiten Elemente Munchen - Wien.

[8] Lusternik,L.A.,Sobolev,V.J.(1974) - Elements of Functional Analysis. J. WILEY & SONS.

[9] Fried,I.(1971) - Condition of finite element matrices generated from nonuniform meshes. Math. Rept. 7 Lakehead Univ. - Canadá.

[10] Van der Sluis, A.(1970) Condition, equilibration, and pivoting in linear algebraic systems. Numer. Math. 15,74 - 86.

[11] Ruas, V. (1979) - Introdução aos Problemas Variacionais - Guanabara Dois.

[12] Courant,R.;Hilbert,D.(1962) Methods of Mathematical Physics. Interscience.

[13] Isaacson,E;Keller,H.B.(1966) Analysis of Numerical Methods,J.Wiley.

[14] Fujii,H.(1972) Finite Element Schemes; Stability and Convergence, Second U.S. - Japan Seminar.