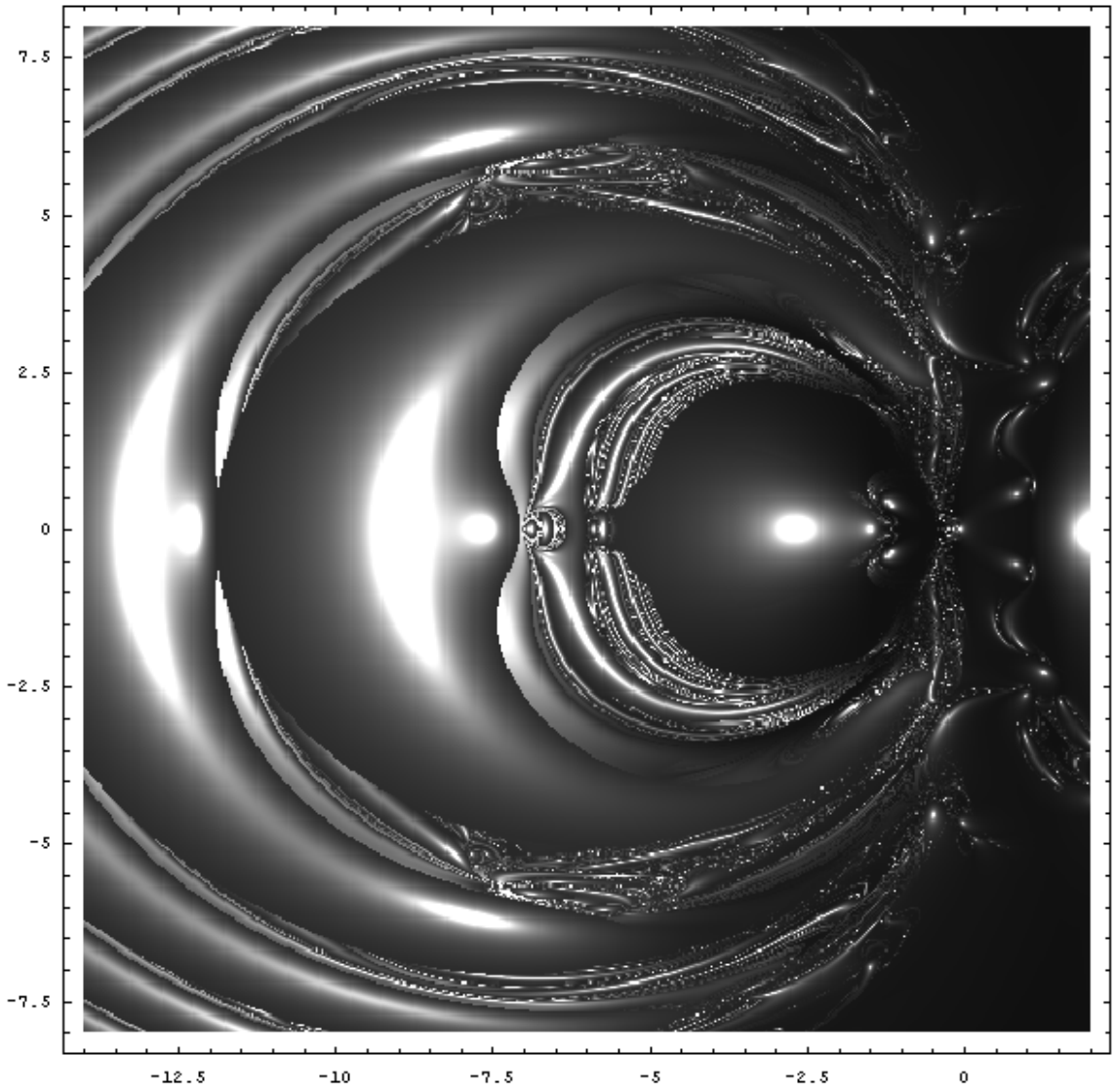


Fundamentos de Análise Numérica (I)

Teoria e Exercícios

Carlos J. S. Alves

Versão de 2001/2002

Departamento de Matemática
Instituto Superior Técnico, Universidade Técnica de Lisboa
Av. Rovisco Pais 1, 1049-001 LISBOA, PORTUGAL

Versão também disponível na AEIST - Associação de Estudantes do Instituto Superior Técnico

Indíce

1 Introdução **1**

1.1 Representação de números . . . 3
1.1.1 Números reais . . . 4
1.1.2 Sistema de ponto flutuante . . . 5
1.1.3 Tipos de arredondamento . . . 6
1.1.4 Erro absoluto e erro relativo . . . 7
1.2 Algoritmos e propagação de erros . . . 8
1.2.1 Propagação do erro em funções de uma variável . . . 10
1.2.2 Propagação do erro em operações . . . 11
1.2.3 Propagação de erros de arredondamento . . . 12
1.2.4 Algoritmos e rotinas elementares . . . 14
1.3 Condicionamento e estabilidade numérica . . . 15
1.4 Exercícios . . . 21

2 Determinação de Raízes Reais e Complexas **23**

2.1 Convergência de sucessões . . . 26
2.1.1 Ordem de convergência . . . 29
2.1.2 Tempo de cálculo . . . 33
2.2 Teoremas elementares e a localização de raízes . . . 35
2.3 Método da Bissecção . . . 37
2.4 Método da Falsa Posição . . . 40
2.4.1 Método da falsa posição modificado . . . 46
2.5 Método do ponto fixo num intervalo limitado . . . 47
2.5.1 Aceleração de convergência . . . 56
2.6 Método de Newton . . . 58
2.6.1 Convergência do método de Newton . . . 59
2.6.2 Fórmula de erro do método de Newton . . . 64
2.6.3 Método de Newton no caso de zeros múltiplos . . . 67
2.6.4 Método da Secante . . . 68
2.6.5 Métodos de ordem superior . . . 72
2.7 Métodos para Equações Algébricas . . . 75
2.7.1 Método de Bernoulli . . . 78
2.7.2 Condicionamento no cálculo de raízes . . . 80
2.8 Generalização a raízes complexas . . . 82

2.8.1 Método de Newton nos complexos 85
2.9 Exercícios 87

3 Teorema do Ponto Fixo de Banach **94**
3.1 Espaços Normados 95
3.1.1 Noções Topológicas em Espaços Normados 97
3.1.2 Normas equivalentes 98
3.2 Espaços de Banach 101
3.2.1 Operadores Contínuos 103
3.2.2 Operadores Lineares 104
3.3 Método do Ponto Fixo e o Teorema de Banach 105
3.4 Derivação de Fréchet 109
3.4.1 Corolário do Teorema do Ponto Fixo 111
3.4.2 Comportamento assimptótico da convergência. 113
3.4.3 Convergência de ordem superior 115
3.4.4 Método de Newton 116
3.5 Exercícios 117

4 Resolução de Sistemas de Equações **122**
4.1 Normas de Matrizes 122
4.2 Métodos Iterativos para Sistemas de Equações Não Lineares 126
4.2.1 Método de Newton para Sistemas de Equações 128
4.3 Métodos Iterativos para Sistemas Lineares 132
4.3.1 Métodos de Jacobi e Gauss-Seidel 132
4.3.2 Convergência dos Métodos de Jacobi e Gauss-Seidel 134
4.3.3 Métodos de Relaxação e SOR 141
4.4 Métodos Directos para Sistemas Lineares 145
4.4.1 Condicionamento de um Sistema Linear 145
4.4.2 Método de Eliminação de Gauss 149
4.4.3 Número de Operações 151
4.4.4 Métodos de Factorização ou Compactos 153
4.4.5 Métodos Iterativos e Métodos Directos 158
4.5 Exercícios 160

5 Determinação de Valores e Vectores Próprios de Matrizes **164**
5.1 Noções básicas 164
5.1.1 Valores próprios e o polinómio característico 168
5.2 Teorema de Gerschgorin 169
5.3 Método das Potências 174
5.4 Método das iterações inversas 181
5.5 Métodos de Factorização 183
5.5.1 Método LR 183
5.5.2 Método QR 184
5.5.3 Método QR com deslocamento 186
5.6 Condicionamento do cálculo de valores próprios 187

5.7 Cálculo de raízes polinomiais . . . 188
5.8 Exercícios . . . 190

6 Introdução a métodos de optimização **193**
6.1 Método dos mínimos quadrados . . . 193
6.1.1 Aproximação no caso discreto . . . 196
6.1.2 Aproximação no caso contínuo . . . 197
6.1.3 Dependência não linear nos coeficientes . . . 197
6.2 Minimização sem restrições . . . 198
6.2.1 Métodos de descida . . . 200
6.2.2 Método do gradiente . . . 201
6.2.3 Aplicação à resolução de sistemas . . . 205
6.2.4 Método do gradiente conjugado . . . 206
6.2.5 Método de Newton . . . 209
6.2.6 Método das coordenadas . . . 210
6.3 Exercícios . . . 211

7 Anexos **214**
7.1 Resultados Elementares de Análise . . . 214
7.1.1 Funções de várias variáveis reais . . . 214
7.1.2 Funções de uma variável complexa . . . 216
7.2 Equações às Diferenças . . . 219
7.2.1 Soluções de uma Equação às Diferenças Homogénea . . . 221
7.2.2 Equações às diferenças não homogéneas . . . 222
7.2.3 Método da variação de constantes . . . 223
7.2.4 Exercícios . . . 224
7.3 Teoria de Erros em Espaços Normados . . . 225
7.3.1 Erro, Absoluto e Relativo . . . 225
7.3.2 Propagação de Erros . . . 225
7.4 Operadores integrais e matrizes . . . 228

8 Exercícios resolvidos **232**
8.1 Exercícios de avaliações anteriores . . . 232
8.1.1 1a. Avaliação (97) . . . 232
8.1.2 2a. Avaliação (97) . . . 234
8.1.3 Teste (98) . . . 235
8.1.4 1a. Avaliação (98) . . . 236
8.1.5 2a. Avaliação (98) . . . 238
8.1.6 Teste (99) . . . 240
8.1.7 1a. Avaliação (99) . . . 241
8.2 Resoluções . . . 243
8.2.1 1a. Avaliação (97) . . . 243
8.2.2 2a. Avaliação (97) . . . 246
8.2.3 Teste (98) . . . 250
8.2.4 1a. Avaliação (98) . . . 252

8.2.5 2a. Avaliação (98) . . . 257
8.2.6 Teste (99) . . . 261
8.2.7 1a. Avaliação (99) . . . 264
8.3 Trabalhos computacionais . . . 268
8.3.1 Trabalhos computacionais (97) . . . 268
8.3.2 Trabalhos computacionais (98) . . . 270
8.3.3 Trabalhos computacionais (99) . . . 274
8.4 Glossário . . . 277

Prefácio

Estas folhas seguem essencialmente os cursos de Análise Numérica I leccionados em 1997, 98 e 99 aos alunos da Licenciatura em Matemática Aplicada e Computação do Instituto Superior Técnico. A disciplina Análise Numérica I é leccionada no segundo ano, segundo semestre, pressupondo-se que os alunos já possuem conhecimentos elementares de Álgebra Linear e de Análise Matemática (cujo quarto semestre decorre em simultâneo). Convém referir que Análise Numérica I é complementada pela disciplina de Análise Numérica II, leccionada no semestre seguinte. Durante estes dois semestres são introduzidos não apenas todos os aspectos fundamentais de Análise Numérica, que são leccionados num único semestre nas licenciaturas em engenharia, mas também algumas noções mais profundas dentro da teoria elementar. Assim, na Análise Numérica I, o objecto de estudo é essencialmente a aproximação de soluções de equações, numa perspectiva abstracta que difere profundamente da apresentada nas licenciaturas em engenharia, sendo ainda adicionados dois capítulos introdutórios noutras matérias, um relativo à aproximação de valores próprios de matrizes e um outro, muito breve, relativo a métodos minimização (que poderia também ser ligado à aproximação de soluções de equações).

A matéria apresentada representa não apenas uma solução de compromisso entre aspectos mais teóricos e aspectos mais numéricos, seguindo tópicos de programas anteriores (cf. L. Loura [21] e P. Lima [20]), mas encerra também uma visão pessoal dos fundamentos da Análise Numérica. Podemos dizer que grande parte do curso está centrado no teorema do ponto fixo de Banach, já que as suas consequências acabam por fundamentar muitos aspectos dos métodos numéricos utilizados neste curso introdutório. Esta perspectiva unificadora leva à introdução de noções abstractas de topologia e de análise funcional elementar. O teorema de Banach surge como uma generalização de um teorema elementar nos números reais, num dos casos notáveis em que não se trata de generalizar por generalizar, aspecto simplista e por vezes nefasto na matemática. A abstracção deve ser acompanhada da clara indicação das ideias elementares que a originaram. Uma vez compreendidas essas noções será mais fácil estabelecer o paralelismo e tentar apreender até que ponto é possível e útil a generalização. Quando a aplicação da teoria abstracta se resume ao caso elementar que a gerou, trata-se de uma abstração estéril. Não é isso que acontece com o teorema de Banach. A ideia do teorema do ponto fixo é aplicável não apenas à solução de equações com funções reais ou complexas, mas também a sistemas de equações lineares ou não lineares, a equações diferenciais ou integrais e a outros problemas. Devemos também ter presente que quando colocamos o problema num contexto abstracto, como os resultados são aplicáveis a todas as estruturas que aí estejam enquadradas, perdemos informação acerca da riqueza da estrutura que pretendemos estudar em particular. Assim, o método do ponto fixo será um utensílio razoavelmente geral, mas haverá outros métodos que se adequam particularmente em cada contexto.

Desta forma, no Capítulo 2 é introduzido o método do ponto fixo no caso real e complexo, como uma motivação para a obtenção dos resultados mais abstractos que encontramos no Capítulo 3, mas também são apresentados outros métodos particulares. É claro que poderia ser apresentado imediatamente o caso mais geral, enunciando o teorema de Banach

e deduzindo depois o resultado no caso real como um exercício. No entanto, há dois pontos em ter em conta. Por um lado, não é esse o processo natural de descoberta em matemática, e da descoberta da matemática, parecendo mais conveniente introduzir o conceito como uma generalização. Por outro lado, será conveniente saber aplicar o teorema na sua generalidade aos diversos casos particulares. Assim, no Capítulo 4 são obtidos um conjunto de resultados que devem ser vistos como corolários (ou exercícios...) da teoria apresentada no Capítulo 3, quando aplicada a espaços vectoriais de dimensão finita. Mesmo os Capítulos 5 e 6, acerca da determinação de valores próprios e introdução a métodos de optimização (respectivamente), contêm alguns resultados que surgem como consequência do Capítulo 3.

Havendo várias linguagens de programação que podem acompanhar um curso de Análise Numérica, a escolha da linguagem *Mathematica* é uma herança da introdução dessa linguagem como base de programação na licenciatura. Trata-se de uma linguagem de alto nível, e poderá tornar-se num modelo para futuras linguagens de programação. O tempo de cálculo é em muitas circunstâncias ainda um problema que é compensado pela facilidade de programação e pela maior fiabilidade dos resultados. Como acontece com linguagens de alto nível, o *Mathematica* pode ser um bom banco de ensaios para programas mais sofisticados que poderão depois ser desenvolvidos em linguagens de mais baixo nível, onde o tempo de cálculo poderá ser muito menor (p.ex. o Fortran ou o C... é claro que um bom conjunto de rotinas nessas linguagens pode servir o mesmo propósito). Ao longo do texto são colocadas várias notas referentes a rotinas do *Mathematica*, às suas limitações e a alguns erros que foram detectados por nós desde a versão 2 até à 4. Para maior detalhe acerca da linguagem e das suas possibilidades, salientamos o aspecto mais prático em [2] e formal em [4].

Uma palavra de agradecimento aos alunos que pela receptividade e empenho permitiram uma maior motivação na compilação destas folhas e também aos meus colegas Ana Silvestre e Mário Graça, pela pertinência em críticas e algumas sugestões. Ao meu colega Mário Graça devo ainda uma valiosa lista de gralhas da primeira versão de 1999.

Carlos J. S. Alves

Capítulo 1

Introdução

A Análise Numérica envolve vários aspectos distintos que a ligam directamente à Álgebra, à Análise Matemática, e à Programação. Normalmente, um problema de física (ou de engenharia), leva a formular um modelo, o qual se torna objecto de estudo da matemática aplicada. Para a resolução desse problema, não é suficiente a matemática apresentar resultados teóricos, por exemplo, acerca da existência de solução. Num problema prático é necessário saber construir, ou aproximar, essa solução. Esse é o principal aspecto da Análise Numérica.Por outro lado, mesmo que a construção seja exacta, ela envolve normalmente cálculos infinitos, impossíveis de realizar, pelo que é necessário saber controlar o erro que cometemos ao considerar aproximações; este é outro aspecto importante dentro da Análise Numérica. Finalmente, a realização dos cálculos deve ser optimizada de forma a obtermos rapidamente um bom resultado, através da programação de algoritmos eficientes. A programação constitui portanto um outro aspecto de relevo na Análise Numérica, pois ela permite a obtenção do resultado final. Todos estes passos têm originado o progresso tecnológico a que nos habituámos. A investigação tecnológica leva a novos modelos que não dispensam uma simulação numérica *a priori*. Esses modelos numéricos estão presentes, por exemplo, na propagação de ondas, no escoamento de fluidos ou na resistência de materiais, cujas aplicações na indústria englobam telecomunicações, engenharia mecânica, engenharia de materiais, engenharia civil, aeroespacial, etc...

Breve apontamento histórico

Para compreender o enquadramento da análise numérica na matemática, será conveniente recordar brevemente o seu percurso histórico.

Na antiguidade, o desenvolvimento da matemática ficou a dever-se sobretudo aos gregos, ainda que haja registo de algumas proezas assinaláveis na Babilónia e no Egipto, como seja a resolução de equações de segundo grau. Os gregos basearam a matemática em relações geométricas, tendo o expoente dessa abordagem sido Euclides, que formulou uma teoria (*Elementos*, séc.III-a.C.) baseada em 5 axiomas que permitem deduzir a geometria euclidiana (curiosamente, só no séc.XIX se confirmou que o quinto axioma: "por um ponto exterior a uma recta passa apenas uma paralela" – era mesmo necessário... ou seja, não se deduzia dos restantes).

A concepção grega que preconizava a construção de entidades geométricas através de régua e compasso deixou vários problemas em aberto que só muito mais tarde, com o desen-

volvimento da Análise Matemática foram resolvidos. Por exemplo, a questão da quadratura do círculo, relacionada com a transcendência de π, só foi resolvida no fim do séc.XIX por Hermite e Lindemann. Na antiguidade, o valor atribuído a π era normalmente grosseiro (por exemplo, na Bíblia há citações que indiciam $\pi = 3$), e o primeiro a tentar determinar por aproximações o valor de π foi Arquimedes, baseado no *método da exaustão*, um método que havia sido introduzido por Eudoxo (discípulo de Platão). Considerando dois polígonos, um interior e outro exterior, Arquimedes determinou que $3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}$. Podemos considerar este resultado como um dos primeiros resultados de Análise Numérica.

Só no Renascimento é que a matemática vai sofrer um novo grande impulso, começando com a escola italiana de Bolonha, no séc.XVI, com Tartaglia, Cardano, Ferrari, os quais encontraram fórmulas resolventes para as equações de terceiro e quarto grau. Este desenvolvimento da álgebra vai ligar-se à geometria com Descartes e Fermat (séc.XVII), e surge então a geometria analítica que está na origem do cálculo infinitesimal de Newton e Leibniz (séc.XVIII).

Com o cálculo infinitesimal começaram algumas bases da Análise Numérica. Até ao séc.XIX ela coabitou indistintamente com a Análise Matemática, pois o formalismo axiomático que mais tarde se pretendeu copiar de Euclides ainda não havia sido adoptado. No séc. XIX, com o esforço de rigor de Cauchy, Bolzano e Weierstrass, e posteriormente com as ideias de Dedekind e Cantor, os matemáticos empenharam-se em formalizar axiomaticamente toda a Análise a partir da teoria de conjuntos, com Zermelo e Fraenkel.

Mas a questão axiomática não foi, nem é, pacífica... desde a aceitação do axioma da escolha até à polémica construtivista no princípio do séc.XX. Convém notar que surgiram duas posições claramente distintas. De um lado Hilbert, defendendo a existência de um *infinito actual*, e sustentando que a partir de um número finito de axiomas seria possível deduzir qualquer proposição matemática, e de outro lado, Brouwer, defendendo apenas a existência de um *infinito potencial*. A posição de Hilbert mostrou-se ser irrealizável quando Gödel, nos anos 30, baseado no método da diagonal de Cantor, demonstrou a existência de proposições semanticamente verdadeiras que no entanto não podem ser deduzidas usando os critérios defendidos por Hilbert (teorema da incompletude de Gödel). Essencialmente por razões práticas, a maioria da Análise ainda faz uso da noção de infinito actual, o que permite a existência de demonstrações não construtivas, através de uma *redução ao absurdo*, e utilizando por vezes o *polémico* axioma da escolha. As demonstrações não-construtivas não fornecem qualquer método para explicitar as entidades cuja existência foi provada no sentido clássico, sendo necessário, na prática, resultados construtivos que permitam calcular essas entidades.

A Análise Numérica contém a componente de construtividade da Matemática, fornecendo métodos de aproximação (que podem constituir eles próprios demonstrações construtivas), e estudando o erro inerente em cada passo da aproximação. A noção de *erro* e o controlo do erro constitui assim um aspecto fulcral na Análise Numérica, que começaremos por abordar no caso mais simples, quando nos apercebemos que a própria representação de números reais tem que ser aproximada quando trabalhamos com máquinas de aritmética finita.

Terminamos o breve apontamento histórico, referindo que a Análise Numérica começou a surgir como disciplina bem identificada recentemente, já que o desenvolvimento de métodos numéricos acentuou-se com o aparecimento dos primeiros computadores, especialmente a

partir dos anos 50. Se anteriormente o processo manual de cálculo tornava irrealizáveis muitos métodos, a automatização permitiu a implementação de processos iterativos simples. Paralelamente a esse desenvolvimento computacional novos problemas teóricos foram colocados, mesmo para a resolução de problemas elementares, já que o tempo de cálculo e a estabilidade numérica tornaram-se factores determinantes na eficácia dos algoritmos. Sendo difícil fazer uma análise objectiva de um passado recente, para o qual muitos contribuiram, talvez mereça um especial realce os trabalhos de Wilkinson nos anos 60, especialmente no que diz respeito a algumas bases do cálculo numérico. É claro que, havendo várias áreas onde é importante o desenvolvimento de métodos numéricos (e a também a sua justificação teórica), a própria Análise Numérica pode ser dividida em múltiplas áreas particulares, que vão desde métodos numéricos para problemas elementares, até ao desenvolvimento e estudo de novos métodos para equações diferenciais ou integrais, por exemplo. A evolução na simulação computacional de fenómenos físicos é um bom exemplo onde os métodos numéricos têm determinado grande parte da evolução tecnológica mais recente.

Alguns dos assuntos que iremos abordar podem parecer mais ou menos elementares, ou mais ou menos actuais. O nosso propósito não é apresentar uma lista dos mais eficazes métodos numéricos que existem actualmente para resolver determinados problemas específicos... até porque essa lista correria o risco de estar desactualizada rapidamente. O nosso objectivo é apenas apresentar várias possibilidades de raciocínio que poderão ser úteis na resolução de problemas mais complicados. Essa possível generalização é exemplificada no texto com a adaptação de um método simples, o método do ponto fixo, para contextos menos triviais. Refira-se que, mesmo assim, a maioria dos métodos apresentados são actuais e são ainda utilizados nos mais diversos contextos.

Por fim, salientamos que na sua parte teórica a Análise Numérica abrange diversas áreas da Matemática, desde a Análise à Álgebra, pelo que os seus resultados e métodos podem ser enquadrados em qualquer um desses campos. A única eventual distinção reside na componente de construtividade (como já foi referido) ou nos propósitos subjacentes, já que na Análise Numérica se procura o cálculo efectivo e o controlo desse cálculo através de estimativas de erro. Na sua parte aplicada a Análise Numérica necessita de um conhecimento prático de programação, de linguagens de baixo e alto nível, bem como alguns conhecimentos de computação gráfica.

1.1 Representação de números

Um dos primeiros aspectos numéricos que apreendemos é o sistema de numeração árabe. Há nesse sistema um grande avanço face aos que o precederam historicamente (ex: gregos, romanos), quer na facilidade de execução de operações, quer no optimizar da extensão da notação. O sistema de numeração árabe é um avanço numérico que passou despercebido à matemática grega, mais preocupada com propriedades conceptuais relacionadas com a geometria. O sistema árabe permitiu que o cálculo numérico fosse facilmente mecanizado no que diz respeito às operações elementares, libertando-se de qualquer interpretação aplicada à geometria ou a qualquer outro modelo intuitivo. O exemplo máximo dessa mecanização teve como pioneiros B. Pascal ou C. Babbage que construiram as primeiras máquinas de

calcular mecânicas e que culminaram, passados quase três séculos, no aparecimento de computadores electrónicos.

Por uma questão de eficácia, a nível interno os computadores trabalham com um sistema de numeração binária ao invés do habitual sistema decimal que utilizamos correntemente. No entanto, o processo de efectuar as operações elementares é exactamente o mesmo, qualquer que seja a base que se considere. Qualquer número natural pode ser representado numa base β escrevendo

$$x = a_n ... a_1 a_0$$

em que cada $a_i \in \{0, .., \beta - 1\}$, significando que $x = \sum_{i=0}^{n} a_i \beta^i$.

Assim, $x = 110.1$ na base 2 representa o número 6.5, pois $x = 1 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} = 4 + 2 + 0 + \frac{1}{2} = 6.5$, mas a mesma representação $x = 110.1$ na base decimal ($\beta = 10$) tem o significado habitual, que designa um número completamente diferente... A utilização destes sistemas numéricos tem a vantagem de permitir representar números não inteiros através da introdução do ponto (ou da vírgula...). No entanto, quando o número de algarismos é finito esses números são apenas alguns dos números racionais.

Vejamos quais as possibilidades da notação decimal servir para uma potencial representação de qualquer número real.

1.1.1 Números reais

Começamos por rever a noção de número real. Para isso vamos considerar a construção de Cantor dos números reais a partir dos racionais (e.g. [14]), não seguindo a perspectiva axiomática dos números reais (e.g. [11]).

Começamos por considerar a noção de sucessão de Cauchy de racionais.

Definição 1.1 *Dizemos que uma sucessão* (x_n) *de racionais é uma* sucessão de Cauchy *se*

$$\forall \delta > 0 \ \textit{(racional)} \ \exists p \in \mathbb{N} : |x_n - x_m| < \delta, \ (n, m \geq p)$$

Definição 1.2 *Um* número real *é uma classe de equivalência de sucessões de Cauchy de racionais, definida pela relação de equivalência*

$$(x_n) \sim (y_n) \Longleftrightarrow \forall \delta > 0 \, \exists p \in \mathbb{N} : |x_n - y_m| < \delta, \ (n, m \geq p)$$

Cada elemento da classe de equivalência pode ser um seu representante, mas normalmente tomamos como representante da classe uma sucessão de racionais que são múltiplos de potências de 10 (base decimal). Ou seja, normalmente escrevemos um número real na *notação científica* (decimal):

$$x = \pm 0.a_1 a_2 \ldots a_n \ldots \times 10^t.$$

Com efeito a sucessão de racionais

$$\begin{gathered} x_1 = \pm 0.a_1 \times 10^t, \\ x_2 = \pm 0.a_1 a_2 \times 10^t, \\ \ldots, \\ x_n = \pm 0.a_1 a_2 \ldots a_n \times 10^t, \\ \ldots \end{gathered}$$

é uma sucessão de Cauchy, e é o representante escolhido na notação científica para a classe de equivalência x.

No caso da notação científica, um número representa-se através do sinal, da mantissa e do expoente, na base decimal. Os dígitos a_n variam entre 0 e 9, mas o primeiro dígito da mantissa deve ser diferente de zero (o número zero é representado à parte). Nesta representação pode haver uma ambiguidade, já que o número real 1 pode ser representado pela sucessão 1.0000... ou por 0.99999..., mas podemos considerar uma representação única se privilegiarmos sempre uma aproximação superior (em valor absoluto).

Ao efectuarmos cálculos e, a menos que estivessemos na posse de uma máquina com memória infinita (uma teórica *máquina de Turing*), a representação de um número tem que ser finita. Consequentemente somos obrigados a considerar um número finito de dígitos na mantissa e uma limitação nos valores dos expoentes admitidos, o que leva à noção de representação num sistema de ponto flutuante[1].

1.1.2 Sistema de ponto flutuante

Um sistema de ponto flutuante $FP(\beta, n, t_1, t_2)$ é um subconjunto finito de racionais, que se caracteriza pela base β (usualmente trabalhamos com base decimal, $\beta = 10$, mas internamente as máquinas usam a base binária, $\beta = 2$), pelo número de dígitos na mantissa n, e por uma limitação nos expoentes que podem tomar valores entre t_1 e t_2.

Definição 1.3 *.Sejam $n \in \mathbb{N}$, $t_1, t_2 \in \mathbb{Z}$. O subconjunto dos racionais,*

$$FP(10, n, t_1, t_2) = \left\{ x \in \mathbb{Q} : \begin{array}{c} x = 0 \quad ou \\ x = \pm 0.a_1 a_2 \ldots a_n \times 10^t,\ a_i \in \{0, ..., 9\}, a_1 \neq 0, t \in \{t_1, ..., t_2\} \end{array} \right\}$$

é designado por sistema de ponto flutuante *em base decimal, com n dígitos na mantissa, e expoentes variando entre t_1 e t_2.*
Quando apenas for importante referir o número de dígitos na mantissa, usamos a notação $FP(n)$.

Este tipo de conjuntos é adequado a uma representação aproximada dos reais já que existe densidade dos racionais nos reais (o que é uma consequência da definição apresentada). Assim, qualquer número real pode ser aproximado 'tanto quanto se queira' através de um sistema de ponto flutuante, desde que consideremos um número de dígitos suficientemente grande na mantissa e expoentes adequados... claro que essa precisão será paga em termos de memória computacional e de tempo de cálculo.

Usualmente, um número em *precisão simples* utiliza 32 *bits* (4 *bytes*) e um dos sistemas usado é $FP(2, 24, -127, 127)$ (conjunto que contém mais de 2 mil milhões de números)

[1]Seguimos a designação usada em [27]. É também usada a designação *vírgula flutuante* já que a tradição europeia insiste em representar os números decimais através da vírgula. Essa tradição provoca frequentes *bugs* na interacção de software europeu e americano. Bases de dados com dígitos escritos com vírgulas não são interpretadas devidamente por programas em que se prevê o uso do ponto. Como a esmagadora maioria do software produzido internacionalmente utiliza o ponto, começa a fazer tanto sentido insistir na questão da vírgula, quanto o uso de medidas imperiais anglo-saxónicas ao invés do Sistema Internacional . Da mesma forma, ao longo do texto irá usar-se a notação internacional sin para seno e não sen.

gerando erros relativos de arredondamento da ordem de 0.6×10^{-7}. Se quisermos erros relativos inferiores a 0.2×10^{-16}, teremos que utilizar *precisão dupla*, por exemplo, $FP(2, 56, -127, 127)$, o que consumirá o dobro de memória: 8 *bytes.*

Observação: Na linguagem *Mathematica* o cálculo numérico utiliza, por defeito, uma precisão de 16 dígitos, o que corresponde a precisão dupla. Esse valor pode ser aumentado usando a rotina SetPrecision. É possível saber o número de *dígitos correctos*[2] num número apresentado usando as rotinas Precision ou Accuracy.

• Note-se que há uma *enorme diferença* entre escrever 1 ou 1.0 no *Mathematica,* o primeiro valor é considerado um símbolo, o segundo é encarado como um valor numérico. A diferença cifra-se especialmente no tempo de cálculo, podendo dar origem a tempos de espera insuportáveis. Por exemplo, consideremos um ciclo em que calculamos $x_{n+1} = \cos(x_n)$. Se introduzirmos o valor 1 em x_0 o resultado será cos(cos(cos(cos(1)))) ao fim de 4 iterações. O *Mathematica* não simplica a expressão porque como é dado um símbolo que ele considera exacto (se fizer Precision[1] obterá ∞), tentará ser coerente e dar um valor com a mesma precisão... portanto não poderia dar um valor decimal (com apenas 16 dígitos). No entanto, se introduzir o valor 1.0 em x_0 o resultado já será diferente, porque o valor 1.0 é considerado numérico e poderá dar um resultado com a mesma precisão, ou seja 16 dígitos (se fizer Precision[1.0] obterá 16, ou seja os 16 dígitos com que trabalha numericamente). Note-se que o valor apresentado tem normalmente 6 dígitos significativos, mas poderá visualizar os restantes usando a rotina N[x_4,16]. Se colocar N[x_4,100] aparecerão 100 dígitos, dos quais apenas os 16 primeiros têm algum significado (com efeito se fizer Precision[N[x_4,100]] continuará a obter 16). Para garantir maior precisão deve declarar os valores numéricos que introduzir com precisão suplementar usando a rotina SetPrecision. A filosofia do *Mathematica* neste aspecto parece simples, cada símbolo tem associada uma precisão, e *o resultado terá que ter a precisão do menor símbolo introduzido* (quando isso não é possível aparece uma mensagem de aviso). Esta filosofia é penosa em termos de cálculo, já que para garantir a precisão, podem ser efectuados cálculos internos com uma imensidão de dígitos (que nunca são apresentados). Tem a vantagem de permitir normalmente resultados rigorosos... ao longo do texto referiremos alguns que não o são e que foram por nós detectados... haverá outros!

1.1.3 Tipos de arredondamento

Dado um número real x, descrito em notação científica por

$$x = \pm 0.a_1 a_2 \ldots a_n a_{n+1} \ldots \times 10^t$$

coloca-se a questão de representá-lo aproximadamente num sistema de ponto flutuante $FP(10, n, t_1, t_2)$.

[2]Por vezes é também introduzida a noção de *algarismo significativo.* Preferimos ao longo do texto usar a noção 'vaga' e intuitiva de dígito correcto. Estas noções podem ser bem definidas, mas sem vantagens adicionais para a compreensão.

Apenas importa referir que apesar da aproximação 0.9999 não apresentar nenhum dígito igual aos dígitos de 1.000, é óbvio que se trata de uma boa aproximação, e é nesse sentido que se introduz a noção de algarismo significativo e que falaremos aqui em dígitos 'correctos'.

Se o valor do expoente t não estiver incluído no conjunto $\{t_1, ..., t_2\}$, ocorrem dois tipos de problemas incontornáveis:

– *overflow* : se o valor do expoente t é superior a t_2,

– *underflow* : se o valor do expoente t é inferior a t_1.

Por exemplo, querendo representar $x = 0.1 \times 10^{151}$ em $FP(10, 8, -100, 100)$, ocorre um problema de *overflow* e um problema de *underflow* se tentarmos representar $1/x = 0.1 \times 10^{-149}$ (aproximar por zero poderia levar a graves erros...)

Ainda que se considere um outro sistema maior, por exemplo, $FP(10, 8, -200, 200)$, apesar desses problemas já não ocorrerem para estes valores de x ou $1/x$, vão ocorrer para x^2 e $1/x^2$, e assim sucessivamente... é nesse sentido que estes problemas são incontornáveis[3].

Suponhamos agora que $t_1 \leq t < t_2$, e portanto não há problema com os expoentes. Há no entanto o problema com o número infinito de dígitos da mantissa que podemos resolver considerando dois tipos de arredondamento:

Arredondamento por Corte

$$fl(x) = \pm 0.a_1 a_2 \dots a_n \times 10^t$$

Arredondamento Simétrico

$$fl(x) = \pm 0.a'_1 a'_2 \dots a'_n \times 10^{t'}$$

Os dígitos a'_i e o expoente t' resultam da representação da soma de $0.a_1 a_2 \dots a_n ... \times 10^t$ com $0.5 \times 10^{t-n}$, considerando depois um arredondamento por corte.

O valor $fl(x)$ é um racional, elemento de $FP(10, n, t_1, t_2)$, que aproxima x. (Repare-se que $fl(x)$ nunca é nulo.)

Interessa-nos controlar os efeitos que essa aproximação pode trazer nos cálculos posteriores, devido ao erro que se comete ao aproximar x por $fl(x)$.

1.1.4 Erro absoluto e erro relativo

Ao considerarmos os arredondamentos necessários pelas constrições do sistema FP da máquina vão ocorrer erros, que vamos distinguir.

Definição 1.4 *Consideremos x um valor exacto e $\tilde{x}$ um valor aproximado de x. Definimos:*
Erro : $e_{\tilde{x}} = x - \tilde{x}$
Erro Absoluto : $|e_{\tilde{x}}| = |x - \tilde{x}|$
Erro Relativo[4] : Se $x \neq 0$, definimos $\delta_{\tilde{x}} = \frac{x-\tilde{x}}{x}$, e designamos por erro relativo este valor ou o seu módulo.

[3]Note-se que as linguagens simbólicas podem parecer resolver este problema, já que se escrevermos $10^{100} \times 10^{900}$, elas podem dar o resultado 10^{1000}, mas se pedirmos de seguida $\sin(10^{100} \times 10^{900})$, sem alterar a precisão interna das rotinas, vão ocorrer erros. Por exemplo, no *Mathematica* o valor devolvido seria sistematicamente 0...

[4]Há autores que consideram $\delta_{\tilde{x}} = \frac{x-\tilde{x}}{\tilde{x}}$.

Estas definições irão ser usadas e generalizadas ao longo do curso, inicialmente estamos interessados no caso $\tilde{x} = fl(x)$.

Devemos ter presente que se conhecessemos o valor exacto do erro para uma dada aproximação, isso implicaria imediatamente o conhecimento do valor exacto, e deixaria de fazer sentido falar em erro... Por isso estas noções serão utilizadas com o objectivo de obter estimativas de forma a controlar o erro, através de majorações ou minorações.

A introdução da noção de *erro absoluto* é assim clara, vamos estar interessados apenas em controlar a diferença entre os dois números, majorando-a, não nos interessando se aproximação é feita por defeito ou excesso relativamente ao valor exacto. Ao longo do curso poderemos mesmo cometer o abuso de chamar *erro* ao *erro absoluto*, já que para efeitos de majorações é claro que a única majoração que interessará será a majoração do erro absoluto.

A introdução da noção de *erro relativo* prende-se com o simples facto intuitivo de que um mesmo erro absoluto pode ter significados diferentes em termos relativos. Assim, quando pretendemos medir a distância de Lisboa ao Porto, se apenas garantirmos que o erro absoluto é inferior a 100 Km podemos considerar que se trata de uma majoração excessivamente grande do erro absoluto, mas quando medimos uma distância da Terra a Marte se garantirmos um valor com um erro absoluto inferior a 100 Km, podemos considerá-lo excelente.

O erro relativo pode ser expresso em termos percentuais, assim, se $|\delta_{\tilde{x}}| = 1$, dizemos que temos um erro relativo de 100%, etc.

Como iremos estar interessados em controlar os erros relativos originados pelo arredondamento, começamos por definir a noção de unidade de arredondamento, que está directamente ligada à precisão do sistema FP utilizado.

• Designamos por *unidade de arredondamento* o valor $\mathbf{u}$ que é o menor majorante do erro relativo de arredondamento,

$$|\delta_{arr}| = \frac{|x - fl(x)|}{|x|} \leq \mathbf{u}$$

Exercício 1.1 *Estando a trabalhar num sistema FP com N dígitos na mantissa, mostre que podemos obter os seguintes majorantes para a unidade de arredondamento:*
No caso de arredondamento por corte, $\mathbf{u} = 10^{1-n}$.
No caso de arredondamento simétrico, $\mathbf{u} = 0.5 \times 10^{1-n}$.
(No caso de se trabalhar com uma base binária, $\mathbf{u} = 2^{1-n}$*)*

1.2 Algoritmos e propagação de erros

Interessa-nos agora saber de que maneira os erros podem ser propagados ao efectuarmos o cálculo de uma função ou de uma operação.

Para ilustrarmos o problema em questão, numa vulgar máquina de calcular, pensemos em extrair m vezes uma raiz quadrada a um valor x e de seguida aplicar o mesmo número de vezes o quadrado ao resultado. Se não houvessem erros de arredondamento o resultado

seria o valor x que havíamos inserido, mas como eles existem, ao fim de um número de vezes razoável ($m > 5$) a maioria das máquinas apresenta um resultado diferente do valor inserido. A resistência das máquinas a este processo pode ser maior ou menor consoante o número de 'dígitos de guarda' que utiliza.

A técnica de utilizar *'dígitos de guarda'* consiste em trabalhar internamente com uma precisão superior à que apresenta no visor. Assim, o utilizador entra e vê valores num sistema FP com n dígitos na mantissa, mas internamente a máquina trabalha numa base binária com N dígitos na mantissa de forma a que a unidade de arredondamento 2^{1-N} em que trabalha seja suficientemente inferior à vísivel pelo utilizador que é $0.5 \times 10^{1-n}$. Desta forma, se para $n = 8$ seria suficiente tomar $N = 26$, uma opção comum é usar $N = 32$, garantindo mais dois dígitos decimais que o utilizador não vê.

Em programas de cálculo mais sofisticados, como o *Mathematica*, este problema também ocorre, mas pode ser contornado se o valor x for declarado com uma precisão suplementar, usando a rotina SetPrecision. Esta vantagem é paga em termos de tempo de computação.

Observação: Informação acerca do funcionamento numérico do *Mathematica* pode ser encontrada no *help* relativo ao *package* NumericalMath‘Microscope‘. Note-se ainda que é possível efectuar cálculos num determinado sistema FP usando o *package* NumericalMath‘ComputerArithmetic‘.

Exemplo 1.1 *Consideremos um valor de M fixo. Sendo $x_0 = a$, calculamos uma lista de M valores dada recursivamente por*

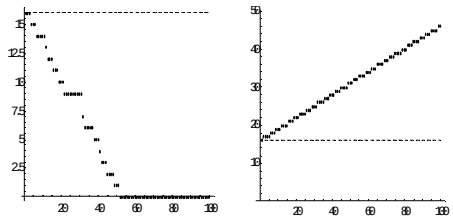
$$\tilde{x}_{n+1} = fl(g(\tilde{x}_n)),$$

ou seja, trata-se do resultado da aplicação sucessiva de uma função g, sujeita a erros de arredondamento. Tendo obtido o valor x_M, tentamos regressar ao valor inicial aplicando a operação inversa. Ou seja, consideramos $y_M = x_M$ e fazemos

$$\tilde{y}_{n-1} = fl(g^{-1}(\tilde{y}_n)).$$

A questão que se coloca é saber se o valor x_0 é ou não igual ao valor y_0. Se os cálculos fossem exactos seria... mas sujeitos a erros de arredondamento, obtemos valores diferentes. Testámos a função Sqrt do Mathematica, começando com $x_0 = 2$. Apesar de apresentar sucessivamente o valor de defeito Precision 16, podemos ver na figura em baixo, à esquerda, que o número de dígitos exactos decresce rapidamente com o valor de M. Ao definir x_0 com SetPrecision[2.0,16] este problema é contornado pelo Mathematica, usando a teoria de propagação de erros, ao aumentar internamente a precisão do número de dígitos dos cálculos intermédios. Como apresentamos na figura em baixo, à direita, o número de dígitos para o cálculo de $\tilde{x}_M$ aumenta e atinge uma precisão de 46 casas decimais quando consideramos $M = 100$. Só desta forma o programa garante os 16 dígitos correctos em $\tilde{y}_0$, não havendo erro assinalável. Como é compreensível, o tempo de cálculo é muito diferente... para calcular um valor $\tilde{x}_n$ o programa aparenta necessitar de aumentar um dígito na mantissa do sistema FP binário. Em alguns dos valores testados a diferença cifrou-se em aproximadamente 6 vezes mais tempo (... o que de certa maneira será análogo a usar para o mesmo

cálculo um processador a 100Mhz em vez de outro a 600Mhz), mas poderá ser muito maior em cálculos mais complexos.



Noutros programas, como o Fortran, Visual Basic ou C, esta possibilidade de alterar o sistema numérico não é normalmente possível, ficando assim limitados a situações semelhantes à apresentada na figura da esquerda, normalmente ainda piores (sendo um caso extremo algumas calculadoras menos sofisticadas).

O estudo básico acerca da propagação de erros, que efectuamos neste parágrafo, é suficientemente exemplificativo de um estudo numa perspectiva mais abstracta, que se poderá encontrar em anexo.

1.2.1 Propagação do erro em funções de uma variável

Se tivermos um valor $\tilde{x}$ que aproxima x, ao calcularmos a imagem por uma função $\phi \in C^2(V_x)$ em que V_x é uma vizinhança de x que contém o ponto $\tilde{x}$, vamos obter um valor aproximado $\phi(\tilde{x})$ que será diferente do valor $\phi(x)$. Para controlarmos o erro que se propaga ao aplicarmos esta função, usamos a expansão em série de Taylor de ϕ em torno de um ponto x :

$$e_{\phi(\tilde{x})} = \phi(x) - \phi(\tilde{x}) = \phi'(x)e_{\tilde{x}} + o(e_{\tilde{x}})$$

quando e_x tende para zero. Desta forma, desprezando o termo $o(e_x)$, podemos definir

$$\tilde{e}_\phi(x) = \phi'(x)e_{\tilde{x}}$$

e para o erro relativo, quando $\phi(x) \neq 0$, obtemos

$$\tilde{\delta}_\phi(x) = \frac{x\phi'(x)}{\phi(x)}\delta_{\tilde{x}}$$

Observação: Estas aproximações pressupõem que se considere erros pequenos, já que o termo desprezado $o(e_{\tilde{x}})$ é, com efeito, $\frac{1}{2}\phi''(\xi_x)e_{\tilde{x}}^2$, e assim assume-se implicitamente, por questões de simplificação, que a função tem segunda derivada e que ela é 'regular', não tomando valores muito altos próximo de x. Quanto maiores forem os valores dessa segunda derivada, entre $\tilde{x}$ e x, piores serão estas estimativas linearizadas.

Definição 1.5 *O módulo do valor* $p_\phi(x) = \frac{x\phi'(x)}{\phi(x)}$ *é normalmente designado* número de condição *de* ϕ *em* x.

A situação de mau condicionamento ocorre para valores de z para os quais $p_\phi(z) = \pm\infty$. Isto significa que ao considerarmos valores $\tilde{z}$ próximos de z vamos obter grandes erros relativos para $\phi(\tilde{z})$, ainda que o erro relativo associado a $\tilde{z}$ seja pequeno. Obtemos, por exemplo:

$$\tilde{\delta}_{x^n} = n\delta_{\tilde{x}};\ \ \tilde{\delta}_{e^x} = x\delta_{\tilde{x}};\ \ \tilde{\delta}_{\cos(x)} = -x\tan(x)\delta_{\tilde{x}};\ \tilde{\delta}_{\sin(x)} = x\cot(x)\delta_{\tilde{x}}$$

Exercício 1.2 *a) Mostre que* $\tilde{\delta}_{(\alpha f+\beta g)} \not\sim \alpha\tilde{\delta}_f + \beta\tilde{\delta}_g$*, ou seja, não há linearidade de* $\tilde{\delta}_f$.
b) Mostre que $\tilde{\delta}_{f\circ g}(x) = p_f(g(x))\tilde{\delta}_g(x)$.

A alínea b) traduz uma propriedade importante, que nos permite separar o cálculo de $\tilde{\delta}_f$ em vários passos. Assim, se quisermos calcular $\tilde{\delta}_{\exp(x^2+1)}$, podemos decompor esse cálculo através de duas funções mais simples, $z_1(x) = x^2+1, z_2(x) = \exp(x)$. Como $\tilde{\delta}_{x^2+1} = \frac{2x^2}{x^2+1}\delta_x$,

$$\tilde{\delta}_{\exp(x^2+1)}(x) = p_{\exp(x)}(x^2+1)\tilde{\delta}_{x^2+1}(x) = (x^2+1)\frac{2x^2}{x^2+1}\delta_{\tilde{x}} = 2x^2\delta_{\tilde{x}}.$$

1.2.2 Propagação do erro em operações

Se considerarmos agora funções com duas ou mais variáveis, $\phi(\mathbf{x})$, em que $\tilde{\mathbf{x}} = (\tilde{x}_1, .., \tilde{x}_N)$ aproxima $\mathbf{x} = (x_1, .., x_N)$, obtemos pela fórmula de Taylor (a duas variáveis)

$$e_{\phi(\tilde{x})} = \nabla\phi(x)\,\mathbf{e}_{\tilde{x}} + \frac{1}{2}\mathbf{e}_{\tilde{x}}\cdot\boldsymbol{\nabla}^2\phi(x+\xi\mathbf{e}_{\tilde{x}})\,\mathbf{e}_{\tilde{x}} = \nabla\phi(x).\mathbf{e}_{\tilde{x}} + o(||\mathbf{e}_{\tilde{x}}||),$$

supondo que $\boldsymbol{\nabla}^2\phi$ (a matriz Hessiana de ϕ) é regular, para $\mathbf{e}_{\tilde{x}} = (e_{\tilde{x}_1}, ..., e_{\tilde{x}_N})$ suficientemente pequenos.

Logo, quando os valores $\tilde{\mathbf{x}}$ são próximos de $\mathbf{x}$, fazemos a mesma linearização, desprezando o termo $o(||\mathbf{e}_{\tilde{x}}||)$ e podemos estabelecer

$$\tilde{e}_\phi(\mathbf{x}) = \nabla\phi(\mathbf{x}).\mathbf{e}_{\tilde{x}} = \sum_{k=1}^{N}\frac{\partial\phi}{\partial x_k}(\mathbf{x})e_{\tilde{x}_k},$$

definindo-se a expressão para a aproximação do erro relativo:

$$\tilde{\delta}_\phi(\mathbf{x}) = \sum_{k=1}^{N}\frac{x_k\frac{\partial\phi}{\partial x_k}(\mathbf{x})}{\phi(\mathbf{x})}\delta_{\tilde{x}_k} = \sum_{k=1}^{N}p_{x_k}\delta_{\tilde{x}_k}$$

em que p_{x_k} são os números de condição

$$p_{x_k} = \frac{x_k\frac{\partial\phi}{\partial x_k}(\mathbf{x})}{\phi(\mathbf{x})}.$$

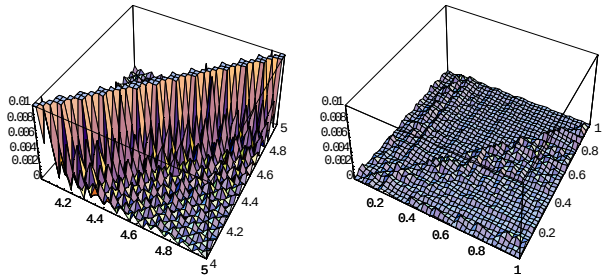
• Obtemos, por exemplo, para o caso das operações elementares,

$$\tilde{\delta}_{xy} = \delta_{\tilde{x}} + \delta_{\tilde{y}}; \qquad \tilde{\delta}_{x/y} = \delta_{\tilde{x}} - \delta_{\tilde{y}}$$

$$\tilde{\delta}_{x+y} = \frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}}; \qquad \tilde{\delta}_{x-y} = \frac{x}{x-y}\delta_{\tilde{x}} - \frac{y}{x-y}\delta_{\tilde{y}}.$$

Vemos imediatamente que quando há subtracção de números muito próximos podemos ter números de condição muito elevados – trata-se do caso de *cancelamento subtractivo.*

Exemplo 1.2 *Vemos na próxima figura como num sistema FP com 4 dígitos na mantissa os erros relativos originados pelo cancelamento subtractivo são significativamente maiores que os erros relativos que advêm de uma divisão, mesmo quando se trata de uma divisão por valores próximos de zero.*



Na figura da esquerda, representa-se o gráfico de $\tilde{\delta}_{x-y}$, para valores $x, y \in]4, 5[$. Nota-se perfeitamente que próximo da diagonal $y = x$ os erros relativos são muito elevados, o que reflecte o problema de mau condicionamento na situação de cancelamento subtractivo. *Na figura da direita apresentamos o gráfico de de $\tilde{\delta}_{x/y}$, para $x, y \in]0, 1[$. Ainda que a divisão por valores próximos de zero coloque problemas numéricos de 'overflow', e um aumento do erro absoluto, o comportamento do erro relativo não é afectado significativamente, conforme previsto.*

1.2.3 Propagação de erros de arredondamento

Ainda que na concepção da matemática clássica o tempo não exista (...ou melhor, todas as relações e expressões matemáticas não dependem do tempo), sabemos que para nós o resultado da expressão $\cos(0.34 + 0.659) + 0.34^2$ não é obtido num único momento. Ou seja, devemos efectuar umas operações antes de outras, de acordo com as prioridades convencionadas. De forma semelhante, as operações efectuadas numa máquina não são todas feitas ao mesmo tempo... é necessário programar um algoritmo.

Um *algoritmo* pode ser visto como uma lista ordenada de tarefas elementares. As tarefas elementares dependem da linguagem de programação. Assim, por exemplo, podemos encarar o *co-seno* como uma função elementar, mas não nos devemos esquecer que isso se deve apenas ao facto da linguagem de programação possuir um algoritmo subjacente que

permite o conforto de evitar uma programação penosa. A própria soma ou multiplicação de números levar-nos-ia à programação de algoritmos que aprendemos na instrução básica[5].

A evolução das linguagens de programação tende a tornar elementares tarefas que anteriormente não o seriam. Assim, as linguagens simbólicas actuais (como o *Mathematica*) contêm subrotinas elementares que não o seriam em linguagens clássicas.

Querendo calcular $\cos(x+y)+x^2$, consideramos o algoritmo:

$$z_1 = x+y; \quad z_2 = x^2; \quad z_3 = \cos(z_1); \quad z_4 = z_2 + z_3.$$

Há assim uma ordenação de tarefas que está subjacente no índice de z, e que nos dita a ordem pela qual podemos decompor a expressão inicial numa sucessão de tarefas elementares. Este é um caso simples, mas podemos pensar de forma semelhante para algo mais complicado.

Actualmente, um outro passo é dado no sentido da programação paralela. Ou seja, ao invés de esperarmos pelo valor de $x+y$ e só depois calcular x^2, podemos pedir a uma outra máquina (ou processador) que o efectue. Isto não evita, no entanto, o uso de um algoritmo semelhante e o escalonamento no tempo, necessitando ainda de uma programação da coordenação entre as várias máquinas. Como é óbvio, neste caso extremamente simples isso não se justificaria, mas podemos retirar a ideia subjacente.

Ainda que os valores dados a x e a y sejam os correctos, não podemos esperar que os valores devolvidos pela máquina sejam exactos, apenas podemos admitir que a rotina seja eficaz e que esse valor esteja só afectado de erros de arredondamento. Basta pensar na rotina *co-seno*... o computador terá que devolver sempre um número num sistema FP, que virá afectado de um erro de arredondamento.

Assim, o resultado obtido em cada uma das operações vem afectado de erros relativos de arredondamento δ_{arr_k}, ou seja,

$$\tilde{\delta}_{z_k} = \tilde{\delta}_f + \delta_{arr_k}$$

onde f é a operação, ou função, calculada no passo k. Os erros relativos $|\delta_{arr_k}|$ podem ser majorados pela unidade de arredondamento $\mathbf{u}$.

Note-se que há uma propagação destes erros de arredondamento ao longo do algoritmo, o que pode também causar grandes erros relativos no resultado (originando um problema de instabilidade numérica).

[5]De referir que, por exemplo, a implementação computacional do algoritmo da divisão não é trivial, e foi assunto de investigação, com o objectivo de diminuir o número de operações elementares e assim aumentar a rapidez de cálculo. Para nos apercebermos da dificuldade inerente a estes algoritmos primários, basta tentar programar as operações elementares com uma base qualquer (que será um parâmetro dado pelo utilizador).

A programação necessária à resolução de um sistema linear, que será aqui abordada, é mais fácil que a programação de um algoritmo de divisão.

No final do algoritmo, com m passos, se tivermos $x_1, ..., x_n$ valores aproximados, obtemos:

$$\tilde{\delta}_z = p_{x_1}\delta_{\tilde{x}_1} + ... + p_{x_n}\delta_{\tilde{x}_n} + q_1\delta_{arr_1} + ... + q_m\delta_{arr_m}$$

Observação: Devido à linearização efectuada (ao desprezar os termos quadráticos), e como consequência da alínea b) de um exercício anterior, os valores $p_{x_1}, ...p_{x_n}$ coincidem, quer considerando um cálculo passo a passo, quer considerando um cálculo directo na função f. Apenas os valores $q_1, ..., q_m$ irão depender do algoritmo considerado.

1.2.4 Algoritmos e rotinas elementares

Da mesma forma que apresentámos um algoritmo para o cálculo da função $\cos(x + y) + x^2$, decompondo esse cálculo em várias operações elementares, podemos fazer o mesmo decompondo o cálculo para problemas mais complicados através de rotinas elementares.

Assim, se quisermos calcular $z = x+y$, em que x é a solução única da equação $\cos(x) = x$ e em que $y = \int_0^1 e^{-t^2}dt$, podemos separar esse cálculo em três etapas

```
x = FindRoot[z == Cos[z], {z, 0}][[1, 2]];
y = NIntegrate[Exp[-t^2], {t, 0, 1}];
z = x + y
```

Não vamos agora entrar em detalhe acerca de pormenores acerca das rotinas FindRoot e NIntegrate, interessa-nos apenas que elas devolvem um valor aproximado do valor correcto que pretendemos calcular, de tal forma que o único erro existente é o erro de arredondamento. Portanto, o único erro que há a calcular é

$$\tilde{\delta}_z = \frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}}$$

em que $\delta_{\tilde{x}}$, $\delta_{\tilde{y}}$ são erros de arredondamento, podendo escrever-se $\tilde{\delta}_z = \frac{x}{x+y}\delta_{arr1} + \frac{y}{x+y}\delta_{arr2}$.

Supondo agora que $y = \int_0^x e^{-t^2}dt$, a segunda atribuição será

```
y = NIntegrate[Exp[-t^2], {t, 0, x}];
```

e no cálculo de y já será necessário considerar o erro de arredondamento obtido no cálculo de x, já que y é entendida como função de x. Assim, $\delta_{\tilde{x}} = \delta_{arr1}$

$$\tilde{\delta}_y = \frac{xy'(x)}{y(x)}\delta_{\tilde{x}} + \delta_{arr2} = \frac{xe^{-x^2}}{y}\delta_{\tilde{x}} + \delta_{arr2}.$$

Neste caso, é possível uma expressão simples já que é fácil calcular $y'(x) = e^{-x^2}$. Finalmente teríamos

$$\tilde{\delta}_z = \frac{x}{x+y}\delta_{\tilde{x}} + \frac{y}{x+y}\delta_{\tilde{y}} + \delta_{arr3} = \frac{x(1+e^{-x^2})}{x+y}\delta_{\tilde{x}} + (\frac{y}{x+y})\delta_{arr2} + \delta_{arr3}$$

Podemos ainda pensar que o valor de y era dado por $y = \int_a^x e^{-t^2} dt$, em que a era um parâmetro de entrada que poderia estar afectado de erro. O cálculo do erro já seria diferente, pois y seria visto como função de a e de x,

$$\tilde{\delta}_y = \frac{x\frac{\partial y}{\partial x}(x,a)}{y(x,a)}\delta_{\tilde{x}} + \frac{a\frac{\partial y}{\partial a}(x,a)}{y(x,a)}\delta_{\tilde{a}} + \delta_{arr2} = \frac{xe^{-x^2}}{y}\delta_{\tilde{x}} - \frac{ae^{-a^2}}{y}\delta_{\tilde{a}} + \delta_{arr2}.$$

1.3 Condicionamento e estabilidade numérica

As noções que encontrámos neste capítulo, referentes a números reais, podem ser generalizadas a espaços abstractos como se apresenta em Anexo. Nesse contexto geral vamos definir a noção de condicionamento para um problema genérico. Iremos ainda falar de estabilidade computacional (ou numérica) já que associado à implementação computacional da resolução de um problema estará subjacente um algoritmo.

• Num problema $\mathcal{P}$ existem dados (de entrada) que podemos agrupar muito geralmente num vector x, e existem os resultados (dados de saída), que genericamente podemos designar por $y = \mathcal{P}(x)$.

Definição 1.6 *Um problema diz-se* bem condicionado *se pequenos erros relativos nos dados produzem pequenos erros relativos no resultado. Caso contrário, diz-se* mal condicionado.

Em concreto, dizemos que o problema $\mathcal{P}$ é bem condicionado para um dado x se existir uma constante $M \geq 0$ tal que[6]

$$||\delta_{\tilde{y}}|| \leq M||\delta_{\hat{x}}||, \forall \hat{x} \in V_x,$$

onde $\tilde{y} = \mathcal{P}(\tilde{x})$, e V_x é uma vizinhança de x.

No caso de dimensão finita, em que os dados de entrada podem ser vistos como $x = (x_1, ..., x_N)$, é claro que se algum dos pesos verificar $p_{x_i} = \infty$, para dados $(x_1, ..., x_N)$, então o problema será mal condicionado para esses dados, pois não será possível encontrar a constante M.

Como já referimos, um problema de mau condicionamento dá-se no caso de *cancelamento subtractivo*. A fórmula

$$\tilde{\delta}_{x_1 - x_2} = \frac{x_1}{x_1 - x_2}\delta_{\tilde{x}_1} - \frac{x_2}{x_1 - x_2}\delta_{\tilde{x}_2},$$

reflecte isso mesmo. Com efeito, para dados $x = (x_1, x_2) \neq 0$ em que $x_1 \to x_2$ é óbvio que $p_{x_1}, p_{x_2} \to \infty$.

[6]No caso em que $y = f(x_1, ..., x_d)$ consideramos simplesmente $||\delta_{\tilde{y}}|| = |\delta_{\tilde{y}}|$ e $||\delta_{\hat{x}}||$ como o erro relativo dado por uma certa norma vectorial, por exemplo :

$$||\delta_{\hat{x}}||_\infty = \frac{\max|\delta_{\hat{x}_i}|}{\max|x_i|} \text{ ou } ||\delta_{\hat{x}}||_1 = \frac{|\delta_{\hat{x}_1}| + ... + |\delta_{\hat{x}_d}|}{|x_1| + ... + |x_d|}.$$

Ver apêndice.

Observação: Podemos observar que o problema de calcular uma função real f poderá ser mal condicionado em $x \neq 0$ se $f'(x) = \infty$ ou se $f(x) = 0$ (excluindo indeterminações)[7].

• Ao traduzirmos um problema $\mathcal{P}$ em termos de um algoritmo $\mathcal{A}$, para além dos erros dos dados temos que considerar também os erros de arredondamento que se irão propagar ao longo da execução do algoritmo. Assim, considerando dados de entrada x e resultados $y = \mathcal{A}(x)$, definimos:

Definição 1.7 *Um algoritmo é* computacionalmente (ou numericamente) estável *se a pequenos erros relativos dos dados, e a pequenos valores da unidade de arredondamento, corresponder um pequeno erro relativo nos resultados. Caso contrário, diz-se* computacionalmente (ou numericamente) instável.

Em concreto, dizemos que um algoritmo $\mathcal{A}$ é computacionalmente estável para um dado x se existir uma constante $M \geq 0$ tal que

$$||\delta_{\tilde{y}}|| \leq M(\mathbf{u} + ||\delta_{\hat{x}}||), \forall \hat{x} \in V_x,$$

onde $\tilde{y} = \mathcal{A}(\tilde{x})$ e V_x é uma vizinhança de x. O valor $\mathbf{u}$ é aqui colocado porque é um majorante de todos os erros relativos de arredondamento $|\delta_{arr_i}|$.

O algoritmo $\mathcal{A}$ é computacionalmente instável para dados $(x_1, ..., x_N)$ se algum dos pesos verificar $p_{x_i} = \infty$ ou $q_i = \infty$ para esses dados.

Observações:

(i) É claro que um algoritmo implementado num problema mal condicionado nunca poderá ser computacionalmente estável.

(ii) A constante M desempenha um papel importante para comparação entre vários algoritmos, já que consoante o algoritmo podemos obter valores de M diferentes, e podemos mesmo obter algoritmos computacionalmente estáveis enquantos outros são instáveis para um mesmo problema.

---

[7] No caso de funções vectoriais (ou operadores A) isto corresponde aos casos em que a matriz jacobiana (ou a derivada de Fréchet) não é limitada.

Ou seja, caso $||A'_x|| = \infty$ ou se $||Ax|| = 0 (\Leftrightarrow Ax = 0)$, isto deve-se à fórmula que obtemos no Anexo,

$$||\tilde{\delta}_A x|| \leq \frac{||x||\ ||A'_x||}{||Ax||} ||\delta_{\tilde{x}}||.$$

Se A for um operador linear contínuo, bijectivo, com inversa contínua, então

$$||\tilde{\delta}_A x|| \leq ||A^{-1}||\ ||A||\ ||\delta_{\tilde{x}}||,$$

e este é o caso das matrizes invertíveis. Como veremos, mais tarde (no capítulo 4), ao valor $||A^{-1}||\ ||A||$ chamaremos *número de condição da matriz* A, e este valor estabelece uma maneira de identificar quais os sistemas em que a exactidão da solução é mais (ou menos) afectada pela exactidão dos dados. Como é claro, um número de condição pequeno garantirá *a priori* melhor exactidão.

Exemplo 1.3 *O exemplo mais trivial surge, por exemplo, ao escrever num algoritmo*

$$z_1 = x - 1$$
$$z_2 = z_1 + 1$$

ao invés de escrever imediatamente $z_2 = x$. *Isto implica que haja um arredondamento na primeira atribuição, o que provoca uma instabilidade para valores de* x *próximos de zero. Com efeito,*

$$\tilde{\delta}_{z_1} = \frac{x}{x-1}\delta_{\tilde{x}} + \delta_{arr_1}$$
$$\tilde{\delta}_{z_2} = \frac{z_1}{z_1+1}\tilde{\delta}_{z_1} + \delta_{arr2}$$

assim,

$$\tilde{\delta}_{z_2} = \frac{x-1}{x}(\frac{x}{x-1}\delta_{\tilde{x}} + \delta_{arr_1}) + \delta_{arr_2} = \delta_{\tilde{x}} + \frac{x-1}{x}\delta_{arr_1} + \delta_{arr_2}$$

e o peso $q_1 = \frac{x-1}{x}$ *tende para infinito quando* x *tende para zero, tornando o algoritmo instável para valores próximos de zero.*
De facto, se estivessemos a trabalhar com arredondamento por corte, com 8 dígitos na mantissa, verificávamos que se $x = 0.1 \times 10^{-11}$ *(que seria representado exactamente) então*

$$z_1 = -0.999\,999\,999\,999 \text{ seria representado } -0.999\,999\,99 \times 10^0$$
$$z_2 = 0.1 \times 10^{-7}$$

isto gera um erro relativo $|\delta_{z_2}| = \frac{|0.1\times10^{-11} - 0.1\times10^{-7}|}{0.1\times10^{-11}} = 9999$, *ou seja um erro relativo de quase 1000%.*
Se não escandaliza que o resultado dê 0.1×10^{-7} *enquanto deveria ser* 0.1×10^{-11}, *isto pode ter consequências mais drásticas em operações subsequentes... basta pensar que fazendo de seguida* $z_3 = z_2/x$, *ao invés de obtermos* 1 *iríamos ter* $10\,000$, *etc.*

(*Nota:* Este exemplo pode não ser observável, em certas máquinas, porque os programas de cálculo, as linguagens de programação, ou as máquinas de calcular, utilizam dígitos suplementares, os designados *dígitos de guarda*, que evitam 'algumas' vezes estes problemas).

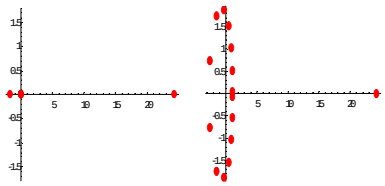
Como é claro, este é um exemplo trivial, mas existem outras situações em que a mudança para um algoritmo computacionalmente estável (só possível se o problema for bem condicionado) não é tão simples!

Exemplo 1.4 *Ilustramos a instabilidade numérica com um problema que ocorria na versão 3 do Mathematica e que foi recentemente corrigido (ou melhor, contornado) na versão.4. Consideramos a seguinte matriz*

$$A = \frac{1}{10}\begin{bmatrix} 1. & 2 & \cdots & N \\ 2 & 3 & \cdots & N+1 \\ \vdots & \ddots & \ddots & \vdots \\ N & N+1 & \cdots & 2N-1 \end{bmatrix}$$

se pretendermos calcular os valores próprios podemos pensar em duas possibilidades:
(i) utilizar a rotina Eigenvalues, que nos dá uma lista com os valores próprios,

(ii) usar a rotina CharacteristicPolynomial, que nos dá o polinómio característico de A, e calcular as raízes desse polinómio, usando a rotina NSolve.
Sem entrar em detalhes sobre as rotinas, o cálculo dos valores próprios por um processo ou por outro é matematicamente equivalente, mas pode não o ser do ponto de vista numérico. Usando a versão anterior do Mathematica, se efectuarmos (i) obtemos o gráfico que se mostra em baixo, à esquerda, onde os pontos correspondem aos valores próprios (no plano de Argand). Efectuando (ii) obtemos o gráfico que mostra à direita. A diferença é enorme... e aqui apenas considerámos uma matriz 15×15*... seria muito maior para matrizes de dimensão superior.*



Após visualizar os gráficos podemos mesmo duvidar se algum deles é correcto. É claro que neste caso, como escolhemos uma matriz simétrica, que terá valores próprios reais, sabemos imediatamente que o segundo gráfico é incorrecto. No entanto, se tivéssemos escolhido outro tipo de matriz, a diferença seria até maior e não poderíamos ajuízar tão facilmente. Estas questões ilustram bem como os problemas numéricos podem ser colocados, mesmo quando se domina a teoria subjacente. Perante estes dois gráficos, o utilizador da versão anterior do Mathematica (ou doutro programa qualquer...) poderia tentar confirmar se cometeu algum erro na introdução dos dados, ou manifestar o seu desagrado ao fabricante... Deverá ter sido esta última hipótese que levou a que fosse contornado o problema na versão mais recente.
Mas analisemos esta diferença face à teoria que acabamos de expor. Iremos ver que o problema do cálculo de valores próprios é bem condicionado (relativamente à variação dos elementos) quando a matriz é simétrica. Portanto, não se trata de uma questão de mau condicionamento do problema. Trata-se, com efeito, de um problema de instabilidade numérica.
Qual o problema com o procedimento (ii)? A anterior rotina CharacteristicPolynomial apresentava o polinómio característico na sua forma canónica, o que não acontece na versão actual em que apresenta apenas a factorização, usando os valores obtidos em Eigenvalues (é nesse sentido que dizemos que contorna o problema). No cálculo efectuado pela anterior rotina, os valores dos coeficientes do polinómio característico eram calculados de forma diferente e sujeitos a sucessivos erros de arredondamento, o que levava a erros apreciáveis nos coeficientes do polinómio característico. Esses erros podem mesmo ser ampliados pelo cálculo das raízes do polinómio, que como veremos (no próximo capítulo) é um problema mal condicionado face à variação dos coeficientes.
Terminamos, notando que esta diferença apenas acontecia se um dos coeficientes da matriz for considerado numérico (...colocámos um pequeno ponto num dos elementos da matriz, para ilustrar esse facto, pois, como já referimos, o programa faz a diferença entre 1 e 1.). Se todos os elementos fossem considerados exactamente, as rotinas do programa devolviam o mesmo resultado.
• *Contudo, não se pode pensar que é sempre possível contornar um problema introduzindo*

valores 'exactos', por três razões diferentes. Primeiro, porque a diferença entre o tempo de cálculo do valor 'exacto' e do valor numérico será abissal... (as diferenças cifram-se entre milissegundos e minutos), já que, não sendo permitadas simplificações numéricas, acumulam-se expressões não simplificadas, que também podem atingir facilmente o limite da memória da máquina... que será sempre finito! Segundo, porque se fosse devolvida uma expressão não simplificada, para a avaliar teriam que ser efectuados cálculos numéricos, podendo de novo haver instabilidade. Terceiro, porque os valores dados (neste caso, os introduzidos nos elementos da matriz) podem resultar de outros cálculos numéricos.

Observação 1:

Realçamos que a questão da instabilidade numérica é um problema do algoritmo, ou seja (como já mencionámos) da necessidade de dividir um cálculo em subcálculos. *O que devemos evitar é que esses subcálculos sejam mal condicionados.* O problema pode ser bem condicionado, mas se para o seu cálculo efectuarmos cálculos em subproblemas mal condicionados, então é natural que haja instabilidade.

Por exemplo, o cálculo da função $f(x) = 1 - \cos(x)$ é um problema bem condicionado. No entanto, se pensarmos em proceder ao seu cálculo efectuando a subtracção, sabemos que estamos a introduzir um problema mal condicionado, devido ao cancelamento subtractivo. Para valores de x próximos de zero (quando ocorre esse cancelamento) devemos pensar numa maneira equivalente de efectuar o cálculo, por exemplo, usando a relação $1-\cos(x) = 2\sin^2(\frac{x}{2})$. A nova maneira de calcular f poderá envolver mais subcálculos, mas nenhum deles é mal condicionado (se x é próximo de zero). Desta forma não se altera o condicionamento do problema inicial. Se decompusermos um problema P em subproblemas $P_1, ..., P_N$, então para que a composição desses problemas mantenha o condicionamento de P cada um deles deverá ser bem condicionado. No exemplo anterior isso não ocorreu quando nos propusémos calcular o polinómio característico com uma rotina mal condicionada.

Observação 2: *(propagação de erro em métodos iterativos)*

No próximo capítulo iremos focar a nossa atenção em métodos iterativos que envolvem algoritmos do tipo

$$z_{n+1} = g(z_n)$$

podemos avaliar o erro relativo, usando a fórmula já estabelecida, e portanto

$$\tilde{\delta}_{z_{n+1}} = \frac{z_n g'(z_n)}{g(z_n)}\delta_{z_n} + \delta_{arr_{n+1}},$$

reparamos que isto origina uma sucessão do tipo $d_{n+1} = p_n d_n + a_{n+1}$, o que no final de m iterações dá

$$d_{m+1} = p_m \cdots p_0 d_0 + p_m \cdots p_1 a_1 + ... + p_m a_m + a_{m+1}.$$

Portanto o erro de arredondamento representado por a_k irá aparecer multiplicado pelos factores $p_k, p_{k+1}, ...$ seguintes. Se a multiplicação destes factores tender para zero, a influência dos erros de arredondamento anteriores será *negligenciada.nas futuras iterações.* Ora para

que a multiplicação desses factores tenda para zero será conveniente que eles sejam em módulo menores que 1, ou seja

$$|p_n| = \left| \frac{z_n g'(z_n)}{g(z_n)} \right| < 1.$$

Reparamos que no caso em que $z_n \approx g(z_n)$ e $|g'(z_n)| < 1$ então a condição verifica-se, e será esse o caso dos métodos de ponto fixo que veremos no próximo capítulo. Portanto, nesses métodos a influência do erro relativo inicial e dos erros de arredondamento anteriores é negligenciável.

Note-se, no entanto, que para valores de $|g'(z_n)|$ próximos de 1, o que acontecerá quando a convergência for lenta, os erros de arredondamento imediatamente anteriores não são tão negligenciáveis.

Como é óbvio, o último erro de arredondamento mantém inalterado, e deve ter-se especial atenção a ele quando atingimos a precisão da máquina, pois nesse caso $|\tilde{\delta}_{z_n}| \approx |\delta_{arr_n}| \approx \mathbf{u}$, e será escusado efectuar mais iterações, pois não obteremos erros mais pequenos!

Observação 3: *(problema inverso)*

Quanto melhor for condicionado um certo problema pior será o condicionamento do problema inverso associado. Com efeito, basta pensar que se os erros relativos dos resultados são muito pequenos face aos erros dos dados, então esses dados passam a ser os resultados, e vice-versa. Portanto, a relação verifica-se no sentido inverso... os pequenos erros relativos dos dados originam os erros relativos apreciáveis nos resultados.

Um exemplo é o caso apresentado na observação anterior. Para a iteração $w_{n+1} = g^{-1}(w_n)$ teremos

$$\tilde{\delta}_{w_{n+1}} = \frac{w_n (g^{-1})'(w_n)}{g^{-1}(w_n)} \delta_{w_n} + \delta_{arr_{n+1}}.$$

e podemos mesmo ver que, como $w_n = g(w_{n+1})$,

$$\tilde{\delta}_{w_{n+1}} = \frac{g(w_{n+1})}{w_{n+1} g'(w_{n+1})} \delta_{w_n} + \delta_{arr_{n+1}},$$

em que o número de condição é o inverso do obtido no cálculo directo de $g(w_{n+1})$.

Seguindo o caso anterior, vimos que o problema era bem condicionado para $w_n \approx g(w_n)$, ou também, $w_n \approx g^{-1}(w_n)$, quando $|g'(w_n)| < 1$. Como $|(g^{-1})'(w_n)| = |\frac{1}{g'(g^{-1}(w_n))}| \approx |\frac{1}{g'(w_n)}| > 1$, ao invés de haver um decréscimo do erro relativo passar a acontecer um incremento, devido à sucessiva multiplicação de factores maiores que 1. Esses factores irão aumentar exponencialmente o efeito dos erros de arredondamento. Podemos perceber isso num exemplo que vimos anteriormente, quando efectuámos raízes sucessivas de um número, e depois pretendemos regressar efectuando quadrados sucessivos. Na primeira etapa, ao considerar $z_{n+1} = \sqrt{z_n}$, os erros relativos comportavam-se na forma

$$\tilde{\delta}_{z_{n+1}} = \frac{1}{2} \delta_{z_n} + \delta_{arr_{n+1}},$$

fazendo decrescer os efeitos dos erros de arredondamento iniciais na proporção $\frac{1}{2^n}$. Na etapa inversa, ao considerar $w_{n+1} = w_n^2$, os erros relativos passam a ter o comportamento

$$\tilde{\delta}_{w_{n+1}} = 2\delta_{w_n} + \delta_{arr_{n+1}},$$

o que incrementa os efeitos dos erros de arredondamento iniciais na proporção 2^n. Assim, ainda que o erro de arredondamento inicial seja inferior a 10^{-16}, ao fim de N iterações teremos $2^N 10^{-16}$, e para $N > 53 \approx 16\frac{\log(10)}{\log(2)}$,o efeito do erro de arredondamento inicial será da grandeza das unidades. Isto justifica os resultados obtidos nesse exemplo, e também a maneira de contornar este problema adicionando um dígito no sistema FP binário, que foi a solução apresentada pelo programa *Mathematica.* Note-se que efectuando o mesmo para uma raiz quarta, e regressando elevando a quatro, a solução começa a ser mais dispendiosa... será necessário adicionar 2 dígitos binários em cada iteração, e ao invés das 46 casas decimais usadas nos cálculos intermédios para $M = 100$, o Mathematica utiliza 76... podemos pensar assim no esforço computacional que poderá ser exigido em cálculos mais complexos.

1.4 Exercícios

1. Considere os valores

$$A = 0.492,\ B = 0.603,\ C = -0.494,\ D = -0.602,\ E = 10^{-5}$$

Com a finalidade de calcular

$$F = \frac{A + B + C + D}{E},$$

dois indivíduos, usando uma máquina com 3 dígitos na mantissa e com arredondamento simétrico, efectuaram esse cálculo de forma distinta, mas aritmeticamente equivalente.

O indivíduo X calculou $A+B$, depois $C+D$, somou os valores, e dividiu por E, obtendo $F = 0$.

Por seu turno, o indivíduo Y calculou $A+C$, depois $B+D$, somou os valores, e dividiu por E, tendo obtido $F = -100$.

Verifique os cálculos efectuados pelos dois indivíduos e comente a disparidade de resultados obtidos, atendendo a que se usaram processos matematicamente equivalentes.

2. Sabendo que $\cos(x_i)$ para $i = 1, ..., 20$ é calculado com um relativo inferior a 10^{-6}, indique uma estimativa para o erro relativo de

$$P = \prod_{k=1}^{20} \cos(x_k)$$

baseando-se nas fórmulas obtidas para a propagação do erro relativo em funções.

3. a) Determine para que valores de x o cálculo de $f(x) = 1 - \cos(x)$ conduz a um problema mal condicionado.

b) Considere o seguinte algoritmo para o cálculo de f

$$z_1 = \cos(x) \quad ; \quad z_2 = 1 - z_1.$$

Mostre que o algoritmo é instável para $x \sim 0$ (apesar de o problema ser bem condicionado).

c) Baseado na fórmula $\frac{1-\cos(x)}{2} = \sin^2(\frac{x}{2})$, escreva um algoritmo equivalente, numericamente estável para $x \sim 0$.

4. Mostre que, usando a aproximação $\cos(x) \sim 1 - \frac{x^2}{2} + \frac{x^4}{24}$, se têm as seguintes estimativas para o erro absoluto e relativo:

$$|e_{\cos(x)}| \leq \frac{1}{2880} \sim 0.35 \times 10^{-3}$$

$$|\delta_{\cos(x)}| \leq \frac{\sqrt{2}}{2880}$$

para qualquer $x \in [-\pi/4, \pi/4]$.

5. Ao calcular-se a expressão $f(x) = x - \sqrt{x^2 - 1}$ numa máquina usando o sistema de virgula flutuante VF(10,6,-30,30) com arredondamento simétrico, verificou-se que para valores de x muito grandes o erro relativo era também grande.

a) Verifique que o erro relativo é 100% para $x = 2000$. Qual o valor do erro relativo para valores de x ainda maiores?

b) Qual a razão desse erro relativo grande: o problema é mal condicionado ou há instabilidade numérica? Justifique e apresente uma forma de calcular $f(x)$ que não apresente erros relativos tão grandes.

6. Determine uma função $f(x)$ que verifique

$$\delta_{f(x)} \sim xe^{-x}\delta_x$$

Sugestão: Verifique primeiro que

$$f(y) \sim C \exp\left(\int^y \frac{\delta_{f(x)}}{e_x} dx\right).$$

e repare que esta fórmula permite obter a função f a partir da expressão do erro relativo.

7. Efectuar o exercício 3 no anexo acerca de equações às diferenças.

Capítulo 2

Determinação de Raízes Reais e Complexas

O objectivo deste capítulo é aproximar soluções reais (ou complexas) de equações da forma:

$$f(x) = 0$$

onde f deverá ser, pelo menos, uma função contínua numa vizinhança da raiz. Os valores x que verificam a equação são também chamados *raízes* da equação, ou ainda, *zeros* da função f.

Para casos particulares de f a determinação das raízes da equação $f(x) = 0$ pode ser considerada trivial, já que a podemos reduzir algebricamente a problemas de resolução elementar. Por exemplo, resolver uma equação linear $ax+b = 0$, reduz-se a efectuar a divisão $-b/a$, enquanto para resolver uma equação do segundo grau $ax^2+bx+c = 0$, podemos usar uma fórmula resolvente que nos dá $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$, valores que são trivialmente calculáveis (admitindo que o cálculo de raízes quadradas também o é!).

A determinação de uma fórmula resolvente para equações de terceiro e quarto grau constituiu um difícil problema algébrico que resistiu dois milénios e foi finalmente resolvido por matemáticos italianos do séc. XVI (Tartaglia, Cardan, Ferrari). No entanto, a determinação da fórmula resolvente para uma equação de grau 5 voltou a causar grandes dificuldades aos matemáticos seguintes, durante quase três séculos. No seguimento de trabalhos de Lagrange, Vandermonde e Ruffini, demonstrou-se no séc. XIX (Galois, Abel), que é impossível obter uma fórmula resolvente geral para equações algébricas de grau maior ou igual a cinco! Este resultado surpreendente apresenta logo grandes restrições à possibilidade de resolução algébrica de equações. Com efeito, se existe dificuldade/impossibilidade em determinar algebricamente zeros de polinómios, essa dificuldade/impossibilidade assume ainda maiores proporções quando consideramos funções f não polinomiais.

Por exemplo, se é simples dizer que os zeros da função $f(x) = \sin(x)$ são os valores $x = k\pi$ para $k \in \mathbb{Z}$, já nos é impossível determinar de forma geral as raízes de equações da forma $\sin(ax^2+b) = cx+d$, etc. Para este tipo de equações não lineares temos, no entanto, a possibilidade de encontrar soluções usando *métodos iterativos*. Os métodos iterativos permitem construir uma sucessão que, em cada termo, tem apenas uma aproximação da solução, mas que, enquanto sucessão, é a própria solução do problema. Importa assim

desmistificar o conceito de solução, ou mais concretamente o conceito de solução exacta, com um exemplo.

Consideremos a equação $x^3 - 3x - 3 = 0$, que tem uma raiz real, dada por

$$z = \frac{1}{\sqrt[3]{2}}(\sqrt[3]{3-\sqrt{5}} + \sqrt[3]{3+\sqrt{5}})$$

e consideremos a sucessão dada por

$$x_0 = 2, \quad x_{n+1} = \frac{2x_n^3 + 3}{3x_n^2 - 3},$$

cujos primeiros termos são $x_1 = \frac{19}{9} = 2.1111...$, $x_2 = \frac{3181}{1512} = 2.1038$. Facilmente diremos que o valor apresentado para z com a fórmula resolvente é a solução exacta e que a sucessão (x_n) apenas nos dará valores aproximados. No entanto, como iremos ver que a sucessão converge para a solução do problema, trata-se de uma sucessão de Cauchy que é um representante da classe de equivalência da solução *e portanto* (x_n) *é a solução exacta*!

Do ponto de vista de cálculo, para obter x_2 foi apenas necessário efectuar algumas multiplicações e uma divisão, e os cinco dígitos apresentados são já os correctos, enquanto que para calcular o valor aproximado de z a partir da fórmula resolvente espera-nos ainda o trabalho de calcular os radicais... É o hábito de ter uma fórmula explícita, em que o valor é dado a partir de funções conhecidas que gera o conceito de fórmula ou solução exacta. Uma fórmula explícita apenas nos transporta de um problema (o de resolver a equação) para outro (o de calcular as raízes). Isso acontece de um modo geral e não apenas neste caso. As fórmulas explícitas remetem-nos para entidades conhecidas, e por isso são úteis teoricamente. No entanto, a nível de cálculo, para obter uma aproximação idêntica, pode ser mais moroso efectuar as operações segundo uma fórmula explícita do que segundo um método iterativo.

Note-se que cada termo da sucessão x_n é apenas uma aproximação racional[1] da solução z. É a própria sucessão (x_n) que se identifica com z. Para além disso, quando calculamos os termos x_n não é possível armazenar todos os dígitos no sistema de ponto flutuante $FP(.)$ que a máquina utiliza. Há um erro de arredondamento, que se traduz em considerar $\tilde{x}_n$, um número pertencente a esse sistema. Mesmo com esse erro de arredondamento, como o método é numericamente estável, os termos seguintes são muito pouco afectados, até que se atinja a máxima precisão do sistema.

Refira-se ainda que o próprio processo empírico de verificar que uma sucessão está a convergir, reparando que os dígitos não se vão alterando (que corresponde ao nosso processo intuitivo de verificar que se trata de uma sucessão de Cauchy... e que por vezes leva a erros de julgamento!), implica que se esteja a vislumbrar a evolução da sucessão e não apenas um termo.

Após esta introdução resumimos o conteúdo do capítulo. Começaremos por apresentar algumas noções acerca da convergência de sucessões e do tempo de cálculo num método

[1] Ainda que os valores x_n sejam reais, podemos sempre tomar racionais $\tilde{x}_n$ (para n suficientemente grande, retirados da sucessão de Cauchy que define x_n), de forma a que a sucessão $(\tilde{x}_n)$ seja uma sucessão de Cauchy de racionais, representante da classe de equivalência do real z.

iterativo. De seguida relembramos alguns teoremas elementares de Análise Matemática, que nos serão úteis e apresentamos dois métodos elementares – os métodos da bissecção e da falsa-posição – que apesar de parecerem pouco sofisticados ou demasiado triviais constituem ainda uma alternativa quando outros não são aplicáveis, ou como métodos para uma primeira aproximação das raízes. De seguida, apresentamos o método do ponto fixo cuja simplicidade e eficácia merece maior relevo. Ao contrário dos anteriores, as possibilidades de generalização deste método para funções de variável complexa ou a casos mais abstractos (que veremos nos capítulos seguintes) são notáveis. Como caso particular do método do ponto fixo, assegurando uma grande rapidez de convergência, salienta-se o método de Newton. Essa rapidez de convergência tem como exigência a diferenciabilidade da função, o que nem sempre se afigura fácil ou possível de efectuar. Uma possibilidade de evitar o cálculo da derivada e obter ainda uma rapidez de convergência razoável pode ser posta em prática usando os métodos da secante ou de Steffensen.

Este capítulo contém ainda uma referência ao método de Bernoulli, aplicável apenas a equações algébricas e não muito eficaz, mas que apresenta ideias completamente distintas dos métodos anteriores (o interesse especial do método de Bernoulli residiu historicamente na possibilidade de aproximar raízes evitando ao máximo o cálculo de divisões).

Finalmente, é incluída uma pequena referência à determinação de raízes complexas através da generalização dos métodos do ponto fixo, de Newton e da secante. Esta generalização antecede e motiva a generalização do método do ponto fixo num contexto mais geral, que constitui o cerne do capítulo seguinte.

Convém fazer uma referência à importância de alguns métodos simples no contexto da Análise Numérica actual. O método mais utilizado para determinar raízes de equações é o método de Newton, devido à sua simplicidade e rapidez de convergência, no entanto não são raros os casos em que a aplicação deste método se afigura impraticável. Com efeito, nem todas as equações que pretendemos resolver envolvem funções elementares, cujo cálculo é quase instantâneo. Nalgumas aplicações os valores para a função f são obtidos após a resolução de problemas complicados, e obter cada um dos valores pode significar várias horas de cálculo, mesmo no computador mais moderno. Para que faça uma ideia, basta pensar que $f(x)$ pode ser o valor de um determinante de uma matriz 1000×1000 cujos elementos dependem de x e podem eles próprios ser determinantes de matrizes semelhantes... Para além do esforço de cálculo para cada um dos valores, explicitar o cálculo da derivada pode ser completamente impraticável. Nestas situações, outros métodos podem ser mais vantajosos... como o método da secante ou o método de Steffensen.

Frequentemente, o próprio esboço do gráfico é irrealizável! O interesse de métodos como o método da bissecção, ou da falsa-posição pode residir na vantagem de se assegurar sempre um intervalo onde se encontra a raiz controlando melhor a sua localização, já que usando outros métodos (se não for assegurada a convergência) podemos ser levados a situações de divergência ou a uma convergência para soluções não pretendidas.

Numa linguagem sofisticada como o *Mathematica* encontramos três tipos de rotinas para a determinação de raízes de equações, as rotinas Solve, NSolve e FindRoot. O utilizador deve ter em mente as limitações e potencialidades destas rotinas, não se devendo influen-

ciar pela tradução literal do nome. As rotinas Solve e NSolve, bastante sofisticadas, têm aplicação restrita a equações algébricas, e a sua diferença reside no facto da primeira permitir explicitar a resolução algébrica (quando é possível a redução a radicais) e a segunda apresentar imediatamente uma aproximação numérica (usando um algoritmo específico para equações algébricas). Quando a função não é polinomial, a única rotina que permite uma aproximação das raízes de equações é a rotina FindRoot, que consiste numa simples implementação numérica do método de Newton e não dispensa a introdução da iterada inicial[2].

2.1 Convergência de sucessões

Um *método iterativo* consiste, de um modo geral, numa aproximação inicial x_0, também designada *iterada inicial*, e num processo de obter sucessivamente novas iteradas x_{n+1} a partir das anteriores x_n, x_{n-1}, ...

Ao executar este procedimento criamos uma sucessão (x_n) que converge para o valor pretendido z.

Será crucial estabelecer de que forma um elemento da sucessão está mais ou menos próximo de z, e para isso definimos em cada iterada o erro

$$e_n = z - x_n.$$

Dessa forma, neste capítulo iremos construir sucessões que convergem para z, solução da equação $f(x) = 0$.

É claro que, havendo convergência, o erro cometido em cada iterada $e_n = z - x_n$ vai tender para zero, permitindo assim considerar os valores x_n como aproximações da raiz z a menos de um valor $\varepsilon > 0$ suficientemente pequeno, majorante do erro absoluto, $|e_n| < \varepsilon$.

Devemos distinguir conceptualmente entre dois tipos de estimativas de erros que podemos estabelecer, as estimativas *a priori* e as estimativas *a posteriori*. Numa estimativa *a priori* não é necessário calcular a iterada x_n para podermos apresentar uma majoração do erro $|e_n|$, ou seja, a partir da função conseguimos prever (sem efectuar iterações) que x_n é um valor suficientemente próximo do resultado. Ao contrário, numa estimativa *a posteriori* apenas podemos estabelecer qual a majoração do erro $|e_n|$ se calcularmos o valor de x_n, através do cálculo efectivo das iterações.

Como é suposto o erro e_n convergir para zero, começamos por relembrar as notações, que permitem comparar a convergência de sucessões.que tendem para zero,

$$\begin{aligned} e_n &= O(a_n), \text{ se } \exists C > 0 : |\frac{e_n}{a_n}| < C \\ e_n &= o(a_n), \text{ se } |\frac{e_n}{a_n}| \longrightarrow 0. \end{aligned}$$

Assim, quando escrevermos $e_n = O(\frac{1}{n^2})$, significa que a sucessão e_n converge para zero de forma semelhante a $\frac{1}{n^2}$. Iremos abordar métodos iterativos para os quais a convergência do

[2]Convém ainda estar atento a possíveis erros, como quando se introduz `FindRoot[x==Sin[x],{x,0.2}]` e nos aparece como resultado `x -> 0.0116915`, que não tem qualquer significado (a única raiz de $\sin(x) = x$ é $x = 0$), não aparecendo qualquer mensagem de aviso.

erro para zero é muito superior a $\frac{1}{n^p}$. Na realidade, poderemos garantir convergências do erro para zero do tipo $O(2^{-n})$ ou mesmo $O(2^{-2^n})$.

Um descuido frequente[3] é supor que se verificarmos que $x_{n+1} - x_n$ converge para zero, isso significa que há convergência. São bem conhecidos exemplos em que isso não se verifica, basta relembrar a série $s_n = \sum_{k=1}^{n} \frac{1}{k}$,que não é convergente[4] e no entanto $s_{n+1} - s_n = \frac{1}{n+1}$ tende para zero. Porém, uma análise mais atenta permite concluir o seguinte lema.

Lema 2.1 *Seja $d_n = x_{n+1} - x_n$. A sucessão (x_n) é convergente se e só se a série*

$$s_n = \sum_{k=0}^{n} d_k$$

for convergente. Em particular, se existir p :

$$|x_{n+1} - x_n| n^p \to c,$$

então (x_n) converge se $p > 1, c < +\infty$, e diverge se $p \leq 1, c > 0$.

Dem:
Basta reparar que

$$x_n - x_m = \sum_{k=m}^{n-1} x_{n+1} - x_n = \sum_{k=m}^{n-1} d_n = s_{n-1} - s_m.$$

Assim (s_n) é sucessão de Cauchy se e só se (x_n) também for. O resultado particular surge de aplicar o critério de comparação com a série $\sum \frac{1}{n^p}$, que é convergente para $p > 1$ e divergente para $p \leq 1$. ■

Aplicando o lema, basta que $x_{n+1} - x_n$ convirja mais rapidamente para zero que $\frac{1}{n^p}$, com $p > 1$, para podermos concluir a convergência. Quando $x_{n+1} - x_n$ converge para zero tão ou mais lentamente que $\frac{1}{n}$, concluímos a divergência. Este facto permite extrair informações importantes em circunstâncias em que não se sabe se uma sucessão converge ou não.

Exemplo 2.1 *Consideremos a sucessão $y_n = \frac{x_n}{10}$, em que x_n é definida recursivamente por*

$$x_1 = 1, \quad x_{n+1} = 1 + \frac{x_n^2}{x_n + 1}.$$

[3]Especialmente porque o critério $|x_{n+1} - x_n| < \varepsilon$ é muitas vezes utilizado como critério de paragem de um método.

[4]Pode obter-se mesmo a fórmula,

$$\sum_{k=1}^{n} \frac{1}{k} = \log(n) + \gamma + O(\frac{1}{n}),$$

em que $\gamma = 0.57721...$ é a denominada *constante de Euler*. Esta fórmula explicita bem a divergência da série.

Calculando os termos da sucessão, obtemos por exemplo

$$y_{998} = 4.37122,\ y_{999} = 4.37345,\ y_{1000} = 4.37569,\ y_{1001} = 4.37792,$$

e reparamos que os termos se estão a aproximar, embora muito lentamente. Uma observação descuidada poderia levar-nos a atribuir uma aproximação 4.37... *para o limite da sucessão. Ora, a sucessão* (x_n) *não converge e consequentemente* (y_n) *também não! Com efeito, façamos uma análise pouco rigorosa, mas instrutiva. Com base na comparação da sucessão* $|x_{n+1} - x_n|$ *com* $\frac{1}{\sqrt{n}}$ *obtemos o gráfico seguinte (ver figura em baixo, à esquerda) em que se torna evidente que* $|x_{n+1} - x_n|\sqrt{n} \sim 0.706...$*, e portanto segundo o critério apresentado* (x_n) *não deverá convergir. Iremos ver que de facto tal sucessão não poderia convergir, reparando que se trata de uma sucessão definida pelo método do ponto fixo, que não converge neste caso. Ainda com base nestes valores e reparando que* $x_{n+1} - x_n = \frac{1}{x_n+1}$*, podemos escrever* $\frac{1}{x_n+1} \approx \frac{0.706}{\sqrt{n}}$*, o que nos dá* $x_n \approx \frac{\sqrt{n}}{0.706} - 1$*. Assim, podemos efectuar uma estimativa para o valor* y_{5000} *com o simples cálculo da raiz e sem fazer as 4000 iterações restantes. O valor obtido pela estimativa é* 9.912*, o que não difere muito do valor correcto* 9.902 *(erro relativo de 0.1%). Se o mesmo fosse efectuado para* $y_{10\,000}$ *obtinhamos pela estimativa* 14.06 *ao invés de* 14.04*, com erro relativo de 0.14%, mas poupando 9000 iterações. Este exemplo ilustra a utilidade da previsão do comportamento assimptótico. Poderá pensar-se que é bem conhecida uma justificação teórica global que permite concluir a estimativa* $x_n = O(\sqrt{n})$*, mas não... Isso não invalida os resultados numéricos, lançando a motivação para os provar*[5].

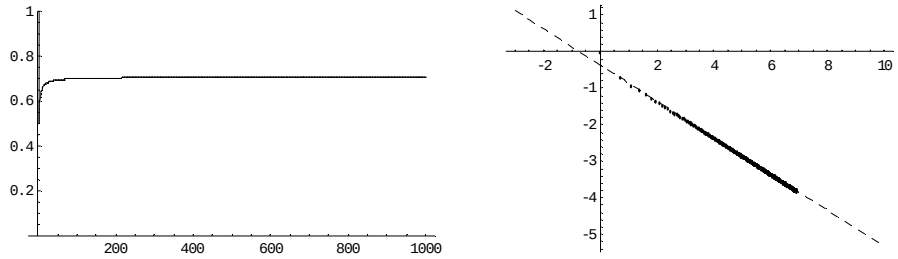


Figura: À esquerda, gráfico com os valores da sucessão $|x_{n+1} - x_n|\sqrt{n}$, onde se constata que, para n grande, o valor é quase constante, próximo de 0.706... À direita, gráfico onde são colocados os pontos $(\log(n), \log|x_{n+1} - x_n|)$. Sendo claro que definem uma recta (que é colocada a tracejado), ela é determinada por regressão linear (mínimos quadrados) como sendo $y = -0.353 - 0.4993x$. O valor da inclinação foi o que nos permitiu especular a relação com $O(n^{-0.5})$, e a exponencial da constante é $e^{-0.353} = 0.703...$ próximo de 0.706. Isto resulta de obter, a partir de $|x_{n+1} - x_n| \sim c\, n^p$, a relação $\log|x_{n+1} - x_n| \sim p\log(n) + \log(c)$. É este o processo adequado para determinar p, que a priori não será conhecido.

[5]Neste caso, o problema 'mais simples' consiste em mostrar que para sucessões definidas recursivamente por $x_{n+1} = x_n + \frac{1}{x_n^p+1}$ o comportamento assimptótico de (x_n) é um $O(n^{\frac{1}{p+1}})$. Repare-se que isto é consistente nos casos limite, $p = 0$, quando se reduz ao somatório, e quando $p = \infty$, em que se reduz à constante inicial.

2.1.1 Ordem de convergência

Consideremos diversos tipos de sucessões que convergem para 1,

$$u_n = 1 + \frac{1}{n^3}, \quad , v_n = 1 + \frac{1}{3^n}, \quad , x_n = 1 + \frac{1}{3^{2^n}}, \quad y_n = 1 + \frac{1}{3^{3^n}}, \quad , z_n = 1 + \frac{1}{3^{3^{n^2}}}.$$

Qualquer uma destas sucessões é um representante do número 1, enquanto sucessões de Cauchy de racionais que convergem para 1. No entanto, reparamos que podem ser distinguidas pela rapidez de convergência, verificando que a mais lenta é a primeira, u_n, e a mais rápida é a última, z_n.

Assim, começamos por estabelecer essa diferença definindo ordem de convergência.

Definição 2.1 *(ordem de convergência linear[6]). Seja (x_n) uma sucessão convergente para z, e seja $e_n = z - x_n \neq 0$. Consideramos a sucessão*

$$K_n = \frac{|e_{n+1}|}{|e_n|}. \tag{2.1}$$

Se $K_n \leq K^+ < 1$, para n suficientemente grande, dizemos que a sucessão tem pelo menos *ordem de convergência 1 ou* linear*.*
Se, para além disso, $K_n \geq K^- > 0$ dizemos que a sucessão tem *ordem de convergência linear.*

• Quando existe o limite finito

$$K_\infty = \lim K_n = \lim \frac{|e_{n+1}|}{|e_n|},$$

este valor é designado designado *coeficiente assimptótico de convergência.*

• Se $K_\infty < 1$, é fácil verificar que estamos numa situação em que há *pelo menos* convergência linear.

• Se $0 < K_\infty < 1$, então podemos mesmo concluir que estamos numa situação em que a convergência é linear.

• Se $K_\infty = 0$ dizemos que a sucessão tem ordem de convergência *supralinear.*

• Falta analisar se a situação $K_\infty \geq 1$ ocorre. Suponhamos que $K_\infty > 1$, então, a partir de certa ordem, teríamos $K_n > 1$, ou seja

$$|e_{n+1}| > |e_n|,$$

e isto contradiria o facto da sucessão convergir. Portanto, apenas podemos ter $K_\infty = 1$.

Quando a sucessão converge e $K_\infty = 1$, então dizemos que a ordem de convergência é *logarítmica* (ou infralinear).

[6]Iremos no capítulo seguinte generalizar esta noção para espaços de Banach.
Notamos ainda que a restrição $K^+ < 1$ é demasiado forte, e há algumas sucessões cuja convergência é claramente linear e que não se enquadram nesta definição. No entanto, caso esta restrição seja suprimida qualquer sucessão convergente apresentaria pelo menos convergência linear, o que também não nos parece apropriado (porém, na generalidade da bibliografia é esta a opção).

Exemplo 2.2 *Vejamos os exemplos apresentados de sucessões que convergem para 1.*
A sucessão (u_n) *tem convergência logarítmica, já que avaliando*

$$\frac{|e_{n+1}|}{|e_n|} = \frac{|1-u_{n+1}|}{|1-u_n|} = \frac{1/(n+1)^3}{1/n^3} = \frac{n^3}{(n+1)^3} \to 1,$$

concluímos que $K_\infty = 1$.
A sucessão (v_n) *tem convergência linear, porque*

$$\frac{|e_{n+1}|}{|e_n|} = \frac{|1-v_{n+1}|}{|1-v_n|} = \frac{3^{-(n+1)}}{3^{-n}} = \frac{1}{3},$$

e assim $K_\infty = \frac{1}{3} \in]0,1[$.
A sucessão (x_n) *tem convergência supralinear, porque*

$$\frac{|e_{n+1}|}{|e_n|} = \frac{|1-x_{n+1}|}{|1-x_n|} = \frac{3^{-2^{(n+1)}}}{3^{-2^n}} = \frac{3^{-2^n}}{3^{-2^n}} 3^{-2^n} = 3^{-2^n} \to 0,$$

e assim $K_\infty \to 0$.
Seria igualmente fácil verificar que as sucessões (y_n) *e* (z_n) *têm ordem de convergência supralinear*

Observação: A definição não é enunciada directamente no caso em que o limite K_∞ existe, porque pode haver situações em que tal limite não exista, e até se verifique *em algumas* iteradas $|e_{n+1}| > |e_n|$. Por exemplo, consideramos a sucessão

$$w_n = 1 + \frac{2+(-1)^n}{4^n}$$

que converge para 1, com rapidez semelhante à de (v_n). Neste caso

$$K_n = \frac{|1-w_{n+1}|}{|1-w_n|} = |\frac{2+(-1)^{n+1}}{4^{n+1}}|\,|\frac{4^n}{2+(-1)^n}| = \frac{2+(-1)^{n+1}}{2+(-1)^n}\frac{1}{4} = \begin{cases} \frac{1}{12}, & n \text{ par} \\ \frac{3}{4}, & n \text{ ímpar} \end{cases}$$

e portanto não há limite. No entanto, como enquadramos

$$0 < \frac{1}{12} \leq K_n \leq \frac{3}{4} < 1,$$

podemos concluir que se trata de convergência linear, de acordo com a definição[7].

[7]Por vezes, não é possível enquadrar K_n convenientemente, ainda que as sucessões tenham convergência semelhante à linear. Por exemplo, tomando

$$w_n = 1 + \frac{2+(-1)^n}{2^n},$$

a sucessão K_n teria como sublimites $\frac{1}{6}$ e $\frac{3}{2}$. Como $\frac{3}{2} \geq 1$, este caso não está incluído na definição.

A inclusão destes casos teria que ser alvo de análise mais detalhada. Por exemplo, se tivermos em conta a sucessão

$$w_n = 1 + \frac{2+(-1)^n}{n^2}$$

que converge para 1, com rapidez semelhante à de (u_n), verificamos que neste caso há também dois sublimites $\frac{1}{3}$ e 3, um deles inferior a 1 e o outro superior.

Iremos ainda classificar mais detalhadamente *algumas* sucessões que têm ordem de convergência supralinear.

Definição 2.2 *(ordem de convergência superior). Seja (x_n) uma sucessão convergente para z. Para $p > 1$ consideremos a sucessão*

$$K_n^{[p]} = \frac{|e_{n+1}|}{|e_n|^p}. \tag{2.2}$$

Se $K_n^{[p]} \leq K^+ < +\infty$, para n suficientemente grande, dizemos que a sucessão tem pelo menos *ordem de convergência p.*
Se, para além disso, $K_n^{[p]} \geq K^- > 0$ dizemos que a sucessão tem *ordem de convergência p.*

• Mais uma vez, quando existe o limite finito,

$$K_\infty^{[p]} = \lim K_n^{[p]} = \lim \frac{|e_{n+1}|}{|e_n|^p},$$

designamos este valor coeficiente assimptótico de convergência de ordem p (quando for clara qual é a ordem, não colocaremos o índice p).

• Se $0 < K_\infty^{[p]} < +\infty$, então podemos concluir a convergência tem ordem p.

• Se $p = 2$ a convergência diz-se *quadrática* e se $p = 3$, *cúbica*. Se $K_\infty^{[p]} = 0$ para qualquer p, então diremos que se trata de uma convergência *exponencial*.

Observação: Note-se que se a sucessão tiver ordem de convergência p, então terá pelo menos ordem de convergência $q < p$, o que torna a designação consistente. Com efeito, como $\lim |e_n|^{p-q} = 0$,

$$\frac{|e_{n+1}|}{|e_n|^q} = \frac{|e_{n+1}|}{|e_n|^p}|e_n|^{p-q} \to K_\infty^{[p]} \lim |e_n|^{p-q} = 0.$$

Se, por outro lado, considerarmos $q > p$, temos $\lim |e_n|^{p-q} = +\infty$, e como $K_{p,\infty} \neq 0$,

$$\frac{|e_{n+1}|}{|e_n|^q} = \frac{|e_{n+1}|}{|e_n|^p}|e_n|^{p-q} \to K_\infty^{[p]} \lim |e_n|^{p-q} = +\infty,$$

donde se conclui que não poderia ter ordem de convergência superior.

Exemplo 2.3 *Analisamos agora os exemplos das sucessões (x_n), $(y_n), (z_n)$ que tinhamos verificado terem convergência supralinear. No caso da sucessão (x_n),*

$$K_n^{[p]} = \frac{|e_{n+1}|}{|e_n|^p} = \frac{3^{-2^{(n+1)}}}{(3^{-2^n})^p} = 3^{-2^{(n+1)}+2^n p} = 3^{-2^n(2-p)},$$

e reparamos que se $p > 2$, temos $K_\infty^{[p]} = \infty$, e se $p < 2$ temos $K_\infty^{[p]} = 0$. Portanto apenas no caso $p = 2$, obtemos $K_\infty^{[2]} = 3$, e concluímos que se trata de uma convergência quadrática. No caso da sucessão (y_n) é fácil ver que a sua convergência é cúbica e que temos $K_\infty^{[3]} = 3$.

Finalmente, no caso da sucessão (z_n) a convergência é considerada exponencial, pois $K_n^{[p]} = 0$, qualquer que seja p, já que

$$K_n^{[p]} = \frac{|e_{n+1}|}{|e_n|^p} = \frac{3^{3^{n^2}p}}{3^{3^{n^2+2n+1}}} = 3^{3^{n^2}(p-3^{2n+1})} \to 0, \forall p.$$

Observação:

Um dos processos muito utilizados para o cálculo de valores de funções baseia-se no desenvolvimento em série de Taylor. Reparamos que se tivermos um desenvolvimento em série de potências em torno de z_0,

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k,$$

o cálculo dos termos da série pode ser considerado num processo iterativo da forma

$$\left.\begin{array}{l} s_0 = a_0, \\ s_{n+1} = s_n + a_{n+1}(z - z_0)^{n+1}. \end{array}\right\} \quad \text{e portanto, } s_n = \sum_{k=0}^{n} a_k (z - z_0)^k.$$

No caso da série de Taylor é bem conhecido que $a_n = \frac{f^{(n)}(z_0)}{n!}$. O caso mais simples ocorre quando os termos a_n são constantes. Por exemplo, se $a_n = 1, z_0 = 0$, obtemos para $|z| < 1$,

$$f(z) = \sum_{k=0}^{\infty} z^n = \frac{1}{1 - z},$$

ou seja, a bem conhecida fórmula da soma geométrica.

Para avaliarmos a rapidez de convergência da série de Taylor, reparamos que

$$\frac{e_{n+1}}{e_n} = \frac{f(z) - s_n + s_n - s_{n+1}}{f(z) - s_n} = 1 - \frac{s_{n+1} - s_n}{f(z) - s_n}.$$

Como

$$\begin{aligned} s_{n+1} - s_n &= \frac{f^{(n+1)}(z_0)}{(n+1)!}(z - z_0)^{n+1}, \\ f(z) - s_n &= \frac{f^{(n+1)}(\xi_{n+1})}{(n+1)!}(z - z_0)^{n+1} \end{aligned}$$

onde $\xi_{n+1} \in]z_0; z[$, pelo resto de Lagrange da série de Taylor.

Efectuando a razão entre os termos, ficamos com

$$\left|\frac{e_{n+1}}{e_n}\right| = \left|1 - \frac{f^{(n+1)}(z_0)}{f^{(n+1)}(\xi_{n+1})}\right|,$$

e como admitimos a continuidade de todas as derivadas, para valores de z próximos de z_0 temos $\frac{f^{(n+1)}(z_0)}{f^{(n+1)}(\xi_{n+1})}$ próximo de 1, o que significa que será possível majorar o valor $\left|\frac{e_{n+1}}{e_n}\right|$

por uma constante inferior a 1, concluindo-se a convergência linear (apenas poderia ser supralinear se $\xi_n \to z_0$).

Quando falarmos do método de Newton, veremos que tem convergência quadrática, e que permite obter excelentes aproximações de raízes de números a partir de simples operações aritméticas elementares. A sua convergência supralinear permite concluir que se trata de um melhor método para aproximar raízes do que utilizar a expansão em série de Taylor. Pode-se então levantar a questão, será que é possível obter a aproximação de senos, exponenciais ou logaritmos utilizando esse método com operações elementares? A resposta é negativa, quer utilizando o método de Newton, quer utilizando qualquer outro método de ponto fixo, e está relacionada com o facto desses números poderem ser transcendentes!

2.1.2 Tempo de cálculo

Notamos que nem sempre uma maior 'rapidez' de convergência é profícua.

Assim, suponhamos que (x_n) é uma sucessão que converge para z mais rapidamente do que (y_n). A priori, podemos pensar que é melhor utilizar a sucessão (x_n), mas se o cálculo de cada y_n envolver um tempo médio t_y menor que t_x (o tempo de cálculo médio de cada x_n), isso poderá não acontecer. Poderá acontecer que os tempos de cálculo aumentem com n, mas não iremos considerar aqui essa situação.

No caso de métodos iterativos, em que o cálculo da iterada necessita do valor de uma ou mais iteradas anteriores. Com efeito, podemos assumir que os tempos totais T_x e T_y, necessários para o cálculo de x_n e y_n (respectivamente), serão $T_x = n\,t_x,\ \ T_y = n\,t_y$.

Suponhamos que pretendemos obter um erro absoluto inferior a ε. Examinemos o número de iteradas necessário para garantir essa estimativa, considerando várias majorações de erro, de acordo com a ordem de convergência.

• Caso de convergência logarítmica.

Suponhamos que $|e_n| \leq an^{-s}$, com $s > 0$. Para que $|e_n| \leq \varepsilon$, obtemos de $an^{-s} \leq \varepsilon$, a relação,

$$\log n \geq \frac{1}{s} \log(\frac{a}{\varepsilon}).$$

• Caso de convergência linear.

Suponhamos agora que $|e_n| \leq ar^{-n}$, com $r > 1$. Da mesma forma, de $ar^n \leq \varepsilon$, obtemos,

$$n \geq \frac{\log(\frac{a}{\varepsilon})}{\log(r)}.$$

• Caso de convergência supralinear, de ordem $p > 1$.

Considerando $|e_n| \leq ar^{-p^n}$, com $r > 1$. De $ar^{-p^n} \leq \varepsilon$, obtemos,

$$p^n \geq \frac{\log(\frac{a}{\varepsilon})}{\log(r)}.$$

Estabelecendo a relação $a = M\varepsilon$, em que M será um valor elevado (pois ε pretende-se pequeno) e assumindo que escrevemos $s = \log r$, obtemos:

- $n \geq M^{1/s}$, no caso de convergência logarítmica.
- $n \geq \log(M^{1/s}) = \frac{\log(M)}{\log(r)}$, no caso de convergência linear.
- $n \geq \log_p(\log(M^{1/s})) = \log_p(\frac{\log(M)}{\log(r)})$, no caso de convergência supralinear (de ordem p).

Desta forma, é bem visível a diferença no número de iterações a calcular para obter uma certa precisão. Nestas três classes o número de iterações cresce de forma bem diferenciada e podemos ver que:

(i) Se (x_n) converge linearmente e (y_n) logaritmicamente, para atingir um erro absoluto inferior a ε, a relação entre os tempos totais será

$$\frac{T_x}{T_y} = \frac{\log(M)t_x}{\log(r)M^{1/s}t_y} \underset{M\to\infty}{\longrightarrow} 0,$$

ou seja, como a é fixo, assumirmos $M \to \infty$ corresponde a considerar $\varepsilon \to 0$. Nesse caso, qualquer que seja a relação entre t_x e t_y (considerados constantes), o tempo total de cálculo T_x será inferior a T_y, quando se pretendem erros cada vez mais pequenos. Isto não invalida, como é óbvio, que para alcançar um certo erro, não possa ser inferior T_y. Por exemplo, sendo $a = 1$,para alcançar $\varepsilon = 10^{-N}$, teremos $T_x = T_y$ se

$$\frac{t_y}{t_x} = \frac{\log_{10}(10^N)}{\log_{10}(r)10^{N/s}} = \frac{N}{\log_{10}(r)}10^{-\frac{N}{s}}.$$

Num caso concreto, $r = 10, s = 4, N = 8$ (precisão simples), obtemos

$$\frac{t_y}{t_x} = \frac{\log_{10}(10^8)}{\log_{10}(10)(10^8)^{1/4}} = \frac{8}{10^2} = 0.08,$$

e portanto se $t_y < 0.08t_x$, o cálculo de 100 iterações para obter y_{100} com a estimativa $|e_n| \leq n^{-4}$ (portanto $|e_{100}| \leq (10^2)^{-4} = 10^{-8}$), irá demorar menos tempo que o cálculo de 8 iterações para obter x_8 com a estimativa $|e_n| \leq 10^{-n}$. É também claro que se mantivessemos esta relação entre t_y e t_x, ao considerar um erro mais pequeno $\varepsilon = 10^{-16}$ (precisão dupla), o tempo total para calcular y_{10000} seria maior que x_{16} (a menos que $t_y < 0.0016\, t_x$).

(ii) De forma semelhante existe uma grande diferença entre considerar (x_n) a convergir supralinearmente e (y_n) linearmente. Cálculos semelhantes levariam ao mesmo tipo de conclusões.

(iii) Será que existe também uma grande diferença entre as várias ordens de convergência supralinear? A resposta é não! Vejamos porquê.

Suponhamos que (x_n) converge com ordem p e (y_n) com ordem q, com $p > q$. Obtemos (para coeficientes iguais, para coeficientes diferentes podemos obter a mesma relação no limite),

$$\frac{T_x}{T_y} = \frac{\log_p(\log(M^{1/s}))t_x}{\log_q(\log(M^{1/s}))t_y} = \frac{\log q}{\log p}\frac{t_x}{t_y}, \tag{2.3}$$

ou seja, neste caso obtém-se uma razão constante. Podemos concluir que se $t_x = t_y$, aumentar a ordem do método de $q > 1$ para p, apenas se traduz na redução de tempo de cálculo

no factor $\frac{\log q}{\log p}$. No caso de passarmos de ordem convergência quadrática para cúbica o salto não é significativo e traduz-se num simples decréscimo de $\frac{\log 2}{\log 3}$, ou seja, aproximadamente 63%. Isto significa que *a priori* pode não compensar executar um algoritmo que nos dê ordem de convergência cúbica ao invés de quadrática se ele demorar o dobro do tempo. Para que houvesse uma grande diferença entre estes valores, o p teria que ser tão grande quanto possível, ou seja entraríamos no caso de convergência exponencial.

(iv) Comparação entre métodos com a mesma ordem de convergência, mas com coeficientes diferentes. Já mencionámos que no caso de convergência supralinear, para os exemplos considerados, a diferença esbate-se no limite. No caso de convergência linear, essa diferença assume um carácter de proporcionalidade tal como acontecia anteriormente, no caso supralinear, ao considerar diferentes ordens p.

Com efeito, se considerarmos $M_x = cM_y$ e $r_x \neq r_y$, obtemos

$$\frac{T_x}{T_y} = \frac{\log(M_x)\log(r_y)}{\log(r_x)\log(M_y)}\frac{t_x}{t_y} = \frac{\log(r_y)}{\log(r_x)}\frac{\log(M_y)+\log(c)}{\log(M_y)}\frac{t_x}{t_y} \underset{M_y\to\infty}{\longrightarrow} \frac{\log(r_y)}{\log(r_x)}\frac{t_x}{t_y},$$

ou seja, verifica-se uma relação semelhante à anterior, mas agora para os coeficientes assimptóticos que são $K_{x,\infty} = \frac{1}{r_x}$ e $K_{y,\infty} = \frac{1}{r_y}$.

2.2 Teoremas elementares e a localização de raízes

Traçando o gráfico da função, podemos ter uma ideia aproximada da localização das raízes, mas para assegurarmos rigorosamente que num intervalo existe uma e uma só raiz recordamos dois teoremas elementares da Análise Matemática.

Teorema 2.1 *(do Valor Intermédio de Bolzano):*
Seja f uma função contínua no intervalo $[a,b]$. Se $f(a)f(b) \leq 0$ então existe pelo menos *uma raiz da equação $f(x) = 0$ no intervalo $[a,b]$.*

Teorema 2.2 *(do Valor Médio de Lagrange)*
Se $f \in C^1([a,b])$, então $\exists \xi \in]a,b[:\ f'(\xi) = \frac{f(b)-f(a)}{b-a}$.$\square$

Corolário 2.1 *(Rolle) Seja $f \in C^1([a,b])$. Se $f(a) = f(b)$ então $\exists \xi \in]a,b[:\ f'(\xi) = 0$.*

O contra-recíproco deste corolário é útil para a unicidade:

• Se $f'(x) \neq 0$ para todo o $x \in [a,b]$, então existe no máximo um $z \in [a,b]$ tal que $f(z) = 0$.

Repare-se que $\forall_{x\in I}\ f'(x) \neq 0$, implica que $\forall_{x\in I}\ f'(x) > 0$ (f estritamente crescente) ou $\forall_{x\in I}\ f'(x) < 0$ (f estritamente decrescente).

É fácil ver que podemos garantir a unicidade desde que a função f seja *estritamente monótona*, não sendo necessário exigir que seja diferenciável.

O teorema do valor médio de Lagrange permite ainda obter um resultado muito importante nas estimativas *a posteriori* de erros de uma raiz de uma função f diferenciável.

Teorema 2.3 *(estimativa elementar de erros de raízes). Seja $\tilde{x}$ uma aproximação da raiz z da função $f \in C^1([\tilde{x}; z])$, então*[8]

$$|e_{\tilde{x}}| = |z - \tilde{x}| \leq \frac{|f(\tilde{x})|}{\min_{\xi \in [\tilde{x};z]} |f'(\xi)|}.$$

Demonstração: Exercício (aplicação imediata do teorema do valor médio).■

Observação 1:
(i) A simplicidade deste resultado torna por vezes esquecida a sua utilidade. Assim, quando aplicarmos qualquer dos métodos que iremos desenvolver nas secções seguintes (ou outros métodos), ao obter uma iterada x_n que aproxima a solução z, podemos apresentar uma estimativa *a posteriori* baseada neste resultado.
(ii) Nem sempre poderá ser fácil calcular o mínimo de f'. No entanto, caso a função seja diferenciável e não seja nula próximo da raiz, tendo obtido uma aproximação $\tilde{x}$ e uma outra $\tilde{y}$, tais que $f(\tilde{x})f(\tilde{y}) < 0$, podemos arriscar a estimativa

$$|e_{\tilde{x}}| \sim \frac{|f(\tilde{x})|\,|\tilde{x} - \tilde{y}|}{|f(\tilde{x}) - f(\tilde{y})|},$$

que será tanto melhor quanto $\tilde{x}$ estiver próximo de $\tilde{y}$ (Exercício!). A utilização deste tipo de aproximações é sistematizada no método da falsa posição e da secante, que iremos expor.

Observação 2:
Quando se tenta estabelecer majorações para os erros em intervalos somos confrontados com a necessidade de minimizar ou maximizar o valor absoluto de uma função num intervalo. Convém por isso relembrar que os extremos de uma função real num intervalo são atingidos nas extremidades do intervalo ou então em pontos interiores que sejam extremos relativos.
O critério $f'(x) = 0$ para encontrar extremos relativos só é aplicável se a função for diferenciável, no entanto isso não acontece com o módulo de uma função (a menos que ela seja sempre positiva ou negativa). Por isso, o melhor processo ao analisar os extremos de $|f(x)|$ é analisar os extremos de $f(x)$. Se uma função contínua f tiver x^- como mínimo e x^+ como máximo nesse intervalo então $|f(x)|$ terá como máximo $\max\{|x^-|, |x^+|\}$, e como mínimo $\min\{|x^-|, |x^+|\}$ (a menos que x^- seja negativo e x^+ positivo, já que nesse caso o mínimo será zero).
Para evitar o incómodo de calcular derivações para determinar se há extremos relativos, podemos usar propriedades acerca da monotonia de funções[9]: se f, g são crescentes $f + g$

[8]Usaremos a notação $[a; b]$ para designar um intervalo que é $[a, b]$ se $b > a$, e $[b, a]$ caso contrário.
[9]Para deduzir estas e outras propriedades, é cómodo usar o facto que uma função diferenciável f é crescente se $f' \geq 0$.

também é, e se ambas forem positivas fg também o será... se f é crescente $-f$, $\frac{1}{f}$ serão decrescentes, etc...

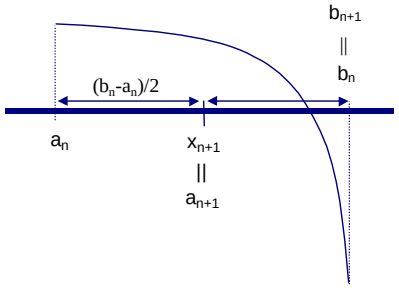
Estes teoremas de análise elementar não explicitam um método para aproximar a solução do problema. No entanto, podemos basear-nos neles para desenvolver alguns métodos iterativos muito simples.

2.3 Método da Bissecção

Vamos supor que no intervalo $[a, b]$ a equação $f(x) = 0$ tem uma só raiz z. A partir do intervalo $[a_0, b_0] = [a, b]$, vamos construir intervalos $[a_n, b_n]$ com metade do comprimento dos anteriores, através do ponto médio

$$x_{n+1} = \frac{1}{2}(a_n + b_n).$$

Assegurando a existência da raiz nesses intervalos através do teorema do valor intermédio, isso permite obter uma sucessão (x_n) que converge para essa raiz.



O *método da bissecção* pode ser esquematizado num algoritmo (ver observação 3):

Intervalo Inicial : $[a_0, b_0] = [a, b]$	
Repetir:	**1)** $x_{n+1} = (a_n + b_n)/2$
	2) Se $f(x_{n+1})f(a_n) < 0$
	Então $a_{n+1} = a_n; b_{n+1} = x_{n+1}$
	Senão $a_{n+1} = x_{n+1}; b_{n+1} = b_n$
Até que:	$f(x_{n+1}) = 0$ **ou** $\lvert x_{n+1} - x_n \rvert < \varepsilon$

Assim,

$$f' \geq 0, g' \geq 0 \Rightarrow (f+g)' = f' + g' \geq 0,$$

mas para obter

$$f' \geq 0, g' \geq 0 \Rightarrow (fg)' = f'g + fg' \geq 0,$$

devemos exigir que f e g sejam positivas.

A composição de funções crescentes é crescente, pois $(f \circ g)' = f'(g)g'$, mas é preciso ter em atenção o domínio.

Usamos o *critério de paragem* $|x_{n+1} - x_n| < \varepsilon$, onde o valor $\varepsilon > 0$ é um valor suficientemente pequeno, já que se poderá garantir que o erro absoluto verifica $|e_{n+1}| < \varepsilon$. Com efeito, como temos

$$|x_{n+1} - x_n| = \frac{1}{2}|a_n - b_n|,$$

basta mostrar que

$$|z - x_{n+1}| \leq \frac{1}{2}|a_n - b_n|.$$

Isto verifica-se se $z \in [a_n, b_n]$ porque

$$|z - x_{n+1}| = |z - \frac{a_n + b_n}{2}| = \frac{1}{2}|z - a_n + z - b_n| \leq \frac{1}{2}(\,|z - a_n| + |z - b_n|\,),$$

e de reparar que $a_n \leq z \leq b_n$, logo $|z - a_n| + |z - b_n| = z - a_n + b_n - z = b_n - a_n$.

Resta saber se $z \in [a_n, b_n]$, o que é a priori garantido por construção. No entanto, apresentamos o resultado na demonstração do seguinte teorema.

Teorema 2.4 *Se $f(a)f(b) < 0$ e f for contínua com apenas uma raiz em $[a, b]$, então o método da bissecção converge e verifica-se a estimativa a priori (independente da função),*

$$|e_n| \leq \frac{1}{2^n}\,|a - b|. \tag{2.4}$$

Demonstração:

Pelo teorema do valor intermédio há uma raiz em $[a, b]$, que por hipótese será única. Iremos agora mostrar por indução que essa raiz está sempre em $[a_n, b_n]$. O caso $n = 0$ já foi justificado. Supondo que a raiz está em $[a_n, b_n]$, e que temos $f(a_n)$ e $f(b_n)$ com sinais diferentes, então qualquer que seja o sinal de $f(x_{n+1})$, será diferente de um deles (a menos que seja nulo e aí o método atinge a raiz). Os valores a_{n+1}e b_{n+1} são escolhidos de forma a serem iguais aos anteriores ou a x_{n+1}, verificando-se $f(a_{n+1})f(b_{n+1}) < 0$, o que garante a existência de raiz em $[a_{n+1}, b_{n+1}]$.

A partir daqui podemos determinar facilmente um majorante do erro para uma iterada x_n a partir do comprimento do intervalo inicial:

$$|e_n| \leq \frac{1}{2}|a_{n-1} - b_{n-1}| \leq \frac{1}{2}(\frac{1}{2}|a_{n-2} - b_{n-2}|) \leq ... \leq \frac{1}{2^n}|a_0 - b_0|,$$

e isto prova a convergência, já que o erro tende para zero. ■

• A partir da estimativa a priori, deduzimos imediatamente que basta um número de iteradas $n > \log_2(\frac{1}{\varepsilon}|a - b|)$ para obter $|e_n| < \varepsilon$.

• Um critério para mostrar que há apenas uma raiz em $[a, b]$ consiste em garantir que a função é estritamente monótona (o que acontece caso a derivada não se anule em $]a, b[$), no entanto poderá haver outros processos. Note-se que $f(a)$ e $f(b)$ terem sinais diferentes não garante, como é óbvio, que haja apenas uma raiz, e exigir que haja apenas uma raiz não garante que $f(a)$ e $f(b)$ tenham sinais diferentes (basta ver o caso de raízes duplas, em que a derivada também é nula na raiz, e que aqui é excluído).

Observação 1 *(representação em binário).*

(a) Consideremos $[a, b] = [0, 1]$. Então é fácil concluir que os possíveis valores para x_n são dados de forma exacta usando a representação em binário (que é a utilizada pelos computadores). Numa iterada m os possíveis valores que x_m pode tomar são todas as 2^m combinações de $(0.c_1c_2...c_m)_2$ (o índice 2 significa base binária) com $c_i \in \{0, 1\}$. Ao aumentar o valor de m estamos a acrescentar algarismos na representação correcta de z na notação binária. Conclui-se assim que o método atinge a raiz num *número finito* de iterações quando ela é representada de forma exacta (i.e. com um número finito de algarismos) na notação binária.

Com efeito, podemos descrever o algoritmo da bissecção (neste intervalo) da seguinte forma:

Sendo $a_n = (0.c_1...c_{n-1}0)_2$, teremos $b_n = (0.c_1...c_{n-1}1)_2$ e $x_{n+1} = (0.c_1...c_{n-1}01)_2$. O valor do algarismo c_n será 0 se a raiz estiver em $[a_n, x_{n+1}]$ e será 1 se a raiz estiver em $]x_{n+1}, b_n]$.

Repare-se que o valor a_n corresponde a efectuar, na base binária, um arredondamento por corte, e b_n corresponde a um arredondamento por excesso.

(b) Note-se que podemos sempre reduzir o problema ao intervalo $[0, 1]$ com uma mudança de variável elementar, que não altera significativamente a função. No entanto isso não reduz os cálculos, que são transportados para a mudança de variável. Sempre que começarmos com um intervalo em que os extremos são exactos na base binária, as iteradas x_n ainda são exactas nessa base.

(c) Poder-se-ia pensar num método da 'deca-secção', dividindo o intervalo em 10 partes, com o intuito de obter o valor exacto na notação decimal, mas isso implicaria ter que calcular 9 vezes a função em cada iterada, o que seria contraproducente, já que ao fim de 4 cálculos efectuados pelo método da bissecção temos a raiz enquadrada em intervalos mais pequenos, porque $2^{-4} < 10^{-1}$. Para qualquer outra base $b > 2$, a questão coloca-se da mesma maneira pois $2^{-b-1} < b^{-1}$. Conclui-se assim que a escolha de dividir em dois é a melhor escolha.

Observação 2 *(ordem de convergência do método da bissecção).*

O caso do método da bissecção é um dos casos em que a aproximação de z pode ser melhor na iterada x_n do que na iterada x_{n+1}, ou seja podemos ter $|e_{n+1}| > |e_n|$, num número infinito de iteradas. Percebemos isto facilmente se interpretarmos as aproximações como arredondamentos por corte ou por excesso, conforme explicado na observação anterior. É claro que, conforme o valor z, que é desconhecido, tanto podemos obter uma melhor aproximação por defeito como por excesso, e isso pode repetir-se um número infinito de vezes (será o caso em mantissas com período p, que representam racionais, já que o problema se repetirá de p em p iterações).

Consequentemente, de acordo com a definição clássica apresentada, não se poderá concluir a convergência linear do método. No entanto, convém realçar que se trata de um problema da definição, já que a estimativa $|e_n| \leq \frac{1}{2^n} |a - b|$ permite concluir que se trata de um método cuja convergência é semelhante à linear. Por isso, é *assumido* normalmente que *a convergência do método da bissecção é linear.*

Observação 3 *(tempo de cálculo do método da bissecção).*

Em cada iteração a função é calculada apenas uma vez (no ponto x_{n+1}), já que os valores $f(a_n)$ e $f(b_n)$ devem estar guardados, não devem ser recalculados! Portanto, sendo

t_f o tempo de cálculo médio da função f, o tempo de cálculo total para obter x_n será aproximadamente

$$T = n\, t_f \ ,$$

desprezando o tempo necessário para verificar se tem o mesmo sinal (a multiplicação é mais lenta e não deve ser efectuada).

Observação 4 *(método da bissecção com várias raízes).*

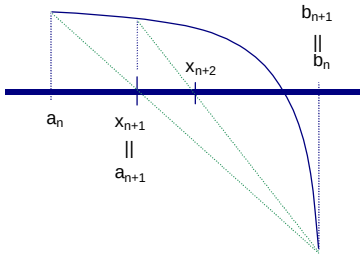
Com uma pequena modificação o método da bissecção poderá ser implementado em situações em que se exija apenas que a função seja contínua, desde que a função mude de sinal no intervalo (o que pode não acontecer, p.ex. raízes duplas, ou pares). Com efeito, se $f(a)f(b) > 0$ e se a função mudar de sinal haverá duas ou mais raízes, que podem ser enquadradas fazendo a divisão sucessiva do intervalo até que se encontre um ponto x_n tal que $f(x_n)$ tenha sinal diferente de $f(a)$ e de $f(b)$.

2.4 Método da Falsa Posição

Este método, também conhecido pela designação *Regula Falsi* (do latim), é um método semelhante ao método da bissecção, e em que também supomos que no intervalo $[a, b]$ a equação $f(x) = 0$ tem uma e uma só raiz. A única diferença reside no cálculo de x_{n+1},

$$x_{n+1} = a_n - f(a_n)\frac{b_n - a_n}{f(b_n) - f(a_n)}, \tag{2.5}$$

onde intervêm os valores de f. Este valor corresponde à raiz do polinómio do primeiro grau representado pela recta que une os pontos $(a_n, f(a_n))$ e $(b_n, f(b_n))$.



Observação 1. O denominador nunca pode ser nulo, porque $f(a_n)$ e $f(b_n)$ têm sinais opostos, e que também podemos escrever

$$x_{n+1} = \frac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)} = b_n - f(b_n)\frac{b_n - a_n}{f(b_n) - f(a_n)}.$$

Sendo também usada a notação,

$$f[a, b] = \frac{f(b) - f(a)}{b - a},$$

a recta que une os pontos é obtida pelo polinómio

$$P_1(x) = f(a_n) + f[a_n, b_n](x - a_n),$$

ficando claro que o ponto em que anula é justamente $x_{n+1} = a_n - \frac{f(a_n)}{f[a_n,b_n]}$, o que confirma a fórmula (2.5).

Também nos será útil o valor

$$f[a,b,c] = \frac{f[c,b] - f[c,a]}{b-a},$$

que permite obter o polinómio de 2^o grau (...parábola) que passa pelas imagens dos pontos a_n, b_n e c, através de

$$P_2(x) = P_1(x) + f[a_n, b_n, c](x - a_n)(x - b_n).$$

Estes valores estão directamente relacionados com derivadas, quando a função é regular, pois pelo teorema de Lagrange obtemos[10]

$$f[a,b] = f'(\xi), \quad f[a,b,c] = \frac{1}{2}f''(\eta),$$

onde $\xi \in]a;b[, \eta \in]a;b;c[$.□

Observação 2. É intuitivo que $x_{n+1} \in]a_n, b_n[$, mas devemos justificar.

Suponhamos que $f(a_n) < 0$ e que $f(b_n) > 0$. Verificar

$$a_n < x_{n+1} = \frac{f(b_n)a_n - f(a_n)b_n}{f(b_n) - f(a_n)} < b_n$$

é equivalente a (porque $f(b_n) > f(a_n)$)

$$f(b_n)a_n - f(a_n)a_n < f(b_n)a_n - f(a_n)b_n < f(b_n)b_n - f(a_n)b_n,$$

ou seja,

$$f(a_n)(b_n - a_n) < 0, \quad 0 < f(b_n)(b_n - a_n).$$

Como $b_n > a_n$, então por hipótese estas condições são verificadas.

Na outra hipótese possível $f(a_n) > 0$, $f(b_n) < 0$, a justificação é semelhante.□

Quando as funções são regulares e os intervalos suficientemente pequenos, a aproximação linear é razoável e este processo justifica-se, especialmente se a inclinação $|f'|$ for superior ao 'valor' da concavidade $|f''|$. No entanto, aparece persistentemente um problema, quando a concavidade é acentuada e há zonas em que a derivada é pequena, que apenas é possível contornar com uma modificação do método (que veremos no próximo parágrafo).

Iremos agora verificar que sob as mesmas condições do método da bissecção, o método da falsa posição converge.

[10]Basta ver que $g(x) = f(x) - P_2(x)$ tem três zeros, os pontos a, b, c onde a função f coincide com o polinómio. Pelo teorema de Rolle, a sua derivada terá dois (um em $]a;b[$ e outro em $]b;c[$. Da mesma forma, a segunda derivada terá um, que designamos η e que está em $]a;b;c[$.

Como $g''(x) = f''(x) - 2f[a,b,c]$, concluímos que

$$0 = f''(\eta) - 2f[a,b,c].$$

Teorema 2.5 *Se $f(a)f(b) > 0$ e f for contínua com apenas uma raiz em $[a, b]$, então o método da falsa posição converge. Quando $f \in C^2[a_n, b_n]$ verifica-se*

$$e_{n+1} = -\frac{f''(\eta_n)}{2f'(\xi_n)}(z - a_n)(z - b_n) \tag{2.6}$$

em que $\xi_n, \eta_n \in]a_n, b_n[$.

Demonstração:

(i) A demonstração de que haverá sempre uma raiz no intervalo $[a_n, b_n]$ é semelhante à efectuada para o método da bissecção, bastando reparar que, de acordo com a fórmula (2.5), $x_{n+1} \in]a_n, b_n[$ (só seria igual a a_n ou b_n se fossem raízes). Portanto ou $a_n < a_{n+1}$ ou $b_{n+1} > b_n$, pelo que temos $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$. Como (a_n) e (b_n) são ambas sucessões monótonas num intervalo limitado, têm limite, que designaremos $\alpha = \lim a_n, \beta = \lim b_n$, e por construção $z \in [\alpha, \beta]$.

Se $\alpha = \beta$, é óbvio que $z = \alpha = \beta$.

Se $\alpha \neq \beta$, então, como assumimos que f é contínua, existe o limite

$$\lim x_{n+1} = \frac{f(\beta)\alpha - f(\alpha)\beta}{f(\beta) - f(\alpha)}$$

e como $x_n = a_n$ ou $x_n = b_n$, existindo limite, ele será α ou β.

Se $\lim x_n = \alpha$, retiramos

$$\alpha = \frac{f(\beta)\alpha - f(\alpha)\beta}{f(\beta) - f(\alpha)} \Leftrightarrow f(\alpha) = 0,$$

ou seja, $\alpha = z$, porque há uma única raiz. Se $\lim x_n = \beta$, é semelhante.

(ii) Com a finalidade de estabelecermos uma fórmula de erro para o método da falsa-posição, consideramos a fórmula (ver observação 1)

$$f(x) = P_1(x) + f[a_n, b_n, x](x - a_n)(x - b_n) \ ,$$

Como $f(z) = 0$,

$$0 = P_1(z) + f[a_n, b_n, z](z - a_n)(z - b_n).$$

Basta reparar agora que $P_1(x_{n+1}) = 0$, e assim

$$P_1(z) = P_1(z) - P_1(x_{n+1}) = f[a_n, b_n](z - x_{n+1}),$$

Juntando as duas igualdades temos

$$0 = f[a_n, b_n](z - x_{n+1}) + f[a_n, b_n, z](z - a_n)(z - b_n),$$

e assim

$$e_{n+1} = -\frac{f[a_n, b_n, z]}{f[a_n, b_n]}(z - a_n)(z - b_n) \ , \tag{2.7}$$

Usando a relação com as derivadas (ver observação 1) obtemos o resultado. ■

Proposição 2.1 *Consideremos f nas condições do teorema anterior, com $f \in C^2(V_z)$, em que V_z é uma vizinhança da raiz. Se $f''(z) \neq 0$, então, a partir de certa ordem, um dos extremos do intervalo $[a_n, b_n]$, definido pelo método da falsa posição, fica fixo.*

Demonstração:

Seja $\alpha = \lim a_n, \beta = \lim \beta_n$, tal como na demonstração anterior, onde vimos que $z = \alpha$ ou $z = \beta$. Quando $\beta \neq \alpha = z$, isso implica que a sucessão (b_n) seja constante a partir de certa ordem, porque senão tomaria valores x_n que convergem para z e teríamos $\beta = z$. Situação semelhante ocorre quando $\alpha \neq \beta = z$, sendo nesse caso (a_n) constante a partir de certa ordem.

Resta ver que a situação $\alpha = \beta = z$ não pode ocorrer sob as hipóteses da proposição.

Como $f(a)f(b) < 0$ e há apenas uma raiz só podemos ter $f'(z) = 0$ se $f''(z) = 0$, caso excluído.

Portanto $f'(z) \neq 0$ e então numa vizinhança $V_{\varepsilon'}(z)$ a função f' (que é contínua) tem o mesmo sinal que $f'(z)$. Da mesma forma, numa vizinhança $V_{\varepsilon''}(z)$, a função f'' terá o mesmo sinal que $f''(z)$. Assim considerando $V_z = V_{\varepsilon''}(z) \cap V_{\varepsilon''}(z)$, o sinal de $\frac{f''}{2f'}$ é fixo em V_z.

Por outro lado, pela fórmula (2.6), o sinal de $e_{n+1} = z - x_{n+1}$ é igual ao sinal de $\frac{f''(\eta_n)}{2f'(\xi_n)}$, porque o valor $(z - a_n)(z - b_n)$ é sempre negativo.

Se, por absurdo, $\alpha = \beta = z$, temos $a_n, b_n \to z$. Notando que $\xi_n, \eta_n \in]a_n, b_n[$, temos $\xi_n, \eta_n \in V_z$, para n suficientemente grande. Isto significa que o sinal de e_{n+1} é fixo, ou seja, ou x_{n+1} é sempre maior que z, ou será sempre menor. Isto significa que um dos extremos fica fixo, não podendo convergir para z, o que contradiz $\alpha = \beta = z$. ■

A condição de f'' se anular na raiz raramente acontecerá, e assim no método da falsa posição, a partir de certa ordem q, um dos extremos de $[a_n, b_n]$ irá ficar imobilizado. Temos assim duas hipóteses:

– Se $f''(z)f'(z) > 0$, ou o que é equivalente, $f''(z)f(b_q) > 0$ (pois o sinal de $f(b_q)$ é igual ao de f'), .então as iteradas $x_{q+1}, x_{q+2}, ...$ vão ficar à esquerda da raiz significando isso que temos $b_q = b_{q+1} = b_{q+2} = ...$

– Se $f''(z)f'(z) < 0$, ou seja se $f''(z)f(a_q) > 0$ (porque o sinal de $f(a_q)$ é oposto ao de f'), então $a_q = a_{q+1} = a_{q+2} = ...$

Resumindo, se f'' tiver sinal constante num intervalo $[a_q, b_q]$ então o extremo do intervalo para o qual f possui esse sinal, permanece fixo.

Por exemplo, se $f''(z) > 0$, então ou temos $f(a_q) > 0$, e fica constante $a_n = a_q$, ou temos $f(b_q) > 0$, e fica constante $b_n = b_q$.

Assim, se $f''(z)f(a_q) > 0$, o método reduz-se a efectuar (para $n \geq q$) :

$$x_{n+1} = x_n - f(x_n)\frac{x_n - a_q}{f(x_n) - f(a_q)}. \tag{2.8}$$

sem haver necessidade de fazer comparações.

Usando (2.6), se $f \in C^2[a_q, b_q]$, a fórmula para o erro fica (para $n \geq q$)

$$e_{n+1} = -\frac{f''(\eta_n)}{2f'(\xi_n)}(z - a_q)e_n, \tag{2.9}$$

com $\xi_n, \eta_n \in]a_q; x_n[$.

Reparamos que neste caso, podemos explicitar

$$K_n = \frac{|e_{n+1}|}{|e_n|} = \frac{|f''(\eta_n)|}{2|f'(\xi_n)|}|z - a_q|$$

e existe o coeficiente assimptótico de convergência $K_\infty = \lim K_n$, porque f' e f'' são contínuas. Podemos mostrar que o coeficiente $K_\infty < 1$, concluindo que se trata de um método com convergência linear (quando $f''(z) \neq 0$). No entanto, para provar isso, não usamos directamente a fórmula (2.9), porque os limites dos valores ξ_n, η_n são desconhecidos (ver observação 3).

No caso $f''(z)f(b_q) > 0$, as fórmulas seriam semelhantes, trocando simplesmente a_q por b_q.

Observação 3 *(convergência do método da falsa posição).*
Consideramos ainda $f''(z) \neq 0$ e $a_n = ... = a_q$.
Como temos $z < x_{n+1} < x_n$, é óbvio que

$$0 < \frac{e_{n+1}}{e_n} < 1,$$

passando ao limite (que existe) podemos concluir $K_\infty \leq 1$, mas o essencial é ver que $K_\infty \neq 1$. Como,

$$\frac{x_{n+1} - x_n}{z - x_n} = \frac{-f(x_n)/f[a_n, x_n]}{-f(x_n)/f[z, x_n]} = \frac{f[z, x_n]}{f[a_q, x_n]},$$

temos

$$\frac{e_{n+1}}{e_n} = 1 - \frac{x_{n+1} - x_n}{z - x_n} = 1 - \frac{f[z, x_n]}{f[a_q, x_n]} \to 1 - \frac{f'(z)}{f'(\xi_z)},$$

onde $\xi_z \in]a_q, z[$. O caso $K_\infty = 1$, apenas poderia ocorrer se $f'(z) = 0$, situação excluída.

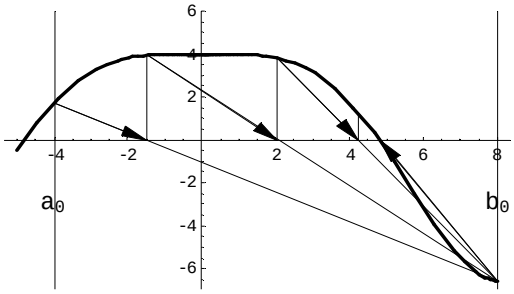
Conclui-se assim que o *método tem convergência linear* e coeficiente assimptótico

$$0 < K_\infty = 1 - \frac{f'(z)}{f'(\xi_z)} < 1,$$

com $\xi_z \in]a_q, z[$. A situação limite em que o método exibiria *convergência supralinear* seria $\xi_z = z$, o que aconteceria se $a_n \to z$, i.e. se $f''(z) = 0$, como aliás também se pode verificar na fórmula (2.9).

Exemplo 2.4 *Consideremos a função* $f(x) = 4 - \frac{x^2}{5}\sin(\frac{x}{4})^2$, *representada na figura em baixo. Começando com o intervalo* $[-4, 8]$ *obtiveram-se as seguintes iteradas,* $x_1 = -1.498, x_2 = 2.058, x_3 = 4.231, x_4 = 4.845, ..., x_7 = 4.79879$. *As três primeiras iteradas dão valores que se situam à esquerda da raiz, mas como há um ponto* $w = 5.63...$ *no qual* $f''(w) = 0$, *na quarta iterada é o extremo à direita que se desloca, e a partir daí o método converge fixando o extremo* $b_4 = x_4$ *e apenas variando* a_n. *No entanto, como temos* $f'(z) = -2.44, f''(z) =$

-0.572, *o método converge de forma razoavelmente rápida para* z, *de forma que em* x_7 *os dígitos apresentados são já os correctos.*



Observação 4 *(estimativa a priori)*
Podemos ainda estabelecer a estimativa *a priori*

$$|e_{n+1}| \leq K^n |b-a|$$

se encontrarmos um intervalo $[a,b]$ (que poderá não ser o inicial) em que

$$K = \frac{\max_{x\in[a,b]} |f''(x)|}{2\min_{x\in[a,b]} |f'(x)|}(b-a) < 1. \tag{2.10}$$

Para que a estimativa de erro seja mais eficaz que a do método da bissecção torna-se importante que K seja inferior a $\frac{1}{2}$,isso pode ser conseguido se o comprimento do intervalo for suficientemente pequeno e, é claro, apenas quando $K_\infty < \frac{1}{2}$.

Observação 5 *(tempo de cálculo).*
Tal como no método da bissecção, em cada iteração a função é apenas calculada uma vez, no ponto x_{n+1}, a partir dos valores $f(a_n), f(b_n)$ guardados. Normalmente, o tempo de cálculo da função é muito maior que o tempo de cálculo das operações elementares (3 subtracções, 1 divisão e uma multiplicação), que será designado ε. Sendo t_f o tempo de cálculo médio da função f, o tempo de cálculo total, ao fim de n iterações, será aproximadamente

$$T = n\,(t_f + \varepsilon) \approx n t_f\ .$$

desprezando o tempo de cálculo para as operações elementares.

Observação 6 *(erros de arredondamento).*
Reparando nas fórmulas (2.5) e (2.8) verificamos que há subtracções que podem indiciar o aparecimento de erros de cancelamento subtractivo, já que, por exemplo em (2.5), a diferença entre os valores a_n e b_n será extremamente pequena, visto que ambos irão convergir para o mesmo valor, a raiz.

2.4.1 Método da falsa posição modificado

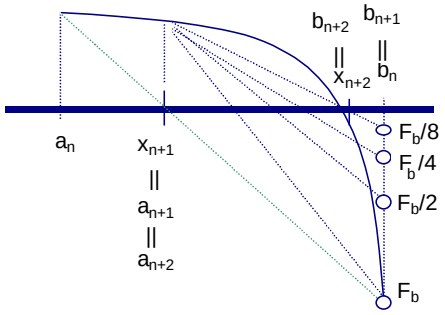
Trata-se de uma pequena modificação no método anterior de forma a evitar que um dos extremos se imobilize permanentemente. A ideia essencial é dividir o valor da função por 2 no extremo que se imobiliza.

Suponhamos que se imobilizava o extremo a_p, tendo-se $a_{p+1} = a_p$, neste caso consideramos $F_a = f(a_p)/2$ mantendo $F_b = f(b_p)$ e calculamos

$$\tilde{x}_{p+1} = b_p - F_b \frac{b_p - a_p}{F_b - F_a}, \tag{2.11}$$

o valor $\tilde{x}_{p+1}$ ainda pertence ao intervalo $[a_p, b_p]$ mas é menor que o valor x_{p+1} calculado usando (2.5) que neste caso seria maior que a raiz, ou seja $z < x_{p+1}$. Portanto, como $\tilde{x}_{p+1} < x_{p+1}$ é provável que $\tilde{x}_{p+1} < z$, passando nesse caso a ter-se $a_{p+1} = \tilde{x}_{p+1}$.

Se isso não acontecer, podemos ainda considerar um $\tilde{F}_a = F_a/2$ que levaria a um $\widetilde{\tilde{x}}_{p+1}$ ainda mais pequeno... e assim sucessivamente até que obtivessemos um valor $\widetilde{\tilde{x}}_{p+1} < z$. Desta forma consegue-se normalmente uma maior rapidez de convergência.



Apresentamos o algoritmo para o *método da falsa posição modificado:*

Inicialização : $[a_0, b_0] = [a, b]$; $F_a = f(a_0), F_b = f(b_0)$	
Repetir :	**1)** $x_{n+1} = b_n - F_b \frac{b_n - a_n}{F_b - F_a}$
	2) Se $f(x_{n+1})f(a_n) < 0$
	Então $a_{n+1} = a_n; b_{n+1} = x_{n+1};$
	$F_b = f(x_{n+1})$; **Se** $f(x_{n+1})f(x_n) > 0$ **Então** $F_a = F_a/2$
	Senão $a_{n+1} = x_{n+1}; b_{n+1} = b_n;$
	$F_a = f(x_{n+1})$; **Se** $f(x_{n+1})f(x_n) > 0$ **Então** $F_b = F_b/2$
Até que :	$f(x_{n+1}) = 0$ **ou** $\lvert x_{n+1} - x_n \rvert < \varepsilon$

Observação:

Não iremos apresentar estimativas de erro específicas para este método, mas poderemos sempre recorrer ao teorema que apresentámos inicialmente e que nos dá estimativas de erro a posteriori elementares, baseadas no teorema do valor médio.

Quanto ao tempo de cálculo, será semelhante ao tempo do método da falsa posição, já que a divisão por 2 não é significativa, compensando em termos de eficácia, já que irá evitar que um dos extremos fique constante, acelerando a convergência.

2.5 Método do ponto fixo num intervalo limitado

Vamos introduzir um método fundamental, que surpreende pela sua simplicidade. Iremos ver, no capítulo seguinte, que esta simplicidade permite que o método seja aplicado em contextos muito mais gerais, revelando-se como um precioso resultado de existência (e que é construtivo) na matemática.

A ideia base consiste na noção de ponto fixo. Dizemos que z é *ponto fixo* de uma função g se

$$g(z) = z.$$

O nome está associado ao facto de que o ponto z não é alterado pela função g.

Podemos transformar qualquer equação $f(x) = 0$ numa equação $x = g(x)$, estabelecendo a equivalência

$$f(x) = 0 \Leftrightarrow x = g(x), \tag{2.12}$$

num certo domínio D. É claro que se $z \in D$ for zero da função f, então será ponto fixo de g e vice-versa!

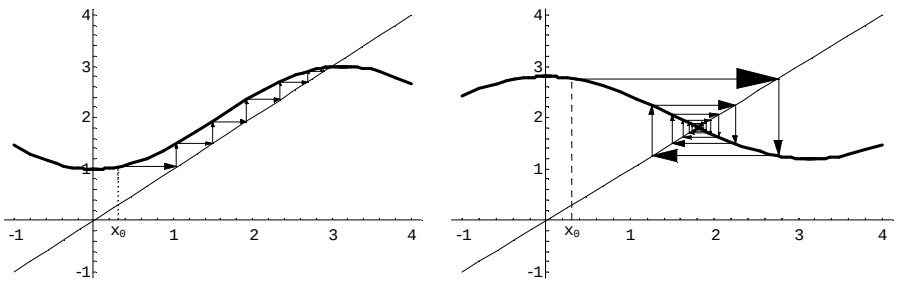
Há infinitas possibilidades para escolher g (que é denominada *função iteradora*) de forma a que a equivalência se verifique... por exemplo, basta pensar que se $\omega \neq 0$ temos $f(x) = 0 \Leftrightarrow x = x + \omega f(x)$. Como iremos ver, algumas escolhas de g serão menos apropriadas que outras, para o objectivo em vista.

Assim, o *método do ponto fixo* consiste simplesmente em[11]

$$\begin{cases} \text{Escolher uma iterada inicial : } x_0 \\ \text{Iterar } x_{n+1} = g(x_n). \end{cases}$$

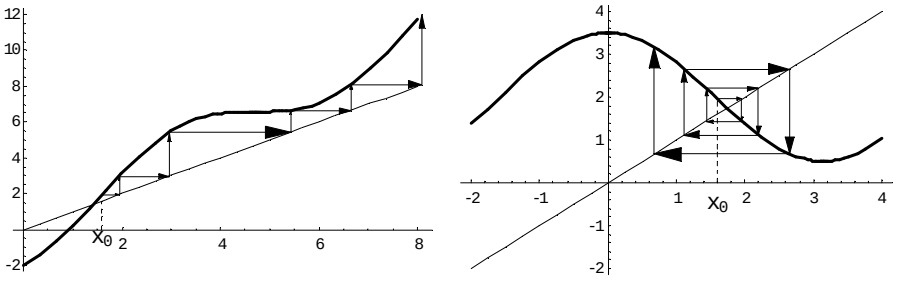
Considerando g uma função contínua, se o método convergir, converge para um ponto fixo z que, pela equivalência estabelecida, será uma raiz da equação, ou seja $f(z) = 0$.

Geometricamente, um ponto fixo corresponde a um ponto de intersecção do gráfico de g com a recta bissectriz, e a iteração do ponto fixo pode ser vista como uma trajectória (órbita). Várias situações podem ocorrer (ver figura em baixo). Quando a derivada da função g (numa vizinhança do ponto fixo) é em módulo inferior a 1 reparamos que há convergência (*monótona* se $0 < g' < 1$, *alternada* se $-1 < g' < 0$) e caso a derivada seja em módulo superior a 1 não se observa convergência para o ponto fixo.



[11] Também é designado *método das aproximações sucessivas.* Noutras situações, esta iteração é designada *iteração de Picard.*

Método do ponto fixo aplicado a funções $g(x) = 2 + \beta\cos(x)$. Nestes dois gráficos a única variação é feita no parâmetro β, em ambos os casos $x_0 = 0.3$. No primeiro gráfico consideramos $\beta = -1$ e no segundo $\beta = 0.8$. Podemos observar facilmente que há convergência para o ponto fixo[12] *nas duas situações, com uma única diferença. No primeiro caso, as iteradas aproximam-se do ponto fixo sempre pelo mesmo lado (convergência monótona, $0 < g' < 1$) e no segundo caso as iteradas aparecem alternadamente à esquerda e à direita do ponto fixo (convergência alternada, $-1 < g' < 0$).*



Método do ponto fixo aplicado às funções $g_1(x) = \frac{1}{2}(-1 + 3x - 3\sin(x))$, representada no primeiro gráfico, e $g_2(x) = 2 + \frac{3}{2}\cos(x)$, representada no segundo. Para ambas as funções começamos com $x_0 = 1.6$, que é um valor próximo do ponto fixo de cada uma delas. No entanto, as iteradas vão afastar-se rapidamente (no segundo caso alternando à esquerda e à direita do ponto fixo), o que reflecte a situação de não convergência. Temos $g_1'(z) > 1$, no primeiro caso, e $g_2'(z) < -1$, no segundo, o que são situações em que não há convergência do método do ponto fixo para esse z. Como curiosidade, notamos que a função g_1 não é limitada, e a sucessão das iteradas irá tender para $+\infty$, no entanto, a função g_2 fica limitada no intervalo $[\frac{1}{2}, \frac{7}{2}]$, e a sucessão de iteradas irá tomar sempre valores nesse intervalo. Sendo uma sucessão limitada, poderá ter sublimites (o que corresponde a situações de 'órbitas periódicas') ou não (o que corresponde a uma situação de 'caos').

Nos casos em que a derivada é em módulo inferior a 1, temos (pelo teorema do valor médio) $|g(x) - g(y)| < |x - y|$, o que significa que a distância entre dois pontos x e y é 'encurtada' pela transformação g, levando à noção de contractividade. É intuitivo que, neste caso de contractividade, a aplicação sucessiva de g, ao 'encurtar' as distâncias, origina a convergência para um ponto – o ponto fixo. No caso oposto, em que a derivada é em módulo superior a 1, verifica-se $|g(x) - g(y)| > |x - y|$, ou seja, as distâncias aumentam, o que provoca a divergência. De seguida provaremos estes resultados.

Definição 2.3 *Uma função g diz-se* Lipschitziana *em $[a, b]$ se existir $L \geq 0$:*

$$|g(y) - g(x)| \leq L|x - y|, \forall x, y \in [a, b]$$

Se $L < 1$ a função g diz-se *contractiva* em $[a, b]$.

Exercício 2.1 *Mostre que uma função Lipschitziana num intervalo, é contínua nesse intervalo.*

[12] *Que é designado um atractor, segundo a terminologia de sistemas dinâmicos.*

Proposição 2.2 *Se $g \in C^1([a,b])$ e temos $|g'(x)| \leq L < 1$, para qualquer x em $[a,b]$, então a função g é contractiva nesse intervalo.*

Demonstração:

Usando o teorema do valor médio de Lagrange, sabemos que, para quaisquer x, y em $[a,b]$

$$|g(y) - g(x)| \leq |g'(\xi)||x - y|,$$

para um certo $\xi \in]x;y[\subset [a,b]$ e podemos concluir, aplicando a hipótese.■

Teorema 2.6 *(Teorema do ponto fixo - caso de um intervalo fechado e limitado).*
Se g é uma função contractiva em $[a,b]$, e $g([a,b]) \subset [a,b]$, então:
i) g tem um e um só ponto fixo z em $[a,b]$.
ii) A sucessão $x_{n+1} = g(x_n)$ converge para esse ponto fixo z, dado qualquer x_0 em $[a,b]$.
iii) Verificam-se as estimativas de erro a priori

$$|z - x_n| \leq L^n |z - x_0| \tag{2.13}$$

$$|z - x_n| \leq L^n/(1-L)|x_1 - x_0| \tag{2.14}$$

e a estimativa de erro a posteriori

$$|z - x_n| \leq 1/(1-L)|x_{n+1} - x_n|. \tag{2.15}$$

Demonstração:

Existência (de ponto fixo). Consideramos uma função auxiliar $h(x) = g(x) - x$, contínua.

Como $g([a,b]) \subset [a,b]$ temos $g(a) \geq a$, $g(b) \leq b$, e assim $h(a)h(b) \leq 0$, logo pelo Teorema do valor intermédio, existe um $z \in [a,b] : h(z) = 0$, logo $g(z) = z$.

Unicidade (do ponto fixo). Supondo que g é contractiva e z e w são pontos fixos de g em $[a,b]$, temos:

$$|z - w| = |g(z) - g(w)| \leq L|z - w|,$$

logo $(1-L)|z-w| \leq 0$ e como $L < 1$ podemos concluir que $|z - w| = 0$, ou seja $z = w$.

Convergência do método.

É fácil ver (por indução) que se x_0 pertence a $[a,b]$, qualquer x_n também pertence.

Basta reparar que $x_{n+1} = g(x_n)$ e que $g([a,b]) \subset [a,b]$.

Por outro lado, temos

$$|z - x_{n+1}| = |g(z) - g(x_n)| \leq L|z - x_n|,$$

logo

$$|z - x_{n+1}| \leq L^{n+1}|z - x_0| \to 0,$$

pois $L < 1$.

Estimativas. A obtenção da estimativa *a posteriori* (2.15) resulta de considerar a desigualdade triangular (somando e subtraindo x_{n+1})

$$|z - x_n| \leq |z - x_{n+1}| + |x_{n+1} - x_n| \leq L|z - x_n| + |x_{n+1} - x_n|$$
$$\Leftrightarrow (1 - L)|z - x_n| \leq |x_{n+1} - x_n|.$$

Finalmente, a estimativa (2.14) resulta de aplicar (2.15) com $n = 0$ e considerar (2.13). ■

Observação: Veremos, no capítulo seguinte, uma demonstração ligeiramente diferente deste teorema, num contexto mais geral, e que permite estabelecer este resultado mesmo para intervalos *ilimitados.*□

Corolário 2.2 *Seja g uma função $C^1([a,b])$, tal que $g([a,b]) \subseteq [a,b]$ e*

$$L = \max_{x \in [a,b]} |g'(x)| < 1$$

então as condições do teorema do ponto fixo anterior são verificadas e portanto temos:
i) g tem um e um só ponto fixo z em $[a,b]$.
ii) A sucessão $x_{n+1} = g(x_n)$ converge para esse ponto fixo z, dado qualquer x_0 em $[a,b]$.
iii) Verificam-se as majorações de erro apresentadas no Teorema.

Demonstração:
É uma consequência imediata dos resultados anteriores. Reparamos apenas que, como se trata de um intervalo compacto, a derivada atinge um máximo, inferior a 1, o que poderia não acontecer se o intervalo fosse ilimitado. ■

Podemos dividir as condições do corolário anterior em dois casos.

Proposição 2.3 *Seja $g \in C^1[a,b]$ tal que $g([a,b]) \subseteq [a,b]$.*
• *Seja $0 < g'(x) < 1, \forall x \in [a,b]$. Se $x_0 \in [a,b]$ há convergência* monótona *do método do ponto fixo (ou seja, as iteradas ficam todas à esquerda [respect. à direita] da raiz, se a iterada inicial estiver à esquerda [respect. à direita] da raiz).*
• *Seja $-1 < g'(x) < 0, \forall x \in [a,b]$. Se $x_0 \in [a,b]$ há convergência* alternada *do método do ponto fixo (ou seja, as iteradas vão ficar alternadamente à esquerda e à direita da raiz).*

Demonstração:
As condições do corolário são imediatamente verificadas em ambos os casos, logo podemos concluir a convergência para um único ponto fixo $z \in [a,b]$.

No primeiro caso, se $0 < g'(x) < 1$, basta reparar que, se $x_n < z$ obtemos $x_{n+1} < z$, porque

$$z - x_{n+1} = g(z) - g(x_n) = g'(\xi)(z - x_n) > 0.$$

Portanto, se $x_0 < z$, por indução, vemos que temos sempre $x_n < z$ e concluímos que a convergência é monótona.

Usando o mesmo argumento, se $-1 < g'(x) < 0$, partindo de $x_n < z$ obtemos $x_{n+1} > z$, ou mais geralmente,

$$\text{sinal}(z - x_{n+1}) = \text{sinal}\,(g'(\xi)(z - x_n)) = -\,\text{sinal}(z - x_n),$$

pelo que se pode concluir que a convergência é alternada. ■

Observação: No caso em que a convergência é alternada, duas iteradas sucessivas definem um intervalo onde se encontra a raiz. □

Exemplo 2.5 *Vejamos que, para* $a > b \geq 1$, *a sucessão*

$$x_0 = 1;\ x_{n+1} = a + \frac{b}{x_n}$$

converge alternadamente para a solução da equação $x^2 - ax - b = 0$ *que se encontra no intervalo* $[a, a+b]$.

Repare-se que esta sucessão define aquilo que se designa por 'fracção contínua'[13], *ou seja,*

$$x = a_0 + \cfrac{b_0}{a_1 + \cfrac{b_1}{a_2 + \cfrac{b_2}{\ddots}}};\ \textit{neste caso } a_n = a, b_n = b,\ \textit{ou seja, } x = a + \cfrac{b}{a + \cfrac{b}{a + \cfrac{b}{\ddots}}}$$

Podemos aplicar a proposição anterior. Basta ver que $g(x) = a + b/x$ *verifica as condições requeridas no intervalo* $I = [a, a+b]$.

Com efeito, $-1 < g'(x) < 0, \forall x \in I$, *porque* $g'(x) = -b/x^2$ *é crescente em* I *e temos* $g'(a) = -b/a^2$, $g'(a+b) = -b/(a+b)^2$ *onde ambos os valores pertencem a* $]-1, 0[$ *pois* $a > b \geq 1$. *Por outro lado, como* g *é decrescente em* I *e temos* $g(a) = a + b/a$, $g(a+b) = a + \frac{b}{a+b}$, *ambos os valores pertencendo a* $[a, a+b]$, *é claro que* $g(I) \subseteq I$. *Para terminar reparamos que, apesar de* $x_0 = 1 \notin I$, *temos* $x_1 = a + b \in I$, *pelo que podemos considerar* x_1 *como sendo a 'nova iterada inicial'. As condições estão verificadas.*
Acabamos de mostrar que, quando $a > b \geq 1$, *a fracção contínua indicada converge para* $\frac{a}{2} + \frac{1}{2}\sqrt{a^2 + 4b} \in [a, a+b]$. *Por exemplo, considerando* $a = 2, b = 1$, *podemos obter*

$$1 + \sqrt{2} = 2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{\ddots}}}\ .$$

Verificar as condições do teorema do ponto fixo num intervalo $[a, b]$ permite garantir a convergência para qualquer iterada inicial nesse intervalo. No entanto, podemos ver que se houver um conhecimento aproximado do ponto fixo z, a ponto de começarmos com uma iterada inicial 'suficientemente próxima', basta saber que $|g'(z)| < 1$, para ficarem asseguradas as condições de convergência.

[13]Tal como as séries, as fracções contínuas foram objecto de estudo, especialmente no séc. XIX. Uma propriedade curiosa das fracções contínuas é que podem originar uma sucessão que pode ser encarada como a 'melhor' aproximação racional para o irracional que seja o seu limite (deve considerar-se nesse caso um desenvolvimento com $b_n \equiv 1$).

Teorema 2.7 *(convergência local). Seja $g \in C^1(V_z)$, em que V_z é uma vizinhança de z ponto fixo de g, tal que $|g'(z)| < 1$. A sucessão $x_{n+1} = g(x_n)$ converge para z, desde que seja considerado um x_0 suficientemente próximo de z.*

Demonstração:
Podemos sempre considerar um intervalo $I = [z - \varepsilon, z + \varepsilon]$ suficientemente pequeno, tal que $|g'(x)| < 1$, $\forall x \in I$. Falta apenas ver que $g(I) \subseteq I$ para aplicar o teorema do ponto fixo. Ora, se $x \in I$, temos $|z - x| \leq \varepsilon$, e portanto

$$|z - g(x)| = |g(z) - g(x)| = |g'(\xi)|\,|z - x| \leq \varepsilon.$$

Logo $g(x) \in I$, e a convergência é assegurada pelo teorema do ponto fixo, considerando um $x_0 \in I$. ■

Proposição 2.4 *Se $0 < |g'(z)| < 1$, a convergência do método do ponto fixo é* linear *e o coeficiente assimptótico de convergência é $|g'(z)|$.*

Demonstração:
Usando o teorema do valor médio de Lagrange, sabemos que

$$z - x_{m+1} = g(z) - g(x_m) = g'(\xi_m)(z - x_m)$$

com $\xi_m \in]x_m; z[$.
Logo, dividindo e passando ao limite, temos:

$$\lim_{m\to\infty} \frac{e_{m+1}}{e_m} = \lim_{m\to\infty} \frac{z - x_{m+1}}{z - x_m} = g'(z) \Rightarrow K_\infty = |g'(z)|$$

pois ξ_m tende para z, porque havendo convergência $x_m \to z$. Como por hipótese $g'(z) \neq 0$, o teorema está provado. ■

Acabámos de ver que na situação em que a derivada é menor que 1 (em módulo) há pelo menos convergência linear. A condição $g'(z) \neq 0$, apenas foi usada para afirmar a convergência linear. Com efeito, se $g'(z) = 0$, conclui-se que a convergência é supralinear. Podemos ver mais especificamente qual a ordem p para essa convergência, admitindo alguma regularidade na função g, de forma a poder usar um resto de Lagrange conveniente para o desenvolvimento em série de Taylor.

Teorema 2.8 *(convergência supralinear). Seja g uma função $C^p(V_z)$, com $p \geq 2$, onde V_z é uma vizinhança de z ponto fixo de g.*
Se

$$g'(z) = ... = g^{(p-1)}(z) = 0, \text{ com } g^{(p)}(z) \neq 0 \tag{2.16}$$

então

$$\lim_{m\to\infty} \frac{e_{m+1}}{e_m^p} = -\frac{(-1)^p}{p!} g^{(p)}(z) \tag{2.17}$$

ou seja, o método do ponto fixo tem convergência de ordem p, e o coeficiente assimptótico é

$$K_\infty^{[p]} = \frac{1}{p!}|g^{(p)}(z)|.$$

Demonstração:
Fazendo o desenvolvimento em série de Taylor, para um $x_m \in I$

$$g(x_m) = g(z) + g'(z)(x_m - z) + \ldots + \frac{g^{(p-1)}(z)}{(p-1)!}(x_m - z)^{p-1} + \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p,$$

com $\xi_m \in]x_m; z[$. Usando as hipóteses, temos

$$x_{m+1} = g(x_m) = z + \frac{g^{(p)}(\xi_m)}{p!}(x_m - z)^p,$$

ou seja,

$$e_{m+1} = -(-1)^p \frac{g^{(p)}(\xi_m)}{p!}(e_m)^p.$$

Concluímos, reparando (como fizemos na proposição anterior) que $\xi_m \to z$, logo $\lim \frac{|e_{m+1}|}{|e_m^p|} = \frac{1}{p!}|g^{(p)}(z)| = K_\infty > 0$. ■

Já vimos casos em que podemos assegurar convergência do método do ponto fixo, vamos agora estabelecer um critério em que se conclui que não pode haver convergência para um determinado ponto fixo z em que se verifique $|g'(z)| > 1$. Note-se que, no entanto, poderá haver convergência para um outro ponto fixo.

Teorema 2.9 *(não convergência para um ponto fixo). Seja V_z vizinhança de um ponto fixo z de $g \in C^1(V_z)$, tal que $|g'(z)| > 1$. Neste caso, a sucessão $x_{n+1} = g(x_n)$ não pode convergir para esse ponto fixo z (excepto se 'excepcionalmente' $x_m = z$ para algum m).*

Demonstração:
Supondo, por absurdo, que (x_n) converge para o ponto fixo $z \in V_z$, então:

$$\forall \varepsilon > 0 \; \exists p : n \geq p \; : \; |z - x_n| < \varepsilon.$$

Como $|g'(z)| > 1$, podemos sempre considerar um $\varepsilon > 0$ suficientemente pequeno tal que $I = [z - \varepsilon, z + \varepsilon] \subset V_z$ em que $|g'(x)| > 1, \forall x \in I$. Aplicando o teorema do valor médio

$$|z - x_{n+1}| = |g'(\xi_n)| \, |z - x_n|,$$

como $x_n, z \in I$ também $\xi_n \in I$, logo $|g'(\xi_n)| > 1$ e temos $|z - x_{n+1}| > |z - x_n|$ para $n \geq p$, pois $z \neq x_n$. Isto significa que a sucessão $|z - x_n|$ é crescente e consequentemente não converge para zero, o que provoca a contradição.■

Exercício 2.2 *Suponha que z e w são pontos fixos consecutivos de $g \in C^1$. Mostre que se $|g'(z)| < 1$ então $|g'(w)| \geq 1$.*
Resolução: *Suponhamos que $z < w$, e portanto não há mais nenhum ponto fixo em $]z, w[$. Consideramos $f(x) = x - g(x)$, e temos $f \in C^1$ e $f(z) = f(w) = 0$.*
Como não existe mais nenhuma raiz de f entre z e w, concluimos que, ou $f(x) > 0$ ou $f(x) < 0$, em $]z, w[$. Como, $f'(z) = 1 - g'(z) > 0$, e como $f'(z) = \lim_{x \to z^+} \frac{f(x)}{x - z}$ (porque f tem derivada contínua e $f(z) = 0$), concluímos que $f(x) > 0$ num intervalo $]z, z + \varepsilon[$. Portanto $f(x) > 0$ em $]z, w[$, logo

$$f'(w) = \lim_{x \to w^-} \frac{f(x)}{x - w} \leq 0$$

pois $\frac{h(x)}{x-w} < 0$ quando $x < w$. Assim, $g'(w) = 1 - f'(w) \geq 1$. O caso $w < z$ seria semelhante.

O resultado deste exercício pode ser melhorado para concluir que – se encontrarmos um ponto fixo usando uma função iteradora g, contínua, será impossível encontrar os pontos fixos que lhe são adjacentes com a mesma função!!

Isto poderá parecer que é uma grande limitação para o método do ponto fixo, já que à partida poderíamos ficar circunscritos a encontrar menos de metade das soluções... no entanto, há soluções que foram encontradas sem que o próprio problema tivesse ocorrido a quem as encontrou! Uma que iremos ver de seguida, é o método de Newton. A função iteradora do método de Newton resolve este problema de forma drástica... a função iteradora irá explodir num ponto entre as raízes, quebrando a continuidade de g.

Observação 1 *(tempo de cálculo).*

O método do ponto fixo envolve apenas o cálculo de g em cada iterada, pelo que podemos escrever

$$T = n\, t_g.$$

É claro que como podemos estabelecer várias funções iteradoras, o tempo de cálculo t_g irá variar com a maior ou menor complexidade de g.

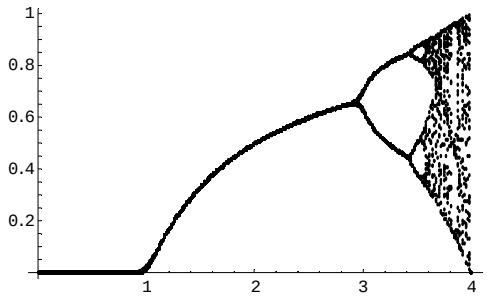
Observação 2 *(erros de arredondamento).*

Ao garantir a convergência do método do ponto fixo num intervalo, é óbvio que a influência dos erros de arredondamento no processo de cálculo é mínima. Basta reparar que se numa determinada iterada n utilizarmos $\tilde{x}_n$ ao invés de x_n, esse valor será próximo e continua dentro do intervalo de convergência, não havendo qualquer problema. O limite deste processo é o limite imposto pela precisão do sistema FP utilizado *(ver observação 2, no final do primeiro capítulo).*

Um critério normalmente utilizado para avaliar o final da iteração é aquele em que os dígitos apresentados não se alteram, ou seja $\tilde{x}_{n+1} = \tilde{x}_n$. Este critério não é exacto e pode prestar-se a confusões. Ao atingir a precisão da máquina, os últimos dígitos não são significativos e podem mesmo ocorrer pequenas variações, periódicas. Visualizar $\tilde{x}_{n+1} = \tilde{x}_n$ não significa, de modo algum, que $x_{n+1} = x_n$, o que significaria erro nulo! Normalmente, nesses casos, o erro é da ordem de grandeza da precisão do sistema FP, por exemplo, será da ordem 10^{-8} no caso de precisão simples e 10^{-16} no caso de precisão dupla.

Observação 3 *(órbitas periódicas).*

Ao aplicar o método do ponto fixo uma das situações de não convergência que pode ocorrer é o caso em que se obtém $x_{n+2} \approx x_n, x_{n+1} \neq x_n$. Nesses casos a sucessão poderá ter dois sublimites z_1 e z_2, e alterna entre valores que irão ficar próximo de z_1 e outros que irão ficar próximo de z_2. Trata-se de uma situação em que se fala de órbitas de período 2. Essa convergência está associada a uma convergência do método do ponto fixo aplicado não à função g mas sim à função $g \circ g$. É claro que se z for ponto fixo de g também será ponto fixo de $g \circ g$, mas no mesmo domínio o método do ponto fixo poderá não convergir usando a função iteradora g e convergir ao usar a função iteradora $g \circ g$. É nestas situações que temos órbitas com período 2. Este raciocínio é facilmente estendido a órbitas de período $p > 2$. Um exemplo típico desta situação ocorre quando se considera a denominada função logística $g(x) = \mu x(1-x)$, variando o coeficiente μ. No gráfico seguinte mostramos o que acontece quando começamos com $x_0 = 0.5$ e variamos μ entre 0 e 4. Para cada valor de μ são colocados os vários pontos x_n para n grande, e para estes valores de μ, é fácil ver que $g([0,1]) \subseteq [0,1]$ Reparamos que há apenas dois pontos fixos $z_0 = 0$ e $z_\mu = \frac{\mu-1}{\mu}$, que só é positivo quando $\mu > 1$.



Portanto, para $\mu \leq 1$, as iteradas vão convergir para $z_0 = 0$, e depois aproximam-se de z_μ desenhando uma parte da hipérbole $\frac{\mu-1}{\mu}$, até que há um primeiro ponto de bifurcação. Repare-se que $g'(z_\mu) = 2 - \mu$ e portanto para $\mu \in]1,3[$, garante-se a convergência local do método do ponto fixo para z_μ. A partir daí entramos na situação em que devemos estudar $g \circ g$, que tem quatro pontos fixos. Dois deles são 0 e $\frac{\mu-1}{\mu}$, pontos onde a derivada da função é maior que um em módulo, e os outros dois novos valores são z_μ^-, z_μ^+ onde a derivada verifica $(g \circ g)'(z_\mu^\pm) = 4 + 2\mu - \mu^2$ e será em módulo inferior a 1 para $\mu \in]3, 1 + \sqrt{6} (= 3.4495)[$. O extremo do intervalo é justamente o valor em que deixa de haver convergência para $g \circ g$ e o gráfico apresenta duas novas bifurcações. Sucessivamente precisaríamos de estudar mais composições de g, mas é conhecido, pela teoria de sistemas dinâmicos discretos que o período irá duplicar, de forma que se irá verificar o seguinte número de bifurcações $2, 2^2, 2^3, ...$ Surpreendentemente, após bifurcações que são potências de 2, ocorrerá uma bifurcação de período 3... Antes de ocorrer uma bifurcação de período 5 irão ocorrer bifurcações cujo período será dado por potências de 3 multiplicadas por potências de 2. Assim, sucessivamente...após 3 e 5 será 7, 11, 13... percorrendo os números primos. Verifica-se uma nova ordenação para os números naturais! Quando estas hipóteses se esgotam, atinge-se uma situação limite comumente designada por *caos*.

2.5.1 Aceleração de convergência

Uma técnica frequentemente utilizada para aumentar a ordem de convergência de um método consiste em prever e utilizar o comportamento assimptótico da sucessão das iteradas. É o caso do método que veremos de seguida.

Foi estabelecido que se o método do ponto fixo convergir linearmente para um ponto fixo z da função $g \in C^1(V_z)$, então $K_\infty = |g'(z)|$, ou melhor,

$$\lim_{n\to\infty} \frac{z - x_{n+1}}{z - x_n} = g'(z).$$

É igualmente fácil estabelecer que (exercício)

$$\lim_{n\to\infty} \frac{x_{n+1} - x_n}{x_n - x_{n-1}} = g'(z).$$

Ora, para n suficientemente grande poderemos então escrever

$$\frac{x_{n+1} - x_n}{x_n - x_{n-1}} \approx g'(z) \approx \frac{z - x_{n+1}}{z - x_n}.$$

Será que substituir o símbolo $\approx$ por $=$ permite obter uma estimativa melhor para z do que calcular x_{n+2}? Iremos ver que sim. A ideia será prever o valor $\tilde{z}$ que verificaria a relação

$$\frac{x_{n+1} - x_n}{x_n - x_{n-1}} = \frac{\tilde{z} - x_{n+1}}{\tilde{z} - x_n},$$

e a partir daí utilizar esse novo valor como iterada. Vamos primeiro resolver a equação anterior em $\tilde{z}$, e obtemos

$$\frac{\tilde{z} - x_n}{x_n - x_{n-1}} = \frac{\tilde{z} - x_{n+1}}{x_{n+1} - x_n} \Leftrightarrow \tilde{z} = \frac{x_{n-1}x_{n+1} - x_n^2}{2x_n - x_{n-1} - x_{n+1}},$$

ou o que é equivalente,

$$\tilde{z} = x_{n-1} - \frac{(x_n - x_{n-1})^2}{(x_{n+1} - 2x_n + x_{n-1})} \tag{2.18}$$

que é a *fórmula de extrapolação de Aitken*[14].

Reparando que $x_n = g(x_{n-1}), x_{n+1} = g(g(x_n))$, a fórmula de Aitken pode ser reescrita de forma a constituir um método iterativo,

$$x_n = x_{n-1} - \frac{(g(x_{n-1}) - x_{n-1})^2}{g(g(x_{n-1})) - 2g(x_{n-1}) + x_{n-1}}, \tag{2.19}$$

[14] *A fórmula (2.18) também pode ser escrita na forma*

$$\tilde{z} = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n}$$

usando a notação de diferenças finitas, $\Delta x_n = x_{n+1} - x_n$, por essa razão é usada a designação método-Δ^2 de Aitken.

Note-se que a fórmula é aplicável a outras sucessões (x_n) com convergência linear, não necessariamente resultantes do método do ponto fixo.

que também é denominado *método de Steffensen* (mas iremos usar preferencialmente essa designação quando aplicado ao cálculo de raízes, com $g(x) = f(x) + x$).

Observação 1 *(convergência quadrática do método de Steffensen).*

Podemos supor que g é uma função iteradora qualquer, com ponto fixo z, e tal que $g'(z) \neq 1$. Ainda que o método do ponto fixo aplicado a g não convirja, consideramos a nova função iteradora obtida G pela fórmula de extrapolação de Aitken,

$$G(x) = x - \frac{(g(x) - x)^2}{g(g(x)) - 2g(x) + x} = x - \frac{A(x)^2}{B(x)},$$

em que $A(x) = g(x) - x$, $B(x) = g(g(x)) - 2g(x) + x$. Esta nova função iteradora G tem os mesmos pontos fixos que g, e tem a particularidade de convergir quadraticamente para z, desde que $g'(z) \neq 1$.

(i) Começamos por ver que $g(z) = z \Leftrightarrow G(z) = z$.

É claro que se $G(z) = z$, então $(g(z) - z)^2 = 0$ e portanto $z = g(z)$.

Supondo agora que $g(z) = z$, notamos que $A(z) = B(z) = 0$, e assim, quando $x \to z$, temos uma indeterminação $\frac{0}{0}$. Mas podemos mostrar que $\lim_{x\to z} G(x) = z$. Com efeito, como $A'(x) = g'(x) - 1$, e $B'(x) = g'(x)g'(g(x)) - 2g'(x) + 1$, podemos calcular o limite

$$\lim_{x\to z} \frac{A(x)}{B(x)} = \frac{A'(z)}{B'(z)} = \frac{g'(z) - 1}{g'(z)^2 - 2g'(z) + 1} = \frac{1}{g'(z) - 1},$$

que é finito, porque assumimos $g'(z) \neq 1$. Portanto,

$$\lim_{x\to z} G(x) = z - A(z)\frac{1}{g'(z) - 1} = z.$$

(ii) Resta ver que $G'(z) = 0$. Como

$$G'(x) = 1 - 2A'(x)\frac{A(x)}{B(x)} + B'(x)\frac{A(x)^2}{B(x)^2},$$

obtemos

$$\lim_{x\to z} G'(x) = 1 - 2\frac{g'(z) - 1}{g'(z) - 1} + (g'(z) - 1)^2 \left(\frac{1}{g'(z) - 1}\right)^2 = 0.$$

Podemos assim concluir que se $g \in C^1(V_z)$, o método aplicado à função G converge pelo menos quadraticamente para z.

Mesmo que o método inicial com g divirja, com $|g'(z)| > 1$, o método de Steffenson irá convergir quadraticamente (convergência local).

Observação 2: Ver exercício 4.

2.6 Método de Newton

O método de Newton pode ser encarado como um caso particular do método do ponto fixo, onde é possível obter uma convergência quadrática. Basta reparar que se $f'(x) \neq 0$ então

$$f(x) = 0 \Leftrightarrow x = x - f(x)/f'(x)$$

definindo a função iteradora $g(x) = x - f(x)/f'(x)$, os pontos fixos de g serão os zeros de f.

Para além disso, podemos ver que

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}, \tag{2.20}$$

ora como $f(z) = 0$ então $g'(z) = 0$. Pelo teorema relativo à convergência supralinear, usando esta função iteradora g é possível arranjar uma vizinhança da raiz onde asseguramos, pelo menos, uma convergência quadrática (desde que $f'(z) \neq 0$).

O *método de Newton* pode assim resumir-se no esquema

$$\begin{cases} \text{Iterada inicial} : x_0 \\ \text{Iterar } x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \end{cases} \tag{2.21}$$

Historicamente, a origem do método de Newton não é esta, é geométrica, e consiste em definir a nova iterada a partir da intersecção do eixo das abcissas com a tangente à função f (calculada na iterada anterior). Basta reparar que a equação da tangente num ponto x_n é

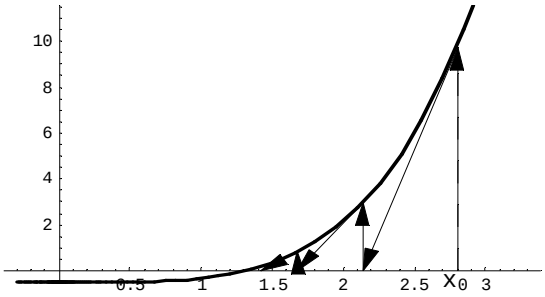
$$y = f(x_n) + f'(x_n)(x - x_n)$$

e a iterada x_{n+1} é a "raiz da tangente"... basta fazer $y = 0$ para verificarmos que o valor de x coincide com o valor obtido para x_{n+1}.

É também claro, mesmo geometricamente, que não podemos ter iteradas em que $f'(x_n) = 0$, pois ficaríamos com tangentes paralelas ao eixo das abcissas, que nunca o intersectariam *(... na geometria euclidiana!)*

Exemplo 2.6 *Apresentamos na figura seguinte um esquema que representa a evolução das iteradas numa situação em que há convergência do método de Newton. A função escolhida foi* $f(x) = \frac{1}{6}x^4 - \frac{1}{2}$, *começando com a iterada inicial* $x_0 = 2.8$. *Este exemplo foi escolhido de forma a que a derivada da função fosse quase nula próxima da raiz, para poder visualizar algumas iteradas do método, já que noutros casos a convergência é tão rápida que não*

permite uma visualização adequada.



Observação *(g é descontínua entre raízes)*

A aparente limitação para o método do ponto fixo, ao considerar funções iteradoras g contínuas, referida na secção anterior, é resolvida de forma curiosa pelo método de Newton. Como $g(x) = x - \frac{f(x)}{f'(x)}$, a função 'explode' sempre que $f'(x) = 0$, e isso acontece sempre entre duas raízes, devido ao teorema de Rolle. Esta característica permite ao método de Newton aproximar raízes consecutivas, o que não seria possível se g fosse contínua (pelo menos desta maneira... mais à frente, quando analisarmos um método de ordem 3 veremos outra solução engenhosa).

Para assegurar a convergência, poderíamos tentar aplicar as condições do teorema do ponto fixo à função $g(x) = x - f(x)/f'(x)$, mas na maioria dos casos isso levaria a longos cálculos, podendo ser estabelecido um critério mais simples.

2.6.1 Convergência do método de Newton

Teorema 2.10 *(condição* suficiente *de convergência para o método de Newton).*
Seja f uma função $C^2[a,b]$ que verifique :
1) $f(a)f(b) \leq 0$
2) f' *não se anula em* $[a,b]$,
e portanto a equação $f(x) = 0$ tem uma solução única z pertencente ao intervalo $[a,b]$.
Consideramos ainda duas outras condições
3) $f'' \geq 0$ *ou* $f'' \leq 0$ *em* $[a,b]$,
4) $f(x_0)f'' \geq 0$ *para* $x_0 \in [a,b]$.
então o método de Newton converge monotonamente para z.
Para além disso, se $|f(a)/f'(a)| \leq |a-b|$ e $|f(b)/f'(b)| \leq |a-b|$,
então, qualquer que seja $x_0 \in [a,b]$, a iterada x_1 verifica a condição 4).

Demonstração:

Consideremos o caso em que $f(a) < 0, f(b) > 0, f'(x) > 0, f''(x) \geq 0$, já que os restantes casos são semelhantes.

Se tivermos $f(x_0) > 0$ então as condições 1), 2), 3) e 4a) são verificadas e podemos provar que $x_n \geq z$ para todo n, porque

i) $x_0 > z$, pois $f(x_0) > 0$.

ii) fazendo o desenvolvimento em série de Taylor em torno de x_n, temos

$$0 = f(z) = f(x_n) + f'(x_n)(z - x_n) + \frac{1}{2}f''(\xi_n)(z - x_n)^2.$$

Como $f''(x) \geq 0$, vem $f(x_n) + f'(x_n)(z - x_n) \leq 0$ e dividindo por $f'(x_n) > 0$, obtém-se

$$z - x_{n+1} = \frac{f(x_n)}{f'(x_n)} + z - x_n \leq 0, \text{ ou seja } x_{n+1} \geq z.$$

Por outro lado, tendo $x_n \geq z$, como f é crescente $f(x_n) \geq f(z) = 0$, logo $\frac{f(x_n)}{f'(x_n)} > 0$ e portanto $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n$.

Mostramos assim que se trata de uma sucessão estritamente descrescente e limitada (porque $z \leq x_n \leq x_0$), logo tem limite.

Esse limite é obrigatoriamente a solução única de $f(z) = 0 \Leftrightarrow z = g(z) = z - \frac{f(z)}{f'(z)}$ nesse intervalo, pois g é contínua.

– Resta ver que se a última condição for verificada temos convergência em todo o intervalo. Reparamos que $g'(x) = \frac{f(x)f''(x)}{f'(x)^2} \leq 0$ se $x \in [a, z]$, logo g é decrescente e se $f(x_0) < 0$ (que é o caso que resta verificar) então $x_0 < z \Rightarrow x_1 = g(x_0) > g(z) = z$. Portanto x_1 verifica $f(x_1) > 0$ e aplica-se 4) desde que $x_1 \in [a, b]$. Como $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \geq a$, basta notar que $x_0 \geq a \Rightarrow x_1 = g(x_0) \leq g(a)$ e que $g(a) \leq b$ porque $g(a) = a - \frac{f(a)}{f'(a)} \leq a + b - a = b$, por hipótese. ■

Note-se que este critério apenas fornece condições suficientes de convergência. Para além disso, se tivermos informações acerca da localização da raiz, a proposição seguinte garante que se $f \in C^2$ numa vizinhança da raiz e se $f'(z) \neq 0$, então escolhendo a iterada inicial *suficientemente próxima* de z, o método de Newton converge com, pelo menos, ordem 2.

Proposição 2.5 *(convergência local). Seja f uma função $C^2(I)$, onde I é um intervalo que é vizinhança da raiz z.*
Se $f'(z) \neq 0$ então a sucessão definida pelo método de Newton converge para z (desde que x_0 seja suficientemente próximo de z) com pelo menos ordem de convergência quadrática,

$$\lim_{m\to\infty} \frac{e_{m+1}}{e_m^2} = -\frac{1}{2}\frac{f''(z)}{f'(z)}.$$

Se $f''(z) \neq 0$ a convergência será quadrática e o coeficiente assimptótico de convergência será $K_\infty^{[2]} = \frac{1}{2}\frac{|f''(z)|}{|f'(z)|}$. Se $f''(z) = 0$, o método terá convergência superior à quadrática.

Demonstração: Já vimos que $g'(z) = \frac{f(z)f''(z)}{(f'(z))^2} = 0$, e portanto as condições do teorema de convergência local para o método do ponto fixo estão estabelecidas (apenas exigiam $|g'(z)| < 1$). O teorema acerca da convergência supralinear do método do ponto fixo, mostra que a convergência será pelo menos quadrática e dá-nos o coeficiente assimptótico de convergência. No caso $f''(z) = 0$, temos $K_\infty^{[2]} = 0$, o que indicia a convergência de ordem superior, que é assegurada se $f \in C^3(I)$.■

• Como $g'(z) = \frac{f(z)f''(z)}{(f'(z))^2}$ e $f(z) = 0$, a situação de não convergência prevista no método do ponto fixo, $|g'(z)| > 1$, apenas se poderia verificar se $f'(z) = 0$. Mas como iremos ver, nem nesse caso isso acontece, podendo garantir-se convergência (linear).

– Localmente, o método de Newton irá sempre convergir!

No entanto, o facto de convergir quando estamos próximo da raiz, não significa que não haja situações de divergência, dependentes da iterada inicial.

Observação 1 *(não convergência, órbitas periódicas).*

Sendo a iteração do método de Newton um caso particular da iteração do ponto fixo ocorrem também situações semelhantes às vistas anteriormente, em que o método não converge.. nomeadamente, o caso de órbitas periódicas. Procurar situações de órbitas de período 2 em que $x_2 = x_0 \neq x_1$ pode aparentemente não ser fácil se encararmos $x_2 = g(g(x_0))$, já que a expressão $g \circ g$ complica-se.

Reparamos, entretanto, que o método de Newton efectua uma soma que pode ser escrita da seguinte forma:

$$x_{n+1} = x_0 - \sum_{k=0}^{n} \frac{f(x_k)}{f'(x_k)}, \text{ e no limite.} z = x_0 - \sum_{k=0}^{\infty} \frac{f(x_k)}{f'(x_k)}.$$

Esta fórmula não tem qualquer interesse prático, já que necessita do cálculo prévio dos valores $x_0, ..., x_n$. Porém, do ponto de vista teórico tem uma utilidade evidente. Avaliar se situações do tipo $x_2 = x_0$ podem ocorrer corresponde a verificar se podem existir x_0, x_1 diferentes e tais que:

$$x_2 = x_0 - \sum_{k=0}^{1} \frac{f(x_k)}{f'(x_k)} \Leftrightarrow \frac{f(x_0)}{f'(x_0)} + \frac{f(x_1)}{f'(x_1)} = 0$$

Se uma tal situação puder ocorrer existe a possibilidade do método cair numa órbita de período 2. Da mesma forma, para avaliar se pode ocorrer uma órbita de período p, bastará encontrar $x_0, x_1, ..., x_{p-1}$, diferentes, tais que

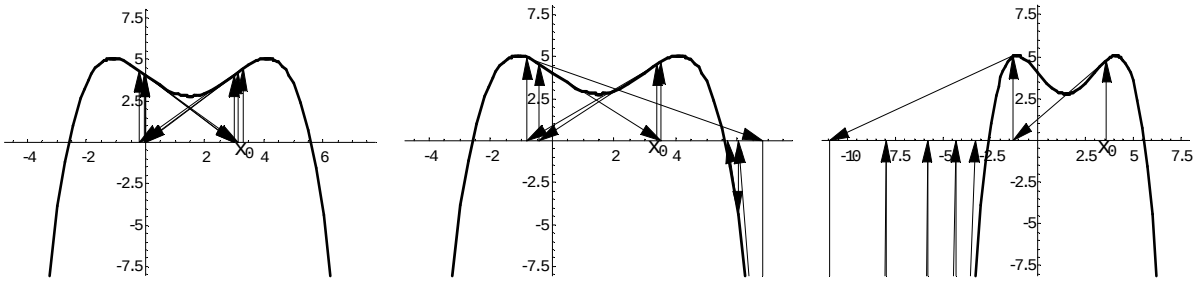
$$\sum_{k=0}^{p-1} \frac{f(x_k)}{f'(x_k)} = 0. \tag{2.22}$$

Exemplo 2.7 *Consideremos a função (que foi obtida de forma a verificar a condição (2.22) para $x_0 = 0, x_1 = 3$)*

$$f(x) = \frac{1}{20}(81 - 27x + 6x^3 - x^4).$$

Esta função tem duas raízes reais $z_1 = -2.6003, z_2 = 5.6003$. Vejamos o que se passa quando variamos ligeiramente a iterada inicial x_0. Consideramos três valores diferentes $x_0 = 3.3, x_0 = 3.4$ e $x_0 = 3.6$. No primeiro gráfico (em baixo, à esquerda) é visualizada a situação em que $x_0 = 3.3$. Nesse caso, as iteradas vão aproximar-se alternadamente de dois valores, originando uma situação de órbita de período 2. No entanto, verificamos que os valores não estão muito longe dos extremos relativos da função (pontos críticos). Assim, é natural que uma pequena variação da iterada inicial provoque grandes alterações, como

iremos ver.
No segundo gráfico (em baixo, ao centro), consideramos $x_0 = 3.4$, e após algumas iterações o valor x_n é enviado para uma zona em que é garantida a convergência do método de Newton para z_2. Da mesma forma, ao considerar $x_0 = 3.6$, na figura em baixo, à direita, ao fim de duas iterações, o valor x_2 está numa zona em que é garantida a convergência do método de Newton para z_1.



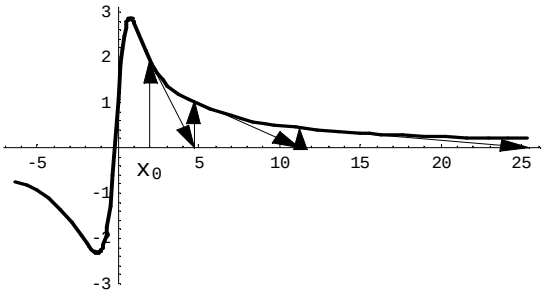
Observação 2 *(não convergência, assímptotas horizontais).*

A situação de órbitas periódicas não é a única situação em que não há convergência para nenhuma raiz. Um outro caso que isso ocorre é aquele em que podemos ter sucessões que tendem para infinito devido ao facto da função possuir uma assímptota horizontal. Com efeito, consideremos o caso em que $f'(x) \to 0^-$ quando $x \to +\infty$, e f é positiva. Se tivermos um x_n suficientemente grande, como f/f' é um valor negativo

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} > x_n,$$

e a sucessão tenderá para infinito. Note-se que como podemos encarar $f(+\infty) = 0$, o método mantém a sua coerência ao tentar aproximar a raiz $z = +\infty$, só que isso é entendido como uma divergência.

Um exemplo, apresentado na figura que se segue, é $f(x) = \frac{5x+\cos(x)}{x^2+1}$, começando com $x_0 = 2$.



É claro que, neste exemplo, usar a função f é despropositado, já que para resolver $f(x) = 0$, bastaria aplicar o método de Newton a $f(x) = 5x + \cos(x)$.

• Quando possível, é importante escolher funções f *equivalentes*, mais simples, ou mais adequadas...

Refira-se que se esta função fosse utilizada, sem uma análise do gráfico de f, dificilmente encontraríamos a raiz, já que o método de Newton não converge fora do intervalo $]-1, 0.5[$.
.

Observação 3 *(supressão de zeros).*

Tendo obtido uma das raízes da equação, podemos sempre procurar outras, através de um processo muito simples, designado *supressão de zeros*[15]. Suponhamos que usando f obtemos a aproximação de uma primeira raiz $\tilde{z}_1$, a ideia consiste em usar de seguida

$$f_1(x) = \frac{f(x)}{x - \tilde{z}_1}.$$

Note-se que se fosse possível considerar o valor correcto, z_1, a função f_1 seria ainda contínua em z_1, porque $\lim_{x\to z_1} \frac{f(x)}{x-z_1} = f'(z_1)$. Tendo uma aproximação $\tilde{z}_1$, a função mantém o zero em z_1 e terá uma assímptota vertical em $\tilde{z}_1$. No entanto, quando $\tilde{z}_1$ é uma aproximação próxima da precisão máxima FP, e como iremos procurar o outro zero noutra região, este comportamento da função f_1 próximo de z_1 será quase negligenciável. O que será menos negligenciável, é o comportamento quando $x \to \pm\infty$, já que podem ser introduzidas assímptotas horizontais, podendo introduzir situações de não convergência, como referidas na observação anterior.

Consoante $\tilde{z}_1$ seja uma aproximação por excesso ou defeito de z_1, deverá colocar-se o novo x_0 do lado em que não esteja z_1, para evitar que encontre a mesma raiz. Este processo poderá ser de novo aplicado para encontrar novas raízes.

Vejamos um exemplo. Consideramos a função $f(x) = x^3 - 3\sin(x) + 1$, que tem 3 raízes, situadas em $[-2, 2]$. Começando com $x_0 = -3$, obtemos $\tilde{z}_1 = -1.19603$. Mesmo usando a aproximação inicial anterior e $f_1(x) = \frac{x^3-3\sin(x)+1}{x-\tilde{z}_1}$, obtemos imediatamente $\tilde{z}_2 = 0.355809$. Fazendo o mesmo para $f_2(x) = \frac{x^3-3\sin(x)+1}{(x-\tilde{z}_1)(x-\tilde{z}_2)}$, obtemos $\tilde{z}_3 = 1.22042$. Todas estas aproximações têm os dígitos correctos.

• A deflação é uma técnica semelhante, aplicada a polinómios, mas a divisão é efectuada de forma algébrica (como iremos ver, pela regra de Ruffini) o que pode levar a um aumento significativo da imprecisão. No caso polinomial, o efeito das assímptotas horizontais não transparece, porque o número de raízes é igual ao grau do polinómio, consequentemente ao efectuar a divisão o grau do numerador será sempre superior ao grau do denominador.

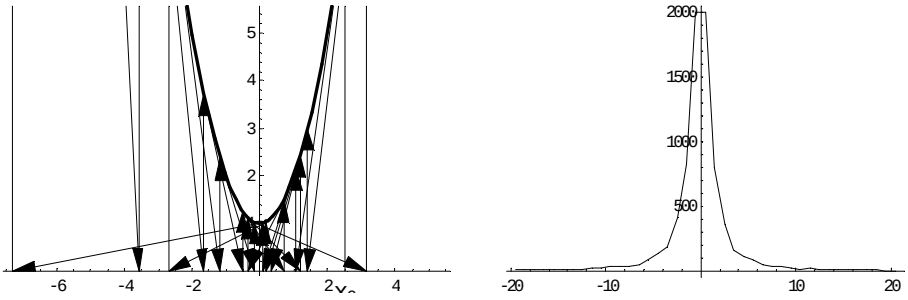
• Como já referimos, o problema com as assímptotas horizontais pode ser facilmente contornado através de uma mudança apropriada da função f. Bastará multiplicar f por uma função sem zeros cujo comportamento assimptótico anule a existência de assímptota horizontal. Uma sugestão será utilizar $f(x)(x^{2p} + 1)$, ou mesmo $f(x)e^{x^2}$. De acordo com a observação anterior isto pode ainda ser encarado como uma supressão do zero, já que estamos a tentar fazer desaparecer a 'raiz infinita'.

Observação 4 *(o método mais utilizado).*

O método de Newton é o método mais difundido para encontrar soluções de equações não lineares. Para além de ser simples e apresentar convergência quadrática, o que permite em poucas iterações encontrar uma boa aproximação, tem outra particularidade atractiva... encontra 'quase sempre' uma das raízes da equação. Vejamos uma possível explicação. Consideremos a função $f(x) = x^2 + 1$. Esta função não tem qualquer raiz real, e portanto o método de Newton não irá convergir se começarmos com iteradas reais. Apesar de se

[15] O termo é usado em [29], a propósito de polinómios, referindo apenas a possibilidade de ser utilizado para outras funções.

tratar de um caso de divergência, é instrutivo ver o que se passa. A função tem um mínimo em zero que servirá de *atractor-repulsor* para as sucessivas iteradas (ver figura em baixo, à esquerda). As iteradas aproximam-se do mínimo, mas o facto de a derivada se anular nesse ponto funciona como um repulsor. De tal forma, que se fizermos uma análise estatística do número de vezes que a função se encontra num determinado intervalo, obtemos o gráfico que se apresenta em baixo, à direita (e que poderá ser encarado como uma medida de probabilidade, cf. teoria ergódica). Reparamos que se trata de um gráfico semelhante ao de uma gaussiana, em que se torna evidente que as iteradas irão passar 'mais tempo' próximo do mínimo da função do que nos restantes pontos.



Este facto permite explicar a característica do método de Newton convergir habitualmente para uma das raízes. Com efeito, mesmo estando longe da raiz, sendo atraído por um extremo relativo (mínimo positivo ou máximo negativo) irá ser repelido para pontos distantes, onde poderá encontrar uma zona próximo de uma raiz, onde é válida a convergência local.

Assim, é a própria característica, aparentemente nefasta, de a derivada se poder anular, que permite ao método encontrar raízes em pontos remotos. Como as iteradas nunca serão verdadeiramente atraídas para os valores exactos em que há um extremo relativo e a derivada se anula, o problema com a derivada nula coloca-se mais ao nível de pretender restringir as iterações a um certo intervalo.

É claro que há dois contrapontos nesta questão:

(i) é também próximo dos extremos relativos que ocorrem situações de órbitas periódicas. Basta ver que, de acordo com a observação 1, se $f(x_0)$ e $f(x_1)$ têm o mesmo sinal, então $f'(x_0)$ e $f'(x_1)$ terão sinal diferente, o que significa que entre x_0 e x_1 há um ponto em que a derivada será nula.

(ii) quando a função tem assímptotas horizontais, *repelir 'para longe'* pode significar enviar para uma zona em que há uma 'raiz infinita', verificando-se a divergência.

2.6.2 Fórmula de erro do método de Newton

Supondo que $f \in C^2[x_n; z]$, pela fórmula de Taylor (em torno de x_n)

$$f(z) = f(x_n) + f'(x_n)(z - x_n) + \frac{1}{2}f''(\xi_m)(z - x_n)^2$$

para um certo $\xi_m \in]z; x_n[$.

Dividindo por $f'(x_n)$, não nulo, ficamos com:

$$0 = \frac{f(x_n)}{f'(x_n)} + z - x_n + \frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}(z - x_n)^2$$

portanto como $-x_{n+1} = \frac{f(x_n)}{f'(x_n)} - x_n$, temos

$$e_{n+1} = -\frac{1}{2}\frac{f''(\xi_n)}{f'(x_n)}e_n^2, \tag{2.23}$$

fórmula que, no limite, confirma o coeficiente assimptótico,

$$K_\infty = \frac{1}{2}\left|\frac{f''(z)}{f'(z)}\right|,$$

que já havia sido deduzido através do método do ponto fixo.

Tendo assegurado as condições de convergência exigidas no teorema 2.10, vemos que o sinal de e_{n+1} é constante porque os sinais de f' e f'' também o são. Nessas condições a convergência é sempre monótona (o que também tinha sido já provado na demonstração desse teorema).

A partir de (2.23) obtém-se imediatamente a estimativa *a posteriori*

$$|e_{n+1}| \leq \frac{\max_{x\in[a,b]}|f''(x)|}{2|f'(x_n)|}|e_n|^2, \tag{2.24}$$

que pode ser transformada numa estimativa *a priori* calculando

$$K = \frac{1}{2}\frac{\max_{x\in[a,b]}|f''(x)|}{\min_{x\in[a,b]}|f'(x)|},$$

obtendo-se

$$|e_n| \leq \frac{1}{K}(K|e_0|)^{2^n}, \tag{2.25}$$

já que de $|e_{n+1}| \leq K|e_n|^2$ obtemos $K|e_{n+1}| \leq (K|e_n|)^2$, ou seja $a_{n+1} \leq a_n^2$, com $a_n = K|e_n|$. Recursivamente temos $a_n \leq (a_0)^{2^n}$, o que implica (2.25).

• Observamos que esta estimativa é apenas útil se tivermos garantido que $K|e_0|$ é suficientemente pequeno (pelo menos menor que 1), majorando o erro inicial $|e_0|$ pelo comprimento de um intervalo suficientemente pequeno onde se garanta a existência da raiz.

Observação 1 *(critério de paragem)*

Se utilizarmos o habitual critério de paragem, $|x_{n+1}-x_n| < \varepsilon$, quando x_n já está próximo da raiz, podemos garantir que um majorante do erro absoluto será aproximadamente ε. Com efeito, de

$$|e_n| = |z - x_n| \leq |z - x_{n+1}| + |x_{n+1} - x_n| \leq K|e_n|^2 + |x_{n+1} - x_n|,$$

concluímos que se $K|e_n|$ for inferior a um certo valor $c < 1$, podemos obter

$$|e_n| \leq \frac{1}{1-c}|x_{n+1} - x_n|.$$

Como o erro irá diminuir, o valor de c também poderá diminuir, pelo que na prática, próximo da raiz, temos

$$|e_n| \leq 1.000...|x_{n+1} - x_n| \approx |x_{n+1} - x_n|,$$

o que constitui uma *fácil estimativa a posteriori* (e que, no fundo, corresponde a substituir z por x_{n+1}).

Se quisermos ser mais rigorosos, reparando que se $K|e_n| \leq c < 1$

$$|e_{n+1}| \leq \frac{c}{1-c}|x_{n+1} - x_n|.$$

bastará ter $K|e_n| \leq \frac{1}{2}$,para garantirmos que

$$|e_{n+1}| \leq |x_{n+1} - x_n|.$$

Exemplo 2.8 *Aplicando o método de Newton à função* $f(x) = x^3 - \cos(x) - 1$, *podemos ver que as condições para o teorema 2.10 são verificadas em* $[1, 2]$,*já que* $f \in C^2[1, 2]$, *e*
1) $f(1)f(2) < 0$, *pois* $f(1) = -\cos(1) = -0.54, f(2) = 7 - \cos(2) = 7.416$.
2) $f'(x) \neq 0, \forall x \in [1, 2]$, *pois* $f'(x) = 3x^2 + \sin(x) \geq 3x^2 - 1 \geq 2 > 0$.
3) $f''(x) \geq 0, \forall x \in [1, 2]$, *pois* $f''(x) = 6x + \cos(x) \geq 6x - 1 \geq 5$.
4) Como $|f(1)/f'(1)| = 0.14$ *e* $|f(2)/f'(2)| = 0.574$ *são ambos menores que* $|2-1| = 1$, *fica provada a convergência para qualquer iterada.*
Escolhemos $x_0 = 1$, *e obtemos* $x_1 = 1.14065, x_2 = 1.126711$.
Qual a confiança que podemos depositar neste último valor? Podemos obter $\max |f''(x)| = f''(2) = 11.58$ *(note-se que* f'' *é crescente em* $[1, 2]$, *pois* $f''' = 6-\sin(x) > 0$), *e* $\min |f'(x)| = f'(1) = 3.84$. *Assim,* $K = \frac{11.58}{2 \cdot 3.84} = 1.51$, *e sabemos também que* $|e_2| \leq 0.127$, *porque como a convergência é monótona a partir de* x_1, *teremos neste caso*

$$1 \leq z < ... < x_n < ... < x_2 < x_1.$$

É claro que $K|e_2| < 0.2$, *e portanto*

$$|e_2| \leq |x_2 - x_1| = \frac{0.2}{1 - 0.2} 0.014 = 0.0035.$$

Efectuando uma nova iterada, $x_3 = 1.126562$, *como* $|e_3| < |e_2| = 0.0035$, *temos* $K|e_3| < 0.0053$, *e obtemos*

$$|e_3| \leq |x_3 - x_2| = \frac{0.0053}{1 - 0.0053} 0.00015 = 0.8 \times 10^{-6}.$$

Na realidade todos os dígitos apresentados são correctos (arredondamento simétrico). Para obter novas estimativas, haveria que considerar um maior número de dígitos... caso contrário, esgotada a precisão, obteríamos valores sem significado... por exemplo, mantendo-nos com 7 ou 8 dígitos, teríamos $x_4 - x_3 = 0$, *o que não faria qualquer sentido.*
Preferiu-se utilizar a fórmula com a diferença entre as iteradas, já que normalmente conduz a melhores estimativas. No entanto, poderíamos aplicar qualquer uma das outras fórmulas, e até a fórmula elementar

$$|e_3| \leq \frac{|f(x_3)|}{\min_{x \in [1,2]} |f'(x)|} = \frac{0.8 \times 10^{-7}}{1.51} = 0.53 \times 10^{-7}$$

se pode revelar como a mais vantajosa... sendo este um caso em que isso acontece!

Observação 2 *(tempo de cálculo)*

O método de Newton, sendo um método de ponto fixo, também envolve apenas o tempo de cálculo $T = n\,t_g$. No entanto, como g assume aqui uma forma particular em que é necessário o cálculo da derivada da função f, devemos considerar (desprezando o tempo da subtracção e da divisão)

$$T = n\,(t_f + t_{f'}).$$

Note-se que para funções polinomiais o cálculo da derivada envolve até menor tempo que o tempo para calculara f, mas de um modo geral passa-se a situação inversa, sendo frequentemente utilizado o subterfúgio de considerar um ponto adicional, próximo da iterada para aproximar f', e nesse caso podemos considerar que $T = 2\ n\,t_f$.

Observação 3 *(método de optimização)*

Apesar de considerarmos o problema de optimização no último capítulo, referimos já que o método de Newton pode ser utilizado para encontrar mínimos de funções regulares. Querendo encontrar um mínimo para uma função regular ϕ, isso corresponde a encontrar um $z : \phi'(z) = 0$, pelo que o método de Newton pode ser aplicado desde que $\phi \in C^2$. Convém apenas referir que neste caso, ao definir

$$x_{n+1} = x_n - \frac{\phi'(x_n)}{\phi''(x_n)},$$

o ponto x_{n+1} é o ponto de mínimo para a parábola definida por

$$y = \phi(x_n) + \phi'(x_n)(x - x_n) + \frac{1}{2}\phi''(x_n)(x - x_n)^2,$$

que corresponde a uma aproximação de ϕ pelo seu desenvolvimento em série de Taylor (de segunda ordem).

2.6.3 Método de Newton no caso de zeros múltiplos

Definição 2.4 *Dizemos que uma função f tem um zero z de multiplicidade $p > 1$ se existir uma função h contínua em z tal que $h(z) \neq 0$ e que*

$$f(x) = (x - z)^p h(x).$$

Se $h \in C^p(V_z)$ então isto significa que

$$f(z) = f'(z) = \ldots = f^{(p-1)}(z) = 0,\ f^{(p)}(z) \neq 0.$$

O facto de, para zeros múltiplos, se ter $f'(z) = 0$ coloca algumas questões relativamente à aplicabilidade do método de Newton. No entanto, podemos ver que mesmo nesse caso o método apresenta convergência local, embora já não seja de ordem quadrática.

Com efeito, calculando

$$f'(x) = (x - z)^p h'(x) + p(x - z)^{p-1} h(x),$$

e considerando

$$g(x) = x - C\frac{f(x)}{f'(x)} = x - C\frac{(x-z)^p h(x)}{(x-z)^p h'(x) + p(x-z)^{p-1}h(x)} = x - C\frac{(x-z)h(x)}{(x-z)h'(x) + ph(x)} \quad ,$$

temos a função iteradora do método de Newton apenas quando $C = 1$.

Calculando a derivada, obtemos facilmente $g'(z) = 1 - \frac{C}{p}$. Ou seja, de acordo com os teoremas que vimos, acerca da convergência do método do ponto fixo, quando $p > 1$, podemos ainda assegurar convergência local para o método de Newton usual ($C = 1$), porque $0 < g'(z) = 1 - \frac{1}{p} < 1$, mas a ordem de convergência será linear e não quadrática.

Uma pequena modificação no método de Newton, usando $C = p$, ou seja,

$$x_{n+1} = x_n - p\frac{f(x_n)}{f'(x_n)} \tag{2.26}$$

permite reobter a convergência quadrática, pois neste caso $g'(z) = 0$. No entanto, há que salientar que isto pressupõe que saibamos à partida p, a multiplicidade do zero que pretendemos aproximar (esse conhecimento pode ser obtido através de resultados teóricos).

Exercício 2.3 *Mostre que, caso haja raízes múltiplas, se aplicar o método de Newton à função $F(x) = \frac{f(x)}{f'(x)}$ obtém um método de ordem 2.*
Resolução: *Nesse caso temos*

$$F(x) = \frac{f(x)}{f'(x)} = (x-z)\frac{h(x)}{(x-z)h'(x) + p\,h(x)} = (x-z)H(x)$$

que tem um zero simples, pois $H(z) = \frac{1}{p} \neq 0$. Isto significa que o método de Newton aplicado a $F(x)$ converge quadraticamente.
Por outro lado, como $F'(x) = 1 - \frac{f(x)f''(x)}{f'(x)^2}$, o método de Newton aplicado a F pode escrever-se na forma:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\,\frac{f'(x_n)^2}{f'(x_n)^2 - f(x_n)f''(x_n)} = x_n - \frac{f(x_n)f'(x_n)}{f'(x_n)f'(x_n) - f(x_n)f''(x_n)}. \tag{2.27}$$

Iremos apresentar uma fórmula semelhante (não igual... aparecem coeficientes 2), em (2.30) para um método de ordem 3, que pode ser usado em conjunção com este, no caso de ser desconhecida a natureza das raízes.

2.6.4 Método da Secante

Sendo muito semelhante ao método de Newton, o *método da secante* (também designado falsa posição, por alguns autores, que atribuem o nome falsa posição clássico ao que vimos anteriormente) substitui o cálculo das derivadas pelo cálculo de uma razão incremental. Geometricamente, corresponde a substituir o papel da recta tangente, no método de Newton, por uma recta secante (de onde vem o nome).

É claro que isto significa que vamos precisar sempre de dois pontos para determinar essa recta secante, o que implica que tenhamos que considerar duas iteradas iniciais, que

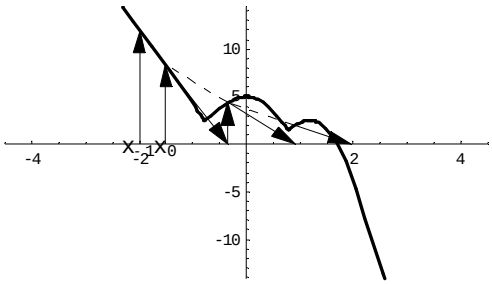
designaremos por x_{-1} e x_0. Mas notamos que a partir daí, em cada iteração, o cálculo da função é efectuado apenas uma vez.

De forma semelhante à que fizemos no método de Newton, calculando agora o ponto de intersecção da secante com o eixo das abcissas, obtemos a fórmula para x_{n+1}, e o *método da secante* resume-se no esquema

$$\begin{cases} \text{Iteradas iniciais} : x_{-1},\ x_0 \\ \text{Iterar} \quad x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \end{cases} \tag{2.28}$$

garantindo que $f(x_n) \neq f(x_{n-1})$, o que pode ser assegurado imediatamente se a função for injectiva[16].

Exemplo 2.9 *Consideramos a função* $f(x) = |3\cos(2x)| - x^3 + 2$, *que não é diferenciável, e começamos com* $x_{-1} = -2, x_0 = -1.5$. *Na figura seguinte podemos ver graficamente a evolução das iterações,* $x_1 = -0.346, x_2 = 0.910, x_3 = 1.97$, *e quando atingimos* $x_{10} = 1.69876$, *já todos os dígitos apresentados são correctos.*



Neste exemplo, a função considerada apresentava descontinuidades na derivada, mas poderia ser calculada seccionalmente, pelo que a não diferenciabilidade da função não é a razão de utilizar o método da secante ao invés do método de Newton. O principal problema na maioria das vezes é a dificuldade de calcular a derivada. É claro que se poderá sempre pensar em aproximá-la usando uma razão incremental, mas esse artifício será mais moroso do que aplicar directamente o método da secante.

Proposição 2.6 *(*condição suficiente de convergência*) É idêntica à enunciada para o método de Newton, apenas a hipótese 4a) deverá ser substituída por*
4a') $f(x_0)f''(x) \geq 0$ *e* $f(x_{-1})f''(x) \geq 0$ *para qualquer* $x \in [a, b]$. ■

• Fórmula de Erro

Usando os mesmos argumentos que levaram à determinação da fórmula de erro (2.6) para o método da falsa-posição, ou reparando que no caso do método da secante podemos substituir os valores a_n e b_n por x_n e x_{n-1}, obtemos

$$e_{n+1} = -\frac{1}{2}\frac{f''(\xi_n)}{f'(\eta_n)} e_n e_{n-1} \tag{2.29}$$

[16]Convém referir que só poderemos ter $x_{n+1} = x_n$ se $f(x_n) = 0$ ou se $x_n = x_{n-1}$. Se $f(x_n) = 0$ isso significa que encontrámos uma raiz, e a condição $x_n = x_{n-1}$ nunca se verifica porque como é óbvio escolhemos inicialmente $x_{-1} \neq x_0$.

com $\xi_n, \eta_n \in]x_{n-1}; z; x_n[$. A fórmula (2.29) permite obter a estimativa

$$|e_{n+1}| \leq K\ |e_n|\ |e_{n-1}|$$

em que o K é o mesmo que o do método de Newton, ou seja, $K = \frac{1}{2}\frac{\max_{x\in[a,b]}|f''(x)|}{\min_{x\in[a,b]}|f'(x)|}$. Esta estimativa pode ser convertida numa estimativa *a priori* por recursividade. Para esse efeito, devemos considerar a equação às diferenças $u_{n+1} = u_n + u_{n-1} + k$, chamada sucessão de Fibonacci, em que $u_n = \log|e_n|$, $k = \log(K)$. (Devemos ter em mente que ao escrever a equação estamos a evitar a desigualdade, já que tratamos com a majoração $u_{n+1} \leq u_n + u_{n-1} + k$).

Esta equação às diferenças[17] tem como solução geral

$$u_n = -C_1(\frac{1}{\rho})^n - C_2\rho^n - k,$$

em que ρ é o 'número de ouro' $\frac{1+\sqrt{5}}{2} = 1.61803...$ (raiz dominante da equação $x^2 = x + 1$) e em que as constantes C_1, C_2 podem ser obtidas a partir de $u_0 = \log|e_0|$ e de $u_{-1} = \log|e_{-1}|$, resolvendo um sistema, cuja solução é

$$C_1 = \frac{1}{\rho}\log(|e_0|K) - \log(|e_{-1}|K),\ \ C_2 = \log(|e_{-1}|K) - \rho\log(|e_0|K).$$

Daqui podemos obter a estimativa *a priori*

$$|e_n| \leq \exp(-C_1(\frac{1}{\rho})^n + -C_2\rho^n - k) \approx \frac{1}{K}\,e^{-C_2\rho^n},$$

ou ainda, aproximadamente

$$|e_n| \leq \frac{1}{K}\left(\frac{(|e_0|K)^\rho}{|e_{-1}|K}\right)^{\rho^n}.$$

Observação 1 *(ordem de convergência).*

Através da expressão de u_n podemos mesmo concluir que a *ordem de convergência* é supralinear, igual ao número de ouro $\rho = 1.61803....$

Com efeito, reparamos que

$$\log(\frac{|e_{n+1}|}{|e_n|^\rho}) = \log|e_{n+1}| - \rho\log|e_n| = u_{n+1} - \rho u_n,$$

e que $u_{n+1} - \rho u_n = C_1(\frac{1}{\rho})^{n+1}(1-\rho^2) + (\rho-1)k \to (\rho-1)k$. Como o valor de $k = \log(K)$ resulta de uma majoração, não podemos concluir que $K_\infty = \exp((\rho-1)k) = K^{\rho-1}$, mas pode verificar-se (cf.[1]) que

$$K_\infty = \left(\frac{|f''(z)|}{2|f'(z)|}\right)^{\rho-1}.$$

[17] Ver Anexo sobre Equações às Diferenças.

Observação 2 *(tempo de cálculo).*

Como já referimos, o método da secante envolve apenas o cálculo de f uma vez, em cada iteração, por isso (negligenciando as operações elementares),

$$T = n\, t_f.$$

Se compararmos o tempo de cálculo relativamente ao método de Newton, para obter um mesmo majorante de erro ε, já vimos em (2.3), assumindo constantes semelhantes, que

$$T_N = \frac{\log \rho}{\log 2} \frac{t_f + t_{f'}}{t_f} T_S = 0.6942(1 + \frac{t_{f'}}{t_f}) T_S$$

pelo que podemos concluir que o método de Newton será mais eficaz se o tempo de cálculo da derivada verificar

$$0.6942(1 + \frac{t_{f'}}{t_f}) < 1 \Leftrightarrow \frac{t_{f'}}{t_f} < 0.4405,$$

ou seja, será conveniente que o cálculo da derivada demore menos que metade do tempo do cálculo de f, o que normalmente nem acontece!... No entanto, a melhor precisão do método de Newton num menor número de iterações é considerada preferencialmente[18].

Observação 3 *(outras modificações do método de Newton).*

Como já foi referido, uma outra possibilidade para evitar o cálculo da derivada no método de Newton consiste em considerar um ε suficientemente pequeno e efectuar

$$f'(x_n) \approx \frac{f(x_n + \varepsilon) - f(x_n)}{\varepsilon}, \text{ e portanto } x_{n+1} = x_n - \frac{f(x_n)\varepsilon}{f(x_n + \varepsilon) - f(x_n)},$$

o que é muitas vezes designado por método de Newton modificado. No entanto, isto implica o cálculo adicional de $f(x_n + \varepsilon)$, o que implica $T = 2n\, t_f$, e facilmente é visível uma maior morosidade face ao método da secante. Para além disso, quando a diferença entre as iterações for menor que ε, já deixa de fazer sentido esta aproximação da derivada.

Um método que tem convergência quadrática, e que pode ser visto como uma variante do método de Newton, é o *método de Steffenson*, quando aplicado a $g(x) = f(x) + x$. Nesse caso a fórmula (2.19) fica

$$x_{n+1} = x_n - \frac{f(x_n)^2}{f(x_n + f(x_n)) - f(x_n)},$$

reparando que isso consiste em considerar ε igual ao valor $f(x_n)$, que será cada vez mais pequeno, quando nos aproximamos da raiz. O método de Steffenson implica também o cálculo adicional de f em $x_n + f(x_n)$. A condição para a convergência quadrática do método de Steffenson, sendo $1 \neq g'(z) = f'(z) + 1$, é equivalente a $f'(z) \neq 0$, ou seja a condição do método de Newton.

[18]Refira-se que o *Mathematica* aplica o método de Newton na rotina FindRoot se for apenas dado um valor inicial. Caso sejam dados dois valores iniciais, aplicará o método da secante (sintaxe FindRoot[f[x]==0,{x,{x_{-1},x_0}}]).

Esta classe de variantes do método de Newton é muitas vezes designada por *métodos quasi-Newton.*

• Tal como no caso do método de Newton, estes métodos podem ser aplicados à pesquisa de mínimos de funções, quando aplicados à derivada da função, tendo a vantagem de não necessitar do cálculo da segunda derivada.

2.6.5 Métodos de ordem superior

Podemos obter facilmente métodos com convergência de ordem superior (usando interpolação de Hermite), tal como se mostra no lema seguinte. O processo consiste em generalizar a ideia do método de Newton de aproximar a função em x_n pela recta tangente. A generalização passa por aproximar a função por uma parábola tangente, por um polinómio de 3^o grau tangente, etc...

Não é necessário restringirmo-nos a polinómios. Por exemplo, um método de ordem cúbica é

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)}, \tag{2.30}$$

onde a função f é aproximada por uma hipérbole da forma $r(x) = \frac{ax+b}{cx+1}$.

A aproximação por parábolas pode trazer o inconveniente de que essa parábola não tenha zeros reais. O método poderá ser adequado se nos dispusermos a trabalhar com números complexos. Uma maneira de evitar esse problema, é considerar a aproximação local por uma parábola invertida, o que nos leva ao método,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f(x_n)^2 f''(x_n)}{2f'(x_n)^3} \tag{2.31}$$

que também tem convergência local cúbica.

Como curiosidade, apresentamos um outro método usando uma aproximação com funções racionais do tipo $\frac{ax+b}{cx^2+dx+1}$, que apresenta convergência local de ordem 4,

$$x_{n+1} = x_n - \frac{6f(x_n)f'(x_n)^2 - 3f(x_n)^2 f''(x_n)}{6f'(x_n)^3 - 6f(x_n)f'(x_n)f''(x_n) + f(x_n)^2 f'''(x_n)}. \tag{2.32}$$

Estes métodos raramente são considerados. Como já se referiu, os métodos de ordem superior raramente compensam o número de operações envolvidas. Se o cálculo de uma derivada já pode ser restritivo, o cálculo de um maior número de derivadas ainda piora a situação! No entanto, poderá haver algum interesse quando aplicados a funções cuja expressão da derivada seja extremamente simples, como a soma de polinómios com exponenciais. Em termos de tempo de cálculo, basta ver que uma iteração de um método de ordem 4 pode ser compensada através de duas iterações de um método com convergência quadrática mais simples (como o método de Newton)... basta reparar que $\log(4)/\log(2) = 2$.

Ou seja, considerar a função iteradora do método de Newton composta com ela própria, o que corresponde a efectuar duas iterações, fornece-nos um método de ordem 4. Como já referimos, esta característica é particular dos métodos supralineares, e não se aplica aos

métodos lineares, já que aplicar duas vezes um método linear apenas permite diminuir o seu coeficiente assimptótico, não passamos para uma ordem superior.

De qualquer forma, podemos enunciar um resultado razoavelmente geral, que permitirá obter métodos de ordem superior, em termos de um lema.

Lema 2.2 *Seja r uma função de interpolação para f tal que*

$$r(x_n) = f(x_n),\ r'(x_n) = f'(x_n),\ ...,\ r^{(p-1)}(x_n) = f^{(p-1)}(x_n)$$

então o método, que consiste em atribuir x_{n+1} ao valor tal que $r(x_{n+1}) = 0$, tem pelo menos ordem de convergência p.

Demonstração:
Basta reparar que da expansão de Taylor,

$$(f-r)(z) = (f-r)(x_n) + (f-r)'(x_n)e_n + ... + \frac{(f-r)^{(p-1)}(x_n)}{(p-1)!}(e_n)^{p-1} + \frac{(f-r)^{(p)}(\xi_n)}{p!}(e_n)^p,$$

com $e_n = z - x_n$, e resulta (por hipótese)

$$-r(z) = \frac{(f-r)^{(p)}(\xi_n)}{p!}(e_n)^p.$$

Como $r(x_{n+1}) = 0$, aplicando o teorema de Lagrange obtemos

$$-r'(\eta_n)e_{n+1} = r(x_{n+1}) - r(z) = \frac{(f-r)^{(p)}(\xi_n)}{p!}e_n^p \Rightarrow \frac{e_{n+1}}{e_n^p} = -\frac{(f-r)^{(p)}(\xi_n)}{p!\ r'(\eta_n)}.\ \blacksquare$$

• Este lema foi aplicado para obter os métodos apresentados anteriormente.
Considerámos $r(x) = \frac{ax+b}{cx+1}$, $r(x) = \frac{ax+b}{cx^2+dx+1}$ (para (2.30), (2.32)), e ainda $r^{-1}(y) = ay^2 + by + c$... usando interpolação inversa (para (2.31)).

Observação 1 *(extensões do método da secante).*
Da mesma maneira que é possível efectuar extensões para o método de Newton, também é possível pensar de forma semelhante para o método da secante. Uma possível generalização do método da secante é o *método de Müller*, que consiste em considerar 3 valores iniciais, x_{-2}, x_{-1}, x_0 e fazer uma interpolação através de uma parábola (ao invés de uma recta). A aproximação x_1 é dada encontrando a raiz da parábola que estiver mais próxima de x_0 e assim sucessivamente. Neste caso pode surgir o problema de não existirem raízes reais, mas o método pode ser adequado para encontrar raízes complexas.

Como curiosidade, a ordem de convergência do método de Müller é $p = 1.83929...$ que é a raiz dominante da equação $x^3 = x^2 + x + 1$. Outras possíveis generalizações, usando uma interpolação com curvas de grau m superior a 2 levariam a ordens de convergência p_m superiores, iguais ao valor da raiz dominante das equações $x^{m+1} = x^m + ... + x + 1$.

No entanto, é fácil verificar que seriam sempre inferiores à convergência quadrática (com efeito obtemos

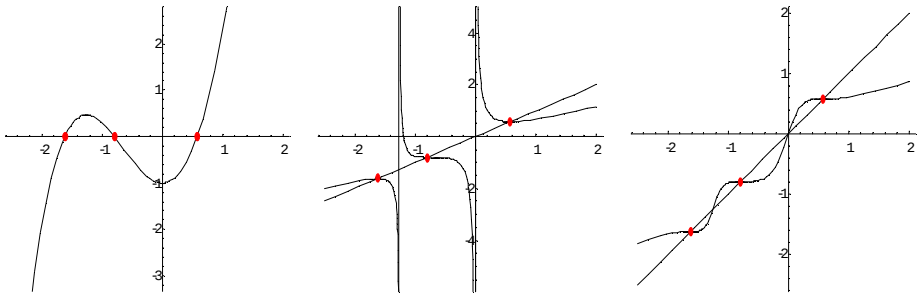
$$p_1 = 1.61803,\ p_2 = 1.83929,\ p_3 = 1.92756,\ p_4 = 1.96595,\ \text{etc...}$$

mas sempre $p_m < 2$, como podemos concluir facilmente pelo teorema 2.12).

Observação 2

O método apresentado em (2.30) apresenta uma solução diferente para contornar a dificuldade com funções iteradoras g contínuas. Se repararmos na expressão de $g(x_n)$ dada por x_{n+1} em (2.30), verificamos que quando $f'(x_n) = 0$ a função g não explode (a menos que também tenhamos $f''(x_n) = 0$), pelo contrário, quando isso acontece temos um ponto fixo!

Vejamos a diferença para a função $f(x) = x^3 - 5\cos(x) + 4$, representado no gráfico em baixo, à esquerda, e onde estão assinalados os três zeros da função. Na figura central, representamos a função iteradora g obtida pelo método de Newton, e os respectivos três pontos fixos (onde é vísivel que g' se anula). Repare-se que as assímptotas verticais separam os pontos fixos (como já foi referido, é uma consequência do teorema de Rolle), quebrando a continuidade de g e permitindo encontrar todas os pontos fixos. Na figura da direita, representamos a função g dada por (2.30). Surpreendentemente, a função g é contínua, e no entanto é possível encontrar também todos os pontos fixos que nos interessam, correspondentes, às raízes de f. A solução engenhosa (e não propositada) resulta de que nos outros dois pontos fixos a derivada é maior que 1 em módulo, e portanto o método nunca irá convergir para esses valores, sendo atraído apenas para os pontos fixos para os quais g' e g'' são nulas... as raízes de f. É claro que isto necessita de justificação, que pode ficar como exercício, mas ilustra bem uma outra possibilidade para encontrar raízes consecutivas.



Como curiosidade adicional, começando com $x_0 = 1$, obtemos $x_4 = 0.576574...$ pelo método de Newton, com todos os dígitos correctos e o mesmo valor em x_3, usando o método (2.30), que tem convergência cúbica. O mesmo se irá passar para encontrar as restantes raízes, verificando-se que a diferença entre os dois métodos se cifra em apenas uma ou duas iterações... com um menor número de operações efectuado pelo método de Newton. Relembre-se que o método (2.27), para zeros múltiplos, é bastante semelhante a (2.30)... apenas difere em factores multiplicados por 2, e resultava de aplicar Newton a uma nova função $\frac{f}{f'}$, o que basicamente equivale em termos de cálculo a efectuar duas vezes Newton.

2.7 Métodos para Equações Algébricas

Vamos agora ver em particular as equações algébricas, isto é, equações da forma:

$$p_m(x) = a_0 + a_1 x + \ldots + a_m x^m = 0.$$

Se os coeficientes do polinómio forem números inteiros, às raízes dessas equações chamamos *números algébricos*. Por exemplo, $\sqrt[4]{3/2}$ é um número algébrico, porque é solução de $2x^4 - 3 = 0$. Nem todos os números reais são algébricos (aos números não-algébricos, chamamos *transcendentes*)! Isto não é uma questão trivial, e só foi esclarecida no final do séc. XIX. O facto de Hermite ter demonstrado que e não era algébrico, permitiu a Lindemann provar que π também não era (graças à fórmula de Euler, $e^{i\pi} + 1 = 0$, os números π e e estão relacionados). Resolveu-se assim um dos problemas mais antigos da geometria – a impossibilidade da quadratura do círculo (ou seja, é impossível, através de regra e compasso, encontrar um quadrado com área igual à de um círculo).

Exercício 2.4 *Mostre que é impossível aproximar um número transcendente através do método do ponto fixo se a função iteradora for uma função racional com coeficientes racionais (ou seja, se $g(x) = \frac{p(x)}{q(x)}$, em que p, q são polinómios com coeficientes racionais).*
Resolução: *Basta reparar que se $z = \frac{p(z)}{q(z)}$ então z verifica $zq(z) - p(z) = 0$, ou seja, z é solução de uma equação algébrica, e portanto nunca poderá ser um número transcendente!*

Recordamos o resultado fundamental acerca das raízes de polinómios, devido a Gauss[19].

Teorema 2.11 *(Fundamental da Álgebra). Uma equação algébrica de grau m:*

$$p_m(x) = c_0 + c_1 x + \ldots + c_m x^m = 0, \quad (c_0, c_1, ..., c_m \in \mathbb{C},\ c_m \neq 0)$$

admite exactamente m raízes complexas (contando com a ordem de multiplicidade).

Demonstração:
Usando o Teorema de Liouville: '*uma função analítica em $\mathbb{C}$ que seja limitada é constante*', provamos que um polinómio não constante tem um zero em $\mathbb{C}$. Se, por absurdo, $p_m(x) \neq 0, \forall x \in \mathbb{C}$, então $f(x) = \frac{1}{p_m(x)}$ seria uma função inteira limitada (porque $f(x)$ seria contínua num compacto, portanto limitada, e fora desse compacto $\lim_{|x|\to\infty} f(x) = 0$), logo pelo Teorema de Liouville seria constante. Ora, como p_m não é constante, chegamos a uma contradição.

Portanto, seja z_m uma raiz de p_m. Pela regra de Ruffini, $p_m(x) = p_{m-1}(x)(x - z_m)$, e assim sucessivamente, até que $p_1(x) = p_0(x)(x - z_1)$ e p_0 é um polinómio de grau 0, ou seja, uma constante. Logo, $z_1, \ldots, z_m$ são as m raízes de p_m. ■

[19]Repare-se que nada há de trivial neste resultado! No séc. XVIII não seria claro que fosse sempre possível escrever

$$p(x) = a_n(x - z_1)...(x - z_n)$$

em que $z_1, ..., z_n$ eram números complexos. Aliás, se não era possível com os reais, não havia garantias de o ser com os complexos... e ainda havia a suspeita (depois confirmada) da inexistência de fórmulas resolventes para equações polinomiais em geral.

Observações:

(i) Se o polinómio tiver coeficientes reais, então se z for uma raiz complexa, o conjugado $\bar{z}$ também será raiz, com a mesma multiplicidade.

(ii) Se os coeficientes forem reais e m for impar, o polinómio tem pelo menos um raiz real.

Vamos agora ver um critério bastante simples que, através de um cálculo imediato nos coeficientes, permite determinar uma bola no plano complexo que contém todas as raízes do polinómio.

Teorema 2.12 *Se z_k é raiz de $p_m(x) = c_0 + c_1x + ... + c_mx^m = 0$, com $c_i \in \mathbb{C}, c_m \neq 0$, temos*

$$|z_k| < 1 + \frac{M}{|c_m|} \tag{2.33}$$

com $M = \max\{|c_0|, \ldots, |c_{m-1}|\}$

Demonstração:

Seja $|x| \geq 1 + M/|c_m|$, temos:

$$|p_m(x)| \geq |c_m|\,|x|^m - (|c_0| + |c_1|\,|x| + \ldots + |c_m|\,|x|^{m-1}) \geq$$

$$\geq |c_m|\,|x|^m - M(1 + |x| + \ldots + |x|^{m-1}) = |c_m|\,|x|^m - M(\frac{|x|^m - 1}{|x| - 1}) > (|c_m| - \frac{M}{|x| - 1})|x|^m$$

como $|c_m| \geq \frac{M}{|x|-1} \Leftrightarrow |x| \geq 1 + \frac{M}{|c_m|}$, temos $|p_m(x)| > 0$, o que significa que $p_m(x) = 0$ só no caso de $|x| \geq 1 + \frac{M}{|c_m|}$ não se verificar. Portanto as raízes estão localizadas para $|x| < 1 + \frac{M}{|c_m|}$. □

Exercício 2.5 *Seja $a \neq 0$ e $\tilde{M} = \max\{|a_1|, \ldots, |a_m|\}$. Mostre que toda a raiz z_k da equação*

$$p(x) = a_0 + a_1x + \ldots + a_mx^m = 0$$

verifica

$$|z_k| > (1 + \frac{\tilde{M}}{|a_0|})^{-1}$$

Consideremos agora polinómios de coeficientes reais. Analisando o sinal desses coeficientes é possível retirar alguma informação acerca da localização das raízes.

Definição 2.5 *Chamamos* número mínimo de variações de sinal *de uma lista de números,*

$$L_a = (a_0, a_1, \ldots, a_m),$$

ao número de variações de sinal que ocorre nessa ordem, excluindo os zeros. Esse número é designado por v_a^-. Chamamos número máximo de variações de sinal *dessa lista quando os zeros são substituídos por valores positivos ou negativos, por forma a obter um número máximo de variações de sinal. Esse número é designado por v_a^+.*

Exemplo 2.10 *Consideremos a lista* $(-2,3,0,0,1,0,0,0)$. *O número mínimo de variações de sinal será dado pelas variações em* $-2,3,1$, *isto é, apenas ocorre uma variação* $(v^- = 1)$. *Um número máximo de variações de sinal ocorre, por exemplo, se escolhermos* $-2,3,-,+,1,-,+,-$, *com 6 variações* $(v^+ = 6)$.

Teorema 2.13 *(Budan-Fourier). Seja* $[a,b] \subseteq \mathbb{R}$ *e* $p(x)$ *um polinómio de coeficientes reais, onde consideramos as listas*

$$L_a = \left(p(a), p'(a), \ldots, p^{(m)}(a)\right)$$

$$L_b = \left(p(b), p'(b), \ldots, p^{(m)}(b)\right).$$

O número de zeros em $[a,b]$ *é igual a* $v_a^- - v_b^+ - 2k$, *para um certo* $k \in \mathbb{N}_0$.

Demonstração: Está fora do âmbito do curso (consultar por exemplo [19]).□

Exercício 2.6 *Baseado no teorema de Budan-Fourier, mostre a regra de Descartes: "*A diferença entre o número mínimo de variações de sinal e o número de zeros positivos de um polinómio é um número maior ou igual a zero e que é par *".*

Exemplo 2.11 *Consideremos* $p(x) = x^5 - 3x + 1 = 0$. *Pela regra de Descartes concluímos imediatamente que ou não há raízes reais positivas ou há apenas duas, e da mesma maneira, considerando* $p(-x) = -x^5 + 3x + 1$ *concluímos que há uma única raiz negativa.*

Como as raízes pertencem à bola $\{|x| < 1 + 3/1 = 4\}$, *vamos analisar onde se situam as raízes, considerando como extremos dos intervalos* $-4,-1,0,1,4$:

	-4	*-1*	*0*	*1*	*4*
$x^5 - 3x + 1 = p(x)$	-	+	+	-	+
$5x^4 - 6 = p'(x)$	+	-	-	-	+
$20x^3 = p''(x)$	-	-	*0*	+	+
$60x^2 = p'''(x)$	+	+	*0*	+	+
$120x = p^{(4)}(x)$	-	-	*0*	+	+
$120 = p^{(5)}(x)$	+	+	+	+	+
v_a^-	*5*	*4*	*2*	*1*	*0*
v_a^+	*5*	*4*	*4*	*1*	*0*
$v_a^- - v_b^+$	–	*0*	*1*	*2*	*0*

Como $v_{-4}^- - v_{-1}^+ = 1$ *concluímos que em* $[-4,-1]$ *há uma e uma só raiz, que é a raiz negativa. Por outro lado, de* $v_0^- - v_1^+ = 1$, *e de* $v_1^- - v_4^+ = 1$, *concluímos que em* $[0,1]$ *existe uma única raiz e em* $[1,4]$ *uma outra. Podemos usar a própria tabela para inspeccionar se as condições suficientes que estabelecemos para a convergência do método de Newton são verificadas. Por exemplo, é fácil verificar que isso acontece no intervalo* $[0,1]$, *nos outros intervalos* $[-4,-1]$ *e* $[1,4]$ *há uma variação de sinal da derivada que requeria uma subdivisão.*

Aplicando os métodos estudados anteriormente, não é dificil aproximar as raízes reais $-1.38879..., 0.334734..., 1.21465...$

(Veremos num próximo parágrafo generalizações de métodos do ponto fixo (p. ex: Newton) que permitem aproximar as duas raízes complexas conjugadas $-0.0802951... \pm 1.32836...i$.)

Observação: *(Esquema de Hörner)*

Ao efectuar o cálculo do polinómio, se o fizermos seguindo a forma canónica reparamos que isso corresponde a efectuar pelo menos $2m-1$ multiplicações e m adições. O número de multiplicações pode ser reduzido a m se efectuarmos o cálculo segundo o esquema de Hörner:

$$p_m(x) = (...((a_m x + a_{m-1})x + a_{m-2})x + ... + a_1)x + a_0$$

que corresponde a considerar $b_m(x) = a_m$ e fazer

$$b_k(x) = a_k + x\, b_{k+1}(x), \text{ para } k = m-1, ..., 0,$$

tendo-se $b_0(x) = p_m(x)$. Definindo $p_{m-1}(x) = b_1(y) + b_2(y)x + ... + b_m(y)x^{m-1}$, reparamos que

$$b_0(y) + (x-y)p_{m-1}(y) = b_0(y) + (x-y)(b_1(y) + b_2(y)x + ... + b_m(y)x^{m-1}) =$$
$$= b_0(y) - yb_1(y) + (b_1(y) - yb_2(y))x + ... + (b_{m-1}(y) - yb_m(y))x^{m-1} + b_m(y)x^m = p_m(x)$$

porque temos $b_k(y) - y\, b_{k+1}(y) = a_k$. Ou seja, mostrámos que $p_m(x) = b_0(y) + (x-y)p_{m-1}(x)$.

O esquema de Hörner permite ainda obter a derivada de p_m pois $p'_m(x) = p_{m-1}(x) + (x-y)p'_{m-1}(x) \Rightarrow p'_m(y) = p_{m-1}(y)$. Isto pode ser útil ao efectuar o método de Newton, considerando $y = x_n$ (pode ser assim evitada a derivação algébrica).

Se $y = z$ for um zero de p_m então $b_0(z) = 0$ e $p_{m-1}(x)$ é o polinómio que resulta da divisão por $x - z$ e pode ser usado para calcular as restantes raízes. Ou seja, neste caso, o processo de supressão de zeros (que aqui é designado por *deflação)* para encontrar novas raízes é efectuado algebricamente, de forma imediata. Esta deflação, sendo efectuada com os erros inerentes à aproximação da raiz, pode acumular alguns erros que podem ser menores se for iniciada na raiz dominante.

• Métodos mais eficazes para encontrar raízes de polinómios foram base de investigação intensa num passado recente, e escapam a um curso introdutório. O *Mathematica* usa na rotina NSolve um método em várias etapas devido a Jenkins e Traub (1970).

2.7.1 Método de Bernoulli

Apresentamos de seguida um método específico para aproximar raízes reais de polinómios – o denominado *método de Bernoulli.* O interesse deste método é essencialmente teórico, e pode visto como um caso particular do método das potências, de que falaremos num capítulo posterior, a propósito da determinação de valores próprios de matrizes. Refira-se que um dos métodos para encontrar raízes de polinómios é encontrar os valores próprios da matriz companheira, como veremos nesse capítulo.

A nível de implementação, o método de Bernoulli caracteriza-se por reduzir o cálculo de iterações a somas, subtracções ou multiplicações, evitando o cálculo de divisões ao máximo – só uma divisão final é necessária. O método converge quando há uma raiz real dominante z_1. Tendo determinado z_1 e efectuando um processo de deflação ou supressão de zeros, poderemos tentar obter as restantes, desde que sejam sucessivamente reais e dominantes, ou seja bastará que $|z_1| > |z_2| > ... > |z_m|$.

A convergência do método é linear, mas como o coeficiente assimptótico será $\left|\frac{z_2}{z_1}\right|$, se a segunda maior raiz z_2 for em módulo próxima do módulo da dominante z_1, a convergência poderá ser muito lenta.

Consideremos a equação algébrica

$$P(x) = a_0 + a_1 x + \ldots + a_{m-1}x^{m-1} - x^m = 0,$$

e supomos que existe uma raiz dominante z_1, ou seja, $|z_1| > |z_2| \geq \ldots \geq |z_m|$.

À equação algébrica associamos a *equação às diferenças*[20]:

$$a_0 x_n + a_1 x_{n+1} \ldots + a_{m-1} x_{n+m-1} - x_{n+m} = 0, \tag{2.34}$$

definindo um processo iterativo, designado por *método de Bernoulli,*

$$\begin{cases} \text{Iteradas iniciais: } x_0 = \ldots = x_{m-2} = 0;\ x_{m-1} = 1, \\ \text{Iterar: } x_{n+m} = a_0 x_n + a_1 x_{n+1} \ldots + a_{m-1} x_{n+m-1}. \end{cases} \tag{2.35}$$

que permite obter a raiz dominante z_1 através do limite[21]

$$\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = z_1.$$

A sucessão (x_n) é definida recursivamente, através da equação às diferenças, e sabe-se que no caso em que o polinómio tem raízes distintas $z_1, \ldots, z_m$, os valores da sucessão são dados por

$$x_n = C_1 z_1^n + \ldots + C_m z_m^n$$

para certas constantes $C_1, \ldots, C_p$ não todas nulas (determinadas a partir dos valores das iteradas iniciais). Devemos escolher os valores iniciais de forma a que $C_1 \neq 0$.

Como $|z_1| > |z_2| \geq \ldots \geq |z_m|$, temos:

$$\frac{x_{n+1}}{x_n} = \frac{C_1 z_1^{n+1} + \ldots + C_m z_m^{n+1}}{C_1 z_1^n + \ldots + C_m z_m^n} = z_1 \frac{C_1 + C_2(\frac{z_2}{z_1})^{n+1} + \ldots + C_m(\frac{z_m}{z_1})^{n+1}}{C_1 + C_2(\frac{z_2}{z_1})^n + \ldots + C_m(\frac{z_m}{z_1})^n} \to z_1 \ .$$

Verifica-se que $\frac{x_{n+1}}{x_n} = z_1 + O(\left|\frac{z_2}{z_1}\right|^n)$, ou seja

$$|e_n| = |z_1 - \frac{x_{n+1}}{x_n}| \leq C \left|\frac{z_2}{z_1}\right|^n ,$$

o que indicia a convergência linear do método. A rapidez de convergência será maior se a razão $\frac{|z_2|}{|z_1|}$ for pequena (razão que corresponde ao factor assimptótico de convergência). Uma análise mais rigorosa da convergência deste método pode ser vista quando falarmos no método das potências para a determinação de valores próprios, do qual este será um caso particular.

[20] Ver Anexo acerca das Equações às Diferenças.

[21] Podem ser escolhidas outras iteradas iniciais. A única condição a verificar é que o coeficiente C_1 (que iremos considerar de seguida, obtido pela resolução da equação às diferenças) não seja nulo!

Exemplo 2.12 *Encontrar a raiz dominante de* $x^5 - 3x + 1 = 0$.
Começando com $x_0 = \ldots = x_3 = 0,\ x_4 = 1$, *iteramos*

$$x_{n+5} = 3x_{n+1} - x_n$$

e ao fim de algumas iterações, calculando $y_n = \frac{x_n}{x_{n-1}}$, *obtemos:*

$$y_{100} = -\frac{44971518385729}{31849964584089} = -1.41198...,\ y_{150} = -1.3908...,\ y_{200} = -1.3889...$$

Apesar de precisarmos de 200 iterações para obter um valor razoavelmente próximo da raiz $-1.38879...$, *notamos que o cálculo das iteradas é extremamente simples! Apenas precisamos de efectuar multiplicações e subtracções com números inteiros, que são operações que um computador efectua com grande rapidez. Fazemos somente uma divisão quando paramos a iteração. Neste caso a convergência é lenta porque* $|\frac{z_2}{z_1}| = 0.874... \sim 1$.
Se aplicássemos o mesmo método à equação

$$x^5 - 10x^4 - 2x^3 - 4x^2 + 2x + 12 = 0$$

obteríamos ao fim de apenas 10 iterações $y_{10} = \frac{1531095181}{149656315} = 10.23074222427...$ *valor que já está muito próximo da raiz dominante* $10.23074222431...$

2.7.2 Condicionamento no cálculo de raízes

O cálculo de raízes de uma função pode ser um problema mal condicionado. Basta reparar que pelo teorema de Lagrange temos

$$f(z) - f(\tilde{z}) = f'(\xi)(z - \tilde{z}) \Rightarrow \delta_{\tilde{z}} = \frac{e_{\tilde{z}}}{z} = \frac{-f(\tilde{z})}{zf'(\xi)},\ \text{com } \xi \in]z; \tilde{z}[$$

se a derivada for nula na raiz (zeros múltiplos) ou mesmo próximo da raiz, o erro relativo pode ser bastante elevado. Outro problema idêntico surge quando modificamos ligeiramente a função. Por exemplo, ao invés de calcularmos o zero z de f, calculamos o zero $\tilde{z}$ de $\tilde{f}$, a fórmula anterior aplica-se de maneira semelhante.

Mas podemos ser mais específicos no caso polinomial. Seguimos[22] [18], onde é provado que se tivermos um polinómio p com um zero simples z, e se considerarmos uma pequena perturbação $p_\varepsilon = p + \varepsilon q$, em que q é um outro polinómio, então para um pequeno ε obtemos um zero z_ε de p_ε que verifica,

$$z - z_\varepsilon \approx \frac{q(z)\,\varepsilon}{p'(z)}.$$

Esta estimativa resulta de considerar $\zeta(\varepsilon) = z_\varepsilon$, que é uma função analítica na vizinhança do zero, portanto

$$\frac{d}{d\varepsilon} p(\zeta(\varepsilon), \varepsilon) = 0 \Leftrightarrow \left(p'(\zeta(\varepsilon)) + \varepsilon q'(\zeta(\varepsilon))\right)\zeta'(\varepsilon) + q(\zeta(\varepsilon)) = 0,$$

[22] Um exemplo idêntico foi pela primeira vez mencionado por Wilkinson (1959).

e para $\varepsilon = 0$ ficamos com a igualdade $\zeta'(0) = -\frac{q(z)}{p'(z)}$.

É ainda dado um exemplo ilustrativo, considerando os polinómios

$$\begin{aligned} p(x) &= (x-1)(x-2)...(x-10) = x^{10} - 55x^9 + ... + 10!, \\ q(x) &= 55x^9. \end{aligned}$$

Considera-se $z = 10$, e para ε pequeno, pela fórmula anterior obtém-se

$$z - z_\varepsilon \approx \frac{q(10)}{p'(10)}\varepsilon = \frac{55 \times 10^9}{9!}\varepsilon \approx 1.5 \times 10^5 \varepsilon,$$

ou seja, basta uma pequena perturbação no coeficiente da nona potência para haver enormes alterações na raiz.

Não entrando em maiores detalhes, em [18], conclui-se que o problema do cálculo de zeros de p é mal condicionado, e que *uma aproximação fiável dos zeros é impossível.*

Mas, vejamos com maior detalhe, de acordo com as definições dadas anteriormente, em que sentido isto é verdade.

Repare-se que, neste caso, o erro relativo dos resultados é:

$$\delta_{z_\varepsilon} = \frac{z - z_\varepsilon}{z} \approx 1.5 \times 10^4 \varepsilon,$$

mas para concluirmos que se trata de um problema mal condicionado, é preciso comparar com o erro relativo dos dados. E aqui convém esclarecer o que se considera dado, o polinómio ou os seus coeficientes?

Para analisarmos a questão do condicionamento em termos do polinómio deveríamos estabelecer uma definição de erro relativo baseada numa norma para a função (ver apêndice). Não querendo entrar nesse detalhe neste momento apenas notamos que, por exemplo, se considerássemos o valor do erro relativo, respeitante ao valor da função num ponto x próximo da raiz, iríamos obter

$$\delta_{p_\varepsilon(x)} = \frac{p(x) - p_\varepsilon(x)}{p(x)} = \frac{-\varepsilon q(x)}{p(x)},$$

o que dará um valor extremamente elevado para $\left|\delta_{p_\varepsilon(x)}\right|$, porque $q(x)$ é elevado e $p(x)$ será próximo de zero. Neste sentido, como o valor do erro relativo dos dados seria elevado, e o problema não seria a priori mal condicionado.

O problema é mal condicionado se considerarmos que os dados são os coeficientes dos polinómios (na sua forma canónica). Atendendo a que o único coeficiente que varia neste caso é o coeficiente a_9, da nona potência, vemos que

$$\delta_{\tilde{a}_9} = \frac{a_9 - \tilde{a}_9}{a_9} = \frac{-55 - (-55 + \varepsilon 55)}{55} = -\varepsilon,$$

ou seja, podemos estabelecer

$$|\delta_{z_\varepsilon}| \approx 1.5 \times 10^4 |\delta_{\tilde{a}_9}|,$$

o que nos indica um número de condição maior que um milhão por cento.

Isto ilustra bem como o problema pode ser *mal condicionado,* considerando *a variação nos coeficientes.*

• Como curiosidade, testámos os resultados apresentados pela rotina NSolve. Para $\varepsilon = 10^{-3}$, em a estimativa apontaria para um valor $|z_\varepsilon - 10| \approx 150$, o valor mais próximo apresentado é $12.1 \pm 3.7i$. Para $\varepsilon = 10^{-5}$, a estimativa apontaria para $|z_\varepsilon - 10| \approx 1.5$, e o valor apresentado é $9.99 \pm 0.92i$. O último teste, para $\varepsilon = 10^{-7}$ deveríamos obter $|z_\varepsilon - 10| \approx 0.015$ e o valor apresentado é 9.98438... finalmente um valor próximo!

Contrariamente ao que se poderia supor numa primeira análise, o problema neste caso não é da rotina NSolve (que apresenta os valores com pelo menos 12 dígitos correctos... ainda que aponte 16 para Precision), mas sim do uso da estimativa. A estimativa admite que a função ζ é analítica na vizinhança do zero simples, e isso pressupõe que na variação de $\zeta(0)$ para $\zeta(\varepsilon)$ não ocorram zeros duplos e o subsequente aparecimento de raízes complexas conjugadas. No exemplo considerado, isso ocorre quando $\varepsilon > 10^{-6}$, portanto os valores apresentados pela estimativa para $\varepsilon = 10^{-3}, 10^{-5}$ não estão a priori correctos. Apenas nos devemos fiar na estimativa quando $\varepsilon = 10^{-7}$, que dá um valor razoavelmente semelhante ao apresentado pela rotina NSolve, ora neste caso $\delta_{z_\varepsilon} = 1.5 \times 10^{-3}$, o que até parece um erro relativo pequeno, mas não é, se comparado com o erro relativo δ_{p_ε} efectuado na perturbação do coeficiente, $|\delta_{\tilde{a}_9}| = \varepsilon = 10^{-7}$, o que nos dá o número de condição enorme já mencionado.

• A este propósito, um exemplo bem mais elucidativo, é aquele que foi apresentado no final do primeiro capítulo, acerca da determinação de valores próprios de uma matriz. Ao calcular o polinómio característico, o erro que aparecer nos coeficientes do polinómio pode gerar erros subsequentes no cálculo da raízes e consequentemente nos valores próprios.

2.8 Generalização a raízes complexas

A generalização de alguns métodos iterativos para a determinação de raízes complexas é possível, especialmente no caso de métodos derivados do método do ponto fixo, como o método de Newton ou os métodos quasi-Newton (...secante, Steffenson). Iremos abordar superficialmente o problema da determinação de zeros de funções complexas analíticas, já que uma outra abordagem requer um conhecimento da teoria de funções de variável complexa (por exemplo, a utilização do príncipio da variação do argumento para determinar o número de zeros existentes numa determinada região do plano complexo).

Consideremos ainda o exemplo do parágrafo anterior, em que aproximámos as raízes reais de

$$x^5 - 3x + 1 = 0,$$

se aplicarmos o método de Newton, mas tomando agora uma iterada inicial complexa, $z_0 = \frac{3}{2}i$, obtemos a sucessão

$$z_0 = 1.5\, i \; ; \quad z_{n+1} = z_n - \frac{1 - 3z_n + z_n^5}{5z_n^4 - 3}$$

cujos termos,

$$z_1 = -0.0448179... - 1.36134...\, i, \;\; z_2 = -0.0762498... - 1.32792\, ...i,$$
$$z_3 = -0.0803047.. - 1.32833...\, i, \text{ etc...}$$

convergem para a solução complexa $-0.080295100117... - 1.3283551098... i$ (também com ordem de convergência quadrática!...)

No entanto, se começarmos com $z_0 = \frac{1}{2}(1+i)$, os termos

$$z_1 = 0.352941... + 0.117647... i, \quad z_2 = 0.336678... + 0.00083115...i,$$
$$z_3 = 0.334734... - 0.41976 \times 10^{-6}... i, \text{ etc...}$$

convergem agora para a solução real 0.334734141943...

Podemos interrogar-nos se, sob condições semelhantes às do caso real, podemos assegurar a convergência de métodos de ponto fixo.

Isto pode ser conseguido, começando por definir que g é uma função contractiva num conjunto $D \subseteq \mathbb{C}$, se existir $L < 1$:

$$|g(a) - g(b)| \leq L|a - b|, \quad \forall a, b \in D,$$

em que o $|.|$ é o módulo nos complexos (que corresponde à norma euclidiana em $\mathbb{R}^2$).

Teorema 2.14 *(do Ponto Fixo em $\mathbb{C}$). Seja $g : D \subseteq \mathbb{C} \to \mathbb{C}$, onde D é um conjunto fechado em $\mathbb{C}$ tal que $g(D) \subseteq D$. Se g é contractiva em D, então:*
i) Existe um e um só $z \in D :\ g(z) = z$
ii) A sucessão $z_{n+1} = g(z_n)$ converge para z desde que $z_0 \in D$.
iii) As estimativas apresentadas no teorema do ponto fixo em $\mathbb{R}$ são ainda válidas (considerando o módulo definido nos complexos).

Demonstração: Iremos ver a demonstração deste resultado num contexto muito mais geral (em espaços de Banach), pelo que se trata de um exercício *a posteriori*. ■

Como a contractividade nem sempre é fácil de verificar, vamos apresentar um resultado semelhante ao que estabelecemos no caso real, em que exigiamos apenas que a derivada fosse inferior a um certo $L < 1$. Para demonstrarmos esse resultado usámos o teorema do valor médio de Lagrange, cuja generalização a funções complexas não é semelhante, com efeito apenas podemos estabelecer:

$$g(a) - g(b) = Re(g'(\xi)(b-a)) + Im(g'(\eta)(b-a))\mathbf{i}$$

para certos ξ, η pertencentes ao segmento no plano complexo que une a e b. Como os valores ξ e η não são necessariamente iguais, o resultado não é igual.

No entanto, podemos estabelecer, para funções complexas diferenciáveis (analíticas):

Proposição 2.7 *Seja g analítica num conjunto D, convexo de $\mathbb{C}$. Se $|g'(x)| \leq L\ \forall x \in D$, então verifica-se*

$$|g(a) - g(b)| \leq L|a - b|, \quad \forall a, b \in D$$

Demonstração:
Consideremos $a, b \in D$ e uma função

$$\begin{array}{rcl} h : [0,1] & \rightarrow & \mathbb{C} \\ t & \rightarrow & a + (b-a)t \end{array}$$

que parametriza o segmento que une a e b.

Se considerarmos um $w \in \mathbb{C}\backslash\{0\}$ constante e

$$\begin{array}{rcl} f : [0,1] & \rightarrow & \mathbb{C} \\ t & \rightarrow & wg(h(t)) \end{array}$$

obtemos $f'(t) = wg'(h(t))(b-a)$ pela regra de derivação da composição em $\mathbb{C}$.

A componente real da função f, que designamos f_1 é uma função real de variável real e aplicando o teorema de Lagrange, obtemos

$$f_1(1) - f_1(0) = f_1'(t_w) = Re(wg'(\xi_w)(b-a))$$

em que $t_w \in]0,1[$ e $\xi_w = h(t_w)$ é um ponto no segmento que une a e b. Como supomos que D é convexo, $\xi_w \in D$.

Por outro lado, como $f_1(1) - f_1(0) = Re(w(g(b) - g(a))$, temos

$$Re(w(g(b) - g(a)) = Re(wg'(\xi_w)(b-a)).$$

Para um qualquer $z \in \mathbb{C}$ não é dificil ver que

$$|z| = \sup_{|w|=1} |Re(\bar{w}z)|$$

aplicando Cauchy-Schwarz (pois usamos a topologia de $\mathbb{R}^2$ e $Re(\bar{w}z)$ corresponde a efectuar o produto interno em $\mathbb{R}^2$ dos vectores w e z) para demonstrar $\geq$, e considerando $w = z/|z|$ para demonstrar $\leq$.

Assim, aplicando as igualdades anteriores, temos

$$|g(b) - g(a)| = \sup_{|w|=1} |Re(\bar{w}(g(b) - g(a))| = \sup_{|w|=1} |Re(\bar{w}g'(\xi_w)(b-a))|,$$

e pela desigualdade de Cauchy-Schwarz,

$$\leq \sup_{|w|=1} |g'(\xi_w)|\,|b-a|.$$

Como $|g'(\xi_w)| \leq L$ por hipótese, porque $\xi_w \in D$, obtemos o resultado. ■

Através deste resultado concluímos que se tivermos uma função analítica, cuja derivada[23] é em módulo inferior a um $L < 1$ num conjunto convexo D, a função é contractiva em D.

Corolário 2.3 *Seja g analítica num conjunto D, convexo e fechado de $\mathbb{C}$, tal que $g(D) \subseteq D$. Se $|g'(x)| \leq L < 1\ \forall x \in D$, então as condições do teorema do ponto fixo verificam-se.*

[23]Para verificar que a derivada de uma função analítica é em módulo menor que 1, pode ser útil usar a estimativa de Cauchy,

$$|g'(w)| \leq \frac{\max_{|x-w|=r} |g(x)|}{r}.$$

(Note-se que a função g tem que ser analítica em $B(w,r)$)

2.8.1 Método de Newton nos complexos

O método de Newton pode ser considerado em $\mathbb{C}$, como um caso particular do método do ponto fixo. Com efeito, se a função for analítica numa vizinhança da raiz z e se $f'(z) \neq 0$, estabelecemos a equivalência

$$f(z) = 0 \Leftrightarrow z = z - \frac{f(z)}{f'(z)}.$$

e podemos aplicar a iteração do ponto fixo, pelo que o método de Newton pode esquematizar-se novamente

$$\begin{cases} \text{Iterada inicial} : z_0 \in \mathbb{C} \\ \text{Iterar } z_{n+1} = z_n - \frac{f(z_n)}{f'(z_n)}, \end{cases}$$

assegurando que $f'(z_n) \neq 0$, o que pode ser conseguido se tomarmos z_0 suficientemente próximo de z e se $f'(z) \neq 0$. Ou seja, pode ser provado um teorema de convergência local, que deixamos como exercício.

Exercício 2.7 *Seja f analítica numa vizinhança V_z duma raiz $z \in \mathbb{C}$, tal que $f'(x) \neq 0, \forall x \in V_z$, mostre que o método de Newton converge para essa raiz se considerarmos z_0 suficientemente próximo de z. Observe ainda que a convergência do método de Newton, nestas condições, é quadrática.*

Também é possível estabelecer condições para assegurar a convergência num certo conjunto (numa bola que contenha a raiz), mas são substancialmente mais complicadas de provar (cf. [10]):

Proposição 2.8 *Se f é uma função analítica na bola $\bar{B}(z_0, R) = \{z \in \mathbb{C} : |z - z_0| \leq R\}$ tal que $\max_{z \in \bar{B}(z_0,R)} |f''(z)| \leq M$,*

$$C = 2M \frac{|f(z_0)|}{|f'(z_0)|^2} \leq 1$$

com $\frac{|f(z_0)|}{|f'(z_0)|} \leq \frac{R}{2}$, então existe um e um só zero de f em $\bar{B}(z_0, R)$ e o método de Newton converge, tendo-se $|e_n| \leq \frac{R}{2^n} C^{2^n - 1}$. ■

Observação *(domínio de convergência).*

Para mostrarmos como pode ser complicado determinar o domínio em que o método do ponto fixo converge, recorremos a exemplos sobejamente conhecidos (visualmente).

(i) *Conjunto de Mandelbrot.* Consideremos a iteração $z_0 = 0, \;\; z_{n+1} = z_n^2 + c$.

A função g é neste caso $g(x) = x^2 + c$.

Vejamos que g está nas condições do teorema do ponto fixo em $D = \bar{B}(0, r)$, com $r < \frac{1}{2}$, desde que $|c| \leq \frac{1}{4}$.

D é fechado e convexo e dado $|x| \leq r$, temos $|g(x)| = |x^2 + c| \leq |x|^2 + |c| \leq r^2 + \frac{1}{4} < \frac{1}{2}$.

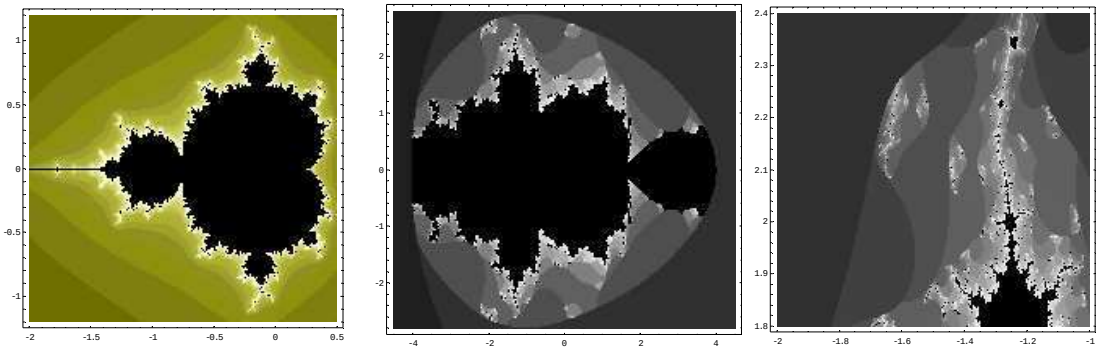
Portanto concluímos que $x \in D \Rightarrow g(x) \in D$, ou seja $g(D) \subseteq D$.

Falta ver que g é contractiva em D. Ora, $|g'(x)| = |2x| \leq 2r < 1$, e portanto podemos concluir pelo teorema do ponto fixo que a iteração apresentada é convergente desde que $|c| \leq 0.25$.

É claro que este resultado é apenas uma condição suficiente, não nos dá todos os valores para os quais há convergência, e podemos ver quão longe estamos desse objectivo se visualizarmos a figura em baixo, à esquerda. Trata-se do conjunto de Mandelbrot, onde se representam a negro os pontos $c \in \mathbb{C}$ para os quais a sucessão z_n não tende para infinito. A complexidade do conjunto é evidente (ainda que se trate de um *fractal,* ie. a estrutura macroscópica é repetida em porções mais pequenas da fronteira, característica de transformações nos complexos).

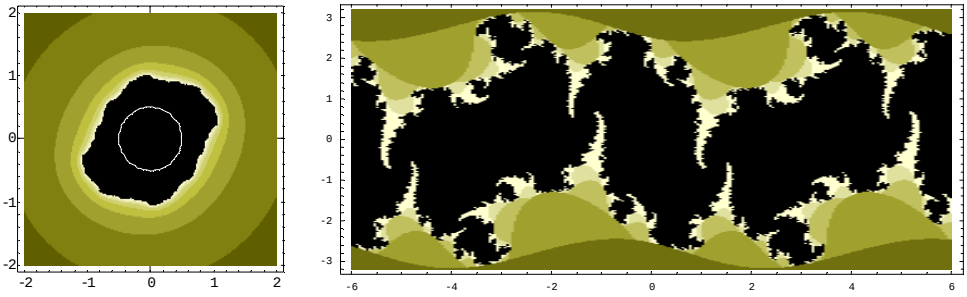
O domínio D obtido é apenas uma pequena bola dentro desse conjunto.

Este tipo de fenómeno, repete-se para a iteração de outras funções. Como exemplo, na figura central representamos o que foi obtido considerando $z_{n+1} = -\cos(z_n) + c$, e na figura da direita aumentamos uma zona da fronteira. O assunto foi muito publicitado durante a última década e encheram-se livros com retratos da fronteira do conjunto de Mandelbrot (com $g(x) = x^2 + c$), como se se tratasse de uma autêntica exploração de paisagens desconhecidas. Uma maior compreensão do assunto, nomeadamente no que diz respeito à fractalidade e à possibilidade de determinar a dimensão da fronteira, são assunto da teoria de sistemas dinâmicos discretos.



(ii) *Conjunto de Julia.* Uma outra possibilidade é manter o valor de c constante e fazer variar o valor da iterada inicial z_0. Na figura em baixo, à esquerda, representamos o conjunto obtido com $c = \frac{i}{4}$. O conjunto poderia ser mais interessante visualmente se tivéssemos escolhido outros valores para c, mas interessou-nos considerar aquele valor porque podemos comparar com o domínio D obtido teoricamente (que é a bola incluída no conjunto).

Na figura em baixo, à direita, representamos então um conjunto mais interessante, que se obteve para $g(x) = -\cos(x) + c$, escolhendo $c = i - 1$.



2.9 Exercícios

1. Considere a equação

$$x^3 + \frac{1}{2^m} = x + \frac{1}{8^m}.$$

a) Mostre que para qualquer $m \geq 0$ há apenas uma raiz em $[0, \frac{1}{2}]$.

b) Mostre que se $m \in \mathbb{N}$ o método da bissecção atinge essa raiz num número finito de iteradas. Quantas?

c) O mesmo que em b) para a equação $x^3+2x+2^{-m} = 3x^2+8^{-m}$, aplicando ao intervalo $[1, \frac{3}{2}]$.

2. Considere a função f dada por

$$f(s) = 1 - \sum_{k=1}^{M} \frac{1}{\pi(k)^s}$$

em que $\pi(k)$ dá o k-ésimo primo $(2, 3, 5, 7, 11,$ etc...).

Sabe-se que $f(1)$ diverge quando $M \to \infty$, e é óbvio que converge quando $s > 1$, pois o valor da soma é majorado pela função zeta de Riemann

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

Para $M = 5000$, foram obtidos os valores $f(1.3) = -0.18, f(1.5) = 0.15$. As funções f, f', f'' são estritamente monótonas, tendo-se, $f'(1.3) = 2.19, f'(1.5) = 1.26, f''(1.3) = -6.85, f''(1.5) = -3.05$.

Pretendendo aplicar o método da falsa posição para determinar o zero de f, determine:

a) Qual o extremo que se iria manter constante.

b) Calcule uma estimativa do número de iteradas necessário para garantir um erro absoluto inferior a 10^{-4}.

c) Qual método seria a priori mais rápido, o da bissecção ou o da falsa posição? O que seria necessário efectuar para aumentar a rapidez do método da falsa posição?

3. Seja g uma função contínua tal que $g(a) = b$ e $g(b) = a$.

a) Mostre que existe pelo menos um ponto fixo de g em $[a, b]$.

b) Mostre que se $g \in C^1[a, b]$ então a derivada de g toma o valor -1 em algum ponto desse intervalo.

4. Pretendendo determinar $z = g(z)$, estabeleceu-se a equivalência, para $\theta \neq 0$,

$$x = g(x) \Leftrightarrow x = (1 - \theta)x + \theta g(x),$$

considerando a nova função $G(x) = (1 - \theta)x + \theta g(x)$ como função iteradora *(método de relaxação)*, que é uma combinação convexa da função identidade com g, quando $\theta \in]0, 1[$.

a) Verifique que quando $g'(z) \neq 1$ é conhecido, se escolher

$$\theta = \frac{1}{1 - g'(z)}$$

então o método do ponto fixo aplicado a G terá ordem de convergência quadrática, local.

(Note que o objectivo é encontrar z, e portanto a priori não pode ser conhecido $g'(z)$).

b) Suponha que o método do ponto fixo aplicado à função iteradora g gera uma sucessão (x_n) que converge lentamente, com $K_\infty < 1$, mas longe de zero. Considere três valores x_{m+1}, x_m, x_{m-1}, para m razoavelmente grande. Mostre que se tomar

$$\theta = \frac{x_{m-1} - x_m}{x_{m+1} - 2x_m + x_{m-1}}$$

a iteração

$$y_{n+1} = (1-\theta)y_n + \theta g(y_n)$$

convergirá mais rapidamente.

c) Aplique este método para acelerar a convergência de $x_{n+1} = \sin(\frac{4}{5}x_n) + \frac{1}{5}$. Considere $x_0 = 0$, e calcule até $m = 10$. Partindo desses valores calcule 5 iteradas com y_n. e compare com o que teria obtido se tivesse calculado novas 5 iteradas com x_n.

d) Reparando que podemos aplicar o mesmo processo a G, vemos que estamos perante um método conhecido. Mostre que se considerar $y_n = x_m$, o valor obtido em b) para y_{n+1} corresponde ao valor obtido na extrapolação de Aitken.

5. Considere um intervalo $I = [a, b]$ que tem um único ponto fixo z de uma função $g \in C^1(I)$. Seja $g'(z) = 1$.

a) Mostre que se $0 < g'(x) < 1, \forall x \in I \backslash \{z\}$, então o método do ponto fixo converge qualquer que seja $x_0 \in I$.

Sugestão: Verifique que a sucessão definida pelo método do ponto fixo é estritamente monótona e limitada.

b) Aplique este resultado para mostrar que $x_{n+1} = \sin(x_n)$ converge para 0, qualquer que seja $x_0 \in \mathbb{R}$.

6. Pretende-se resolver a equação $f(x) = 0$ em que $f \in C^1(\mathbb{R})$.

a) Mostre que se a função verificar

$$-b \leq f'(x) \leq -a < 0 \text{ ou } 0 < a \leq f'(x) \leq b$$

tem um único zero em $\mathbb{R}$, e que podemos encontrar ω tal que o método iterativo

$$x_{n+1} = x_n + \omega f(x_n)$$

converge para esse zero, qualquer que seja $x_0 \in \mathbb{R}$.

b) Aplique o método anterior a $f(x) = 3x - \cos(2x)/2 + \sin(4x)/4$

7. Seja f uma função diferenciável em $\mathbb{R}$. Pretende-se estabelecer uma equivalência em todo $\mathbb{R}$,

$$f(x) = 0 \Leftrightarrow x = g(x),$$

em que g é diferenciável em $\mathbb{R}$ e $g'(x) < 0, \forall x \in \mathbb{R}$.

Mostre que existem funções f para as quais é impossível estabelecer essa equivalência.

Sugestão: Verifique nessas condições g tem um único ponto fixo.

(Isto significa que é escusado procurar um procedimento geral que encontre funções iteradoras cuja derivada é sempre negativa – em particular, funções iteradoras que levem

a uma convergência alternada. Esta procura teria sentido já que isso permitiria encontrar intervalos que contivessem a raiz!)

8. O processo iterativo

$$x_0 = 1, \quad x_{n+1} = \frac{(p-1)}{p} x_n + \frac{a}{p x_n^{p-1}}$$

permite obter em poucas iteradas uma boa aproximação de

$$\sqrt[p]{a}$$

em que $p, a > 1$. Escolha um intervalo adequado em que estejam satisfeitas as condições suficientes para a convergência do método de Newton.

Calcule uma aproximação de $\sqrt[3]{231}$ garantindo um erro absoluto inferior a 10^{-3}.

Observação: Claro que a aproximação inicial $x_0 = 1$ não é boa e pode ser melhorada... por exemplo, considerando $x_0 = 1 + \frac{a}{p}$, ou simplesmente encontrando um x_0 tal que $x_0^p < a < (x_0+1)^p$... (para $\sqrt[3]{231}$ bastaria reparar que $6^3 = 216 < 231 < 343 = 7^3$ e considerar $x_0 = 6$).

Uma utilidade deste resultado é a de permitir calcular aproximações de raízes de ordem p utilizando apenas operações elementares (... por exemplo, quando estamos na posse de uma calculadora rudimentar que não calcula raízes!).

Acontece, por vezes, que as raízes calculadas pela máquina aparecem só em precisão simples, podendo servir esse valor como aproximação inicial para este processo iterativo, que ao fim de muito poucas iterações (normalmente 1 ou 2) garante todos os dígitos em precisão dupla, ou mesmo superior!

Com efeito, como nesse caso a aproximação já é bastante boa, $|e_{n+1}| \sim \frac{p(p-1)a^{(p-2)/p}}{2pa^{(p-1)/p}}|e_n|^2$, ou seja $|e_{n+1}| \sim \frac{p-1}{2\sqrt[p]{a}}|e_n|^2$ é possível duplicar a precisão numa única iteração, desde que $\frac{p-1}{2\sqrt[p]{a}}$ não seja muito maior que 1, o que acontece normalmente para valores não muito elevados de p.

9. Mostre que se tivermos uma equação algébrica

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n = 0,$$

onde os coeficientes $a_i \in \mathbb{R}$, não é possível verificar as condições do teorema do ponto fixo usando um conjunto $D = \{z \in \mathbb{C} : 0 \leq \alpha \leq |z| \leq \beta\}$ para determinar uma raiz não real de $p(x) = 0$.

10. É evidente que se $x \neq 0$ é solução da equação algébrica

$$P(x) = a_0 + a_1 x + \ldots + a_n x^n = 0,$$

então $y = 1/x$ é solução da equação

$$Q(y) = a_0 y^n + \ldots + a_{n-1} y + a_n = 0.$$

a) Supondo que $a_0, a_n \neq 0$, mostre que o número de zeros reais (respect. complexos) de P é igual ao de Q.

b) Conclua que basta analisar P e Q no circulo $\{z \in \mathbb{C} : |z| \leq 1\}$ para localizarmos todos os zeros desses polinómios em $\mathbb{C}$.

c) Quantas raízes complexas tem a equação $x^6 + ax^4 + ax^2 + 1 = 0$, no caso de possuir uma única raiz real em $[-1, 1]$?

d) Mostre que a equação $4 - (1 + 2i)x - 64x^2 + (16 + 32i)x^3 = 0$ tem uma e uma só raiz em $\{z \in \mathbb{C} : |z| > 1\}$.

11. Considere a sucessão em $\mathbb{C}$,

$$z_{n+1} = (i\, z_n^p + 1)/(p + 1)$$

em que $z_0 = 0$ e p é um número natural.

a) Mostre que esta sucessão converge para um valor em $\{z \in \mathbb{C} : |z| \leq 1\}$.

b) Se quisermos calcular uma aproximação do limite da sucessão, ao fim de quantas iterações podemos garantir um erro $|e_n| \leq 0.01$?

Considere $p = 7$ e determine uma aproximação que satisfaça essa majoração.

c) Usando o exercício anterior, e as alíneas anteriores, mostre que

$$i - (p + 1)z^{p-1} + z^p = 0$$

tem uma única raiz tal que $|z| > 1$ e indique um valor aproximado da raiz para $p = 7$.

---

12. Mostre que a equação $z - \sin(zX) - \sin(Y) = 0$ tem um único ponto fixo para os parâmetros $X \in]-1, 1[,\ \forall Y \in \mathbb{R}$, e que a sucessão

$$z_{n+1} = \sin(z_n X) + \sin(Y)$$

converge se $z_0 = 0.1$

b) Considere $X = 2.5$ e $Y = 0$. Experimentalmente verifique que a sucessão (z_n) vai ter dois sublimites: $\alpha = 0.997934...$ e $\beta = 0.602602...$

Justifique a inexistência de limite através do comportamento da função iteradora e mostre que existe uma raiz da equação em $[\alpha, \beta]$.

13. Considere $f(x) = 0 \Leftrightarrow x = g(x)$ uma equação em $\mathbb{R}$ que tem pelo menos duas raízes z_1 e z_2 consecutivas (ou seja, não existe nenhuma outra raiz entre elas).

a) Mostre que se $g \in C^1(\mathbb{R})$ e $|g'(z_1)| < 1$ então $g'(z_2) \geq 1$.

b) Suponha que $z_2 \in I = [a, b]$, que $|g'(x)| > 1, \forall x \in I$ e que $I \subseteq g(I)$. Mostre que o método $x_{n+1} = g^{-1}(x_n)$ converge para z_2 qualquer que seja $x_0 \in I$.

c) Seja $f \in C^p(\mathbb{R})$, tal que a raiz z_2 tem multiplicidade $p \geq 1$, e seja g tal que $g'(z_2) > 1$. Indique uma função iteradora que assegure uma convergência local linear para z_2, e uma outra que assegure convergência quadrática, para cada caso de p.

14. Considere o intervalo $I = [a, b] \subset \mathbb{R}$, e as funções $g, h \in C^1(I)$ tais que $g \circ h \neq h \circ g$.

Sabemos que $g(I) \subseteq I, h(I) \subseteq I$, e que

$$|g'(x)| \leq L_1,\ \ |h'(x)| \leq L_2,\ \ \forall x \in I,$$

com $L_1 L_2 < 1$.

a) Se estabelecermos em I:

$$x = g(x) \Leftrightarrow x = h(x) \Leftrightarrow f(x) = 0,$$

mostre que existe uma única raiz $z \in I$ de $f(x) = 0$ e indique (justificando) três funções iteradoras distintas que assegurem uma convergência para essa raiz, qualquer que seja $x_0 \in I$.

b) Supondo que $a < 0$ e $b \geq 0$, mostre que o zero z da função f em I verifica:

$$|z| \leq \frac{\min\{|h(g(0))|, |g(h(0))|\}}{1 - L_1 L_2}$$

c) Suponha agora que os pontos fixos de g e h em I são diferentes. A equação $x = g(h(x))$ tem uma solução única em I? Justifique.

d) Mostre que as equações:

$$2x - \cos(\frac{0.5}{x^2+1}) = \sin(\frac{1}{x^2+1})$$

$$x = \frac{4}{(\cos(x/2) + \sin(x))^2 + 4}$$

têm uma única solução no intervalo $[0, 1]$ e indique funções iteradoras convergentes para a solução.

Sugestão: Usar funções g e h apropriadas e aplicar o anterior.

15. Considere uma função g contínua no intervalo $[0, 1]$, diferenciável, que verifica:

$$g(0) = \frac{1}{2}, \qquad g'(x) = \frac{1}{2} x \cos^2(g(x))$$

Mostre que a função g tem um único ponto fixo no intervalo $[0, 1]$.

16. Sabendo que $h(x), h'(x) \in C^1([-1, 1])$ são crescentes, e que h tem uma raiz em $[-1, 1]$, pretende-se determinar a raiz da equação

$$F(x) = x + h(x) = 0$$

usando o método

$$\begin{cases} x_0 = a, \ x_1 = b \\ x_{n+1} = x_n - \frac{(x_n - x_{n-1}) F(x_n)}{F(x_n) - F(x_{n-1})} \end{cases}$$

Mostre que F tem uma única raiz em I e que existem $a, b \in I$ para os quais há convergência. Qual a ordem de convergência?

17. Considere a equação

$$4z^3 - (1 + 2i)z^2 - 64z + 16 + 32i = 0$$

Escreva-a na forma $z = g(z)$ por forma a que se verifiquem as condições do teorema do ponto fixo no conjunto $D = \{z \in \mathbb{C} : |z| \leq 1\}$.

Determine z_1 e z_2 a partir de $z_0 = 0$ e calcule a raiz da equção com menor módulo.

18. Considere a equação

$$\cos(z) - 2z - \frac{e^{-iz}}{2} = 0$$

Aplique o Teorema do Ponto Fixo, a uma função iteradora conveniente que convirja para a solução situada no subconjunto $D = \{z \in \mathbb{C} : Im(z) \geq -1\}$.

19. Considere a equação

$$a + \sin(x) + x^2 = 0$$

a) Seja $a = -1$. Verifique que as condições suficientes para a convergência do Método de Newton estão asseguradas no intervalo $[-2, -1]$.

b) Indique um intervalo para valores de a em que essas condições estejam asseguradas.

c) Seja c a solução da equação $\cos(x) = -2x$. Mostre que se considerarmos $a = -\sin(c) - c^2$, o método de Newton converge linearmente para a raiz da solução, se considerarmos uma iterada inicial suficientemente próxima. Indique uma modificação do método, por forma a obter convergência quadrática.

20. Ao utilizar o método do ponto fixo para determinar uma raiz de uma equação, foram obtidos os seguintes valores

$$x_3 = -0.914260304;\ x_4 = -0.825329540;\ x_5 = -0.884002249;\ x_6 = -0.847330076$$

a) Sabendo que a função iteradora era um polinómio do quarto grau, da forma $p(x) = \alpha x^4 + \beta x^2 + \gamma$ determine aproximadamente as duas raízes reais da equação.

b) Determine os valores possíveis para x_2.

c) Determine uma estimativa para a majoração do erro absoluto em x_{20}.

21. Considere a equação polinomial

$$p(x) = x^5 - 6x^4 + 4x^3 + 12x^2 - 5x - 5 = 0$$

a) Indique a coroa no plano complexo que contem todas as raizes.

b) Usando a regra de Budan-Fourier, determine intervalos onde se encontrem as raizes reais, e conclua que existe uma raiz dominante.

c) Determine aproximadamente o valor da raiz dominante, usando o método de Bernoulli.

22. Admitindo que o corte topográfico da costa é dado pela função $p(x) = (x^5 + x^3 - 8x^2 + 3)/4$, pretende-se determinar se um veleiro com a vigia a uma altura $h = 0.1$ consegue observar sempre o farol com altura= 0.25, situado no topo da ilha.

Para isso comece por determinar:

a) A primeira raiz positiva de p com um erro relativo inferior a 0.005.

b) Determine um ponto $x \in [0, 1]$ tal que a recta tangente a p nesse ponto passe também pelo ponto $(0, 1)$ (com um erro inferior a 0.001).

c) Determine a intersecção da recta tangente da alínea b) com a recta $y = h$ e conclua acerca da questão colocada no início do enunciado. *(Caso não tenha resolvido a alínea anterior, considere $x \sim 0.37$).*

d) Sabendo que $z_1 \sim -0.58$, $z_2 \sim 0.65$, $z_3 \sim 1.74$ são raízes de p, determine aproximadamente as suas raízes complexas. *Sugestão: Lembre que $p(x) = a_n(x - z_1)\ldots(x - z_n)$.*

23. Considere $g \in C^1(\mathbb{R})$ uma função limitada e estritamente decrescente. Mostre que existe um único ponto fixo de g em $\mathbb{R}$. Conclua que também existe um único ponto fixo de $-g$ quando g é limitada e estritamente crescente. É necessário exigir que a função seja estritamente monótona em $\mathbb{R}$ ou bastará exigir apenas num intervalo? Qual?

Capítulo 3

Teorema do Ponto Fixo de Banach

O objectivo deste capítulo é generalizar a aplicação do método do ponto fixo à resolução de equações num contexto muito mais geral do que aquele visto nas situações precedentes, em que as incógnitas eram apenas números reais ou complexos.

Por exemplo, sendo a incógnita uma função f contínua tal que

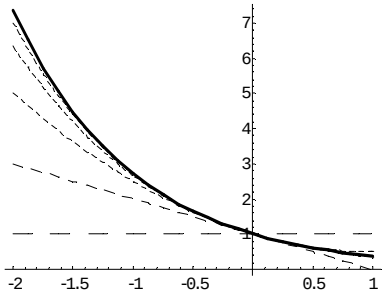
$$f(x) = 1 - \int_0^x f(t)dt$$

podemos pensar em aplicar o método do ponto fixo começando com $f_0 = 0$, fazendo

$$f_{n+1}(x) = 1 - \int_0^x f_n(t)dt,$$

o que dá $f_1(x) = 1,\ f_2(x) = 1 - x,\ f_3(x) = 1 - x + \frac{x^2}{2}$, etc... Neste caso, curiosamente, a sucessão de funções f_n vai reproduzir a expansão em série de Taylor da função e^{-x}, ou seja, a sucessão de funções f_n vai convergir para a solução $f(x) = e^{-x} = 1 - x + ... + \frac{(-x)^n}{n!} + ...$

Vemos na figura seguinte o resultado das 5 primeiras iterações, onde é vísivel a aproximação das funções $f_1, f_2, f_3, ...$ (representadas a tracejado) à função limite, que neste caso é $f(x) = e^{-x}$ (representada a cheio).



Por outro lado, sendo e^x ponto fixo da derivação, pois $e^x = (e^x)'$, e começando com $f_0 = 0$, ao efectuarmos $f_{n+1} = (f_n)'$ vamos obter sempre 0, que é um outro ponto fixo. Não é dificil ver que funções do tipo Ce^x serão os pontos fixos operador derivação. Portanto, podemos ter sucessões que convergem para os diferentes pontos fixos se fizermos $f_0 =$

$p_k(x) + Ce^x$, onde $p_k(x)$ é um polinómio de grau k, já que ao fim de $k+1$ iterações as derivações sucessivas anulam o polinómio. No entanto, se começarmos com $f_0 = \sin(x)$ vamos 'orbitar' entre co-senos e senos, não havendo convergência!

Interessa pois saber sob que condições poderemos garantir convergência, considerando equações mais gerais, em que as incógnitas podem ser números, vectores, sucessões, funções, etc... A liberdade para as incógnitas não é total! Para garantirmos a existência de um teorema do ponto fixo, num contexto tão geral, precisamos de ter uma estrutura (...um espaço) com propriedades mínimas que permitam um resultado de existência *construtivo*, como será o teorema do ponto fixo de Banach que iremos apresentar.

Uma estrutura suficientemente geral que permite obter esse resultado é a noção de *Espaço de Banach* (que são espaços vectoriais normados e completos), noção que iremos definir neste capítulo. Também poderíamos obter o teorema do ponto fixo de Banach para espaços métricos completos (como foi originalmente apresentado), mas preferimos considerar apenas espaços de Banach, já que a dedução é semelhante e mais simples, permitindo ainda apresentar resultados relativos à derivação de Fréchet.

3.1 Espaços Normados

Começamos por recordar a noção de espaço normado. Um espaço vectorial normado é, como o nome indica, um espaço vectorial a que associamos uma *norma*. A norma, sendo uma aplicação que toma valores reais, vai permitir introduzir no espaço vectorial uma topologia que resulta indirectamente da topologia de $\mathbb{R}$. Assim, a noção de norma, que é crucial neste contexto, vai generalizar o papel desempenhado pelo módulo nos reais (ou nos complexos).

Definição 3.1 *Seja E um espaço vectorial, em que o corpo dos escalares é $\mathbb{R}$ ou $\mathbb{C}$. Uma aplicação $||.|| : E \to [0, +\infty[$ designa-se* norma *se verificar:*
i) $||\alpha x|| = |\alpha| \, ||x|| \ , \ \forall x \in E, \ \forall \alpha \in \mathbb{R}$ *(ou $\mathbb{C}$),*
ii) $||x + y|| \leq ||x|| + ||y||, \ \forall x, y \in E$ *(desigualdade triangular),*
iii) $||x|| = 0 \Leftrightarrow x = 0$.

A um espaço vectorial E munido de uma norma $||.||$, chamamos *espaço vectorial normado* e indicamos $(E, ||.||)$ apenas em caso de ambiguidade. Normalmente apenas indicamos E, subentendendo qual a norma em questão. Quando indicarmos $||.||_E$ referimo-nos à norma no espaço $(E, ||.||)$.

Observações:
(i) A partir de uma norma, podemos definir imediatamente uma distância $d(x, y) = ||x - y||$, que nos permite quantificar uma certa proximidade entre dois elementos do espaço vectorial (beneficiando da relação de ordem existente nos reais). Consequentemente, fica estabelecida uma noção de vizinhança, que definirá a topologia.
(ii) É importante notar que estando definidas várias normas sobre um mesmo espaço vectorial, elas podem estabelecer um critério de proximidade diferente ... *ou seja, é importante estar subjacente qual a "norma" usada! Quando, ao longo do capítulo, escrevemos*

$x_n \to x$, *é fulcral termos presente segundo que norma isso acontece.* Iremos ver que se o espaço vectorial tiver dimensão finita, todas as normas aí definidas são *equivalentes*, mas isso não é válido para espaços com dimensão infinita... poderá acontecer que uma sucessão convirja segundo uma norma, mas não segundo outra!

(iii) Se no espaço vectorial estiver definido um produto interno $x \cdot y$, então a norma natural associada a esse produto interno é $||x|| = \sqrt{x \cdot x}$, e podemos usar a importante desigualdade de Cauchy-Schwarz:

$$|x \cdot y| \leq ||x||\ ||y||$$

(iv) Reparamos que a generalização da noção de módulo é explícita na propriedade $||\alpha x|| = |\alpha|\ ||x||$, pois precisamos da noção de módulo no corpo, para que esta propriedade se verifique.

Exercício 3.1 .
a) Mostre que em $\mathbb{R}^N$ *ou* $\mathbb{C}^N$ *são normas as aplicações*

$$||x||_\infty = \max\{|x_1|, ..., |x_N|\}$$
$$||x||_p = (|x_1|^p + ... + |x_N|)^{1/p}$$

b) Verifique que se considerarmos $u = (u_n)_{n\in\mathbb{N}}$ *uma sucessão de reais (ou complexos), a aplicação*

$$||u||_p = (|u_1|^p + ... + |u_n|^p + ...)^{1/p}$$

é uma norma no subespaço das sucessões

$$l^p = \{(u_n)_{n\in\mathbb{N}} : ||u||_p < +\infty\},$$

e que a aplicação

$$||u||_\infty = \sup\{|u_1|, ..., |u_n|, ...\}$$

é uma norma no sub-espaço das sucessões

$$l^\infty = \{(u_n)_{n\in\mathbb{N}} : ||u||_\infty < +\infty\}.$$

c) Da mesma forma, considerando o espaço das funções f *definidas num intervalo* I*, a menos de um conjunto com medida de Lebesgue nula (i.e: conjunto numerável), mostre que a aplicação*

$$||f||_p = \left(\int_I |f(x)|^p dx\right)^{1/p}$$

é uma norma no subespaço

$$L^p = \{f(x) : ||f||_p < +\infty\},$$

e que a aplicação

$$||f||_\infty = \sup_{x\in I} |f(x)|$$

é uma norma no espaço das funções contínuas $C(I)$ *quando* I *é compacto. (No contexto das funções* L^p*, a norma é definida usando o* supremo essencial).
Nota*: Admita a desigualdade triangular para as normas* $||x||_p$ *(conhecida como desigualdade de Minkowski).*□

3.1.1 Noções Topológicas em Espaços Normados

A norma define uma certa topologia num espaço normado. Devemos ter presente que num mesmo espaço vectorial podemos trabalhar com várias normas, e consequentemente com noções de proximidade diferentes, ou seja com topologias diferentes! A noção fundamental que define a topologia do espaço é a noção de vizinhança.

Definição 3.2 . *Designamos por ε-vizinhança de x ao conjunto* $V_\varepsilon(x) = \{y \in E : ||x - y|| < \varepsilon\}$.

– Um conjunto A é *aberto* em E se $\forall x \in A\ \exists \varepsilon > 0 : V_\varepsilon(x) \subseteq A$
– Um conjunto A é *fechado* se $E \backslash A$ for aberto.

Exercício 3.2 *Mostre que os conjuntos* $B(a,r) = \{x \in E : ||x - a|| < r\}$ *são conjuntos abertos, designados por* bolas abertas*, e que são conjuntos fechados* $\bar{B}(a,r) = \{x \in E : ||x - a|| \leq r\}$*, chamados* bolas fechadas.

Para além disso, reuniões de abertos são abertos e intersecções de fechados são fechados, mas só podemos garantir que intersecções de abertos são abertos, ou que reuniões de fechados são fechados se forem em número finito.

Nota importante: os conjuntos E e $\emptyset$ são abertos e fechados!

– Um conjunto A diz-se *limitado* se $\exists R \geq 0 : \forall x \in A, ||x|| \leq R$.
– Um conjunto A é *compacto* se toda a sucessão em A tem uma subsucessão convergente, com limite pertencente a A.

Num espaço de dimensão finita, se um conjunto for fechado e limitado é um *compacto*, mas em espaços de dimensão infinita isso nem sempre acontece[1].

Definição 3.3 *Uma sucessão* (x_n) *num espaço normado* E converge *para* $x \in E$*, e escrevemos* $x_n \to x$*, se*

$$||x_n - x|| \overset{n\to\infty}{\longrightarrow} 0$$

É claro que o limite, a existir, é único. Basta reparar que se x e y fossem limites da sucessão (x_n), então para qualquer $\delta > 0$ existe um n suficientemente grande tal que $||x - x_n|| < \delta, ||x_n - y|| < \delta$, logo $||x - y|| \leq ||x - x_n|| + ||x_n - y|| < 2\delta$. Ou seja, $\forall \delta > 0, ||x - y|| < 2\delta$, o que implica $x = y$.

[1]Basta pensar na bola $\bar{B}(0,1)$ no espaço l^∞. As sucessões $u^{(k)} \equiv (u_n^{(k)})$ tais que $u_n^{(k)} = \delta_{kn} = \begin{cases} 1 \text{ se } n=k \\ 0 \text{ se } n\neq k \end{cases}$ pertencem a $\bar{B}(0,1)$ mas não é possível extraír nenhuma subsucessão convergente da sucessão $u^{(k)}$ porque os elementos constituem uma base do espaço l^∞.

3.1.2 Normas equivalentes

Duas normas distintas dão geralmente valores diferentes para um mesmo elemento do espaço, no entanto, esta diferença quantitativa pode não reflectir uma diferença qualitativa, já que as propriedades topológicas podem revelar-se equivalentes. É neste quadro que iremos introduzir a noção de normas equivalentes e verificar que as normas em espaços de *dimensão finita* (p. ex. $\mathbb{R}^N$ ou $\mathbb{C}^N$) são equivalentes.

Definição 3.4 . *Duas normas $||.||$ e $|||.|||$, num mesmo espaço vectorial E, dizem-se equivalentes se existirem $C_1, C_2 > 0$ tais que:*

$$C_1||x|| \leq |||x||| \leq C_2||x|| \ , \ \forall x \in E \tag{3.1}$$

• Como é claro, esta noção de equivalência entre normas significa que as topologias também serão equivalentes, ou seja, que os abertos e fechados serão os mesmos, que um conjunto sendo limitado para uma norma também o será para outra, que a continuidade numa norma implica continuidade na outra, etc. (exercício).

Lema 3.1 *Seja E um espaço normado de dimensão finita. Então, qualquer que seja a norma $|||.|||$ em E, existe $R > 0$:*

$$|||x||| \leq R, \ \ \forall x \in \{||x||_\infty \leq 1\}$$

Demonstração:
Seja $\mathbf{e}^{(1)}, ..., \mathbf{e}^{(N)}$ uma base do espaço vectorial E, sabemos que sendo $x = x_1\mathbf{e}^{(1)} + ... + x_n\mathbf{e}^{(n)}$ então

$$|||x||| = |||x_1\mathbf{e}^{(1)} + ... + x_N\mathbf{e}^{(N)}||| \leq (|||\mathbf{e}^{(1)}||| + ... + |||\mathbf{e}^{(N)}|||) \max_{i=1,...,N} |x_i|.$$

Como $||x||_\infty = \max_{i=1,...,N} |x_i| \leq 1$ basta tomar $R = |||\mathbf{e}^{(1)}||| + ... + ||\mathbf{e}^{(N)}||| > 0$.■

Teorema 3.1 *As normas em espaços de dimensão finita são equivalentes.*

Demonstração:
Basta ver que qualquer norma $|||.|||$ é equivalente à norma $||.||_\infty$, devido à transitividade da relação de equivalência.

Consideremos o conjunto $S = \{x \in E : ||x||_\infty = 1\}$. S é um compacto na topologia de $|||.|||$, porque no lema anterior vimos que era limitado e, sendo fechado, isto é suficiente, num espaço de dimensão finita!

Como a norma é um operador contínuo, e S é compacto, vai existir um máximo e um mínimo (generalização do T. Weierstrass):

$$C_1 \leq |||x||| \leq C_2, \ \ \forall x \in S$$

e $C_1 > 0$ pois $|||x||| = 0$ sse $x = 0 \notin S$.

Ora, qualquer que seja $y \in E, y \neq 0$ podemos escrever $y = ||y||_\infty \frac{y}{||y||_\infty}$, onde $x = \frac{y}{||y||_\infty} \in S$. Portanto:

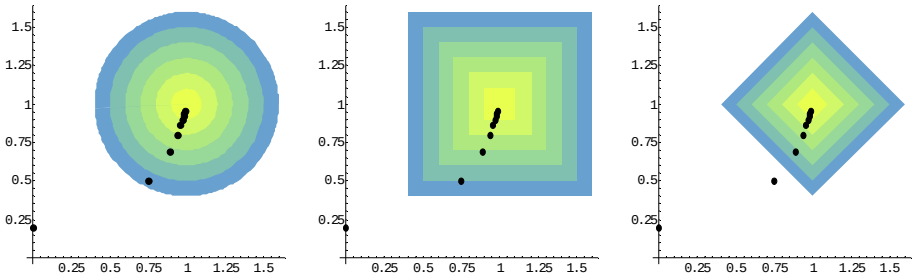
$$C_1 \leq |||\frac{y}{||y||_\infty}||| \leq C_2, \ \ \forall y \in E\backslash\{0\}$$

e obtemos, como pretendiamos,

$$C_1||y||_\infty \leq |||y||| \leq C_2||y||_\infty \ , \ \forall y \in E,$$

incluindo trivialmente o caso $y = 0$.■

Exemplo 3.1 *Consideremos a sucessão* $x_n = (1 - \frac{1}{n^2}, \frac{n^2}{n^2+4})$ *cujos pontos representamos nas três figuras em baixo. É óbvio que esta sucessão tende para o ponto* $x = (1, 1)$*, o que se pretende pôr em evidência é que isso acontece segundo qualquer uma norma em* R^2*, já que foi isso que acabou de ser demonstrado. Considerámos três normas diferentes (a que correspondem as três figuras), a norma euclidiana* $||.||_2$*, a norma do máximo* $||.||_\infty$ *e a norma da soma* $||.||_1$*, que são as mais usuais, e em torno do ponto limite* $x = (1, 1)$ *foram consideradas bolas (o nome não se aplica apenas à primeira... *$B(a, r) = \{x : ||x - a|| \leq r\}$*) com raios entre 0.5 e 0.1. É fácil perceber que qualquer que seja a norma, por mais pequeno que seja o raio, é sempre possível encontrar um dos elementos da sucessão dentro dessa bola (vizinhança). Isto significa que a sucessão converge segundo qualquer uma das normas.*



Por outro lado, também é claro que estabelecer a equivalência entre as normas $||.||_2$ *e* $||.||_\infty$*, é fácil, podemos mesmo explicitar as constantes. Com efeito, como (dimensão=*N*)*

$$\begin{aligned} ||x||_2^2 &= x_1^2 + ... + x_N^2 \geq \max\{|x_1|^2, ..., |x_N|^2\} = ||x||_\infty^2, \\ ||x||_2^2 &= x_1^2 + ... + x_N^2 \leq N \max\{|x_1|^2, ..., |x_N|^2\} = N||x||_\infty^2, \end{aligned}$$

concluímos que

$$||x||_\infty \leq ||x||_2 \leq \sqrt{N}||x||_\infty.$$

Isto corresponde a dizer, em dimensão 2, que se um quadrado contém o círculo com o mesmo raio, um círculo com $\sqrt{2}$ *vezes esse raio já irá conter o quadrado. A equivalência é simplesmente isto, e permite concluir que bolas numa certa norma vão estar incluídas em bolas noutra norma e vice-versa*[2]*. Para terminar, referimos que explicitar as constantes*

[2]Refira-se a este propósito que quando a dimensão do espaço aumenta, seria necessário uma hiperesfera com $\sqrt{N}$ vezes o hipercubo para que ele estivesse contido nela. Isto indicia que em espaços de dimensão infinita as coisas irão passar-se de forma diferente. Com efeito, quando a dimensão tende para infinito os hipercubos unitários serão infinitamente maiores que as hiperesferas unitárias.

para a equivalência entre $||.||_1$ e $||.||_\infty$ é igualmente fácil. Como,

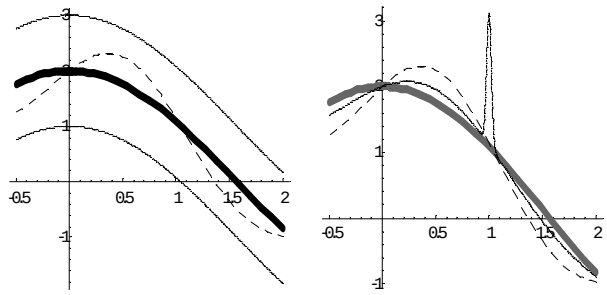
$$\begin{aligned} ||x||_1 &= |x_1| + ... + |x_N| \geq \max\{|x_1|, ..., |x_N|\} = ||x||_\infty, \\ ||x||_1 &= |x_1| + ... + |x_N| \leq N \max\{|x_1|, ..., |x_N|\} = N||x||_\infty, \end{aligned}$$

concluímos que

$$||x||_\infty \leq ||x||_1 \leq N||x||_\infty.$$

(o que significa que o losango irá estar incluído num quadrado com o mesmo raio e que esse quadrado estará incluído num losango com o dobro do raio).

Exemplo 3.2 *No caso em que trabalhamos em espaços de funções, as bolas tomam um aspecto menos trivial, porque o espaço deixa de ter dimensão finita.*
Na figura seguinte, à esquerda, representamos a função $f(x) = 2\cos(x)$ (curva a cheio) no intervalo $[0, 2]$. A bola centrada em f e de raio 1, segundo a norma das funções contínuas $||.||_\infty$, será o conjunto das funções g tais que $||f - g||_\infty = \max_{[0,2]} |f(x) - g(x)| < 1$, ou seja será definida pelas funções g que verifiquem $2\cos(x) - 1 < g(x) < 2\cos(x) + 1$, limites que estão representados por curvas contínuas mais finas e que formam uma banda em redor de f. Um exemplo de função g que pertence a essa bola é $g(x) = 2\cos(x) + \frac{1}{2}\sin(3x)$, função representada a tracejado. No gráfico da direita, voltamos a considerar as mesmas funções f e g e uma outra, $h(x) = 2\cos(x) + \frac{1}{m}\sin(3x) + 2e^{-200m(x-1)^2}$, com $m = 5$ (em que a última parcela, uma gaussiana pronunciada, é responsável pelo pico vísivel no gráfico). Mesmo não estando representada a banda, é perceptível que o pico estará fora dos limites, e portanto fora da bola de raio 1 definida pela norma do máximo.
No entanto, este exemplo foi escolhido por outra razão. Se virmos a diferença entre f e h em termos de área, ou seja em termos da norma L^1, $||.||_1$,essa diferença é menor do que a diferença entre f e g. Mais, podemos mesmo considerar uma sucessão de funções h que, quando o parâmetro m tende para infinito, irá aproximar-se da função f. Bom... excepto no ponto $x = 1$, já que o pico irá persistir. O limite pontual será uma função $\tilde{f}$, idêntica a f, mas que no ponto $x = 1$ irá valer $2cos(1) + 2 \sim 3.08$. Do ponto de vista da norma $||.||_\infty$, a sucessão não converge, porque o pico irá sempre ficar fora de qualquer bola de raio menor que 1. Do ponto de vista da norma $||.||_1$ (definida pelo integral do módulo) a diferença entre as áreas irá tender para zero, pelo que a sucessão irá convergir. Isto é bem conhecido da teoria do integral de Lebesgue, que identifica f e $\tilde{f}$ a menos de conjunto de medida nula (neste caso, o ponto $x = 1$).



Este exemplo torna também claro que não há equivalência entre as normas $||.||_\infty$ e $||.||_1$, o que também poderá ser compreendido se pensarmos que a função $f(x) = \frac{1}{\sqrt{x}}$ que está na

bola $\bar{B}(0,2)$ definida pela norma em $L^1(]0,1[)$, já que o integral existe, tendo-se $||f||_1 = 2$. No entanto, sendo uma função ilimitada não há qualquer bola definida pela norma $||.||_\infty$ que contenha essa função.
Também fica claro que, para desenhar os limites duma bola para a norma $||.||_\infty$ basta considerar uma banda circundante, mas para desenhar os limites duma bola para a norma $||.||_1$ é-nos simplesmente impossível...

3.2 Espaços de Banach

O facto de introduzirmos uma topologia num espaço normado não significa que exista um elemento (pertencente a esse espaço) que seja limite de sucessões de Cauchy (... como no exemplo anterior). É nesse sentido que iremos introduzir a noção de espaço completo, e consequentemente a noção de espaço de Banach (espaço normado completo). Começamos por ver que as sucessões convergentes são sucessões de Cauchy, mas que o recíproco pode não ser válido.

Definição 3.5 *Uma sucessão (x_n) num espaço normado E diz-se* sucessão de Cauchy *em E se*

$$||x_m - x_n|| \overset{m,n\to\infty}{\longrightarrow} 0.$$

Proposição 3.1 *Se $x_n \to x$ em E, então (x_n) é sucessão de Cauchy em E.*

Demonstração: $||x_m - x_n|| \leq ||x_m - x|| + ||x - x_n|| \longrightarrow 0$, quando $m, n \to \infty$.■

Observação: *O recíproco desta proposição nem sempre será válido.* Ou seja, podemos ter sucessões cujos termos se aproximam indefinidamente, mas que não têm limite em E. Isto é análogo ao que se passa com os racionais... uma sucessão de Cauchy de racionais pode não ter limite nos racionais, basta pensar na sucessão $x_0 = 1, x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}$ cujos termos são sempre racionais, mas que converge para $\sqrt{2}$, ou na sucessão definida por $y_n = (1 + \frac{3}{n})^n \in \mathbb{Q}$ e cujo limite é e^3.

A solução foi considerar essa sucessões como sendo números, constituindo os números reais, completando assim o espaço dos racionais, como vimos no início do texto. A partir daí o nosso espaço comum de trabalho é o dos números reais. No caso das funções irá passar-se algo semelhante, mas com a grande diferença de podermos num mesmo espaço considerar várias normas. Assim, se sucessões de Cauchy de funções contínuas segundo a norma do máximo são ainda funções contínuas, devido à continuidade uniforme, o mesmo não irá acontecer se considerarmos outra norma, por exemplo a norma L^1.

Isto poderia significar que tendo obtido uma sucessão de Cauchy com o método do ponto fixo, esta não teria ponto fixo num simples espaço normado. Torna-se por isso conveniente trabalhar num espaço em que isso não aconteça – num *espaço de Banach*:

Definição 3.6 *Um espaço vectorial normado E diz-se* espaço de Banach *se for* completo*, ou seja, se toda a sucessão de Cauchy em E for uma sucessão convergente para um certo $x \in E$.*

Exemplo 3.3 *Os exemplos mais simples de espaços de Banach, são os próprios corpos* $\mathbb{R}$ *ou* $\mathbb{C}$*. ou ainda* $\mathbb{R}^N$ *ou* $\mathbb{C}.^N$.
Um espaço vectorial com produto interno que seja completo é normalmente designado espaço de Hilbert*. Como é óbvio, usando a norma* $||x|| = (x.x)^{1/2}$*, um espaço de Hilbert será sempre um caso particular de um espaço de Banach, sendo válidas todas as propriedades que iremos deduzir de seguida.*

Num espaço normado, um conjunto fechado tem a importante propriedade de conter o limite de qualquer sucessão convergente, cujos termos lhe pertençam.

Proposição 3.2 *Se o conjunto* A *é fechado em* E *então*

$$\forall (x_n) \subseteq A : \; x_n \to x \quad \Rightarrow \quad x \in A$$

Demonstração:
Se, por absurdo, $x \notin A$, teríamos $x \in E \backslash A$ que é um aberto, existindo assim uma vizinhança $V_\varepsilon(x) \subseteq E \backslash A$ e portanto $||x - x_n|| > \varepsilon$ para qualquer n, contrariando a hipótese de convergência. ■

Portanto um subespaço vectorial fechado de um espaço de Banach, é ainda um espaço de Banach para a mesma norma.

Exercício 3.3 *Mostre que o espaço de funções* $C^m[a,b]$*, munido da norma*

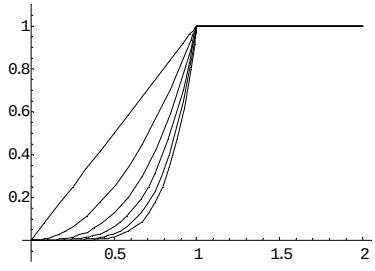
$$||f||_{\infty,m} = \sup_{x\in[a,b]} |f(x)| + \sup_{x\in[a,b]} |f'(x)| + ... + \sup_{x\in[a,b]} |f^{(m)}(x)|$$

é um espaço de Banach. Se $m = 0$*, temos a norma já apresentada para as funções contínuas, e a completude resulta do facto da convergência uniforme de funções contínuas ser uma função contínua.*

Exercício 3.4 *Verifique que qualquer um dos espaços normados apresentados no exercício3.1 é um espaço de Banach.*

Observação: *(incompletude das funções contínuas em* L^p*).*
Retomamos a ideia enunciada no último exemplo da secção anterior, que iremos agora analisar mais detalhadamente, num exemplo clássico. Consideremos a sucessão de funções contínuas em $[0,2]$

$$f_n(x) = \begin{cases} x^n & \text{se } x \in [0,1[\\ 1 & \text{se } x \in [1,2], \end{cases}$$



Vemos que $f_n \in C([0,2])$ é uma sucessão de Cauchy para a norma L^1, porque

$$||f_m - f_n||_1 = \int_0^1 |x^m - x^n| dx = |\frac{1}{m+1} - \frac{1}{n+1}| \overset{m,n\to\infty}{\longrightarrow} 0.$$

No entanto verificamos que, pontualmente, a sucessão (f_n) converge para uma função que é nula em $[0,1[$ e é igual a 1 em $[1,2]$, ou seja, uma função que é descontínua! Concluímos que $C([0,2])$ não é completo para a norma L^1. Vejamos que para a norma habitual de $C[a,b]$ que é $||.||_\infty$, a sucessão em causa não é de Cauchy. Com efeito,

$$||f_m - f_n||_\infty = \sup_{x\in[0,1]} |x^m - x^n|$$

e se considerarmos $m = 2n$, temos $(x^{2n} - x^n)' = 0 \Leftrightarrow x = 0$ ou $x^n = \frac{1}{2}$. O máximo é assim atingido no ponto $(\frac{1}{2})^{1/n}$, logo

$$\sup_{x\in[0,1]} |x^{2n} - x^n| = |\frac{1}{4} - \frac{1}{2}| = \frac{1}{4} \not\to 0$$

portanto, como se previa, a sucessão não é de Cauchy para a norma $||.||_\infty$.

Concluímos assim que o espaço das funções contínuas não é completo para a norma L^1 (nem o será, para nenhuma das outras normas L^p).

3.2.1 Operadores Contínuos

Como um espaço normado é uma estrutura muito geral, que pode ter como elementos funções, é costume designar as funções definidas em espaços normados por *operadores*. Como abreviatura, é também habitual designar a imagem de x pelo operador A por Ax, ao invés de $A(x)$. Na realidade, este tipo de notação será também coerente com a notação matricial[3], quando os operadores a considerar forem operadores lineares em espaços de dimensão finita.

Definição 3.7 *Sejam E, F espaços normados. Dizemos que um operador $A : X \subseteq E \to F$ é contínuo em X, se para qualquer $x \in X$ tivermos*

$$\forall (x_n) \subseteq X,\ x_n \to x \ \Rightarrow\ Ax_n \to Ax.$$

[3]Com a precaução devida, encarar os operadores como matrizes pode também ser uma boa maneira de olhar para esta teoria pela primeira vez. De facto os resultados obtidos para operadores em espaços de Banach serão em particular válidos para matrizes (que são operadores lineares e contínuos em espaços de dimensão finita)... o contrário nem sempre será válido – essa é a principal precaução que se deve ter sempre!

Exemplo 3.4 *A própria norma é um operador contínuo de E em $\mathbb{R}$. Com efeito, se $x_n \to x$ em E, então*

$$| \, ||x_n|| - ||x|| \, | \, \leq ||x_n - x|| \to 0,$$

porque se $||x_n|| \geq ||x||$ temos $|||x_n|| - ||x||| \, = ||x_n + x - x|| - ||x|| \leq ||x_n - x|| + ||x|| - ||x||$. Da mesma maneira, se $||x_n|| \leq ||x||$, temos $|||x_n|| - ||x||| \, = ||x - x_n + x_n|| - ||x_n|| \leq ||x - x_n|| + ||x_n|| - ||x_n||$.

Exercício 3.5 *Sejam E, F, G espaços normados.*
a) Mostre que se $A, B : X \subseteq E \to F$ forem operadores contínuos, $A + B$ é um operador contínuo, e que para qualquer $\alpha \in \mathbb{R}$ (ou $\mathbb{C}$) o operador αA é contínuo.
b) Mostre que se $A : X \subseteq E \to Y \subseteq F$, $B : Y \subseteq F \to G$ são operadores contínuos, então $B \circ A$ também é contínuo (em X).

(Quando não há perigo de confusão, é normalmente adoptada a notação multiplicativa para designar a composição, ou seja $BA = B \circ A$, tal como nas matrizes)

3.2.2 Operadores Lineares

De entre os operadores contínuos, são especialmente importantes aqueles que sejam *lineares*, ou seja, que verifiquem as propriedades

$$A(x + y) = Ax + Ay, \ \forall x, y \in E$$
$$A(\alpha x) = \alpha Ax, \ \forall \alpha \in \mathbb{R}, \ \forall x \in E.$$

Os operadores lineares[4] são contínuos se e só se forem *limitados*, ou seja, se verificarem

$$\sup_{||x||_E \leq 1} ||Ax||_F < +\infty.$$

Como se tratam de operadores lineares, isto significa que transformam qualquer conjunto limitado num conjunto limitado[5].

• Sejam E, F espaços de Banach. Podemos considerar um espaço associado aos operadores, o espaço dos operadores lineares contínuos, $\mathcal{L}(E, F)$, que com a norma

$$||A||_{\mathcal{L}(E,F)} = \sup_{||x||_E \leq 1} ||Ax||_F = \sup_{x \neq 0} \frac{||Ax||_F}{||x||_E} \tag{3.2}$$

é um espaço de Banach. É claro que, para qualquer $x \in E$,

$$||Ax||_F \leq ||A||_{\mathcal{L}(E,F)} ||x||_E \ .$$

[4]Esta propriedade é válida apenas para operadores lineares!

[5]Reparamos que se A for linear e contínuo em 0, então A é contínuo em qualquer x, pois $x_n \to x \Rightarrow Ax_n - Ax = A(x_n - x) \to 0$.

Logo, quando o operador é linear e limitado, se considerarmos $||x||_E \leq \varepsilon$ temos $||Ax||_F \leq C\varepsilon$, logo se $x_n \to 0$ temos $Ax_n \to 0$, o que significa que A é contínuo em 0.

Exercício 3.6 *Mostre que se $A, B \in \mathcal{L}(E,E)$, temos*

$$||AB||_{\mathcal{L}(E,E)} \leq ||A||_{\mathcal{L}(E,E)}||B||_{\mathcal{L}(E,E)}$$

A introdução de operadores lineares é importante já que, em muitos casos tenta linearizar-se o operador para simplificar o seu estudo. *Em certos exemplos esta técnica pode ser vista como uma generalização da aproximação local de uma função através da tangente, que utilizaremos quando falarmos de derivação de Fréchet.*

3.3 Método do Ponto Fixo e o Teorema de Banach

Iremos agora concretizar a generalização do método e do teorema do ponto do fixo a espaços de Banach.

Seja A um operador qualquer definido num subconjunto X (designado *domínio*) de um espaço de Banach E,

$$A : X \subseteq E \to E.$$

Pretendemos encontrar os pontos fixos de A, ou seja $z \in X$:

$$z = Az$$

e para esse efeito vamos usar o *método do ponto fixo* (também designado método de Picard),

$$\begin{cases} x_0 \in X \\ x_{n+1} = Ax_n \end{cases} .$$

Como o método implica repetições sucessivas do operador A, é natural exigir que imagem ainda esteja no domínio, ou seja $A(X) \subseteq X$.

Como vimos em $\mathbb{R}$ e em $\mathbb{C}$, para assegurar a convergência do método foi usada a noção de contractividade, que neste contexto se define da seguinte forma:

Definição 3.8 *Um operador $A : X \subseteq E \to E$, num espaço de Banach E diz-se* contractivo *em X, se existir $0 \leq L < 1$ (chamada* constante de contractividade*):*

$$||Ax - Ay|| \leq L||x - y|| \ , \ \forall x, y \in X.$$

Proposição 3.3 *Se A é contractivo em X, conjunto fechado, então A é contínuo em X.*

Demonstração:
Com efeito, basta considerar $(x_n) \in X$ tal que $x_n \to x$

$$||Ax_n - Ax|| \leq L||x_n - x|| \to 0 \ \Rightarrow \ Ax_n \to Ax. \ \blacksquare$$

Estamos agora nas condições de demonstrar o teorema do ponto fixo de Banach.

Teorema 3.2 *(Teorema do ponto fixo de Banach).*
Seja X um conjunto fechado não vazio[6] *num espaço de Banach E, e seja A um operador contractivo em X tal que $A(X) \subseteq X$. Então*
i) Existe um e um só ponto fixo $z \in X : \ Az = z$
ii) A sucessão $x_{n+1} = Ax_n$ converge para o ponto fixo z, qualquer que seja $x_0 \in X$.
iii) Verificam-se as desigualdades:
$||z - x_n|| \leq L||z - x_{n-1}|| \leq L^n||z - x_0||$,
$||z - x_n|| \leq \frac{1}{1-L}||x_{n+1} - x_n||$,
$||z - x_n|| \leq \frac{L^n}{1-L}||x_1 - x_0||$,
onde $L < 1$ é a constante de contractividade.

Demonstração:

1^o) Prova-se por indução que qualquer $x_n \in X$, porque assumimos $x_0 \in X$, e se $x_n \in X$, temos $x_{n+1} = Ax_n \in X$, pois $A(X) \subseteq X$.

2^o) (x_n) é sucessão de Cauchy.

Como A é contractivo em X e $x_n \in X, \forall n \in \mathbb{N}$ temos

$$||x_{n+1} - x_n|| = ||Ax_n - Ax_{n-1}|| \leq L||x_n - x_{n-1}||,$$

portanto $||x_{n+1} - x_n|| \leq L^n||x_1 - x_0||$, e introduzindo somas e subtrações sucessivas, obtemos assim:

$$||x_{n+m} - x_n|| \leq ||x_{n+m} - x_{n+m-1}|| + ... + ||x_{n+1} - x_n|| \leq L^{n+m-1}||x_1 - x_0|| + ... + L^n||x_1 - x_0|| =$$

$$= L^n(L^{m-1} + ... + 1)||x_1 - x_0|| = L^n \frac{1 - L^m}{1 - L}||x_1 - x_0|| \leq \frac{L^n}{1 - L}||x_1 - x_0||$$

que converge para zero quando $n, m \to \infty$.

3^o) Existência e convergência.

Como E é completo e (x_n) é sucessão de Cauchy, existe $z \in E$ tal que $x_n \to z$. Por outro lado, como X é fechado, concluímos que $z \in X$.

Como $x_n \to z$ e A contínuo (porque é contractivo), então $x_{n+1} = Ax_n \to Az$. Pela unicidade do limite, temos $z = Az$, o que prova a existência de um ponto fixo em X.

4^o) Unicidade:

Supondo que existiam $z, w \in X$ tais que $z = Az$ e que $w = Aw$, então

$$||z - w|| = ||Az - Aw|| \leq L||z - w|| \Rightarrow (1 - L)||z - w|| \leq 0$$

ora como $L < 1$ temos $||z - w|| \leq 0$, ou seja, $||z - w|| = 0 \Leftrightarrow z = w$.

5^o) Estimativas:

$$||z - x_n|| \leq ||Az - Ax_{n-1}|| \leq L||z - x_{n-1}|| \leq ... \leq L^n||z - x_0||$$

[6] Uma precaução... por vezes podem demonstrar-se todas as hipóteses e esquecermo-nos de mostrar que o conjunto X tem elementos. Isso acontece quando X não é um conjunto concreto, e é definido de forma a verificar certas propriedades... que por vezes nenhum elemento verifica.

$$||z - x_n|| \leq ||z - x_{n+1}|| + ||x_{n+1} - x_n|| \leq L||z - x_n|| + ||x_{n+1} - x_n||$$

e daqui saiem facilmente as restantes. ■

Observação:

(i) Nesta demonstração, ao provarmos que a sucessão é de Cauchy, asseguramos imediatamente a existência de ponto fixo o que difere da demonstração apresentada para o caso de intervalos limitados em que assegurámos existência através do teorema do valor intermédio[7].

(ii) Note-se que ainda que esteja estabelecida a equivalência entre normas (como entre todas as normas no caso de dimensão finita), provar a contractividade para uma norma não significa que ela seja válida para as normas equivalentes. A contractividade é uma propriedade quantitativa e não qualitativa, e poderá haver diferenças. Por exemplo, em dimensão finita, é muitas vezes possível demonstrar a contractividade, num certo conjunto, para a norma $||.||_\infty$ e não para a norma $||.||_1$ ou vice-versa. É claro que isso não invalida que haja convergência nas duas normas, e se considerarmos um conjunto mais pequeno será mesmo possível mostrar a contractividade em qualquer das normas equivalentes.

Exemplo 3.5 *Consideremos o operador*

$$Af(x) = 1 - \int_0^x f(t)dt$$

no espaço de Banach $E = C[0, \frac{1}{R}]$ *com a norma* $||f||_\infty = \max |f(x)|$, *em que* $R \geq 2$. *O subconjunto fechado que consideramos é* $X = \{f \in C[0, \frac{1}{R}] : ||f||_\infty \leq R\}$, *constatando que, sendo* f *contínua em* $I = [0, \frac{1}{R}]$, Af *é ainda uma função em contínua em* I. *Vejamos que* $A(X) \subseteq X$. *Ora, supondo* $||f||_\infty \leq R$,

$$||Af||_\infty = \max_{x \in I} |1 - \int_0^x f(t)dt| \leq 1 + \frac{1}{R} \max_{x \in I} |f(x)| \leq 2.$$

Para assegurarmos a convergência, falta apenas verificar a contractividade:

$$||Af - Ag||_\infty = \max_{x \in I} |1 - 1 - \int_0^x f(t) - g(t)dt| \leq \frac{1}{R} \max_{x \in I} \max_{t \in [0,x]} |f(t) - g(t)| \leq \frac{1}{R}||f - g||_\infty.$$

Estão assim asseguradas as hipóteses do teorema do ponto fixo de Banach, e a convergência para o ponto fixo está provada. Como já tinhamos mencionado, trata-se da função e^{-x}.

[7]Em espaços de dimensão finita podemos usar o teorema do ponto fixo de Brouwer para garantir existência em conjuntos convexos e limitados. Em espaços de dimensão infinita é utilizado um teorema de Schauder que exige que o operador seja 'compacto'.

Exemplo 3.6 *Consideremos o sistema em* $\mathbb{R}^3$

$$\left\{\begin{array}{c} 3x_1+x_2=1 \\ 2x_1+4x_2+x_3=0 \\ x_2+2x_3=2 \end{array}\right. \Leftrightarrow \left\{\begin{array}{c} x_1=1/3-x_2/3 \\ x_2=-x_1/2-x_3/4 \\ x_3=1-x_2/2 \end{array}\right.$$

Podemos pois considerar $A:\mathbb{R}^3\to\mathbb{R}^3$ *definido por*

$$A(x_1,x_2,x_3)=(1/3-x_2/3,-x_1/2-x_3/4,1-x_2/2)$$

Vejamos que A *é contractivo em* $\mathbb{R}^3$ *para a norma* $||.||_\infty$:

$$||Ax-Ay||_\infty = || \left[\begin{array}{c} \frac{1}{3}(x_2-y_2) \\ \frac{1}{2}(x_1-y_1)+\frac{1}{4}(x_3-y_3) \\ \frac{1}{2}(x_2-y_2) \end{array}\right] ||_\infty$$

designando $M=||x-y||_\infty=\max\{|x_1-y_1|,|x_2-y_2|,|x_3-y_3|\}$, *obtemos assim*

$$||Ax-Ay||_\infty \leq \max\{\frac{1}{3}M,\frac{3}{4}M,\frac{1}{2}M\}\leq\frac{3}{4}M$$

e portanto uma constante de contractividade é $\frac{3}{4}$, *e sendo contractiva em* $I\!R^3$, *que é fechado, qualquer aproximação inicial permite, através do método do ponto fixo, obter a solução única* $x\sim(0.5294,-0.5882,1.2941)$.

Vemos assim que o teorema do ponto fixo é tão geral que pode ser aplicado a equações que envolvem integrais, a sistemas de equações, ou simplesmente a equações em $\mathbb{R}$ ou $\mathbb{C}$.

É claro que quanto mais pequena for a constante de contractividade L, mais rápida será a convergência. Como no caso real, podemos falar em *ordem de convergência*. Convém assim restabelecer a definição

Definição 3.9 *Dizemos que* x_n *converge para* z *com* pelo menos *ordem de convergência* linear *na norma* $||.||$ *se existir* $K<1$:

$$K_n=\frac{||e_{n+1}||}{||e_n||}\leq K.$$

Quando $K_n\to 0$, *diremos que a ordem de convergência é* supralinear.

No caso de aplicação do teorema do ponto fixo, como mostrámos que $||e_{n+1}||\leq L||e_n||$, com $L<1$, podemos concluir que a convergência é pelo menos linear. Para prosseguirmos com a análise, avaliando se o limite de K_n existe, precisamos de introduzir a noção de derivação aplicada aos espaços de Banach. Havendo duas possibilidades, optamos por introduzir a noção de derivação de Fréchet e não a de Gateaux, que nos parece mais adequada para os nossos objectivos. Essa noção de diferenciabilidade permitirá estender muitos dos critérios observados no caso real, e apresentar o método de Newton.

3.4 Derivação de Fréchet

A derivação em espaços abstractos tem aspectos não triviais, que omitiremos deliberadamente (para uma compreensão aprofundada ver, por exemplo [6]). Iremos concentrar-nos no objectivo principal que é estabelecer resultados análogos aos que existem em $\mathbb{R}$, e que depois possam ser aplicados em $\mathbb{R}^N$. Estas noções são imediatamente reconhecidas no caso em que o cálculo diferencial em $\mathbb{R}^N$ foi apresentado recorrendo à noção de forma diferencial.

Definição 3.10 *Sejam E, F espaços normados e A um operador $A : X \subseteq E \to F$, cujo domínio X é um aberto*[8]*. Dizemos que A é Fréchet-diferenciável (ou F-diferenciável) no ponto $x \in X$ se existir um operador linear $T \in \mathcal{L}(E, F)$ tal que:*

$$||A(x+h) - Ax - Th||_F = o(||h||_E) \quad \textit{quando } ||h||_E \to 0 \tag{3.3}$$

Caso o operador T exista, é chamado "derivada de Fréchet" *em x e escrevemos A'_x, tendo-se*

$$\begin{array}{rccl} A' : & X & \longrightarrow & \mathcal{L}(E,F) \\ & x & \longmapsto & A'_x : E \to F \quad \textit{(operador linear)} \end{array}$$

Se A for F-diferenciável em todos os pontos $x \in X$ diremos que A é F-diferenciável em X.

Observação: Uma função $f : I \subseteq \mathbb{R} \to \mathbb{R}$ é *diferenciável* em $x \in I$, se existir o limite

$$\lim_{y \in I,\, y \to x} \frac{f(y) - f(x)}{y - x},$$

que designamos por derivada de f de x, ou abreviadamente $f'(x)$. Reparamos que fazendo $h = y - x$, isto é equivalente a

$$\frac{f(x+h) - f(x)}{h} \to f'(x), \text{ quando } h \to 0.$$

Ou seja, é equivalente a dizer que existe um número $f'(x)$:

$$|f(x+h) - f(x) - f'(x)h| = o(|h|), \text{ quando } |h| \to 0,$$

o que corresponde à noção de derivação de Fréchet.

Proposição 3.4 *Se A'_x existir é único.*

Demonstração:

[8]Quando X é fechado, diremos que A é F-diferenciável em X se existir um aberto $\tilde{X} \supset X$ onde A é F-diferenciável. Para esse efeito, é claro que é necessário que A esteja definido em $\tilde{X}$.

Seja $A'_x = T$ e consideremos outro operador U nas condições da definição. Então teríamos, para qualquer $y \in E\backslash\{0\}$,

$$\begin{aligned}\frac{||(T-U)y||_F}{||y||_E} &= \frac{||T(\varepsilon y) - U(\varepsilon y)||_F}{||\varepsilon y||_E} \\ &\leq \left(\frac{||T(\varepsilon y) - A(x+\varepsilon y) + Ax||_F}{||\varepsilon y||_E} + \frac{||A(x+\varepsilon y) - Ax - U(\varepsilon y)||_F}{||\varepsilon y||_E}\right) \xrightarrow{\varepsilon \to 0} 0\end{aligned}$$

(somando e subtraindo $A(x+\varepsilon y) - Ax$ com $\varepsilon > 0$), onde εy corresponde ao h da definição.

Usando a definição de norma em $\mathcal{L}(E,F)$, concluimos que $||T-U||_{\mathcal{L}(E,F)} = 0$, i.e: $T = U$. ■

Exercício 3.7 *Verifique que se A é Fréchet-diferenciável, então A é contínuo.*

Exemplo 3.7 *O exemplo mais simples, para além da derivação vulgar em $\mathbb{R}$ ou $\mathbb{C}$, aparece em $\mathbb{R}^N$. Com efeito, se considerarmos uma função $f : \mathbb{R}^N \to \mathbb{R}^N$, a derivada de Fréchet corresponde a considerar a matriz jacobiana*[9]*,*

$$\boldsymbol{\nabla} f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_N}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1}(x) & \dots & \frac{\partial f_N}{\partial x_N}(x) \end{bmatrix},$$

que é uma aplicação linear $\mathbb{R}^N \to \mathbb{R}^N$. Isto é uma consequência da fórmula de Taylor em $\mathbb{R}^N$,

$$f(y) = f(x) + \boldsymbol{\nabla} f(x)(y-x) + o(||y-x||)$$

– Por exemplo, se $A(x_1, x_2) = (x_1^2 + x_2, x_1 e^{x_2})$ temos

$$A'_{(x_1,x_2)} = \begin{bmatrix} 2x_1 & 1 \\ e^{x_2} & x_1 e^{x_2} \end{bmatrix}$$

Observações:

(i) – A derivada de Fréchet de qualquer operador constante é o operador nulo.

(ii) – Seja A um operador linear, a sua derivada de Fréchet em qualquer ponto x é sempre o próprio operador linear (sendo assim constante em x)!

É bom interpretar correctamente esta afirmação... Como $A(x+h) - A(x) - Ah = 0$, para qualquer ponto x, a derivada é sempre $A'_x = A$, ou seja é constante relativamente a x. Assim a 'segunda derivada' seria o operador nulo, já que $A'_{x+h} - A'_x = A - A = 0$. Vemos assim que tudo se mantém coerente com as propriedades habituais.

[9]Por vezes também é designada matriz jacobiana a transposta desta. Essa escolha implicaria escrever $[\boldsymbol{\nabla} f]^\top v$, quando quisessemos efectuar o produto por v, o que tornaria as notações mais pesadas.

(iii) – A derivação, assim definida, possui algumas das propriedades habituais, como a linearidade: $(A+B)' = A' + B'$ e $(\alpha A)' = \alpha A'$; ou a propriedade para a composição:

Sendo E, F, G espaços de Banach, e $A : X \subseteq E \to Y \subseteq F$, $B : Y \subseteq F \to G$, diferenciáveis, a aplicação $B \circ A : X \subseteq E \to G$ é diferenciável e temos

$$(B \circ A)'_x = B'_{Ax} \circ A'_x \tag{3.4}$$

onde $(B \circ A)'_x \in \mathcal{L}(E, G)$. Para além disso, é claro que

$$||(BA)'_x||_{\mathcal{L}(E,G)} \leq ||B'_{Ax}||_{\mathcal{L}(F,G)} ||A'_x||_{\mathcal{L}(E,F)}.$$

3.4.1 Corolário do Teorema do Ponto Fixo

Com o intuito de aplicar o Teorema do Ponto Fixo de Banach, reparamos que se exigirmos que o conjunto seja convexo[10] podemos obter um resultado, semelhante ao do caso real (ou complexo), que relaciona a norma da derivada inferior a $L < 1$ à contractividade.

Definição 3.11 *Um conjunto não vazio $X \subseteq E$ diz-se* convexo *se verificar*

$$x, y \in X \Rightarrow \forall t \in [0,1],\ x + t(y - x) \in X. \tag{3.5}$$

Observação:

Usando a definição, é fácil ver que as bolas são conjuntos convexos: porque se $x, y \in B(a, r) = \{w \in E : ||w - a|| < r\}$, então

$$||x+t(y-x)-a|| = ||(1-t)(x-a)+t(y-a)|| \leq (1-t)||x-a||+t||y-a|| \leq (1-t)r+tr = r.$$

Teorema 3.3 *Sejam E, F espaços de Banach e seja A um operador Fréchet-diferenciável num* convexo *X*

$$A : X \subseteq E \to F$$

Se tivermos

$$||A'_x||_{\mathcal{L}(E,F)} \leq L < 1,\ \ \forall x \in X$$

então

$$||Ax - Ay||_F \leq L||x - y||_E\ ,\ \forall x, y \in X.$$

[10] Mesmo no caso de $\mathbb{R}$ não basta que o módulo da derivada seja inferior a 1. A convexidade é garantida nesse caso porque trabalhamos com intervalos (contendo eles próprios os segmentos que definem a convexidade).

Reforçamos assim a observação de que a passagem de contractividade para norma da derivada menor que 1 é aqui obtido usando a hipótese de que o conjunto é convexo.

Demonstração:

Consideramos $B(t) = A(x + t(y - x))$, com $t \in [0, 1]$. Se $x, y \in X$, como é convexo, temos $x + t(y - x) \in X$. Usando a regra de derivação da função composta (3.4), obtemos:

$$B'_t = A'_{x+t(y-x)}(y - x)$$

e vamos usar a seguinte generalização da fórmula dos acréscimos finitos (ver p.ex: [6]):

Lema: Seja F um espaço de Banach e $f : [a, b] \to F$ tal que $||f'_t||_{\mathcal{L}(\mathbf{R},F)} \leq K, \forall t \in [a, b]$. Então $||f(b) - f(a)||_F \leq K(b - a)$.□

Aplicando este resultado a $B : [0, 1] \to F$, como

$$||B'_t||_{\mathcal{L}(\mathbf{R},F)} = ||A'_{x+t(y-x)}(y - x)||_{\mathcal{L}(\mathbf{R},F)} \leq ||A'_{x+t(y-x)}||_{\mathcal{L}(E,F)}||y - x||_{\mathcal{L}(\mathbf{R},E)}$$

e sendo X convexo temos $x + t(y - x) \in X$, logo $||A'_{x+t(y-x)}||_{\mathcal{L}(E,F)} \leq L$.

Portanto, $||B'_t||_{\mathcal{L}(\mathbf{R},F)} \leq L||y - x||_E$ (reparando que $||y - x||_{\mathcal{L}(\mathbf{R},E)} = ||y - x||_E$). Isto implica

$$||B(1) - B(0)||_F \leq L||y - x||_E$$

e como $B(1) = Ay,\ B(0) = Ax$, o resultado fica provado.■

Corolário 3.1 (*do Teorema do Ponto Fixo de Banach*). *Seja A um operador Fréchet-diferenciável em X, um conjunto não vazio,* convexo e fechado *num espaço de Banach E. Se $A(X) \subseteq X$ e tivermos*

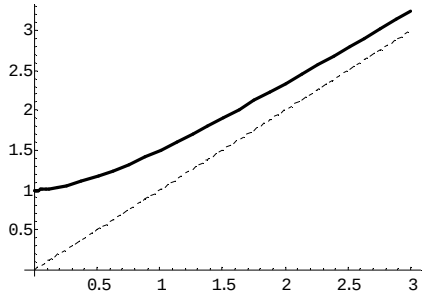
$$||A'_x||_{\mathcal{L}(E,F)} \leq L < 1, \quad \forall x \in X$$

as condições do Teorema do Ponto Fixo de Banach estão verificadas.■

Observações:

(i) Se considerarmos os espaços $\mathbb{R}^N$ ou $\mathbb{C}^N$ e se o conjunto X for limitado então, sendo fechado, é um compacto (porque são espaços de dimensão finita) e basta exigir $||A'_x|| < 1$. Com efeito $||A'_x||$ é uma função contínua de $\mathbb{R}^N$ em $\mathbb{R}$ e pelo teorema de Weierstrass atinge um máximo $L < 1$.

No caso de se tratar de um conjunto ilimitado, exigir $||A'_x|| < 1$ não basta! Podemos pensar como contra-exemplo a função $A(x) = 1 + x^2/(x + 1)$ que verifica $|A'(x)| < 1$ no intervalo $X = [0, +\infty[$ e $A(X) \subseteq X$, no entanto, esta função não tem qualquer ponto fixo em X. Se traçarmos o gráfico,



reparamos que a bissectriz é uma assímptota do gráfico de g, e portanto, apesar de se aproximar da bissectriz, nunca a intersecta. Isto já não acontece para uma função que verifique $|A'(x)| \leq L < 1$, pois esta condição obriga a que haja intersecção! (Este foi um exemplo que encontrámos no início do capítulo 2, ficando agora claro que o método do ponto fixo nunca poderia convergir).

(ii) Mesmo ao aplicarmos este resultado em $\mathbb{R}$ vemos como a convexidade é importante. No caso de $\mathbb{R}$ a convexidade traduz-se em conexidade e significa podermos aplicar o resultado a um único intervalo fechado (que pode ser ilimitado), já que se considerassemos X como sendo a reunião de dois intervalos fechados, em que o módulo da derivada era inferior a 1, poderíamos ter um ponto fixo em cada um deles, contrariando a unicidade.

(iii) Se considerarmos uma função $g(x)$ definida em $\mathbb{R}$ tal que $|g'(x)| \leq L < 1$ então existe um e um só ponto fixo em $\mathbb{R}$. Um exemplo é considerar $g(x) = a\cos(x)$ com $|a| < 1$.

3.4.2 Comportamento assimptótico da convergência.

Estamos agora em condições de estudar o comportamento do erro obtido pela iteração do ponto fixo.

Proposição 3.5 *Seja A um operador F-diferenciável numa vizinhança do ponto fixo z. Se (x_n) é a sucessão obtida pela aplicação do método do ponto fixo convergente para z, então o erro $e_n = z - x_n$ verifica*

$$\frac{e_{n+1}}{||e_n||} - A'_z \frac{e_n}{||e_n||} \longrightarrow 0.$$

Demonstração: Sendo $z = Az$, $x_{n+1} = Ax_n$, temos

$$||x_{n+1} - z - A'_z(x_n - z)|| = ||Ax_n - Az - A'_z(x_n - z)|| = o(||x_n - z||),$$

portanto

$$||e_{n+1} - A'_z e_n|| = o(||e_n||),$$

o que significa que

$$\frac{||e_{n+1} - A'_z e_n||}{||e_n||} \longrightarrow 0. \ \blacksquare$$

• Este resultado significa que a razão $\frac{e_{n+1}}{||e_n||}$ se aproxima de $A'_z(\frac{e_n}{||e_n||})$. Se no caso real, foi imediato estabelecer que o coeficiente assimptótico de convergência era $|g'(z)|$, aqui não poderemos dizer que é $||A'_z||$.

Com efeito, o limite de $\frac{||e_{n+1}||}{||e_n||}$ pode não existir. Isto compreende-se pois pode acontecer que a sucessão $\frac{e_n}{||e_n||}$ não convirja. Qual a diferença com o caso real? No caso real, quando temos convergência alternada, o valor $\frac{e_n}{|e_n|}$ também não converge, pode ser ± 1, mas ao calcular o módulo, o valor $|g'(z)\frac{e_n}{|e_n|}|$ seria sempre $|g'(z)|$.

No entanto, podemos retirar algumas informações acerca do comportamento de $\frac{||e_{n+1}||}{||e_n||}$.

Corolário 3.2 *Nas condições da proposição anterior, temos*

$$\limsup \frac{||e_{n+1}||}{||e_n||} \leq ||A'_z||.$$

Se $A'_z = 0$, *então o método do ponto fixo tem convergência supralinear.*

Demonstração:

Usando a proposição anterior, é imediato que se $A'_z = 0$, então a convergência é supralinear.

Para obter a estimativa, designamos

$$\varepsilon_n = \frac{||e_{n+1} - A'_z e_n||}{||e_n||},$$

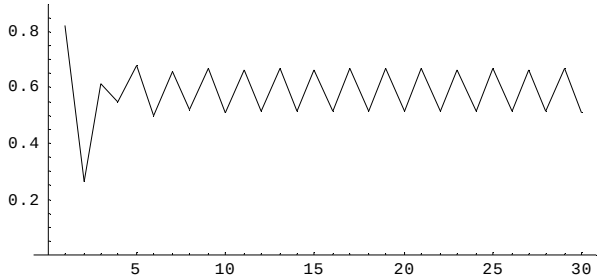
que tende para zero, de acordo com a proposição anterior, e obtemos

$$\frac{||e_{n+1}||}{||e_n||} \leq \frac{||e_{n+1} - A'_z e_n|| + ||A'_z e_n||}{||e_n||} = \varepsilon_n + ||A'_z(\frac{e_n}{||e_n||})|| \leq \varepsilon_n + ||A'_z||. \ \blacksquare$$

• Concluímos assim que a razão $K_n = \frac{||e_{n+1}||}{||e_n||}$ pode oscilar, mas no limite os seus valores não devem ser superiores a $||A'_z||$.

A noção de coeficiente assimptótico de convergência pode ser generalizada considerando $\tilde{K}_\infty = \limsup K_n$ e assim podemos concluir que no método do ponto fixo, quando há convergência linear, $\tilde{K}_\infty \leq ||A'_z||$.

Exemplo 3.8 *Consideremos a função* $g(x_1, x_2) = 0.9(\cos(x_2), \sin(x_1))$, *que tem apenas um ponto fixo* $z = (0.7395, 0.6065)$. *Começando com* $x^{(0)} = (0, 0)$, *colocamos no gráfico seguinte os valores de* $K_n = \frac{||e_{n+1}||_\infty}{||e_n||_\infty}$ *e verificamos que eles oscilam entre os valores* 0.513 *e* 0.665.



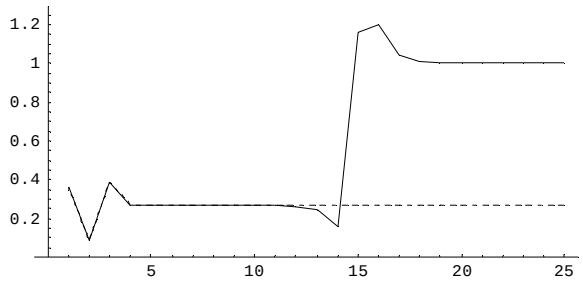
Este exemplo ilustra o facto de não se poder falar no coeficiente assimptótico de convergência como o limite, mas apenas como o limite superior. Reparando que

$$||\boldsymbol{\nabla} g(z)||_\infty = \max\{0.9\,|\sin 0.6065|\,, 0.9\,|\cos 0.7395|\} = \max\{0.513, 0.665\} = 0.665...$$

concluímos que a estimativa para K_∞ *coincide com o valor da norma.*

• *Num outro exemplo consideramos* $g(x_1, x_2, x_3) = \frac{1}{2}(\cos(x_1 x_3), \sin(x_3), \sin(x_1 + x_3))$, *com ponto fixo* $z = (0.4911, 0.1872, 0.3837)$. *Começamos com* $x^{(0)} = (1, 0, 0)$ *e reparamos que a*

razão K_n fica constante, próximo de 0.27 até $n < 12$, depois sofre um incremento súbito e para $n > 18$ vai ficar próximo de 1 (ver figura em baixo, curva contínua). Quando temos convergência linear e o valor K_n fica muito próximo ou maior que 1, significa que o método deixou de convergir, normalmente porque foi esgotada a precisão nos cálculos. Aumentando a precisão, verificamos que o salto desaparece (curva a tracejado), tendo sido corrigida a imprecisão numérica.



Neste exemplo $K_n \to 0.2727$, valor que é mais baixo que $||\nabla g(z)||_\infty = 0.641$, como previsto pela teoria.

3.4.3 Convergência de ordem superior

Pelo que vimos no parágrafo precedente,

$$||e_{n+1} - A'_z e_n|| = o(||e_n||),$$

e assim, quando a F-derivada A' é nula no ponto fixo z, obtivémos $\frac{||e_{n+1}||}{||e_n||} = o(1)$,o que significa que a convergência é supralinear. Resta saber se podemos especificar essa convergência em termos de ordem p, definindo:

Definição 3.12 *Dizemos que x_n converge para z com* pelo menos *ordem de convergência p se*

$$K_n^{[p]} = \frac{||e_{n+1}||}{||e_n||^p} \leq K.$$

Quando $K_n^{[p]}$ não tende para zero, podemos dizer que a ordem de convergência é exactamente p. No entanto, o que nos interessa neste momento saber é se o facto de $A'_z = 0$ implica uma convergência pelo menos quadrática, como acontecia no caso real, quando a função era regular.

Aqui também será necessário considerar uma maior regularidade para A, de forma a que possa ser estabelecido um desenvolvimento de segunda ordem,

$$A(x+h) = Ax + A'_x h + \frac{1}{2} A''_x(h,h) + o(||h||^2),$$

em que A''_x é uma função bilinear contínua correspondente à segunda derivada (no caso de $\mathbb{R}^N$ corresponde a considerar as matrizes hessianas).

Desta forma, obtemos

$$x_{n+1} - z = Ax_n - Az = A'_z(x_n - z) + \frac{1}{2}A''_z(x_n - z, x_n - z) + o(||x_n - z||^2),$$

e portanto, como supômos $A'_z = 0$,

$$||e_{n+1} + \frac{1}{2}A''_z(e_n, e_n)|| = o(||e_n||^2) \Leftrightarrow ||\frac{e_{n+1}}{||e_n||^2} + \frac{1}{2}A''_z(\frac{e_n}{||e_n||}, \frac{e_n}{||e_n||})|| = \varepsilon_n = o(1),$$

o que significa que[11]

$$\frac{||e_{n+1}||}{||e_n||^2} \leq \frac{1}{2}||A''_z|| + \varepsilon_n \leq K,$$

ou seja, a convergência é pelo menos quadrática e temos $\tilde{K}_\infty \leq \frac{1}{2}||A''_z||$.

3.4.4 Método de Newton

Neste contexto geral dos espaços de Banach, apenas fazemos uma breve referência ao método de Newton, já que iremos ver no próximo capítulo a aplicação a sistemas não-lineares, que nos irá interessar particularmente.

Tal como vimos, no estudo em $\mathbb{R}$ ou em $\mathbb{C}$, o método de Newton aparece como um caso particular do método do ponto fixo, tendo uma convergência quadrática desde que a função seja diferenciável e que a derivada não se anule.

No caso dos espaços de Banach, fazemos aparecer de forma semelhante o método de Newton, exigindo que o operador seja F-diferenciável, e que a derivada de Fréchet seja invertível numa vizinhança da solução. Nessas condições, podemos estabelecer a equivalência:

$$Ax = 0 \Leftrightarrow (A'_x)^{-1}(Ax) = 0,$$

porque o inverso do operador linear contínuo A'_x será um operador linear contínuo, e portanto só será nulo quando o seu argumento for nulo (neste caso o argumento é Ax).

Assim, $Ax = 0$ é equivalente a

$$x = x - (A'_x)^{-1}(Ax)$$

e, dado x_0, obtemos o método de Newton

$$x_{n+1} = x_n - (A'_{x_n})^{-1}(Ax_n), \tag{3.6}$$

que nesta generalização também é designado como *método de Newton-Kantorovich.*

Observação:

[11] A norma $||A''_z||$ é a norma das aplicações bilineares contínuas, definida por

$$||B|| = \sup_{v,w \neq 0} \frac{||B(v,w)||}{||v||\,||w||}.$$

Podemos verificar que o método de Newton-Kantorovich tem convergência supralinear. Sendo $Gx = x - (A'_x)^{-1}(Ax)$, e como $z = Gz$, podemos ver que $G'_z = 0$.

Com efeito,

$$G(z+h) - G(z) = z + h - (A'_{z+h})^{-1}(A(z+h)) - z,$$

e reparando que $A(z+h) = A(z) + A'_{z+h}h + o(||h||) = A'_{z+h}h + o(||h||)$, temos

$$G(z+h) - G(z) = h - (A'_{z+h})^{-1}(A(z+h)) = h - (A'_{z+h})^{-1}(A'_{z+h}h + o(||h||)).$$

Usando a linearidade de $(A'_{z+h})^{-1}$,

$$G(z+h) - G(z) = h - h - (A'_{z+h})^{-1}(o(||h||)) = o(||h||),$$

porque $||(A'_{z+h})^{-1}(o(||h||))|| \leq ||(A'_{z+h})^{-1}||\, o(||h||) = o(||h||)$, admitindo que $(A'_{z+h})^{-1}$ são limitados.

Podemos mesmo ver que se trata de convergência quadrática, se admitirmos que $o(||h||) = O(||h||^2)$, o que poderia ser obtido considerando um desenvolvimento de segunda ordem em A, como referido antes.

3.5 Exercícios

1. Seja E um espaço de Banach, e A um operador linear contínuo em E, isto é $A \in \mathcal{L}(E,E)$.

a) Mostre que se $||A||_{\mathcal{L}(E,E)} < 1$, então o operador $(I-A)^{-1}$ existe e pertence a $\mathcal{L}(E,E)$.

Para além disso, mostre que se verifica a igualdade com a *série de Neumann* (que é uma generalização da série geométrica),

$$(I-A)^{-1} = \sum_{k=0}^{\infty} A^k, \quad \text{e que } ||(I-A)^{-1}||_{\mathcal{L}(E,E)} \leq \frac{1}{1 - ||A||_{\mathcal{L}(E,E)}}$$

Sugestão: A partir da equivalência $(I-A)X = I \Leftrightarrow X = I + AX$, escolha $G(X) = I + AX$ como função iteradora e aplique o teorema do ponto fixo de Banach.

b) Mostre que se $A \in \mathcal{L}(E,E)$ for invertível e se tivermos

$$||B - A|| < \frac{1}{||A^{-1}||}$$

então B é invertivel e temos:

$$||B^{-1}|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}||\, ||B - A||}$$

Sugestão: Prove que B^{-1} é a solução única de $X = A^{-1} + (I - A^{-1}B)X$, e aplique o teorema do ponto fixo de Banach.

Nota: *Deve distinguir-se invertibilidade de invertibilidade à direita ou à esquerda! Dizemos que T é invertível à direita se $\exists X : TX = I$, e que T é invertível à esquerda se*

$\exists X : XT = I$. Por exemplo, se D é o operador derivação e P o operador de primitivação, sabemos que $DP = I$, mas $PD \neq I$ (a igualdade só é válida a menos da constante de primitivação).

Apenas quando se verifica invertibilidade à esquerda e à direita dizemos que T é invertível.

Neste exercício, nas sugestões, apenas são dadas indicações para mostrar a invertibilidade à direita.

2. Considere a equação de Fredholm

$$f(x) = \lambda \int_a^b K(x,y) f(y) dy + \phi(x)$$

em que $K \in C([a,b] \times [a,b])$, e $\phi \in C[a,b]$.

a) Mostre que se

$$|\lambda|(b-a) \max_{x,y \in [a,b]} |K(x,y)| < 1$$

então existe uma e uma só solução f contínua em $[a,b]$ e que o método do Ponto Fixo define uma sucessão de funções que converge uniformemente para f.

b) Usando a alínea anterior, indique um intervalo para λ tal que

$$\lambda \int_{-1}^{1} (x^2 + y^2) f(y) dy + x = f(x)$$

tenha solução única em $C[-1,1]$.

c) Considere $\lambda = 1/5$ em b). Usando como iterada inicial $f_0(x) = 0$, determine as duas primeiras iteradas e conclua acerca da solução.

Se utilizar $f_0(x) = 1$, determine um majorante do erro para f_{10} na norma $||.||_\infty$, calculando apenas a primeira iterada.

3. Considere $I = [d - \varepsilon, d + \varepsilon]$, e o espaço das funções contínuas $C(I)$ munido da norma

$$||f||_e = \max_{x \in I} |f(x) e^{-M|x-d|}|$$

com $M \geq 0$ fixo. Mostre que esta norma é equivalente á norma $||f||_\infty$.

4. *(Teorema de Picard-Lindelöf)*

a) Consideremos o problema de Cauchy

$$\begin{cases} y' = f(x,y) \\ y(x_0) = y_0 \end{cases}$$

e o conjunto $Q = [x_0 - a, x_0 + a] \times [y_0 - b, y_0 + b]$, onde f é uma função Lipschitziana:

$$|f(x,y) - f(x,w)| \leq M|y - w|, \ \ \forall (x,y),(x,w) \in Q$$

verificando $|f(x,y)| \leq K, \ \forall (x,y) \in Q$.

Se $0 < c < \min\{a, \frac{b}{K}, \frac{1}{M}\}$, então existe uma e uma só solução contínua no intervalo $I = [x_0 - c, x_0 + c]$ e a *iteração de Picard* definida por $y_0(x) = y_0$,

$$y_{n+1}(x) = y_0 + \int_{x_0}^{x} f(t, y_n(t))dt$$

converge uniformemente para a solução, no intervalo I.

Sugestão: Utilizar o conjunto fechado $S = \{\phi \in C(I) : \ ||\phi - y_0||_\infty \leq b\}$.

b) Mostre que pode exigir apenas $0 < c < \min\{a, \frac{b}{K}\}$ se considerar em a) a norma $||f||_e$ utilizada no exercício anterior.

c) Considere a equação diferencial

$$y' - y^2/2 = x^2/2$$

com $y(0) = 0$. Aplique o resultado de a) usando $a = 1, b = 1$. Calcule 2 iteradas e determine um majorante do erro na norma $||.||_\infty$.

5. Considere os operadores $A_m : C[0,1] \to C[0,1]$, com $m \in \mathbf{N}$

$$A_m f(x) = \int_0^x (f(t))^m dt$$

a) Mostre que a derivada de Fréchet de A_m é $A'_{m,f}(h) = m \int_0^x f(t)^{m-1} h(t)dt$.

b) Mostre que no espaço $X = \{f \in C[0,1] : ||f||_\infty \leq \frac{3}{4}\}$ existe uma e uma só solução f de

$$\int_0^x (f(t))^5 dt + \int_0^x (f(t))^2 dt + 4f(x) = \phi(x), \tag{3.7}$$

onde $||\phi||_\infty < 1$. Indique um método iterativo que convirja, qualquer que seja a iterada inicial em X.

c) Determine um número de iteradas suficiente pelo método do ponto fixo começando com $f_0 = 0$, para que f_n, uma aproximação da solução da equação (3.7) verifique $||e_n||_\infty < 10^{-2}$.

6. **a)** Seja E um espaço de Banach e X um conjunto não vazio, fechado. Mostre que um operador contínuo $A : E \to E$ que verifica $X \subseteq A(X)$ (pode assumir que $A(X)$ é fechado) e

$$||Ax - Ay|| \geq L||x - y||, \forall x, y \in X$$

para um certo $L > 1$, tem um e um só ponto fixo $z \in X$, e que o método $x_n = Ax_{n+1}$ converge para esse ponto fixo, qualquer que seja $x_0 \in X$.

b) Seja A uma contracção num aberto não vazio X, onde se sabe que existe um ponto fixo z de A.

Mostre que existe um subconjunto de X onde são verificadas as condições do Teorema do Ponto Fixo de Banach, e conclua que pode assegurar convergência local do método do ponto fixo.

---

7. Considere $E = \mathbb{R}^N$. Mostre que uma matriz A com a diagonal estritamente dominante por linhas é invertível, escrevendo $A = DC$ em que D é a matriz diagonal e, verificando que $||I - C||_\infty < 1$. Conclua que o processo iterativo

$$X_0 = I; \;\; X_{n+1} = I + X_n - C\, X_n$$

permite obter uma matriz X tal que $A^{-1} = D^{-1}X$.

8. Considere a sucessão de funções em $C([a,b])$,

$$f_{n+1}(x) = g(f_n(x))$$

para um qualquer $f_0 \in X = \{f \in C([a,b]) : f([a,b]) \subseteq [a,b]\}$.

a) Mostre que X é fechado em $C([a,b])$ para a norma $||.||_\infty$.

b) Mostre que se $g([a,b]) \subseteq [a,b]$ e g for contractiva em $[a,b]$, então a sucessão (f_n) converge uniformemente para $f(x) = z$, em que z é o ponto fixo de g em $[a,b]$.

c) Aplique o resultado anterior para determinar a função que é o limite da sucessão de funções em $C([0,1])$,

$$f_{n+1}(x) = \cos(f_n(x))$$

para um qualquer f_0 que verifique $f_0([0,1]) \subseteq [0,1]$.

9. a) Verifique que a sucessão $w_k = \frac{1}{k(k+1)}$ está em l^1, ou seja,

$$||w||_1 = \sum_{k=1}^{\infty} |w_k| < +\infty$$

e determine $||w||_1$.

b) Mostre que se a iterada inicial $x^{(0)} \in l^1$, a sucessão de sucessões definida por

$$x^{(n+1)} = (x^{(n)})/2 + w$$

converge para a sucessão $2w$ na norma $||.||_1$.

10. Considere a equação integral

$$f(x) = x + \int_0^x (f(t))^2 dt$$

que pretendemos resolver em $Y = C([0,a])$, para $a < 1/2$, usando o operador iterativo

$$Af(x) = x + \int_0^x (f(t))^2 dt$$

a) Mostre que a derivada de Fréchet de A é o operador dado por

$$(A'(f)h)(x) = \int_0^x 2h(t)f(t)dt$$

b) Mostre que $||A'(f)||_{\mathcal{L}(Y,Y)} \leq 2a||f||_\infty$.

c) Conclua que, se considerarmos o subconjunto $X = \{f \in C([0,a]) : \; ||f||_\infty \leq M\}$ com $1 \leq M < \frac{1}{2a}$, existe uma e uma só função $f \in X$ que verifica a equação integral.

d) Considere $a = 1/3, M = 1$ e $f_{n+1} = Af_n$, com $f_0(x) = 0$. Determine f_3 e um majorante para o erro absoluto $||e_3||_\infty$.

11. Seja A um operador linear contínuo e invertível num espaço de Banach E. Considere o seguinte método iterativo para determinar A^{-1}:

$$X_{n+1} = 2X_n - X_nAX_n$$

a) Mostre que se considerarmos um X_0 inicial, tal que $||I - AX_0||_{\mathcal{L}(E,E)} < \frac{1}{2}$, e que $||I - X_0A||_{\mathcal{L}(E,E)} < \frac{1}{2}$, então o método converge para A^{-1}.

Sugestão: Para verificar uma das condições do teorema do ponto fixo, num conjunto fechado apropriado, utilize, por exemplo, a igualdade $I - A(2X - XAX) = (I - AX)^2$.

b) Mostre que a convergência do método é quadrática, ou seja, verifica-se:

$$||A^{-1} - X_{n+1}||_{\mathcal{L}(E,E)} \leq K||A^{-1} - X_n||^2_{\mathcal{L}(E,E)}$$

onde $K = ||A||_{\mathcal{L}(E,E)}$.

c) Seja $L = ||I - \omega A||_{\mathcal{L}(E,E)} < 1$, para um certo ω. Mostre que A^{-1} pode também ser aproximado pelo método de relaxação $Y_{n+1} = Y_n + \omega(I - AY_n)$, e verifique que

$$||A^{-1} - Y_n|| \leq \omega\frac{L^n}{1 - L}.$$

Conclua que se $||A|| < 1$ e $n > \log(\frac{1-L}{2\omega})/\log(L)$, então podemos considerar Y_n como iterada inicial X_0, no método da alínea a), e assim obter uma rapidez de convergência quadrática.

Capítulo 4

Resolução de Sistemas de Equações

Vamos agora nos concentrar na resolução de sistemas de equações em $\mathbb{R}^N$, aplicando os resultados obtidos no capítulo anterior. Começaremos por ver o caso dos sistemas de equações não lineares, e depois iremos ver mais especificamente o caso dos sistemas lineares. Iniciamos este capítulo estudando as normas de operadores lineares em $\mathbb{R}^N$, ou seja, normas de matrizes.

4.1 Normas de Matrizes

Como estaremos interessados em trabalhar em $\mathbb{R}^N$, vamos começar por relembrar algumas normas vectoriais, notando que todas as normas em $\mathbb{R}^N$ são equivalentes, pois trata-se de um espaço de dimensão finita. A partir dessas normas iremos introduzir normas matriciais, já que os operadores lineares em $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ são representados por matrizes, e podemos caracterizar de forma fácil essas normas, a partir das normas vectoriais definidas em $\mathbb{R}^N$. Estaremos apenas interessados no caso de matrizes quadradas, mas com as devidas adaptações estas normas e as suas expressões podem ser generalizadas a qualquer tipo de matrizes.

Consideremos uma norma qualquer em $\mathbb{R}^N$ e seja $A : \mathbb{R}^N \to \mathbb{R}^N$ um operador linear (que será uma matriz quadrada).

Observação: *(uma matriz define um operador contínuo para qualquer norma).*

Basta ver que é contínuo em zero para uma norma, porque as outras são equivalentes – escolhemos a euclidiana $||.||_2$.

Com efeito, designando por $\mathbf{a}^i$ as linhas de A, e por $\mathbf{e}^i$ a base, se $||\mathbf{x}^{(k)}||_2 \to 0$ temos

$$||A\mathbf{x}^{(k)}||_2 = ||\sum_{i=1}^{N} \mathbf{a}^i.\mathbf{x}^{(k)}\mathbf{e}^i||_2 \leq \sum_{i=1}^{N} |\mathbf{a}^i.\mathbf{x}^{(k)}| \; ||\mathbf{e}^i||_2 \leq$$

e aplicando a desigualdade de Cauchy-Schwarz temos

$$\leq ||\mathbf{x}^{(k)}||_2 \sum_{i=1}^{N} ||\mathbf{a}^i||_2 \; ||\mathbf{e}^i||_2 \to 0. \; \square$$

Concluímos que as matrizes definem operadores lineares contínuos em $\mathbb{R}^N$, e podemos definir a norma de uma matriz *induzida* pela norma $||.||$ aplicando a definição geral, usada para operadores lineares contínuos [1],

$$||A|| = \sup_{||x||\leq 1} ||Ax|| = \max_{||x||=1} ||Ax|| = \sup_{||x||\neq 0} \frac{||Ax||}{||x||} \tag{4.1}$$

Proposição 4.1 *Sejam A, B matrizes quadradas, x um vector e $||A||$ uma norma matricial de A induzida por $||.||$, em $\mathbb{R}^N$. Temos:*
a) $||Ax|| \leq ||A||\ ||x||$ (o que significa que a norma matricial é compatível *com a norma vectorial)*
b) $||AB|| \leq ||A||\ ||B||$ (o que significa que a norma matricial é regular*)*

Demonstração:
a) Para um qualquer vector $x \neq 0$,

$$||A|| = \sup_{||y||\neq 0} \frac{||Ay||}{||y||} \geq \frac{||Ax||}{||x||}$$

logo $||Ax|| \leq ||A||\ ||x||$ e para $x = 0$ é trivial.
b) Como $||AB|| = \sup_{||x||\leq 1} ||ABx||$ usando a) temos:

$$||AB|| \leq \sup_{||x||\leq 1} ||A||\ ||Bx|| = ||A|| \sup_{||x||\leq 1} ||Bx|| = ||A||\ ||B||. \blacksquare$$

Exemplo 4.1 *Consideremos a norma ("da soma") $||x||_1 = \sum_{i=1}^N |x_i|$. Queremos obter uma expressão para a norma induzida. Ora,*

$$||Ax||_1 = \sum_{i=1}^N |\sum_{j=1}^N a_{ij}x_j| \leq \sum_{i=1}^N \sum_{j=1}^N |a_{ij}|\ |x_j| \leq \sum_{j=1}^N \left(|x_j| \sum_{i=1}^N |a_{ij}| \right) \leq$$

$$\leq \sum_{j=1}^N |x_j| \max_{j=1,\ldots,N} \sum_{j=1}^N |a_{ij}| \leq \alpha ||x||_1$$

[1] As normas induzidas não são as únicas normas matriciais que se podem definir em $\mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$. Por exemplo, a norma de Fröbenius

$$||A||_{Fr} = \sqrt{\sum_{i,j=1}^N |a_{ij}|^2}$$

não é induzida por nenhuma norma de $\mathbb{R}^N$.

Esta norma corresponde a considerar matriz como um vector e aplicar a norma euclidiana. No entanto, como iremos ver, a expressão da norma matricial induzida pela norma euclidiana é diferente. Pode mostrar-se que

$$||A||_2 \leq ||A||_{Fr}$$

e portanto $||A||_1 \leq \alpha$, *definindo* $\alpha = \max_{j=1,\ldots,N} \sum_{i=1}^{N} |a_{ij}|$.

Com efeito, também temos $\alpha \leq ||A||_1$, *pois escolhendo* $v = \mathbf{e}_k$ *(vector da base canónica), onde* k *é o índice que dá o* j *máximo na definição de* α, *isto é* $\alpha = \sum_{i=1}^{N} |a_{ik}|$, *e como* $||v||_1 = 1$, *obtemos:*

$$||Av||_1 = |\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}\delta_{jk}| = \sum_{i=1}^{N} |a_{ik}| = \alpha,$$

em que δ_{ij} *é o delta de Kronecker (*$\delta_{ij} = 0$ *se* $i \neq j$, *e* $\delta_{ij} = 1$ *se* $i = j$*).*

Assim, $||A||_1 \geq ||Av||_1 = \alpha$ *e concluímos a igualdade, ou seja:*

$$||A||_1 = \max_{j=1,\ldots,N} \sum_{i=1}^{n} |a_{ij}|. \tag{4.2}$$

*– Para calcular esta norma matricial basta, portanto, somar em cada coluna os módulos dos elementos, e encontrar o máximo desses valores. Este processo leva a que esta norma também seja conhecida como "*norma das colunas*".*

Exemplo 4.2 *Analogamente, considerando a norma ("do máximo")*

$$||x||_\infty = \max\{|x_1|, \ldots, |x_N|\}$$

obtemos a norma matricial

$$||A||_\infty = \max_{i=1,\ldots,N} \sum_{j=1}^{N} |a_{ij}|, \tag{4.3}$$

*ou seja, para calcular esta norma somamos em cada linha os módulos dos elementos e encontramos o máximo. Neste caso, é costume designar esta norma, como "*norma das linhas*".*

Podemos justificar a expressão dada para a norma induzida. Com efeito sendo $\alpha = ||A||_\infty$, *obtemos*

$$||Ax||_\infty = \max_{i=1,\ldots,N} |\sum_{j=1}^{N} a_{ij}x_j| \leq \max_{i=1,\ldots,N} \sum_{j=1}^{N} |a_{ij}| \max_{j=1,\ldots,N} |x_j| = \alpha||x||_\infty,$$

e isto justifica $\frac{||Ax||_\infty}{||x||_\infty} \leq \alpha$. *Para obtermos a igualdade basta considerar um valor particular. No caso em que a matriz tem elementos positivos esse valor será* $v = (1, ..., 1)$, *porque como* $||v||_\infty = 1$, *obtemos:*

$$||Av||_\infty = \max_{i=1,\ldots,N} |\sum_{j=1}^{N} a_{ij}1| = \max_{i=1,\ldots,N} \sum_{j=1}^{N} a_{ij} = \alpha.$$

Para uma matriz qualquer, poderíamos considerar $v = (\pm 1, ..., \pm 1)$, *em que os sinais seriam escolhidos de forma a que na linha* m, *em que se atingisse o máximo, o produto* $a_{mj}v_j$ *fosse positivo... para isso basta escolher sinais iguais aos sinais de* a_{mj}. *No caso complexo, o mesmo efeito seria obtido considerando o conjugado do argumento de* a_{mj}.

Outra norma importante em $\mathbb{R}^N$ é a norma euclidiana

$$||x||_2 = \sqrt{x_1^2 + \ldots + x_N^2}\,,$$

mas já não é tão fácil de calcular a norma que induz nas matrizes. Para obter $||A||_2$ necessitamos de calcular um raio espectral.

Definição 4.1 *Dada uma matriz C, com valores próprios $\lambda_1, \ldots, \lambda_N$, definimos* raio espectral *de C como sendo o valor*

$$\rho(C) = \max_{i=1,\ldots,N} |\lambda_i| \tag{4.4}$$

Exemplo 4.3 *A norma euclidiana induz a norma matricial*

$$||A||_2 = \sqrt{\rho(A^*A)} \tag{4.5}$$

onde $A^* = \bar{A}^T$ e onde ρ é o raio espectral.
É claro que no caso de matrizes hermitianas, ou seja $A^* = A$, temos:

$$||A||_2 = \rho(A)$$

pois os valores próprios de A^2 serão os valores próprios de A ao quadrado, reparando que $\det(\lambda^2 I - A^2) = \det(\lambda I - A)\,\det(\lambda I + A)$.
• *No caso de matrizes hermitianas, podemos deduzir este resultado usando a decomposição espectral $A = U^*DU$, em que U é uma matriz unitária e D é uma matriz diagonal (forma de Schur). Notando que $||x||_2^2 = x^*x$, e que $||U||_2 = ||U^*||_2 = 1$, obtemos*

$$||Ax||_2^2 = (Ax)^*Ax = x^*A^*Ax = x^*UD^*U^*U^*DUx \leq ||x^*||_2||D||_2^2||x||_2 = ||D||_2^2||x||_2^2.$$

Como para matrizes diagonais é fácil ver que $||D||_2 = \max_{i=1,\ldots,N} |d_{ii}|$, e como neste caso a decomposição espectral diz-nos que a matriz D tem os valores próprios de A, concluímos que $||D||_2 = \max_{i=1,\ldots,N} |\lambda_i| = \rho(A)$. Fica assim demonstrado que $||A||_2 \leq \rho(A)$. Para obter a igualdade, basta considerar v como sendo um vector próprio unitário associado ao valores próprio dominante, $|\lambda_m|$. Nesse caso

$$||Av||_2^2 = ||\lambda_m v||_2^2 = |\lambda_m|^2\,||v||_2^2 = \rho(A)^2.$$

• *No caso de uma matriz qualquer o raciocínio é semelhante. Basta reparar que $||Ax||_2^2 = x^*A^*Ax$, e que A^*A é uma matriz hermitiana com valores próprios $\mu_1, ..., \mu_N$ não negativos. Portanto, obtemos $A^*A = \tilde{U}^*\tilde{D}\tilde{U}$, em que $\tilde{D}$ terá na diagonal os valores próprios de A^*A. Assim,*

$$||Ax||_2^2 = x^*\tilde{U}^*\tilde{D}\tilde{U}x \leq ||x^*||_2||\tilde{D}||_2||x||_2 = ||\tilde{D}||_2||x||_2^2$$

e como $||\tilde{D}||_2 = \max_{i=1,\ldots,N} \mu_i = \rho(A^*A)$, concluímos que $||A||_2^2 \leq \rho(A^*A)$. Para verificar a igualdade também é semelhante, reparando que $||Av||_2^2 = v^*(A^*Av) = v^*(\mu_m v) = \mu_m = \rho(A^*A)$.

O raio espectral está relacionado com as normas induzidas, podendo ser encarado como o *ínfimo de todas as normas induzidas*, como podemos concluir pelo seguinte resultado.

Teorema 4.1 *(i) Qualquer que seja a norma matricial* $||.||$ *(induzida por uma norma vectorial) temos, para qualquer matriz* A,

$$\rho(A) \leq ||A||. \tag{4.6}$$

(ii) Dada uma qualquer matriz A*, para todo o* $\varepsilon > 0$*, existe*[2] *uma norma induzida* $||.||$ *tal que:*

$$||A|| \leq \rho(A) + \varepsilon \tag{4.7}$$

Demonstração:

i) Como $\rho(A) = |\lambda|$, para um certo λ valor próprio de A, temos $Av = \lambda v$ para certo $v \neq 0$, e portanto

$$||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} \geq \frac{|\lambda| \, ||v||}{||v||} \geq |\lambda|.$$

ii) Seguimos a demonstração de [25]. Consideramos a decomposição na forma canónica de Jordan $A = PJP^{-1}$ em que $J = J_D + J_E$, sendo J_D uma matriz diagonal e J_E a matriz contendo os 1's na subdiagonal superior, respeitantes aos blocos de Jordan.

Tomando a matriz diagonal $D_\varepsilon = [\varepsilon^{i-1}\delta_{ij}]$ em que δ_{ij} é o delta de Kronecker, obtemos $D_\varepsilon^{-1} J D_\varepsilon = J_D + \varepsilon \, J_E$, e portanto

$$||D_\varepsilon^{-1} J D_\varepsilon||_\infty \leq ||J_D||_\infty + \varepsilon \, = \rho(A) + \varepsilon \, .$$

Resta ver que existe uma norma induzida $||.||$ tal que $||A|| = ||D_\varepsilon^{-1} J D_\varepsilon||_\infty$. Para isso consideramos como norma vectorial $||x|| = ||D_\varepsilon^{-1} P^{-1} x||_\infty$.

Basta verificar que

$$||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} = \sup_{x \neq 0} \frac{||D_\varepsilon^{-1} P^{-1} A x||_\infty}{||D_\varepsilon^{-1} P^{-1} x||_\infty} = \sup_{x \neq 0} \frac{||D_\varepsilon^{-1} J P^{-1} x||_\infty}{||D_\varepsilon^{-1} P^{-1} x||_\infty} =$$
$$= \sup_{x \neq 0} \frac{||(D_\varepsilon^{-1} J D_\varepsilon) P^{-1} D_\varepsilon^{-1} x||_\infty}{||P^{-1} D_\varepsilon^{-1} x||_\infty} = \sup_{\substack{y = P^{-1} D_\varepsilon^{-1} x \\ y \neq 0}} \frac{||D_\varepsilon^{-1} J D_\varepsilon \, y||_\infty}{||y||_\infty} = ||D_\varepsilon^{-1} J D_\varepsilon||_\infty. \ \blacksquare$$

4.2 Métodos Iterativos para Sistemas de Equações Não Lineares

Iremos agora apresentar métodos iterativos que permitem aproximara a solução de sistemas de equações. Começamos por apresentar o caso geral, em que se supõe que o sistema pode ou

[2] A norma obtida em (ii) depende *a priori* da matriz A.

Note-se que $\forall A, \forall \varepsilon > 0 \, \exists ||.|| : ||A|| \leq \rho(A) + \varepsilon$, não é equivalente a $\forall \varepsilon > 0 \, \exists ||.|| : \forall A, \ ||A|| \leq \rho(A) + \varepsilon$.

Por outro lado, convém referir que o facto de se garantir que se trata de uma norma induzida permite saber que se trata de uma norma compatível e regular.

não ser linear. No caso de se tratar de um sistema linear, os métodos iterativos constituem apenas um complemento aos métodos directos conhecidos da Álgebra Linear (p.ex: método de eliminação de Gauss, de que falaremos mais à frente).

Tendo discutido as normas matriciais em $\mathbb{R}^N$, estamos nas condições de aplicar o corolário do teorema do ponto fixo usando a matriz jacobiana, que corresponde à derivada de Fréchet em $\mathbb{R}^N$.

Sendo assim, dado um sistema de equações em $\mathbb{R}^N$

$$\begin{cases} f_1(x_1, \dots, x_N) = 0, \\ \vdots \\ f_N(x_1, \dots, x_N) = 0, \end{cases}$$

que podemos escrever abreviadamente $F(x) = 0$, estabelecemos uma equivalência com o sistema na forma $x = G(x)$, ou seja,

$$\begin{cases} x_1 = g_1(x_1, \dots, x_N), \\ \vdots \\ x_N = g_N(x_1, \dots, x_N), \end{cases}$$

Sendo $\boldsymbol{\nabla} G(x)$ a matriz jacobiana de G calculada no ponto x,obtemos como consequência imediata do que vimos no capítulo anterior, o seguinte *teorema do ponto fixo em* $\mathbb{R}^N$:

Corolário 4.1 *(do Teorema de Ponto Fixo de Banach). Seja D um conjunto não vazio, fechado e convexo de $\mathbb{R}^N$.*
Se $G \in C^1(D)$ e $||.||$ é uma norma qualquer em $\mathbb{R}^N$, tal que:
i) $||\boldsymbol{\nabla} G(x)|| \leq L < 1, \ \forall x \in D$
ii) $G(D) \subseteq D$
então estamos nas condições do Teorema do Ponto Fixo de Banach, logo:
i) Existe um e um só ponto fixo $z \in D : z = G(z) \ (\Leftrightarrow F(z) = 0)$
ii) O método do ponto fixo $x^{(n+1)} = G(x^{(n)})$ converge para z, qualquer que seja $x_0 \in D$.
iii) São válidas as estimativas
$||z - x^{(n)}|| \leq L||z - x^{(n-1)}|| \leq L^n||z - x^{(0)}||$
$||z - x^{(n)}|| \leq \frac{1}{1-L}||x^{(n+1)} - x^{(n)}||$
$||z - x^{(n)}|| \leq \frac{L^n}{1-L}||x^{(1)} - x^{(0)}||$

Exemplo 4.4 *Se retomarmos o sistema que já vimos num exemplo anterior:*

$$\begin{cases} 3x_1 + x_2 = 1 \\ 2x_1 + 4x_2 + x_3 = 0 \\ x_2 + 2x_3 = 2 \end{cases} \Leftrightarrow \begin{cases} x_1 = 1/3 - x_2/3 \\ x_2 = -x_1/2 - x_3/4 \\ x_3 = 1 - x_2/2 \end{cases}$$

em que considerámos $G : \mathbb{R}^3 \to \mathbb{R}^3$ *definido por*

$$G(x) = \begin{bmatrix} 1/3 \\ 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & -1/3 & 0 \\ -1/2 & 0 & -1/4 \\ 0 & -1/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{b} + \mathbf{A}\mathbf{x}$$

temos $||\boldsymbol{\nabla} G(\mathbf{x})||_\infty = ||\mathbf{A}||_\infty = 5/6 < 1$, *e garantimos a existência e unicidade de solução em* $\mathbb{R}^3$, *bem como a convergência do método. Alternativamente, com a norma* $||.||_1$, *também obteriamos a contractividade, pois* $||\mathbf{A}||_1 = 3/4 < 1$. *Mas como já foi referido, pode haver casos em que seja possível obter contractividade com uma das normas e não com a outra, o que não impede haver convergência em ambas.*

Exemplo 4.5 *Consideremos agora o sistema não-linear:*

$$\begin{cases} 2x - \cos(x+y) = 2 \\ 3y - \sin(x+y) = 6 \end{cases}$$

Vamos ver que existe uma e uma só solução em $\mathbb{R}^2$ *e que ela está em* $X = [1/2, 3/2] \times [5/3, 7/3]$. *Com efeito, se considerarmos*

$$G(x, y) = (\cos(x+y)/2 + 1, \sin(x+y)/3 + 2),$$

a matriz jacobiana de G *vem*

$$\boldsymbol{\nabla} G(x, y) = \begin{bmatrix} -\sin(x+y)/2 & -\sin(x+y)/2 \\ \cos(x+y)/3 & \cos(x+y)/3 \end{bmatrix}$$

Aplicando o corolário do T. Ponto Fixo, vemos que $||\boldsymbol{\nabla} G(x, y)||_1 \leq 5/6 < 1$ *e concluímos que existe uma e uma só solução em* $\mathbb{R}^2$ *(repare-se que se escolhessemos a norma* $||.||_\infty$ *teríamos apenas* $||\boldsymbol{\nabla} G(x, y)||_\infty \leq 1$, *o que revela bem que as condições são apenas suficientes e não necessárias). Por outro lado, reparando que* $G(\mathbb{R}^2) \subseteq X$ *porque*

$$1/2 \leq \cos(x+y)/2 + 1 \leq 3/2, \text{ e } 5/3 \leq \sin(x+y)/3 + 2 \leq 7/3$$

concluímos que a solução está em X. *Com efeito, poderiamos aplicar directamente o corolário usando este* X, *que é fechado e convexo, mas nesse caso apenas concluíamos a existência e unicidade em* X *e não em* $\mathbb{R}^2$. *Ao fim de algumas iterações (*~ 40*) obtemos como solução aproximada* $(0.549322733, 2.144360661)$.

4.2.1 Método de Newton para Sistemas de Equações

Como já referimos, uma possível escolha de função iteradora do método do ponto fixo em $\mathbb{R}$ (ou em $\mathbb{C}$) é a do método de Newton, que tem de um modo geral convergência mais rápida, sendo necessário que a função fosse diferenciável e que a derivada não se anulasse.

No caso de $\mathbb{R}^N$, vamos estabelecer um método semelhante, exigindo que a função seja C^1 e que a matriz jacobiana tenha inversa, numa vizinhança da solução. Assim, podemos estabelecer as equivalências

$$F(x) = 0 \Leftrightarrow [\boldsymbol{\nabla} F(x)]^{-1} F(x) = 0 \Leftrightarrow x = x - [\boldsymbol{\nabla} F(x)]^{-1} F(x)$$

e a função iteradora será, portanto, $G(x) = x - [\boldsymbol{\nabla} F(x)]^{-1} F(x)$.

Dado $x^{(0)} \in \mathbb{R}^N$, o método consistiria na iteração

$$x^{(n+1)} = x^{(n)} - [\boldsymbol{\nabla} F(x^{(n)})]^{-1} F(x^{(n)}). \tag{4.8}$$

No entanto, como iremos ver, o cálculo de uma matriz inversa é mais moroso que a resolução de um sistema, pelo que o método de Newton para sistemas não lineares consiste em, dada uma iterada inicial $x^{(0)} \in \mathbb{R}^N$, resolver, em cada iterada n, o sistema linear:

$$[\boldsymbol{\nabla} F(x^n)]v = -F(x^{(n)}) \tag{4.9}$$

e definir a próxima iterada $x^{(n+1)} = x^{(n)} + v$.

Desta forma, a resolução de um sistema não-linear pode ser conseguida (... se o método convergir!) através da resolução sucessiva de sistemas lineares.

Exemplo 4.6 *Consideremos o sistema do exemplo anterior. A matriz jacobiana de F vem*

$$\boldsymbol{\nabla} F(x,y) = \begin{bmatrix} 2 + \sin(x+y)/2 & \sin(x+y)/2 \\ -\cos(x+y)/3 & 3 - \cos(x+y)/3 \end{bmatrix}$$

inicializando com $x^{(0)} = (1,1)$ ao fim de 10 iterações obtemos um resultado com uma precisão semelhante ao obtido no exemplo para o método do ponto fixo.

Proposição 4.2 *(convergência local). Seja $F \in C^1(V_z)$, em que V_z é uma vizinhança de uma solução z, onde $\det(\boldsymbol{\nabla} F(x)) \neq 0$, $\forall x \in V_z$. Então o método de Newton converge para z, desde que a vizinhança seja suficientemente pequena e $x_0 \in V_z$.*

Demonstração: Exercício. ■

Teorema 4.2 *Seja $F \in C^2(V_z)$, em que a solução z não é um ponto crítico*[3]*. O método de Newton quando converge para z tem convergência pelo menos quadrática, ou seja, existe um $K > 0$ tal que*

$$||z - x^{(n+1)}|| \leq K||z - x^{(n)}||^2.$$

Demonstração:

Relembramos a fórmula de Taylor para uma função $f : \mathbb{R}^N \to \mathbb{R}$:

$$f(x+h) = f(x) + \nabla f(x) \cdot h + \frac{1}{2} h \cdot \boldsymbol{\nabla}^2 f(x + \xi h)\, h, \text{ para um certo } \xi \in]0,1[$$

onde $\boldsymbol{\nabla}^2 f(y) = [\frac{\partial^2 f}{\partial x_i \partial x_j}]$ é a matriz Hessiana de f calculada no ponto y.

No caso de uma função $F : \mathbb{R}^N \to \mathbb{R}^N$, $F = (f_1, ..., f_N)$ obtemos

$$F(x+h) = F(x) + \boldsymbol{\nabla} F(x) \cdot h + \frac{1}{2} h \cdot \boldsymbol{\nabla}^2 f_i(x + \xi_i h)\, h, \text{ para certos } \xi_i \in]0,1[,$$

[3] Ou seja, $\det(\boldsymbol{\nabla} F(z)) \neq 0$.

onde o termo $\frac{1}{2}h \cdot \boldsymbol{\nabla}^2 f_i(x+\xi_i h) \cdot h$ é um vector, que está apresentado na componente i.

Aplicando este resultado ao método de Newton, obtemos

$$0 = F(z) = F(x^{(n)}) + \boldsymbol{\nabla} F(x^{(n)}) \cdot e^{(n)} + \frac{1}{2} e^{(n)} \cdot \boldsymbol{\nabla}^2 f_i(x^{(n)} + \xi_i e^{(n)})\, e^{(n)}$$

em que $e^{(n)} = z - x^{(n)}$ é o erro na iterada n. Reparando que, no método de Newton, $\boldsymbol{\nabla} F(x^{(n)}) \cdot (x^{(n+1)} - x^{(n)}) = -F(x^{(n)})$, ao somar e subtrair z, ficamos com

$$-F(x^{(n)}) = \boldsymbol{\nabla} F(x^{(n)}) \cdot (x^{(n+1)} - z + z - x^{(n)}) = -\boldsymbol{\nabla} F(x^{(n)}) \cdot e^{(n+1)} + \boldsymbol{\nabla} F(x^{(n)}) \cdot e^{(n)},$$

obtendo-se

$$\boldsymbol{\nabla} F(x^{(n)}).e^{(n+1)} = -\frac{1}{2} e^{(n)} \cdot \boldsymbol{\nabla}^2 f_i(x^{(n)} + \xi_i e^{(n)})\, e^{(n)}.$$

Como $f \in C^2(V_z)$, supomos agora que $||\nabla^2 f_i(x)|| \leq M_2$, e que $||\, [\boldsymbol{\nabla} F(x_n)]^{-1}|| \leq \frac{1}{M_1}$, numa vizinhança da solução[4]. Obtemos a estimativa pretendida,

$$||e^{(n+1)}|| \leq \frac{M_2}{2M_1} ||e^{(n)}||^2. \blacksquare$$

Observação 1: *(estimativa de erro)*

No resultado do teorema não explicitamos que a constante K seria $\frac{M_2}{2M_1}$, como foi deduzido na demonstração, porque na prática não é um valor facilmente cálculável. No entanto, quando se executa o método de Newton procedendo ao cálculo de $[\boldsymbol{\nabla} F(x_n)]^{-1}$, a sua norma pode ser facilmente calculada, e nesse caso podemos escrever a estimativa

$$||e^{(n+1)}|| \leq \frac{1}{2} \max_{x \in V} ||\boldsymbol{\nabla}^2 F(x)||\; ||\, [\boldsymbol{\nabla} F(x_n)]^{-1}||\; ||e^{(n)}||^2, \tag{4.10}$$

tendo em atenção que a estimativa faz apenas sentido quando estamos muito próximo da solução, e portanto a vizinhança V deverá ser uma bola $B(z, \varepsilon)$ com ε pequeno. Por outro lado o valor da norma $||\boldsymbol{\nabla}^2 F(x)||$ deve ser entendido como o máximo das normas matriciais $\max_i ||\boldsymbol{\nabla}^2 f_i(x)||$.

• Há ainda a possibilidade de apresentar uma condição suficiente para a convergência, semelhante à que foi obtida no caso escalar, e que *também poderá servir de critério* em $\mathbb{R}$. Enunciamos apenas o resultado, cuja demonstração pode ser encontrada em [18]:

Teorema 4.3 *(Kantorovich). Seja $D \subset \mathbb{R}^N$ um conjunto aberto e convexo e $F \in C^1(D)$. Se*
(i) $\exists M_1 > 0 : ||\, [\boldsymbol{\nabla} F(x)]^{-1}\, || \leq \frac{1}{M_1},\ \forall x \in D,$
(ii) $\exists M_2 > 0 : ||\boldsymbol{\nabla} F(x) - \boldsymbol{\nabla} F(y)|| \leq M_2 ||x - y||,\ \forall x, y \in D,$
(iii) existe $x_0 \in D$, *tal que* $\varepsilon_0 = 2\, ||\, [\boldsymbol{\nabla} F(x_0)]^{-1} F(x_0)\, ||$ *verifica* $\frac{M_2}{M_1}\varepsilon_0 < 1,$
(iv) $\bar{B}(x_0, \varepsilon_0) \subset D,$

[4]Como assumimos $F \in C^2(V_z)$, e como ∇F é invertível em z (que não é ponto crítico), então, por continuidade, o determinante de ∇F também não é nulo numa vizinhança suficientemente pequena de z.

então há uma única solução $z \in \bar{B}(x_0, \varepsilon_0)$, para a qual o método de Newton converge (começando com a iterada inicial x_0), e verifica-se a estimativa de erro a priori,

$$||e^{(n)}|| \leq \frac{1}{K}(K\varepsilon_0)^{2^n},$$

em que escrevemos $K = \frac{M_2}{2M_1}$ (para pôr em evidência a semelhança com o caso real).

Notamos que a condição (i) implica a existência de inversa para a matriz jacobiana (equivalente no caso real a $f'(x) \neq 0$), e serve ao mesmo tempo para definir M_1 (que corresponde no caso real a $\min |f'(x)|$). A condição (ii) implica a limitação dos valores da matriz Hessiana (caso $f \in C^2$) e define M_2 (que corresponde no caso real a $\max |f''(x)|$). A terceira condição permite garantir que as iteradas vão ficar na bola $\bar{B}(x_0, \varepsilon_0)$, note-se que, por exemplo, $||x_1 - x_0|| = \frac{1}{2}\varepsilon_0 \leq \varepsilon_0$ (e corresponde à condição no caso real $|f(x_0)/f'(x_0)| \leq |b - a|$). A quarta condição é óbvia, e podemos mesmo considerar $\bar{D} = \bar{B}(x_0, \varepsilon_0)$.

Observação 2: *(métodos quasi-Newton)*
No caso de sistemas, há ainda um maior número de variantes do método de Newton que podem ser utilizadas. Um dos objectivos destes métodos é evitar a repetida resolução de sistemas (ver observação seguinte), outro é evitar o cálculo da matriz jacobiana. Uma maneira de evitar esse cálculo é considerar uma aproximação das derivadas parciais usando um cálculo suplementar a uma distância ε (para cada derivada) tal como foi feito no caso unidimensional. É ainda possível generalizar o método da secante (cf. [26]).

Observação 3: *(tempo de cálculo)*
Enquanto que no método do ponto fixo, o tempo de cálculo será apenas $T = n\, t_G$, em que t_G é o tempo médio necessário para avaliar a função G, no caso do método de Newton, devido à forma particular de G, há que considerar não apenas o tempo de cálculo de F, ou o tempo de cálculo de $\boldsymbol{\nabla} F$, como se passava no caso real, mas também devemos considerar um novo tempo de cálculo em cada iteração, t_S, o tempo médio para a resolução de um sistema linear. Assim teremos

$$T = n\,(t_F + t_{\nabla F} + t_S).$$

• Pode acontecer que o tempo de resolução do sistema seja muito maior que o tempo do cálculo da função e das suas derivadas, pelo que é habitual implementar técnicas alternativas que podem consistir em manter a matriz $\boldsymbol{\nabla} F(x^{(n)})$ durante algumas iteradas subsequentes, actualizando-a espaçadamente. Isso permite reduzir consideravelmente o tempo de cálculo, já que sendo a matriz a mesma, podemos guardar a sua factorização para resolver mais rapidamente o sistema, como veremos na secção seguinte.

Observação 4: O *Mathematica* implementa o método de Newton na rotina FindRoot, desde que se inclua uma lista com as equações e se prescreva o valor inicial para cada componente.

Como acabamos de ver, para implementarmos o método de Newton necessitamos à partida de saber resolver sistemas lineares, pelo que vamos agora concentrar-nos na resolução desse tipo de sistemas, mais simples.

4.3 Métodos Iterativos para Sistemas Lineares

Vamos começar por aplicar, mais uma vez, o método do ponto fixo à resolução deste tipo de problemas.

Consideremos uma matriz quadrada $\mathbf{A} \in \mathbb{R}^N \times \mathbb{R}^N$ e um vector $\mathbf{b} \in \mathbb{R}^N$, quaisquer[5]. Pretendemos resolver o sistema

$$\mathbf{Ax} = \mathbf{b}.$$

Para aplicar o método do ponto fixo, eficazmente, precisamos de estabelecer uma equivalência $\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{G}(\mathbf{x})$ em que a função iteradora $\mathbf{G}$ seja muito mais simples de calcular do que a solução... por exemplo, não adiantaria nada escrever $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, porque o cálculo da inversa seria até mais complicado do que resolver o próprio sistema!

Apenas nos iremos interessar por funções $\mathbf{G}$ do tipo

$$\mathbf{G}(\mathbf{x}) = \mathbf{w} + \mathbf{Cx}, \tag{4.11}$$

em que $\mathbf{w}$ é um vector e $\mathbf{C}$ uma matriz. Iremos ver algumas possibilidades distintas para $\mathbf{w}$ e $\mathbf{C}$.

4.3.1 Métodos de Jacobi e Gauss-Seidel

Começamos por escrever a matriz como soma de duas outras, mais simples, $\mathbf{A} = \mathbf{M} + \mathbf{N}$, e ficamos com:

$$\mathbf{Ax} = \mathbf{Mx} + \mathbf{Nx} = \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{M}^{-1}\mathbf{b} - \mathbf{M}^{-1}\mathbf{Nx},$$
$$\text{ou seja, } \mathbf{G}(\mathbf{x}) = \mathbf{M}^{-1}\mathbf{b} - \mathbf{M}^{-1}\mathbf{Nx}$$

desde que a matriz $\mathbf{M}$ seja invertível. Neste caso $\mathbf{w} = \mathbf{M}^{-1}\mathbf{b}$ e $\mathbf{C} = -\mathbf{M}^{-1}\mathbf{N}$.

As escolhas de $\mathbf{M}$ e $\mathbf{N}$ que dão origem aos Métodos de Jacobi e de Gauss-Seidel baseiam-se em escrever a matriz $\mathbf{A}$ na forma $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$:

$$\overbrace{\begin{bmatrix} a_{11} & \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix}}^{\mathbf{A}} = \overbrace{\begin{bmatrix} 0 & \dots & 0 \\ a_{21} & \ddots & \vdots \\ \vdots & & \\ a_{N1} & \dots\ a_{N,N-1} & 0 \end{bmatrix}}^{\mathbf{L}} + \overbrace{\begin{bmatrix} a_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{NN} \end{bmatrix}}^{\mathbf{D}} + \overbrace{\begin{bmatrix} 0 & a_{12} \dots & a_{1N} \\ \vdots & \ddots & \vdots \\ & & a_{N-1,N} \\ 0 & \dots & 0 \end{bmatrix}}^{\mathbf{U}}$$

Método de Jacobi

Corresponde a considerar $\mathbf{M} = \mathbf{D}$, $\mathbf{N} = \mathbf{L} + \mathbf{U}$, garantindo que a diagonal de $\mathbf{D}$ não tenha elementos nulos. Desta forma obtemos o método

$$\begin{cases} \mathbf{x}^{(0)} \in \mathbb{R}^N \\ \mathbf{x}^{(n+1)} = \mathbf{D}^{-1}\mathbf{b} - \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(n)} \end{cases} \tag{4.12}$$

que é o chamado *método de Jacobi*.

[5]Nesta secção iremos usar a notação em **negrito** para vectores e matrizes com o intuito de não causar confusão, especialmente entre o N da dimensão do espaço e o $\mathbf{N}$ que irá ser uma matriz. Essa notação será posteriormente abandonada, quando for claro a qual nos referimos.

Exemplo 4.7 *Com efeito, já usámos o Método de Jacobi num exemplo anterior. Consideremos o sistema linear*

$$\begin{cases} 10x_1 + 3x_2 + x_3 = 14 \\ 2x_1 - 10x_2 + 3x_3 = -5 \\ x_1 + 3x_2 + 10x_3 = 14 \end{cases}$$

A aplicação do método de Jacobi fica

$$\begin{cases} x_1^{(k+1)} = \frac{1}{10}(14 - 3x_2^{(k)} - x_3^{(k)}) \\ x_2^{(k+1)} = \frac{-1}{10}(-5 - 2x_1^{(k)} - 3x_3^{(k)}) \\ x_3^{(k+1)} = \frac{1}{10}(14 - x_1^{(k)} - 3x_2^{(k)}) \end{cases}$$

ou seja, basta escrever na primeira equação a componente x_1 em função das outras. Na segunda equação, escrevemos x_2 em função das outras, etc... Neste exemplo, começando com $\mathbf{x}^{(0)} = (0, 0, 0)$ obtemos sucessivamente

$$\mathbf{x}^{(1)} = (1.4, 0.5, 1.4),\ ...,\ \mathbf{x}^{(6)} = (1.000251, 1.005795, 1.000251)$$

que já está bastante próximo[6] *da solução $\mathbf{x} = (1, 1, 1)$.*

• Podemos reparar que o cálculo da segunda componente $x_2^{(k+1)}$ ao invés de utilizar o valor $x_1^{(k)}$ poderia utilizar o valor, entretanto já calculado, $x_1^{(k+1)}$. E, da mesma forma, no cálculo de $x_3^{(k+1)}$, podiamos já utilizar os valores $x_1^{(k+1)}$ e $x_2^{(k+1)}$.

Se fizermos isso estamos a considerar um outro método:

Método de Gauss-Seidel

Neste caso consideramos $\mathbf{M} = \mathbf{L} + \mathbf{D}$ e $\mathbf{N} = \mathbf{U}$, assumindo de novo que a diagonal de D não possui elementos nulos (e consequentemente, M, matriz triangular inferior, será invertível).

No entanto, não vamos inverter a matriz $\mathbf{M}$. A partir de

$$\mathbf{x}^{(k+1)} = \mathbf{M}^{-1}\mathbf{b} - \mathbf{M}^{-1}\mathbf{N}\mathbf{x}^{(k)},$$

obtemos $\mathbf{M}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{N}\mathbf{x}^{(k)}$ e escrevemos

$$(\mathbf{L} + \mathbf{D})\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(k)} \Leftrightarrow \mathbf{D}\mathbf{x}^{(k+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(k)} - \mathbf{L}\mathbf{x}^{(k+1)}$$

e daqui surge então o método

$$\begin{cases} \mathbf{x}^{(0)} \in \mathbb{R}^N \\ \mathbf{x}^{(k+1)} = \mathbf{D}^{-1}\mathbf{b} - \mathbf{D}^{-1}\mathbf{U}\mathbf{x}^{(k)} - \mathbf{D}^{-1}\mathbf{L}\mathbf{x}^{(k+1)} \end{cases} \tag{4.13}$$

que é designado *método de Gauss-Seidel*.

---

[6] Neste caso, usando (4.14),

$$C = -D^{-1}(L + U) = \frac{1}{10}\begin{bmatrix} 0 & -3 & -1 \\ -2 & 0 & -3 \\ -1 & -3 & 0 \end{bmatrix},\ \text{temos } ||C||_\infty = 0.5$$

e podemos obter a estimativa *a priori* $||e^{(n)}||_\infty \leq \frac{0.5^n}{1-0.5}||x_1 - x_0||_\infty = 1.4 \times 0.5^{n-1}$, prevendo-se que $||e^{(6)}||_\infty \leq 0.04375$, o que acontece, pois vimos que o erro absoluto é $||e^{(6)}||_\infty = 0.005795$.

Exemplo 4.8 *Considerando o mesmo sistema obtemos agora*

$$\begin{cases} x_1^{(k+1)} = \frac{1}{10}(14 - 3x_2^{(k)} - x_3^{(k)}) \\ x_2^{(k+1)} = \frac{-1}{10}(-5 - 2x_1^{(k+1)} - 3x_3^{(k)}) \\ x_3^{(k+1)} = \frac{1}{10}(14 - x_1^{(k+1)} - 3x_2^{(k+1)}) \end{cases}$$

começando com $\mathbf{x}^{(0)} = (0, 0, 0)$ *iremos obter sucessivamente*

$$\mathbf{x}^{(1)} = (1.4, 0.78, 1.026),\ ..., \mathbf{x}^{(6)} = (1.000039, 1.000028, 0.999988)$$

que já ao fim das mesmas 6 iterações está mais próximo[7] *da solução* $\mathbf{x} = (1, 1, 1)$.

4.3.2 Convergência dos Métodos de Jacobi e Gauss-Seidel

Como já foi dito, podemos encarar estes métodos como métodos de ponto fixo, em que a função iteradora é

$$\mathbf{G}(\mathbf{x}) = \mathbf{M}^{-1}\mathbf{b} - \mathbf{M}^{-1}\mathbf{N}\mathbf{x}$$

(admitindo que $\mathbf{M}$ é invertível). Como o espaço $\mathbb{R}^N$ é fechado, convexo e a derivada de Fréchet é a matriz jacobiana, temos para qualquer norma matricial induzida $||.||$,

$$||\boldsymbol{\nabla}\mathbf{G}(\mathbf{x})|| = ||-\mathbf{M}^{-1}\mathbf{N}||,\ \forall \mathbf{x} \in \mathbb{R}^N$$

escrevendo $\mathbf{C} = -\mathbf{M}^{-1}\mathbf{N}$, bastará exigir que

$$||\mathbf{C}|| < 1,\ \text{para uma certa norma induzida,}$$

para garantirmos a convergência dos métodos de Jacobi e Gauss-Seidel.

Proposição 4.3 *Se existir uma norma matricial induzida para a qual a matriz* $\mathbf{C}$ *definida em (4.11) verifique*

$$||\mathbf{C}|| < 1,$$

então a matriz $\mathbf{A}$ *é invertível e o método iterativo (p.ex: Jacobi ou Gauss-Seidel) converge linearmente para a solução* $\mathbf{z}$ *do sistema* $\mathbf{Ax} = \mathbf{b}$, *qualquer que seja* $\mathbf{x}^{(0)} \in \mathbb{R}^N$. *Temos ainda as estimativas apresentadas no teorema do ponto fixo com* $L = ||\mathbf{C}||$, *por exemplo:*

$$||\mathbf{z} - \mathbf{x}^{(k)}|| \leq \frac{||\mathbf{C}||^k}{1 - ||\mathbf{C}||}||\mathbf{x}^{(1)} - \mathbf{x}^{(0)}|| \tag{4.14}$$

[7] Neste caso, usando (4.14),

$$C = -(L+D)^{-1}U = -\begin{bmatrix} 10 & 0 & 0 \\ 2 & 10 & 0 \\ 1 & 3 & 10 \end{bmatrix}^{-1}\begin{bmatrix} 0 & 3 & 1 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.3 & -0.1 \\ 0 & 0.06 & -0.28 \\ 0 & 0.012 & 0.094 \end{bmatrix},\ \text{e temos } ||C||_\infty = 0.4$$

o que dá um valor da norma de C inferior ao do método de Jacobi, confirmando a maior rapidez de convergência. A estimativa *a priori* daria, neste caso, $||e^{(6)}|| \leq \frac{1}{1-0.4}(0.4)^6 1.4 = 0.009557...$

Demonstração:

Como já dissemos, nestas condições podemos aplicar o T. Ponto Fixo a $\mathbf{G}$ em $\mathbb{R}^N$, o que nos dá existência e unicidade de solução, o que implica a invertibilidade de $\mathbf{A}$. Repare-se que podemos construir a inversa de $\mathbf{A}$ resolvendo sistemas $\mathbf{A}\mathbf{x}^i = \mathbf{e}^i$ em que os vectores $\mathbf{e}^i$ são as colunas da matriz identidade, já que as soluções destes sistemas $\mathbf{x}^i$ irão corresponder às colunas da matriz inversa de $\mathbf{A}$.

Por outro lado, garante-se a convergência com qualquer $\mathbf{x}^{(0)} \in \mathbb{R}^N$, bem como as estimativas de erro apresentadas no teorema do ponto fixo.

Tal como no caso do método do ponto fixo, para uma matriz $\mathbf{A}$ genérica, a convergência é linear. Iremos ver que no caso de matrizes e métodos em $\rho(\mathbf{C}) = 0$ há uma convergência num número finito de iteradas. ■

Corolário 4.2 *O método iterativo* $\mathbf{x}^{(n+1)} = \mathbf{G}(\mathbf{x}^{(n)})$ *em que* $\mathbf{G}$ *é dado por (4.11) converge linearmente para a solução do sistema, dada qualquer iterada inicial* $\mathbf{x}^{(0)} \in \mathbb{R}^N$, *se e só se* $\rho(\mathbf{C}) < 1$.
Se $\rho(\mathbf{C}) = 0$ *então esse método atinge a solução após um número finito de iterações* N *(ou menor).*

Demonstração:

i) Se $\rho(\mathbf{C}) < 1$, se considerarmos $\varepsilon = (1 - \rho(\mathbf{C}))/2 > 0$ vimos que existe uma norma $||.||$ tal que

$$||\mathbf{C}|| \leq \rho(\mathbf{C}) + \varepsilon = \frac{1}{2} + \frac{\rho(\mathbf{C})}{2} < 1$$

e pela proposição concluímos a convergência.

ii) Falta provar que $\rho(\mathbf{C}) < 1$ é condição necessária.

Supondo que $\rho(\mathbf{C}) \geq 1$ vamos concluir que existiria pelo menos um $x^{(0)}$ para o qual não haveria convergência.

Seja λ um valor próprio cujo módulo é o máximo, i.e: $|\lambda| = \rho(\mathbf{C}) \geq 1$, e consideremos $\mathbf{v}$ um vector próprio associado, portanto temos $\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$.

Sucessivamente, iremos obtendo $\mathbf{C}^2\mathbf{v} = \lambda\mathbf{C}\mathbf{v} = \lambda^2\mathbf{v}$, etc... $\mathbf{C}^k\mathbf{v} = \lambda^k\mathbf{v}$, portanto:

$$||\mathbf{C}^k\mathbf{v}|| = ||\lambda^k\mathbf{v}|| = |\lambda|^k||\mathbf{v}||$$

Ora, escrevendo o método, a partir de $\mathbf{x} = \mathbf{G}(\mathbf{x}) = \mathbf{w} + \mathbf{C}\mathbf{x}$ (no caso dos métodos de Jacobi e Gauss-Seidel $\mathbf{w} = \mathbf{M}^{-1}\mathbf{b}$ e $\mathbf{C} = -\mathbf{M}^{-1}\mathbf{N}$), temos $\mathbf{x}^{(k+1)} = \mathbf{w} + \mathbf{C}\mathbf{x}^{(k)}$, e, subtraindo as igualdades, obtemos:

$$\mathbf{x} - \mathbf{x}^{(k+1)} = \mathbf{C}(\mathbf{x} - \mathbf{x}^{(k)}).$$

Aplicando sucessivamente, ficamos com

$$\mathbf{x} - \mathbf{x}^{(k)} = \mathbf{C}^k(\mathbf{x} - \mathbf{x}^{(0)}). \tag{4.15}$$

Assim, se escolhermos $\mathbf{x}^{(0)}$ tal que $\mathbf{v} = \mathbf{x} - \mathbf{x}^{(0)}$, vemos que

$$||\mathbf{x} - \mathbf{x}^{(k)}|| = ||\mathbf{C}^k(\mathbf{x} - \mathbf{x}^{(0)})|| = ||\mathbf{C}^k\mathbf{v}|| = |\lambda|^k||\mathbf{v}|| \not\to 0$$

pois $\lambda \geq 1$.

iii) Quando $\rho(\mathbf{C}) = 0$, a matriz $\mathbf{C}$ apenas pode ter o valor próprio nulo, com multiplicidade N. O seu polinómio característico será $p(\lambda) = \lambda^N$. Pelo teorema de Hamilton-Cayley verifica-se $\mathbf{C}^N = 0$. Assim, por (4.15) temos a seguinte fórmula para o erro $\mathbf{e}^{(N)} = \mathbf{C}^N \mathbf{e}^{(0)} = 0$, ou seja, o erro é nulo na iterada N, pelo menos.■

Observações:

(i) Da demonstração (i) conclui-se, obviamente, que se $\rho(\mathbf{C}) < 1$, a matriz $\mathbf{A}$ é invertível. É uma consequência imediata do teorema do ponto fixo, pois $||\mathbf{C}|| < 1$, para qualquer norma induzida.

(ii) A condição é necessária apenas se pretendermos que haja convergência para qualquer iterada inicial $\mathbf{x}^{(0)}$.

Com efeito, a demonstração de que $\rho(\mathbf{C}) < 1$ é condição necessária, permite concluir também que, mesmo quando $\rho(\mathbf{C}) \geq 1$, basta existir um valor próprio tal que $|\lambda| < 1$ para que haja convergência dado um certo $\mathbf{x}^{(0)}$ apropriado (por exemplo, escolhendo-o igual à diferença entre o vector solução e vector próprio associado). Como esse $\mathbf{x}^{(0)}$ não poderá ter componentes segundo os vectores próprios associados aos valores próprios cujo módulo é superior ou igual a 1, então muito dificilmente esse $\mathbf{x}^{(0)}$ poderá ser encontrado. Com efeito, devido a erros de arredondamento, só em casos muito particulares (...em que se conhece a solução exactamente) seria possível evitar essa componente.

(iii) O resultado permite concluir que no caso do método de Jacobi aplicado a matrizes triangulares a solução exacta é atingida ao fim de N iterações, porque $\mathbf{C}$ será uma matriz triangular com zeros na diagonal, e consequentemente os seus valores próprios serão nulos. No caso do método de Gauss-Seidel, ao ser aplicado a matrizes triangulares superiores, reparamos que $\mathbf{M}$ será uma matriz diagonal, e portanto será idêntico a Jacobi, atingindo a solução ao fim de N iterações (no máximo). Ao ser aplicado a matrizes triangulares inferiores, a matriz $\mathbf{N}$ será nula, portanto $\mathbf{C}$ também será nula, atingindo-se a solução numa única iterada... Repare-se que, neste caso, o método de Gauss-Seidel corresponderá a efectuar as substituições sucessivas descendentes, por isso é perfeitamente natural esta superconvergência.

(iv) Repare-se que se considerarmos o valor $||\mathbf{C}||$ para avaliar a rapidez de convergência, esse valor varia com a norma considerada, e no entanto o método será o mesmo. O valor $\rho(\mathbf{C})$, enquanto ínfimo das normas induzidas, determina a rapidez de convergência do método independentemente da norma, e portanto é uma medida mais fiável para avaliar essa rapidez.

Existe uma condição mais simples de verificar para assegurar a convergência dos métodos de Jacobi e Gauss-Seidel, que envolve a comparação dos módulos da diagonal da matriz com a soma dos módulos dos outros elementos, mas salientamos que é apenas uma condição *suficiente* de convergência.

Definição 4.2 *Uma matriz quadrada* $\mathbf{A}$ *diz-se que tem* diagonal estritamente dominante

por linhas *se verificar*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{N} |a_{ij}| \ \ \forall i \in \{1, \ldots, N\}, \tag{4.16}$$

e diz-se que tem diagonal estritamente dominante por colunas *se verificar*

$$|a_{jj}| > \sum_{i=1, i\neq j}^{N} |a_{ij}| \ \ \forall j \in \{1, \ldots, N\}. \tag{4.17}$$

Proposição 4.4 *Se a matriz quadrada* $\mathbf{A}$ *tiver diagonal estritamente dominante por linhas, ou por colunas, então é invertível e para qualquer* $\mathbf{x}^{(0)} \in \mathbb{R}^N$ *os métodos de Jacobi e Gauss-Seidel convergem para a solução única do sistema* $\mathbf{Ax} = \mathbf{b}$.

Demonstração:

Veremos apenas o caso em que tendo diagonal estritamente por linhas, o método de Jacobi converge.

No caso do método de Jacobi $\mathbf{C} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ e temos

$$c_{ij} = \begin{cases} 0 & \text{se } i = j \\ \frac{-a_{ij}}{a_{ii}} & \text{se } i \neq j \end{cases}$$

Portanto

$$||\mathbf{C}||_\infty = \max_{i=1,\ldots,N} \sum_{j=1}^{N} |c_{ij}| = \max_{i=1,\ldots,N} \sum_{j=1, j\neq i}^{N} |\frac{-a_{ij}}{a_{ii}}|$$

e assim

$$||\mathbf{C}||_\infty < 1 \Leftrightarrow \max_{i=1,\ldots,N} \sum_{j=1, j\neq i}^{N} |\frac{a_{ij}}{a_{ii}}| < 1$$

o que é equivalente a

$$\forall i = 1, \ldots, N \quad \sum_{j=1, j\neq i}^{N} |a_{ij}| < |a_{ii}|.$$

Portanto, a matriz ter diagonal estritamente dominante por linhas é equivalente a $||\mathbf{C}||_\infty < 1$, o que (como vimos) implica a invertibilidade da matriz $\mathbf{A}$ e a convergência do método para qualquer $\mathbf{x}^{(0)} \in \mathbb{R}^N$.■

Observação 1: *(invertibilidade de matrizes)*

Reparamos que através deste resultado, um critério extremamente simples para assegurar a invertibilidade de uma matriz é observar que tem a diagonal estritamente dominante por linhas, ou por colunas. Isto poderá ser conseguido, em muitos casos, através de uma simples troca de linhas ou colunas. Considerando o sistema linear

$$\begin{bmatrix} 2 & -10 & 3 \\ 10 & 3 & 1 \\ 1 & 3 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -5 \\ 14 \\ 14 \end{bmatrix}$$

basta uma troca da primeira com a segunda linha para obtermos a matriz do exemplo anterior, que tem a diagonal estritamente dominante por linhas, assegurando também a convergência dos métodos de Jacobi e Gauss-Seidel.

Observação 2: *(matrizes irredutíveis)*

No caso em que a diagonal é dominante, mas não estritamente (ou seja, a desigualdade da definição não é estrita), não se pode concluir a invertibilidade, basta pensar no caso da matriz $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. No entanto, se houver dominância estrita *por um* elemento da diagonal e se a matriz for *irredutível*[8], então obtemos invertibilidade e a convergência do método de Jacobi ou Gauss-Seidel. Um exemplo, é o caso em que

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & 1 \\ 0 & -2 & 2 \end{bmatrix}.$$

Reparamos que a desigualdade estrita é verificada para o primeiro elemento da diagonal, $|2| > |-1|$, mas não é verificada para o segundo elemento da diagonal, $|2| = |-1|+|1|$, nem para o terceiro $|2| = |-2|$. Como a matriz é irredutível (ver nota de rodapé) concluímos a sua invertibilidade e a convergência dos dois métodos iterativos. Este assunto é tratado com maior detalhe, por exemplo, em [25].

Observação 3: *(Jacobi vs. Gauss-Seidel)*

É possível mostrar que, para um certo tipo de matrizes gerais, o método de Gauss-Seidel tem convergência mais rápida que o método de Jacobi, podendo ser duas vezes mais rápido. Um exemplo instrutivo é o que se passa para matrizes 2×2. Reparamos que, no caso do método de Jacobi, se

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow \mathbf{C}_{Jac} = -\begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix}^{-1} \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{b}{a} \\ -\frac{c}{d} & 0 \end{bmatrix}$$

e portanto, os valores próprios serão $\lambda = \pm\sqrt{\frac{bc}{ad}}$, e para que o método convirja (dada qualquer iterada inicial) é necessário que $|bc| < |ad|$. Situação semelhante ocorre com o método de Gauss-Seidel. Temos

$$\mathbf{C}_{GS} = -\begin{bmatrix} a & 0 \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{b}{a} \\ 0 & \frac{bc}{ad} \end{bmatrix}$$

[8]Uma matriz $\mathbf{A}$ diz-se *redutível* se existir uma matriz de permutação $\mathbf{P}$ tal que

$$\mathbf{PAP}^{-1} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{bmatrix},$$

em que $\mathbf{B}$ e $\mathbf{D}$ são matrizes quadradas e $\mathbf{O}$ é um bloco com zeros ($\mathbf{C}$ é um bloco qualquer). Diz-se *irredutível* se não for redutível. Há vários critérios para estabelecer se uma matriz é ou não irredutível, que podem ser encontrados em [25]. É óbvio que matrizes sem zeros são irredutíveis. Também é possível mostrar que matrizes em que as subdiagonais principais não têm zeros (p. ex. matrizes tridiagonais completas) são irredutíveis.

e portanto, $\lambda = 0$ ou $\lambda = \frac{bc}{ad}$, sendo ainda necessário que $|bc| < |ad|$ para que $\rho(\mathbf{C}_{GS}) < 1$ e assim haja convergência. Conclui-se que no caso de quaisquer matrizes 2×2 os métodos convergem sob as mesmas condições, mas com uma diferença apreciável... reparamos que $\rho(\mathbf{C}_{Jac})^2 = \rho(\mathbf{C}_{GS})$, o que traduz a convergência mais rápida do método de Gauss-Seidel. Por exemplo, no caso de matrizes simétricas temos mesmo $||\mathbf{C}_{Jac}||_2^2 = ||\mathbf{C}_{GS}||_2$, o que implica a estimativa de erro para o método de Gauss-Seidel $||\mathbf{e}^{(n)}||_2 \leq ||\mathbf{C}_{GS}||_2^n \, ||\mathbf{e}^{(0)}|| = ||\mathbf{C}_{Jac}||_2^{2n} \, ||\mathbf{e}^{(0)}||$, ou seja, são necessárias o dobro das iteradas do método de Jacobi para obter o mesmo resultado.

Apesar de numa grande quantidade de casos o método de Gauss-Seidel ser mais rápido que o método de Jacobi, isso nem sempre acontece e há mesmo casos em que o método de Jacobi converge e o de Gauss-Seidel não.

Exemplo 4.9 *Um possível exemplo em que o método de Jacobi converge e o de Gauss-Seidel não converge, acontece para matrizes da forma*

$$\mathbf{A} = \begin{bmatrix} a & -b & a \\ b & -b & -a \\ b & -b & a \end{bmatrix},$$

em que temos, para $a, b \neq 0$,

$$\mathbf{C}_{Jac} = \begin{bmatrix} 0 & \frac{b}{a} & -1 \\ 1 & 0 & -\frac{a}{b} \\ -\frac{b}{a} & \frac{b}{a} & 0 \end{bmatrix}, \quad \mathbf{C}_{GS} = \begin{bmatrix} 0 & \frac{b}{a} & -1 \\ 0 & \frac{b}{a} & -1 - \frac{a}{b} \\ 0 & 0 & -1 \end{bmatrix},$$

e os valores próprios de $\mathbf{C}_{Jac}$ *verificam* $\lambda(\lambda^2 + 1 - 2\frac{b}{a}) = 0$, *enquanto os de* $\mathbf{C}_{GS}$ *verificam* $\lambda(\lambda - \frac{b}{a})(\lambda + 1) = 0$.
Isto significa que $\rho(\mathbf{C}_{Jac}) = \left|\sqrt{2\frac{b}{a} - 1}\right|$, *e que* $\rho(C_{GS}) = \max\{1, \left|\frac{b}{a}\right|\} \geq 1$.
No caso em que $0 < b < a$, *o método de Jacobi converge e o de Gauss-Seidel não, porque* $\rho(\mathbf{C}_{Jac}) < 1$ *e* $\rho(\mathbf{C}_{GS}) = 1$, *reparando que para* $a = 2b$, *temos mesmo* $\rho(\mathbf{C}_{Jac}) = 0$.

Como curiosidade verificámos o que se passava com 10 000 matrizes 3×3 dando valores aleatórios nas entradas. Para essa amostra obtivémos os seguintes valores, não rigorosos, completamente empíricos: 80% das matrizes não apresentava convergência com qualquer dos métodos, e de entre as 20% convergentes, apenas um décimo apresentava maior velocidade de convergência para o método de Jacobi (ou seja, 2% do total). Quando o método de Gauss-Seidel convergia, obtinha normalmente a solução em metade das iterações do método de Jacobi (85% dos casos, i.e. aprox 15% do total).

Observação 4: *(complementos para o método do ponto fixo em* $\mathbb{R}^N$*)*

(i) *Iteração de Gauss-Seidel.* Reparamos que em $\mathbb{R}^N$ existe uma possibilidade, que é aplicada frequentemente e que consiste em efectuar a iteração do ponto fixo (ou de Newton) utilizando a substituição das novas componentes pelas calculadas mais recentemente, em analogia ao que se faz no método de Gauss-Seidel.

(ii) *Avaliar a contractividade com o raio espectral.* Da mesma forma que utilizamos aqui o raio espectral ao invés da norma, também o poderemos fazer quando avaliamos a

condição de contractividade $||\boldsymbol{\nabla} G(x)|| \leq L < 1$, no método do ponto fixo em $\mathbb{R}^N$. Nesse caso bastará mostrar que $\rho(G(z)) < 1$, para garantirmos a convergência local, pois sabemos que haverá uma norma induzida que verificará a condição enunciada. No entanto, da condição $\rho(G(x)) \leq L < 1$, num domínio D não se pode inferir o mesmo já que a matriz muda.

4.3.3 Métodos de Relaxação e SOR

A ideia dos métodos iterativos de relaxação consiste em fazer aparecer um parâmetro ω, não nulo, na função iteradora G. Esse parâmetro pode ser controlado de forma a que o método convirja, e até de forma a que convirja mais rapidamente!

• **Método de Relaxação Linear**
Para $\omega \neq 0$, escrevemos:

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{x} = \mathbf{x} - \omega(\mathbf{Ax} - \mathbf{b}),$$
$$\text{ou seja, } \mathbf{G}(\mathbf{x}) = \mathbf{x} - \omega(\mathbf{Ax} - \mathbf{b})$$

e portanto o método da relaxação linear consiste na iteração

$$\begin{cases} \mathbf{x}^{(0)} \in \mathbb{R}^N \\ \mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - \omega(\mathbf{Ax}^{(n)} - \mathbf{b}) \end{cases} . \tag{4.18}$$

Neste caso, temos $\mathbf{w} = \omega\mathbf{b}$ e $\mathbf{C}_\omega = \mathbf{I} - \omega\mathbf{A}$, portanto podemos assegurar a convergência do método quando $||\mathbf{C}_\omega|| = ||\mathbf{I} - \omega\mathbf{A}|| < 1$, ou ainda, quando $\rho(\mathbf{C}_\omega) < 1$.

Exercício 4.1 *Mostre que o método de relaxação linear converge, qualquer que seja* $\mathbf{x}_0 \in \mathbb{R}^N$, *se e só se escolhermos* $|\omega|$ *suficientemente pequeno tal que qualquer valor próprio de* $\mathbf{A}$ *pertence à bola* $B(\frac{1}{\omega}, \frac{1}{|\omega|})$. *Em particular, mostre que se* $\mathbf{A}$ *for definida positiva, basta que* $0 < \omega < \frac{2}{\rho(\mathbf{A})}$ *para que haja convergência.*
Resolução: *Trata-se de verificar que nessas condições* $\rho(\mathbf{I} - \omega\mathbf{A}) < 1$, *ou seja que os valores próprios* μ *de* $\mathbf{I} - \omega\mathbf{A}$ *são em módulo menores que 1. Se* λ *é valor próprio de* $\mathbf{A}$, *temos*

$$(\mathbf{I} - \omega\mathbf{A})\mathbf{v} = \mathbf{v} - \omega\lambda\mathbf{v} \Leftrightarrow \mu = 1 - \omega\lambda$$

e como por hipótese $|\lambda - \frac{1}{\omega}| \leq |\frac{1}{\omega}|$, *então* $|\mu| = \left|\lambda - \frac{1}{\omega}\right| |\omega| < 1$. *No caso de ser definida positiva,* $B(\frac{1}{\omega}, \frac{1}{|\omega|}) =]0, \frac{2}{\omega}[$, *e portanto* $0 < \lambda < \frac{2}{\omega}$, *o que resulta de* $0 < \omega < \frac{2}{\rho(\mathbf{A})} < \frac{2}{\lambda}$. □

Uma escolha razoável para $|\omega|$ é um valor próximo de $\frac{1}{\rho(\mathbf{A})}$, que pode ser aproximado através do teorema de Gerschgorin (que veremos no próximo capítulo).

Exercício 4.2 *Efectue um estudo semelhante para uma função* $\mathbf{G}(\mathbf{x}) = (1-\omega)\mathbf{x} - \omega(\mathbf{Ax} - \mathbf{b})$, *com* $\omega \neq 0$.

• **Método das Relaxações Sucessivas (SOR)**
Trata-se de uma variação do método de Gauss-Seidel, bastante eficaz em termos de aceleração de convergência. Enquanto que no método de Gauss-Seidel considerávamos a matriz $\mathbf{M} = \mathbf{L} + \mathbf{D}$, neste caso vamos diminuir um pouco a contribuição de $\mathbf{D}$ em $\mathbf{M}$, colocando

$$\mathbf{M}_\omega = \mathbf{L} + \frac{1}{\omega}\mathbf{D}$$

(normalmente $\omega > 1$) e compensamos esse facto passando a parte restante da matriz $\mathbf{D}$ para $\mathbf{N}$, fazendo

$$\mathbf{N}_\omega = (1 - \frac{1}{\omega})\mathbf{D} + \mathbf{U}.$$

Assim, para $\omega \neq 0$, temos $\mathbf{C}_\omega = -\mathbf{M}_\omega^{-1}\mathbf{N}_\omega$, ou seja

$$\mathbf{C}_\omega = -(\omega\mathbf{L} + \mathbf{D})^{-1}((\omega - 1)\mathbf{D} + \omega\mathbf{U}),$$

e será esta quantidade que determinará a rapidez de convergência do método.

Portanto, da decomposição

$$\mathbf{Ax} = \mathbf{b} \Leftrightarrow \mathbf{Lx} + \frac{1}{\omega}\mathbf{Dx} + (1 - \frac{1}{\omega})\mathbf{Dx} + \mathbf{Ux} = \mathbf{b},$$

retiramos

$$(\frac{1}{\omega}\mathbf{D} + \mathbf{L})\mathbf{x} = (\frac{1}{\omega} - 1)\mathbf{Dx} + \mathbf{b} - \mathbf{Ux},$$

obtendo

$$(\frac{1}{\omega}\mathbf{D} + \mathbf{L})\mathbf{x}^{(n+1)} = (\frac{1}{\omega} - 1)\mathbf{Dx}^{(n)} + \mathbf{b} - \mathbf{Ux}^{(n)}.$$

Ao invés de inverter $\frac{1}{\omega}\mathbf{D} + \mathbf{L}$, a iteração processa-se de forma semelhante ao método de Gauss-Seidel, ou seja,

$$\mathbf{x}^{(n+1)} = (1 - \omega)\mathbf{x}^{(n)} + \omega\mathbf{D}^{-1}(\mathbf{b} - \mathbf{Lx}^{(n+1)} - \mathbf{Ux}^{(n)}), \tag{4.19}$$

reparando que no caso $\omega = 1$ coincide com o método de Gauss-Seidel.

Resta analisar em que situações há ou não convergência do método SOR.

Teorema 4.4 *(condição necessária, Kahan). Mostre que é necessário que $\omega \in]0, 2[$ para que haja convergência do método, qualquer que seja a iterada inicial.*

Demonstração: Exercício (*sugestão:* provar primeiro que $\rho(\mathbf{C}_\omega)^N \geq |\lambda_1| \cdots |\lambda_N|$ e que $|\det(\mathbf{C}_\omega)| = |1 - \omega|^N$).■

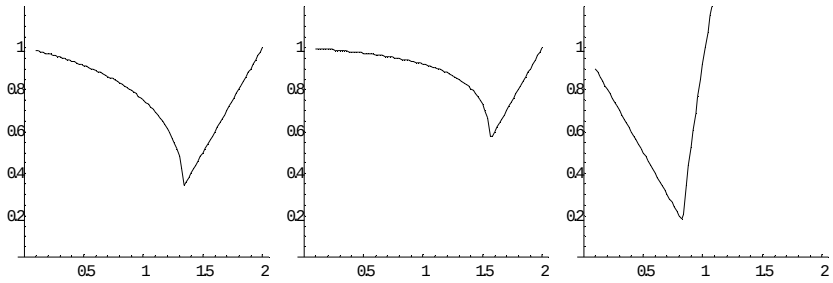
Este resultado dá-nos apenas condições necessárias. Vejamos alguns exemplos, para podermos compreender o resultado seguinte.

Exemplo 4.10 *Consideramos uma matriz $N \times N$, tridiagonal, da forma,*

$$A = \begin{bmatrix} a & b & 0 & \cdots \\ c & a & b & \ddots \\ 0 & c & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Vamos considerar inicialmente $a = 2, b = c = -1$. Analisando o valor de $\rho(\mathbf{C}_\omega)$, que determina a rapidez de convergência, variando $\omega \in]0, 2[$, obtemos o gráfico que apresentamos

em baixo, à esquerda, para $N = 5$. É visível que há um valor de ω, próximo de 1.35 para o qual a convergência será mais rápida, e reparamos também que para os outros valores de $\omega \in]0, 2[$ se verifica $\rho(\mathbf{C}_\omega) < 1$. Na figura seguinte, ao centro, colocámos o mesmo exemplo alterando apenas $N = 10$. Verificamos o mesmo tipo de comportamento, mas o valor de ω para o qual há um mínimo foi agora deslocado para próximo de 1.55. Por outro lado, alterando agora para $c = 1$, obtemos uma situação completamente diferente, visível na figura da direita. Continua a haver convergência para valores de $\omega \leq 1$, mas para alguns valores de ω superiores a 1, o método não converge (é uma situação de antissimetria que não se enquadra no teorema que veremos de seguida e que nos dá uma condição suficiente para a convergência).



Teorema 4.5 *(condição suficiente, Ostrowski-Reich). Se a matriz A for hermitiana e definida positiva então para $\omega \in]0, 2[$ o método SOR converge qualquer que seja a iterada inicial.*
Em particular, o caso $\omega = 1$ corresponde à convergência do método de Gauss-Seidel para matrizes hermitianas definidas positivas.

Demonstração: (Ver e.g. [7], [25]). ■

Observação 1 *(parâmetro optimal ω^*).*
Embora não seja possível a priori, para qualquer matriz definir qual o ω^*, parâmetro optimal que irá permitir uma maior rapidez de convergência, nalguns casos particulares é possível explicitá-lo por estimativas teóricas (ver observação seguinte). Em muitos casos de aplicação prática (sistemas resultantes da discretização de equações diferenciais), esse valor optimal é obtido para $\omega^* > 1$, o que deu origem ao nome de sobre-relaxação (por oposição a $\omega^* < 1$, que é designada sub-relaxação, e que também pode oferecer um parâmetro optimal, veja-se a terceira figura do exemplo anterior). Mas não é necessário conhecer exactamente o valor optimal, para valores ω próximos de ω^*, o método SOR converge mais rapidamente que o método de Gauss-Seidel.

Observação 2 *(equações diferenciais).*
Os métodos iterativos, e em particular o método SOR, são especialmente eficazes para resolver sistemas que resultam da discretização de equações com derivadas parciais. Nesses sistemas é frequente ter como incógnita um vector cujas componentes u_{ij} dependem de dois indíces (no caso de problemas em R^2) e que estão apenas relacionadas com componentes de

indíces vizinhos $u_{i\pm1,j\pm1}$. Isto significa que a matriz do sistema será essencialmente nula, mas os elementos não nulos da matriz estão dispersos, devido à ordenação de dois indíces num único (que permitirá definir um vector como incógnita, e assim estabelecer o sistema). No entanto, essa ordenação é dispensável se utilizarmos um método iterativo, de tal forma que nunca será preciso explicitar qual a matriz do sistema. Como exemplo, apresentamos um sistema (correspondente à discretização da equação de Laplace usando diferenças finitas) em que os valores u_{ij} estão relacionados por equações

$$u_{ij} = \frac{1}{4}(u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}),$$

conhecendo-se alguns valores de u_{ij} correspondentes a valores dados na fronteira do domínio (...problema de Dirichlet). Repare-se que esta relação define um sistema, que é apenas difícil de explicitar para alguns domínios, onde a ordenação dos indíces (i, j) é menos trivial... no caso de um rectângulo é simples (basta ordenar linha a linha... ou coluna a coluna). No entanto, a aplicação do método de Jacobi é neste caso muito simples,

$$u_{ij}^{(n+1)} = \frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n)}),$$

e esta iteração é efectuada para todos os índices (i, j) correspondentes a pontos no interior do domínio (os pontos na fronteira não são alterados, pois são os valores dados... e serão eles que definem a solução). O uso de métodos iterativos é ainda justificado pelo facto de a matriz ser esparsa (a maior parte dos elementos será nulo), e pode mostrar-se que se trata da implementação de um sistema em que a matriz tem a diagonal dominante (não estritamente em todos... repare-se que $1 = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4}$), e que é irredutível. Como curiosidade, notamos que a aplicação do método SOR se traduz nas iterações

$$u_{ij}^{(n+1)} = (1-\omega)u_{ij}^{(n+1)} + \omega\frac{1}{4}(u_{i+1,j}^{(n)} + u_{i-1,j}^{(n+1)} + u_{i,j+1}^{(n)} + u_{i,j-1}^{(n+1)}),$$

admitindo que os valores $u_{i-1,j}^{(n+1)}$ e $u_{i,j-1}^{(n+1)}$ são calculados antes dos valores $u_{i+1,j}^{(n+1)}$ e $u_{i,j+1}^{(n+1)}$. Para o caso específico de um domínio que é um quadrado com $N \times N$ pontos interiores, é possível obter teoricamente o valor optimal ω^* para o método SOR, tendo-se (cf.[1]):

$$\omega^* = \frac{2}{1 + \sin(\frac{\pi}{N+1})},$$

e a sua utilização permite aumentar substancialmente a rapidez de convergência do método SOR.

Observação 3 *(método de Jacobi modificado).*

Tal como o SOR é uma modificação de Gauss-Seidel, o método de Jacobi também pode ser modificado introduzindo um parâmetro de relaxação. Neste caso, para $\omega \neq 0$, escrevemos

$$\mathbf{x} = \mathbf{x} - \omega\mathbf{D}^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}),$$

em que $\mathbf{D}$ se refere à parte diagonal da matriz $\mathbf{A}$,reparando que se $\omega = 1$ obtemos o método de Jacobi usual. Aqui $\mathbf{C}_\omega = \mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{A}$,e assim teremos convergência se $||\mathbf{I} - \omega\mathbf{D}^{-1}\mathbf{A}|| < 1$. A iteraç

ão processa-se da mesma forma

$$\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)} - \omega \mathbf{D}^{-1}(\mathbf{A}\mathbf{x}^{(n)} - \mathbf{b}), \tag{4.20}$$

e reparamos que neste caso as matrizes $\mathbf{M}$ e $\mathbf{N}$ têm a forma:

$$\mathbf{M}_\omega = \tfrac{1}{\omega}\mathbf{D}$$
$$\mathbf{N}_\omega = \mathbf{L} + (1 - \tfrac{1}{\omega})\mathbf{D} + \mathbf{U}.$$

A iteração através do método de Jacobi modificado é optimal se $\omega = 2/(2 - \lambda_C^+ - \lambda_C^-)$, em que λ_C^+ é o maior e λ_C^- o menor valor próprio de $\mathbf{C}_\omega$, supostos reais e em módulo inferiores a 1 (ver [18]) . Neste caso a convergência será sempre mais rápida que a do método clássico. De um modo geral, com escolha apropriada de ω o método modificado permite acelerar a convergência quando para o método clássico se tem $\rho(\mathbf{C}_1) \sim 1$.

4.4 Métodos Directos para Sistemas Lineares

Tendo já estudado alguns métodos iterativos para sistemas lineares, fazemos agora uma breve análise de alguns métodos directos, especialmente no que diz respeito ao condicionamento, à estabilidade, e à optimização do número de operações envolvidas.

Independentemente do método escolhido, a resolução de um sistema linear $Ax = b$ em cujo vector b é dado com erros leva a uma propagação desses erros à solução do sistema, ou seja a um problema de condicionamento.

4.4.1 Condicionamento de um Sistema Linear

Interessa-nos identificar quais as matrizes que podem trazer problemas de mau condicionamento.

Supondo que nos era dado, não o vector b exacto, mas apenas uma aproximação $\tilde{b}$, obtemos um valor aproximado $\tilde{x}$, solução do sistema: $A\tilde{x} = \tilde{b}$.

Pretendemos ver qual a influência que o erro $e_{\tilde{b}} = b - \tilde{b}$ tem no erro do resultado $e_{\tilde{x}} = x - \tilde{x}$, e para isso recuperamos algumas noções já vistas no primeiro capítulo.

Para estabelecermos a relação entre os erros relativos[9] dos dados $||\delta_{\tilde{b}}|| = \frac{||b-\tilde{b}||}{||b||}$ e os erros relativos dos resultados $||\delta_{\tilde{x}}|| = \frac{||x-\tilde{x}||}{||x||}$ vai ser importante estabelecer uma noção que envolve a norma de matrizes :

[9]No caso vectorial, usamos a notação

$$\delta_{\tilde{x}} = \frac{e_{\tilde{x}}}{||x||} = \frac{x - \tilde{x}}{||x||}.$$

É também utilizada por vezes a notação $\delta_{\tilde{x}}$ para designar o valor $\frac{||e_{\tilde{x}}||}{||x||}$. Preferimos usar nesse caso $||\delta_{\tilde{x}}||$, porque põe em evidência qual a norma usada, e que se trata de um escalar.

Definição 4.3 *Designa-se por* número de condição *de uma matriz* A, *relativamente à norma* $||.||$, *o valor :* $cond(A) = ||A||\ ||A^{-1}||$

Proposição 4.5 *Temos as seguintes desigualdades:*
(i) para o erro absoluto:

$$||e_{\tilde{b}}||\ ||A||^{-1} \leq ||e_{\tilde{x}}|| \leq ||A^{-1}||\ ||e_{\tilde{b}}||$$

(ii) para o erro relativo :

$$\frac{||\delta_{\tilde{b}}||}{cond(A)} \leq ||\delta_{\tilde{x}}|| \leq cond(A)||\delta_{\tilde{b}}|| \tag{4.21}$$

Demonstração:
Usando os dois sistemas, retiramos $A(x - \tilde{x}) = b - \tilde{b}$, portanto:

$$||A||\ ||x - \tilde{x}|| \geq ||b - \tilde{b}||, \text{ e } ||x - \tilde{x}|| \leq ||A^{-1}||\ ||b - \tilde{b}||.$$

Assim, a alínea a) fica provada, pois : $||A||^{-1}||b - \tilde{b}|| \leq ||x - \tilde{x}|| \leq ||A^{-1}||\ ||b - \tilde{b}||$.
Por outro lado, de $Ax = b$, retiramos, de forma análoga :

$$||A||^{-1}||b|| \leq ||x|| \leq ||A^{-1}||\ ||b||$$

'dividindo' as desigualdades de (i), usando este resultado, obtemos (ii). ■

Observações:
i) Como é óbvio, podemos concretizar estes resultados para qualquer uma das normas. Por exemplo, podemos retirar a majoração:

$$||\delta_{\tilde{x}}||_1 \leq cond_1(A)||\delta_{\tilde{b}}||_1$$

onde $cond_1(A)$ designa o número de condição relativamente à norma da soma, ou seja : $cond_1(A) = ||A||_1||A^{-1}||_1$
ii) Se a norma da matriz identidade é $||I|| = 1$ (o que acontece sempre para as normas induzidas), e como $||I|| = ||AA^{-1}|| \leq ||A||\ ||A^{-1}||$, obtemos $cond(A) \geq 1$.
iii) Os maiores problemas de condicionamento surgem devido à possibilidade algébrica de encontrar matrizes inversas cujos elementos não são pequenos face aos elementos da matriz original. Assim, não há uma compensação entre as normas de $||A||$ e de $||A^{-1}||$ podendo o seu produto dar números muito elevados. Uma tal situação é retratada no exemplo seguinte.

Exemplo 4.11 *Consideremos o caso em que*

$$A = \begin{bmatrix} 1 & a & -a \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \text{ e temos } A^{-1} = \begin{bmatrix} 1 & 0 & a \\ 1 & 1 & a \\ 1 & 1 & 1+a \end{bmatrix}.$$

Se $a > 1$, *obtemos* $||A||_\infty = 2a+1, ||A^{-1}||_\infty = a+3$. *Portanto* $cond_\infty(A) = (2a+1)(a+3)$ *será um valor elevado quando a também for. Num caso concreto, em que* $a = 100$, *obtemos* $cond_\infty(A) = 20703$, *e podemos ver o efeito deste condicionamento na resolução de um sistema* $Ax = b$.
Consideramos $b = (-100, -0.5, 1)$ *e* $\tilde{b} = (-100.1, -0.51, 1.1)$. *A solução exacta é* $x = (0, -0.5, 0.5)$, *e a solução com o valor aproximado é* $\tilde{x} = (9.9, 9.39, 10.49)$... *valores surpreendentes se não tivéssemos encontrado um número de condição elevado. Neste caso,*

$$||\delta_{\tilde{b}}||_\infty = \frac{0.1}{100} = 0.001 \text{ e temos } ||\delta_{\tilde{x}}||_\infty = \frac{9.99}{0.5} = 19.98,$$

o que significa que $||\delta_{\tilde{x}}||_\infty = 19980||\delta_{\tilde{b}}||_\infty$, *valores que se encontram dentro das estimativas previstas.*
Repare-se que este efeito pode ser apenas vísivel para certos valores de b. Considerando $b = (1, 1, 1)$ *e* $\tilde{b} = (1, 1, 0.99)$. *A solução exacta é* $x = (101, 102, 103)$, *e a solução com o valor aproximado é* $\tilde{x} = (101, 101, 101.99)$... *Apesar de alguma diferença, neste caso,*

$$||\delta_{\tilde{b}}||_\infty = \frac{0.01}{1} = 0.01 \text{ e temos } ||\delta_{\tilde{x}}||_\infty = \frac{1.01}{103} \approx 0.01.$$

Ou seja, quando dizemos que a matriz é mal condicionada, isso reflecte-se mais para valores de b particulares.

- Vejamos agora o caso em que é a matriz que tem erros, e o sistema a resolver será

$$\tilde{A}\tilde{x} = b.$$

Definindo o erro relativo da matriz,

$$||\delta_{\tilde{A}}|| = \frac{||A - \tilde{A}||}{||A||},$$

podemos obter o seguinte resultado.

Proposição 4.6 *Seja* $Ax = b$, $\tilde{A}\tilde{x} = b$. *Se o erro relativo da matriz é suficientemente pequeno, verificando* $||\delta_{\tilde{A}}|| < \frac{1}{cond(A)}$, *temos:*

$$||\delta_{\tilde{x}}|| \leq \frac{cond(A)}{1 - cond(A)||\delta_{\tilde{A}}||}||\delta_{\tilde{A}}||. \tag{4.22}$$

Demonstração:
Esta estimativa clássica pode também ser deduzida pelo teorema do ponto fixo.
Por hipótese,

$$||\delta_{\tilde{A}}|| = \frac{||A-\tilde{A}||}{||A||} < \frac{1}{||A||\,||A^{-1}||} \Leftrightarrow ||A-\tilde{A}||\,||A^{-1}|| < 1.$$

Portanto, $||I - A^{-1}\tilde{A}|| = ||A^{-1}(A-\tilde{A})|| \leq ||A^{-1}||\,||A-\tilde{A}|| < 1$.
Considerando a função iteradora $G(v) = v - A^{-1}\tilde{A}v$, concluímos que tem um único ponto fixo $z \in \mathbb{R}^N$, que é obrigatoriamente zero, porque o vector nulo é ponto fixo, ie. $G(0) = 0$.
Tomando $v_0 = \tilde{x}$, temos $v_1 = G(v_0)$, e ficamos com

$$v_1 = \tilde{x} - A^{-1}\tilde{A}(\tilde{x}) = \tilde{x} - A^{-1}b = \tilde{x} - x.$$

Agora usamos a estimativa de erro do ponto fixo,

$$||0 - v_1|| \leq \frac{||I - \tilde{A}A^{-1}||}{1 - ||I - \tilde{A}A^{-1}||}||v_1 - v_0||,$$

e como $v_1 = \tilde{x} - x$,e $v_1 - v_0 = -x$, obtemos

$$||x - \tilde{x}|| \leq \frac{||I - \tilde{A}A^{-1}||}{1 - ||I - \tilde{A}A^{-1}||}||x||,$$

ou seja,

$$||\delta_{\tilde{x}}|| \leq \frac{||I - \tilde{A}A^{-1}||}{1 - ||I - \tilde{A}A^{-1}||} \leq \frac{cond(A)||\delta_{\tilde{A}}||}{1 - cond(A)||\delta_{\tilde{A}}||}.$$

A última desigualdade resulta de $||I - \tilde{A}A^{-1}|| \leq ||A^{-1}||\,||A-\tilde{A}|| = cond(A)||\delta_{\tilde{A}}||$ (notando também que a função $\frac{t}{1-t}$ é crescente, com $t < 1$).■

Note-se que na demonstração anterior não exigimos que a matriz $\tilde{A}$ fosse invertível, no entanto a própria condição $||A^{-1}||\,||A-\tilde{A}|| < 1$ implica isso (ver exercício 1.b) no final do capítulo anterior).
Apresentamos ainda uma estimativa para quando há erro na matriz e no vector.

Corolário 4.3 *Seja* $Ax = b$, $\tilde{A}\tilde{x} = \tilde{b}$. *Se* $||\delta_{\tilde{A}}|| < \frac{1}{cond(A)}$,

$$||\delta_{\tilde{x}}|| \leq \frac{cond(A)}{1 - cond(A)||\delta_{\tilde{A}}||}\,(||\delta_{\tilde{A}}|| + ||\delta_{\tilde{b}}||). \tag{4.23}$$

Destes resultados podemos concluir que um número de condição elevado não nos permite estabelecer boas majorações para o erro relativo (mas não podemos inferir o mau condicionamento). Quanto maior for o número de condição, pior será a majoração de erro relativo obtida. Consequentemente, para matrizes cujo número de condição seja elevado,

um pequeno erro relativo no vector b, 'pode provocar' um grande erro relativo na solução do sistema. Se o número de condição for baixo (nunca será inferior a 1...) podemos concluir acerca do bom condicionamento da resolução do sistema.

Observação: Para contornar problemas com o mau condicionamento de matrizes é usual estabelecer uma equivalência $Ax = b \Leftrightarrow LAR(R^{-1}x) = Lb$, de forma a que a matriz LAR seja melhor condicionada. Assim, resolvemos o sistema $A^*x^* = b^*$, com $A^* = LAR, b^* = Lb$, e a solução x é obtida a partir de x^* notando que $x^* = R^{-1}x$ e portanto $x = Rx^*$. A esta técnica chama-se *pré-condicionamento* de um sistema.

Terminamos este parágrafo observando que podemos também definir um número de condição associado ao raio espectral :

$$cond_\rho(A) = \rho(A)\rho(A^{-1})$$

que verifica a propriedade (resultante do raio espectral ser o ínfimo das normas induzidas) :

$$cond_\rho(A) \leq cond_{||.||}(A)$$

para qualquer norma induzida $||.||$. Atendendo a que os valores próprios de A^{-1} são os inversos de A, temos :

$$cond_\rho(A) = \frac{\max_{i=1,..,N} |\lambda_i|}{\min_{i=1,..,N} |\lambda_i|} \tag{4.24}$$

onde λ_i designam os valores próprios da matriz A.

4.4.2 Método de Eliminação de Gauss

O método clássico de resolução de um sistema linear é o método de eliminação de Gauss, introduzido em qualquer curso elementar de Álgebra Linear, que iremos relembrar sucintamente. Notamos que este método é ainda considerado como o método mais eficaz para resolver um sistema genérico. No entanto, para certo tipo de matrizes podemos apresentar métodos alternativos, mais adequados.

Consideremos o sistema $Ax = b$ em que A é uma matriz quadrada $N \times N$:

$$\begin{bmatrix} a_{11} & \dots & a_{N1} \\ \vdots & \ddots & \vdots \\ a_{N1} & \dots & a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}$$

O objectivo deste método é eliminar os elementos a_{ij} de forma a obter um sistema equivalente com uma matriz triangular superior. Depois bastará usar substituições sucessivas para chegarmos à solução pretendida.

O método consiste em $N - 1$ passos, onde construimos elementos $a_{ij}^{(k+1)}$ a partir dos elementos $a_{ij}^{(k)}$ considerando $[a_{ij}^{(1)}]$ como sendo a matriz inicial.

Passo k (para $k = 1, ...N-1$)
– Se o pivot for nulo, i.e: $a_{kk}^{(k)} = 0$, há que efectuar troca de linhas.
– Se $a_{kk}^{(k)} \neq 0$ calculamos

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \text{ para } i = k+1, ..., N$$

e atribuímos

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, \quad \text{para } i, j = k+1, ..., N,$$

$$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)}, \text{ para } i = k+1, ..., N.$$

Ao fim dos $N-1$ passos obtemos o sistema triangular superior equivalente:

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{N1}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N}^{(N-1)} \\ 0 & \dots & 0 & a_{NN}^{(N)} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1^{(1)} \\ \vdots \\ \vdots \\ b_N^{(N)} \end{bmatrix}$$

que se pode resolver facilmente por substituição ascendente:

$$\begin{cases} x_N = \frac{b_N^{(N)}}{a_{NN}^{(N)}} \\ x_k = \frac{1}{a_{kk}^{(k)}} (b_k^{(k)} - \sum_{j=k+1}^{n} a_{ij}^{(k)} x_j), \quad \text{para } k = N-1, ..., 1 \end{cases}$$

Armazenando os coeficientes m_{ik} podemos obter uma factorização da matriz A na forma:

$$A = LU = \begin{bmatrix} 1 & 0 & \dots & 0 \\ m_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ m_{N1} & \dots & m_{N,N-1} & 1 \end{bmatrix} \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{N1}^{(1)} \\ 0 & a_{22}^{(2)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{N-1,N}^{(N-1)} \\ 0 & \dots & 0 & a_{NN}^{(N)} \end{bmatrix},$$

caso não sejam efectuadas trocas de linhas.

Caso existam trocas de linhas, a factorização é da forma $PA = LU$ em que P é uma matriz de permutação. Ao resolver o sistema obteriamos $LUx = P^T b$

Observação: A factorização $A = LU$ em que L é uma matriz triangular inferior com diagonal principal unitária, e U é uma matriz triangular superior, é obtida de forma única se os *pivots* verificarem $a_{kk}^{(k)} \neq 0$.

4.4.3 Número de Operações

Analisemos agora qual o número de operações (+ - ou * /) envolvido na resolução de um sistema:

- **Factorização da Matriz**
  – Em cada Passo k:
  Cálculo dos m_{ik} :
  $(N-k)$ divisões – correspondentes a um total de $\sum_{k=1}^{N-1}(N-k)$ operações.
  Cálculo dos a_{ij} :
  $(N-k)^2$ multiplicações e subtracções – correspondentes a um total de $\sum_{k=1}^{N-1}(N-k)^2$ operações.

- **Cálculo dos $b^{(k)}$**
  – Em cada Passo k :
  $N-k$ multiplicações e subtracções correspondentes a um total de $\sum_{k=1}^{N-1}(N-k)$ operações.

- **Substituição:**
  – No total teremos : $N+\sum_{k=1}^{N-1}k = \frac{1}{2}N(N+1)$ multiplicações e divisões, $\sum_{k=1}^{N-1}k = \frac{1}{2}N(N-1)$ subtracções

Como $\sum_{k=1}^{N-1}(N-k) = \frac{1}{2}N(N-1)$ e também $\sum_{k=1}^{N-1}(N-k) = \frac{1}{2}N(N-1)(2N-1)$, obtemos a seguinte tabela,

	$(+;-)$	$(\times;/)$
Factorização	$\frac{1}{6}N(N-1)(2N-1)$	$\frac{1}{3}N(N^2-1)$
Cálculo de $b^{(k)}$	$\frac{1}{2}N(N-1)$	$\frac{1}{2}N(N-1)$
Subtituição	$\frac{1}{2}N(N+1)$	$\frac{1}{2}N(N-1)$
Total	$\sim \frac{N^3}{3}$	$\sim \frac{N^3}{3}$

é fácil ver que o número total de operações, ao considerarmos uma dimensão da matriz elevada, é assimptoticamente equivalente a $\frac{2}{3}N^3$. Normalmente, como as operações $(*,/)$ consomem maior tempo de cálculo que as $(+,-)$, considera-se que o tempo de cálculo para a resolução de um sistema será

$$T_S \approx \frac{N^3}{3}t^*,$$

em que t^* será o tempo médio para efectuar uma multiplicação/divisão.

Este valor é bastante reduzido se comparado com o número de operações que seria necessário efectuar se resolvessemos o sistema pela Regra de Cramer (nesse caso teríamos $\sim (N+1)!$ operações, o que por exemplo, para $N=10$ corresponderia a efectuar aproximadamente 40 milhões de operações $(*,/)$ ao invés de aproximadamente 400 pelo método de Gauss).

Observação:*(Pesquisa de Pivot)*

Já vimos que ao resolver um sistema $Ax = b$ podemos ter problemas de condicionamento, mas mesmo que esses problemas não ocorram podemos ter problemas de instabilidade numérica. Para minorar esses problemas, considerámos técnicas de pesquisa de pivot. No entanto, se o problema for mal condicionado, essas técnicas de pesquisa de pivot têm uma utilidade limitada, já que um problema mal condicionado será sempre numericamente instável.

Da mesma forma que quando o pivot é nulo (i.e: $a_{kk}^{(k)} = 0$) somos obrigados a efectuar uma troca de linhas, no caso de valores próximos de zero, se não fôr efectuada uma troca de linhas ou colunas, os erros de arredondamento (surgidos na factorização da matriz) podem provocar grandes erros nos resultados. Isto acontece se houver um grande desiquilibrio de grandezas nos elementos da matriz – muito maiores face ao pivot (o que é equivalente, dividindo, a que ele seja próximo de zero). Para contornar este problema de estabilidade numérica, usam-se as seguintes estratégias:

(i) Pesquisa parcial de pivot (normalmente por linhas)

Em cada Passo k da eliminação de Gauss, troca-se a linha k com a linha r , onde r é tal que :

$$|a_{rk}^{(k)}| = \max_{i=k,\ldots,N} |a_{ik}^{(k)}|$$

isto, como é claro, só no caso de $k \neq r$.

(ii) Pesquisa total de pivot

Em cada Passo k da eliminação de Gauss, troca-se a linha k com a linha r e a coluna k com a coluna s , onde r, s são tais que :

$$|a_{rs}^{(k)}| = \max_{i,j=k,\ldots,N} |a_{ij}^{(k)}|$$

isto, como é claro, só no caso de $k \neq r$ ou $k \neq s$.

• *Interpretação da pesquisa de pivot*

Vejamos o caso de um sistema 2×2, que é exemplificativo. Temos,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

e suponhamos que a_{11} é muito pequeno relativamente a a_{21}, então $m = \frac{a_{21}}{a_{11}}$ é um valor elevado.

Ao efectuar um passo da eliminação de Gauss obtemos o sistema equivalente

$$\begin{bmatrix} a_{11} & a_{12} \\ 0 & a_{22} - m\, a_{12} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 - m\, b_1 \end{bmatrix}$$

Como m é muito elevado, acontece que ma_{12} e mb_1 podem ser muito superiores aos valores de a_{22} e b_2.

Então iremos obter $x_2 = \frac{b_2 - mb_1}{a_{22} - ma_{12}} \approx \frac{b_1}{a_{12}}$ (quando m é grande), devido aos erros de arredondamento. Este valor não será significativamente diferente do valor correcto, no entanto ao calcular

$$x_1 = \frac{1}{a_{11}}(b_1 - a_{12}x_2)$$

reparamos que como $x_2 \approx \frac{b_1}{a_{12}}$ estamos perante um caso de cancelamento subtractivo, o que provocará grandes erros relativos.

Pelo contrário, se o valor de m for pequeno, o que corresponde a efectuar uma pesquisa de pivot trocando as linhas, então $x_2 \approx \frac{b_2}{a_{22}}$ e o problema de cancelamento subtractivo ficado reduzido a casos particulares $b_1 a_{22} \approx a_{12} b_2$, que podem ser contornados com uma pesquisa total.

Exemplo 4.12 *Consideramos o sistema*

$$\begin{bmatrix} 1 & 5 \\ 500 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

e temos $m = 500$, *portanto*

$$\begin{bmatrix} 1 & 5 \\ 0 & 1 - 2500 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 - 2500 \end{bmatrix}.$$

Obtemos $x_2 = 0.99959984$, *mas supondo que apenas podemos trabalhar num sistema FP com 3 dígitos na mantissa, ficaria arredondado para* $\tilde{x}_2 = 1$. *Portanto*

$$\tilde{x}_1 = 5 - 5\tilde{x}_1 = 0.,$$

o que difere com erro relativo de 100% face ao valor correcto, $x_1 = 0.00200008...$
Reparamos que neste exemplo não há qualquer problema de condicionamento na matriz, pois

$$A^{-1} = \frac{1}{24999} \begin{bmatrix} -1 & 5 \\ 500 & -1 \end{bmatrix} \text{ e assim } cond_\infty(A) = ||A||_\infty \, ||A^{-1}||_\infty = 500 \frac{500}{24999} \approx 1.$$

Se efectuássemos a troca de linhas, correspondente à pesquisa de pivot (parcial ou mesmo total, neste caso), então

$$\begin{bmatrix} 500 & 1 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

e portanto $m = 1/500$, *logo* $\tilde{x}_2 = 1 (\approx \frac{5 - 0.002}{5})$, *e teremos* $\tilde{x}_1 = \frac{2-1}{500} = 0.002$, *ambos os valores com erros relativos inferiores a* 0.05%.

Observação: No *Mathematica* este problema com os erros de arredondamento não é normalmente visível, porque internamente efectua as operações com um maior número de dígitos (o que leva mais tempo...) de forma a garantir a precisão no resultado (ver capítulo inicial).

4.4.4 Métodos de Factorização ou Compactos

Vimos que, usando o método de eliminação de Gauss, no caso de não haver troca de linhas, podemos obter uma factorização da matriz A na forma $A = LU$, onde U seria a matriz triangular superior obtida no final da factorização e L a matriz triangular inferior com diagonal unitária, cujos elementos seriam os multiplicadores m_{ik}. Vamos agora ver uma maneira alternativa de obter essa mesma factorização através do Método de Doolittle.

Método de Doolittle

Pretendemos obter as matrizes L e U tais que $A = LU$:

$$A = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{N1} & \dots & l_{N\,N-1} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & \dots & \dots & u_{1N} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{NN} \end{bmatrix}$$

Logo, efectuando o produto, podemos obter as fórmulas correspondentes ao método de Doolittle. Fixando um índice k, temos:

$$a_{ij} = \sum_{r=1}^{N} l_{ir}\, u_{rj} = \sum_{r=1}^{k-1} l_{ir}\, u_{rj} + l_{ik} u_{kj} + \sum_{r=k+1}^{N} l_{ir}\, u_{rj}\ .$$

Como $l_{kr} = 0$ se $k < r$, o segundo somatório fica nulo e obtemos imediatamente (para $j \geq k$) :

$$a_{kj} = \sum_{r=1}^{k-1} l_{kr}\, u_{rj} + u_{kj} \Leftrightarrow u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr}\, u_{rj}\ .$$

Da mesma forma, como $u_{rk} = 0$ se $k < r$, obtém-se (para $i > k$) :

$$a_{ik} = \sum_{r=1}^{k-1} l_{ir}\, u_{rk} + l_{ik} u_{kk} \Leftrightarrow l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}\, u_{rk} \right)$$

Podemos esquematizar o algoritmo:

Passo 1 :

$$\begin{cases} u_{1j} = a_{1j} & (j = 1, ..., N) \\ l_{i1} = \frac{a_{i1}}{u_{11}} & (i = 2, ..., N) \end{cases}$$

Passo k : $(k = 2, ..., N)$

$$\begin{cases} u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr}\, u_{rj} & (j = k, ..., N) \\ l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}\, u_{rk} \right) & (i = k+1, ..., N) \end{cases}$$

Vemos assim que a decomposição $A = LU$ pretendida é única, porque a partir das condições impostas construimos explicitamente a solução. Como é claro, caso algum pivot seja nulo, $u_{kk} = 0$, será necessário efectuar uma troca de colunas (com uma coluna j em que $u_{kj} \neq 0$) antes de calcular os l_{ik}. Essa troca é sempre aconselhável, fazendo *pesquisa de pivot*, trocando a coluna k com a coluna s : $|u_{ks}| = \max_{j=k,...,N} |u_{kj}|$.

Para factorizar a matriz, usando o método de Doolittle são necessárias o mesmo número de operações que no método de eliminação de Gauss. Há, no entanto, vantagens computacionais apreciáveis no que diz respeito ao armazenamento dos elementos da matriz relativamente ao método de Gauss.

Tendo obtido a decomposição $A = LU$, para resolvermos um sistema $Ax = b$, consideramos dois passos

$$\begin{cases} Ly = b \\ Ux = y. \end{cases}$$

Os dois sistemas são resolvidos facilmente por substituição pois L é uma matriz triangular inferior e U é triangular superior.

Observação 1: Notamos que, *caso exista já a factorização,* o tempo de cálculo para a resolução do sistema será

$$T_s \approx N^2 t^*,$$

que corresponde a $\approx \frac{1}{2}N^2$ operações $(*, /)$ para cada uma das substituições nos dois sistemas $Ly = b$ e $Ux = y$.

Observação 2: De forma semelhante, podemos pensar numa factorização em que ao invés de L, será a matriz U que terá a diagonal principal unitária. Esse outro processo, em tudo semelhante a este, é denominado usualmente por *método de Crout.*

Vamos agora ver alguns métodos particulares para certo tipos de matrizes, em que podemos reduzir o números de operações. Começamos pelas matrizes simétricas e depois vamos ver o caso das matrizes tridiagonais.

Método de Cholesky

Pode ser encarado como uma simplificação do método de Doolittle para matrizes simétricas $(A = A^T)$, de forma a que decomposição seja $A = LL^T$.

Se pensarmos numa matriz unidimensional, reparamos imediatamente que isso corresponde a encontrar $l_{11} : a_{11} = l_{11}^2$, e caso consideremos apenas números reais, isto será apenas possível se $a_{11} \geq 0$. Por outro lado, para resolver $a_{11}x_1 = b_1$ devemos considerar sempre $a_{11} \neq 0$. Vemos assim que no caso unidimensional isso corresponde a exigir que $a_{11} > 0$. No caso de uma matriz real de qualquer dimensão, isso corresponde à noção de matriz definida positiva[10]...

Como é claro, verificar que a matriz é definida positiva é mais moroso do que resolver o sistema... Assim, o método só é aplicado a matrizes que sabemos, por resultados teóricos, serem definidas positivas e simétricas. Veremos, no próximo capítulo, que uma condição suficiente para que a matriz seja definida positiva é ter a diagonal positiva e estritamente dominante (por linhas ou colunas).

Para este tipo de matrizes é válida a factorização: $A = LL^T$, e o método consiste nos seguintes passos:

[10]Relembramos alguns critérios para verificar que uma matriz é definida positiva:
(**i**) $x^T Ax > 0, \forall x \neq 0$,
(**ii**) os valores próprios são todos positivos,
(**iii**) os menores principais são todos positivos.

Passo 1 :

$$\begin{cases} l_{11} = \sqrt{a_{11}} \\ l_{i1} = \frac{a_{i1}}{l_{11}} \quad (i = 2, ..., N) \end{cases}$$

Passo k : $(k = 2, ..., N)$

$$\begin{cases} l_{kk} = \sqrt{a_{kk} - \sum_{r=1}^{k-1} l_{kr}^2} \\ l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir}\, l_{kr} \right) \quad (i = k+1, ..., N) \end{cases}$$

Esta construção da matriz L não é única. No caso dos reais, depende apenas do sinal escolhido para as raizes, sendo única a menos de multiplicação por uma matriz diagonal em que os elementos são 1 ou -1. No caso dos complexos depende do ramo escolhido para as raizes.

No método de Cholesky, o número de operações é aproximadamente metade do efectuado nos métodos de Gauss e Doolittle, porque aproveitamos o facto de a matriz ser simétrica, tendo-se

$$T_S \approx \frac{N^3}{6}\, t^* + N\, t_{\sqrt{\cdot}}$$

em que são apenas consideradas significativas as $\approx N^3/6$ operações $(\times; \div)$, e o cálculo das N raízes quadradas que assumimos demorar um tempo $t_{\sqrt{\cdot}}$. Normalmente, para N grande, o tempo de cálculo da raiz quadrada, não dependendo de N, pode ser considerado menos significativo que as restantes operações (note-se ainda que, com poucas iterações, é possível uma boa aproximação da raiz pelo método de Newton, usando apenas multiplicações e divisões).

Observação: Caso se esteja a trabalhar com complexos, não é necessário exigir que a matriz seja definida positiva! Uma outra possibilidade para evitar o problema de exigir que a matriz seja definida positiva consiste em considerar a decomposição

$$A = LDL^\top,$$

o que também evita o cálculo de raízes quadradas

Cálculo de matrizes inversas

Para calcular uma matriz inversa através de um método de factorização, resolvemos N sistemas

$$Ax = \mathbf{e}_i \ \ (i = 1, ..., N)$$

em que $\mathbf{e}_i$ é um vector $(0, ..., 0, 1, 0, ..., 0)$ da base canónica. Repare-se que a factorização da matriz A é apenas efectuada uma única vez, sendo depois re-utilizada. no cálculo dos outros sistemas. Isto envolve o cálculo de N substituições para resolver $Ly = b$ e ainda N substituições para resolver $Uy = b$. Como cada substituição envolve um número $\sim \frac{N^2}{2}$ de operações, $(+; -)$ ou $(\times; \div)$, isto dá um total de $\sim \frac{4}{3}N^3$ operações, $(+; -)$ ou $(\times; \div)$.

O processo de diagonalização completa (método de Gauss-Jordan) envolve o mesmo número de operações. Na primeira fase, em que obtemos a matriz triangular superior, efectuamos $\sim \frac{N^3}{3}$ operações para a factorização, mas também N operações no segundo membro (cálculo dos $b^{(k)}$ é $\sim \frac{N^2}{2}$), o que dá $\sim \frac{N^3}{2}$. Depois, para efectuarmos a diagonalização completa eliminamos no sentido inverso, mas aí é apenas necessário calcular os multiplicadores, o que envolve apenas $\sim \frac{N^2}{2}$ operações, porque temos zeros na parte triangular inferior. No entanto, voltamos a ter N operações no segundo membro o que dá mais $\sim \frac{N^3}{2}$. Na realidade, este processo é equivalente a resolver cada um dos sistemas triangulares superiores. No total, teremos o mesmo número de operações $\sim \frac{4}{3}N^3$.

– Reparando que as componentes dos vectores $\mathbf{e}_i$ são essencialmente zeros, podemos reduzir significativamente a contagem do número de operações, se repararmos que o cálculo de $Ly = \mathbf{e}_N$ não envolve qualquer operação, já que $y = \mathbf{e}_N$. Da mesma forma $Ly = \mathbf{e}_k$ envolverá apenas $\sum_{i=k}^{N}(N-i) = \frac{(N-k)(N-k+1)}{2}$ operações, que somadas (de $k = 1$ até N) dão apenas $\sim \frac{N^3}{6}$ operações. Concluímos que a soma total é $\sim \frac{N^3}{3} + \frac{N^3}{6} + \frac{N^3}{2} = N^3$.

Sendo o número de operações $\sim N^3$, o triplo do necessário para a resolução de um sistema, nunca é aconselhável inverter uma matriz, sendo preferível armazenar as matrizes L e U, já que de qualquer forma, quer a multiplicar $A^{-1}b$, quer a resolver $Ly = b$ e $Ux = y$, o número de operações envolvido é o mesmo: $\sim N^2$.

Um outro problema computacional que pode ocorrer é o armazenamento durante os cálculos da matriz inversa. No caso destes métodos isso implica $2N^2$ entradas, provindo N^2 das matrizes L e U, e mais N^2 do resultado. No entanto, se não houver necessidade de guardar A, podemos reduzir o número para $N^2 + N$ entradas usando o método de Gauss-Jordan.

Matrizes Tridiagonais

Este é o caso de matrizes em que, à excepção das três diagonais principais, todos os outros elementos são nulos. Ou seja:

$$a_{ij} = 0, \text{ se } |i-j| > 1.$$

Estas matrizes aparecem em muitos problemas práticos (por exemplo, no caso de interpolação por *splines*, ou na resolução de equações diferenciais)

Devido à sua estrutura simples, o número de operações necessário para resolver um sistema, pode ser consideravelmente reduzido, já que podemos escrever $A = LU$, onde L será uma matriz triangular inferior, bidiagonal, com diagonal unitária, e U uma matriz

triangular superior, bidiagonal:

$$\begin{bmatrix} a_{11} & a_{12} & 0 & \dots & 0 \\ a_{21} & a_{22} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & a_{N-1,N} \\ 0 & \dots & 0 & a_{N,N-1} & a_{NN} \end{bmatrix} = \begin{bmatrix} 1 & 0 & & \dots & 0 \\ l_{21} & 1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \dots & 0 & l_{N\,N-1} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & 0 & \dots & 0 \\ 0 & u_{22} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & u_{N-1,N} \\ 0 & \dots & \dots & 0 & u_{NN} \end{bmatrix}$$

O método de factorização de Doolittle reduz-se então a:

Passo 1:

$$\begin{cases} u_{11} = a_{11}; \;\; u_{12} = a_{12} \\ l_{21} = \frac{a_{21}}{u_{11}} \end{cases}$$

Passo $k : (k = 2, ..., N)$

$$\begin{cases} u_{kk} = a_{kk} - l_{k,k-1} u_{k-1,k} \\ u_{k,k+1} = a_{k,k+1} \\ l_{k+1,k} = \frac{a_{k+1,k}}{u_{kk}} \end{cases}$$

Podemos contabilizar o número de operações efectuado na factorização e na resolução dos sistemas triangulares (neste caso, a substituição ainda será mais simples):

Factorização: envolve $N-1$ operações $(+;-)$, e $2N-1$ operações $(\times;/)$.

Resolver $Ly = b$: envolve $N-1$ operações $(+;-)$, e $N-1$ operações $(\times;/)$.

Resolver $Ux = y$: envolve $N-1$ operações $(+;-)$, e $2N-1$ operações $(\times;/)$.

• Corresponde a um total de $(3N-3)+(5N-3) = 8N-6$ operações!

Considerando, de novo, apenas como significativas as operações $(*, /)$ obtemos, para um *sistema tridiagonal*

$$T_S \approx 5N\, t^*.$$

Observação: O *Mathematica* tem implementada a rotina LinearSolve que permite resolver um sistema introduzindo a matriz e o vector. A rotina LUDecomposition permite obter uma lista com três elementos, em cujo primeiro é uma matriz em que está concatenada a matriz L e a matriz U, o segundo elemento é um vector que indica as permutações efectuadas aquando da pesquisa de pivot e o terceiro elemento é uma estimativa do número de condição $cond_\infty(A)$. Note-se que isto pressupõe a entrada de uma matriz com valores numéricos, não exactos (mais uma vez se alerta para a diferença entre 1 e 1.0, por exemplo).Para matrizes tridiagonais refere-se a existência da rotina TridiagonalSolve, onde é apenas necessário introduzir uma lista com as três diagonais e o vector de dados. Para isso é necessário adicionar o *package* LinearAlgebra‘Tridiagonal‘.

4.4.5 Métodos Iterativos e Métodos Directos

Para finalizar o capítulo iremos considerar de novo métodos iterativos, mas agora relacionados com métodos directos.

• *Método de correcção residual.*

Podemos contornar eventuais erros (devidos a um mau condicionamento ou a instabilidade numérica) resultantes da resolução de um sistema linear com um método directo, usando um método iterativo. Trata-se do *método de correcção residual,* que passamos a descrever.

– Supondo que ao resolver o sistema $Ax = b$ obtinhamos um vector impreciso x_0, má aproximação de x, podemos considerar resolver

$$A\varepsilon_1 = r_0$$

em que $r_0 = b - Ax_0$. Se o valor $x_1 = x_0 + \varepsilon_1$ ainda não é suficientemente bom, repetimos sucessivamente o processo, obtendo

$$A\varepsilon_{n+1} = r_n$$

em que $\varepsilon_{n+1} = x_{n+1} - x_n,\ \ r_n = b - Ax_n$.

Cada iteração necessita de apenas $\sim N^2$ operações, porque guardamos as matrizes L e U da factorização.

Observação 1 *(correcção residual, um método de ponto fixo...)*

Vejamos que, em certo sentido, se trata de um método de ponto fixo.

Como vemos, o problema está na resolução do sistema, pelo que podemos encarar que a matriz A usada para resolver o sistema é ligeiramente perturbada. Vamos designá-la por $\tilde{A}$, diferente da matriz usada para calcular $Ax - b$.

Sendo assim, o processo iterativo construído pelo método da correcção residual é o seguinte

$$\tilde{A}(x_{n+1} - x_n) = b - Ax_n,$$

ou seja,

$$x_{n+1} = x_n - \tilde{A}^{-1}b + \tilde{A}^{-1}Ax_n.$$

Designando por x a solução exacta, temos $Ax = b$ e portanto

$$x_{n+1} - x = x_n - x + \tilde{A}^{-1}A(x - x_n).$$

Vemos assim que se trata de um método equivalente a um método do ponto fixo aplicado a $e_n = x - x_n$, tendo-se

$$e_{n+1} = (I - \tilde{A}^{-1}A)e_n.$$

Como já vimos, basta que $||I - \tilde{A}^{-1}A|| < 1$, para que haja convergência do método, ora isso acontece se $\tilde{A}$ for suficientemente próxima de A, ficando provada a convergência do método (veja-se também a demonstração efectuada antes, para a estimativa do erro relativo devido a $\tilde{A}$).

Não podemos dizer que se trata exactamente de um método do ponto fixo porque a perturbação da matriz A na resolução do sistema não será igual a uma matriz fixa $\tilde{A}$, essa perturbação pode variar com x_n de forma não determinada, a não ser experimentalmente. Isso não invalida a demonstração que acabamos de fazer, porque ainda que assim fosse, teríamos

$$||e_n|| \leq ||I - \tilde{A}_n^{-1}A|| \cdots ||I - \tilde{A}_0^{-1}A||\ ||e_0||,$$

e naturalmente assumindo $||I - \tilde{A}_n^{-1}A|| < 1$, obtemos uma convergência para a solução.

Observação 2 *(comparação quanto ao número de operações entre métodos directos e iterativos).*

Os métodos iterativos implicam normalmente um cálculo de $\sim N^2$ operações em cada iteração, o que os torna ineficazes para $n > \frac{N}{3}$. Para além disso, a precisão atingida ao fim de $N/3$ iteradas não é normalmente muito boa (será $\leq L^{N/3}||z - x^{(0)}||$), pelo que só se tornam realmente eficazes para matrizes de grandes dimensões, e especialmente quando a matriz é *esparsa* (ou seja, possui poucos elementos diferentes de zero); nesse caso os métodos directos não se simplificam muito (excepto em casos particulares... como as matrizes tridiagonais), enquanto que os métodos iterativos apresentam uma redução apreciável do número de operações.

4.5 Exercícios

1. Para encontrar as raízes de uma equação algébrica

$$p(x) = a_0 + a_1 x + ... + a_m x^m = 0$$

podemos desenvolver a factorização $p(x) = a_m(x - z_1)...(x - z_m)$ estabelecendo um sistema de equações não lineares em $\mathbb{C}^m$

$$(S) \begin{cases} a_0 = a_m(-z_1)...(-z_m) \\ \vdots \\ a_{m-1} = -a_m(z_1 + ... + z_m) \end{cases}$$

que tem solução única (devido ao teorema fundamental da álgebra). Este processo leva a um método rápido e eficaz para calcularmos todas as raízes se aplicarmos o método de Newton à resolução deste sistema não linear.

a) Suponha que existem soluções complexas para uma equação algébrica cujos coeficientes são reais. Haverá possibilidade de convergência do método de Newton para a solução do sistema (S) se considerarmos todas as iteradas iniciais reais? Porquê?

b) Para o caso de equações do terceiro grau, escreva o sistema em $\mathbb{C}^3$ que deve resolver em cada iteração se pretender aplicar o método de Newton. Aplique esse método para determinar aproximadamente as soluções de $x^3 + 3x + 1 = 0$, calculando três iterações, após ter escolhido uma iterada inicial conveniente.

2. Considere o sistema de equações não lineares

$$\begin{cases} x = f(x + y) \\ y = g(x + y) \end{cases}$$

em que as funções f e g verificam $|f'(t)| < \alpha, |g'(t)| < \beta$, para qualquer $t \in [a, b]$, e em que $f(\mathbb{R}) \subseteq [a, b], g(\mathbb{R}) \subseteq [a, b]$.

a) Mostre que existe uma única solução do sistema em $\mathbb{R}^2$ se $\alpha + \beta < 1$, que essa solução se encontra em $[a, b]^2$, e que o método do ponto fixo converge, quaisquer que sejam os valores iniciais em $\mathbb{R}$.

b) Reduza o sistema anterior à resolução de duas equações em $\mathbb{R}$, e mostre o mesmo resultado que em a).

c) Concretize os resultados anteriores para o sistema

$$\begin{cases} x = \frac{1}{2}\cos(x+y) - \cos^2(\frac{1}{5}(x+y)) \\ y = \sin(\frac{1}{3}(x+y)) + \frac{1}{4}\sin^2(x+y) \end{cases}$$

d) Começando com $(0,0)$, determine uma iterada pelo método de Newton em $\mathbb{R}^2$ para a aproximação da solução do sistema anterior.

3. Pretende-se resolver um sistema linear $Ax = b$ em que os elementos da matriz A são definidos da seguinte forma:

$$a_{ij} = \begin{cases} \frac{1}{i(i+1)} & \text{se} \quad i \neq j \\ C_i & \text{se} \quad i = j \end{cases}$$

a) Indique um intervalo para valores para C_i de forma a que o sistema tenha solução única.

Sugestão: Calcule um majorante para a soma dos elementos das linhas.

b) Considere $C_i = -2 + 1/i$, explicite um método iterativo que convirja para a solução.

4. Pretende-se resolver um sistema linear $Ax = b$ em que os elementos da matriz A são definidos da seguinte forma:

$$a_{ij} = \begin{cases} \frac{M}{C^i + C^j} & \text{se} \quad i \neq j \\ \frac{C|M|}{C-1} & \text{se} \quad i = j \end{cases}$$

em que $C > 1, M \neq 0$.

a) Mostre que a matriz é definida positiva e conclua que é possível decompô-la na forma $A = LL^T$.

b) Considere uma matriz 3×3, com $M = 16, C = 2$. Determine a inversa, usando o método de Cholesky.

5. Considere um sistema $Ax = b$ em que o segundo membro é dado com um erro relativo $||\delta_b||_1 < 0.1$.

a) Sabendo que a matriz é simétrica e que $||A||_\infty \leq 7$, $||A^{-1}||_1 \leq 1$, determine um majorante para $||\delta_x||_\infty$

b) Se a matriz for

$$\begin{bmatrix} 6 & 1 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

determine um majorante para $cond_\rho(A)$, baseado na localização dos valores próprios.

---

6. Considere o sistema de equações

$$\begin{cases} x - y\cos(x)/4 = 0 \\ 1 - y + |x-1| = 0 \end{cases}$$

a) Mostre que existe uma e uma só solução $(x, y) \in [0, 1] \times [1, 2]$.

b) Determine uma aproximação da solução de forma a que o erro absoluto verifique $||e||_\infty < 0.05$.

7. Considere um sistema de equações escrito na forma $F(x) = 0$, e seja $J_F(x)$ a matriz jacobiana de F calculada em x.

a) Mostre que se existir um $\omega \in \mathbb{R}$ tal que $||I + \omega J_F(x)|| \leq L < 1,\ \forall x \in \mathbb{R}^N$, então o sistema possui uma única solução em $\mathbb{R}^N$.

b) Conclua que o sistema

$$\begin{cases} 4x + y + \sin(z) = 1 \\ x + 4y + \cos(z) = 1 \\ \sin(x) + \cos(y) + 4z = 1 \end{cases}$$

tem uma única solução em $\mathbb{R}^N$, que está no conjunto $[-\frac{1}{4}, \frac{1}{2}] \times [-\frac{3}{8}, \frac{3}{8}] \times [-\frac{3}{8}, \frac{3}{8}]$.

c) Determine uma aproximação dessa solução calculando duas iterações pelo método de Newton, começando com a iterada inicial $x^{(0)} = 0$.

8. Considere o sistema

$$\begin{bmatrix} 2 & 10 & 0 \\ 10 & 2 & 2 \\ 0 & 2 & 2 \end{bmatrix}$$

a) Mostre que as condições necessárias e suficientes para que o método de Jacobi convirja dado qualquer $x^{(0)} \in \mathbb{R}^3$ não se verificam.

b) Considerando $x^{(0)} = (0, -1, 5)$ verifique que o método diverge, e considerando $x^{(0)} = (0, 1, -5)$, ou mais em geral, $x^{(0)} = (-1, 1, 0) + \alpha(-1, 0, 5)$ o método converge. Justifique.

9. Considere o sistema de equações:

$$\begin{cases} 2x + y + \varepsilon \cos(z) = 0 \\ x + 3y - 3\varepsilon xz = 0 \\ \varepsilon x^2 + y + 3z = 0 \end{cases}$$

a) Mostre que para $0 < \varepsilon < \frac{1}{2}$ o sistema tem solução única no conjunto

$$S = \{(x, y, z) \in \mathbb{R}^3 : |x|, |y|, |z| \leq \frac{1}{2}\}.$$

b) Usando a função iteradora

$$G(x, y, z) = -(y/2 + \varepsilon \cos(z)/2, x/3 - \varepsilon xz, \varepsilon x^2/3 - y/3),$$

mostre que, aplicando o método do ponto fixo, se tem:

$$|z - z_n| \leq \frac{5^n}{6^{n-1}} ||(x_1 - x_0, y_1 - y_0, z_1 - z_0)||_\infty$$

qualquer que seja o vector inicial $(x_0, y_0, z_0) \in S$.

c) No caso $\varepsilon = 0$, mostre que o método de Jacobi converge e que temos

$$||(x_k, y_k, z_k)||_\infty \leq (\frac{1}{2})^k ||(x_0, y_0, z_0)||_\infty$$

para qualquer $(x_0, y_0, z_0) \in \mathbb{R}^3$.

10. Pretende-se resolver o sistema

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} 0.905 \\ 0.421 \\ 0.265 \end{bmatrix}$$

a) Aplique o método de Cholesky, verificando as condições para uma matriz real.

b) Supondo que o segundo membro foi obtido com um erro absoluto que verifica $||e_b||_\infty \leq 0.01$, determine um majorante para o erro relativo da solução.

11. Considere um sistema $Ax = b$ em que o segundo membro é dado com um erro relativo $||\delta_b||_1 < 0.1$.

a) Sabendo que a matriz é simétrica e que $||A||_\infty \leq 7$, $||A^{-1}||_1 \leq 1$, determine um majorante para $||\delta_x||_\infty$

b) Se a matriz for

$$\begin{bmatrix} 6 & 1 & 0 \\ 1 & 3 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

determine um majorante para $cond_\rho(A)$, baseado na localização dos valores próprios.

12. Considere os dois algoritmos seguintes para resolver $A^2x = b$

Algoritmo I:	*Algoritmo II*:
i) Calcular $B = A^2$	i) Factorizar A
ii) Factorizar B	ii) Resolver $Ay = b$
iii) Resolver $Bx = b$	iii) Resolver $Ax = y$

a) Qual destes algoritmos envolve um menor número de operações quando N é grande?

b) Encontre algoritmos semelhantes para resolver um sistema $A^px = b$, e indique uma estimativa do número de operações envolvidas.

Sugestão: Note que $A^px = b \Leftrightarrow A(A^{p-1}x) = b$.

Capítulo 5

Determinação de Valores e Vectores Próprios de Matrizes

5.1 Noções básicas

Seja E um espaço vectorial. Dizemos que $\lambda \in \mathbb{C}$ é um *valor próprio* de uma aplicação *linear* A se:

$$\exists v \in E, v \neq 0 : Av = \lambda v,$$

e a $v \in E$ chamamos *vector próprio* de A associado a λ. Um mesmo valor próprio λ pode ter associados varios vectores próprios, que geram um subespaço vectorial, designado *subespaço próprio* S_λ associado a λ. Para qualquer $u \in S_\lambda$ é óbvio que $Au = \lambda u$.

Podemos considerar sempre uma base ortonormada em S_λ. Ao longo de cada elemento da base u a aplicação A fica invariante e comporta-se como uma aplicação linear a uma dimensão (i.e: como uma "recta" de inclinação λ). Quando um dos valores próprios é $\lambda = 0$, o subespaço próprio associado é o próprio núcleo (*kernel*) da aplicação A. No caso geral, $S_\lambda = Ker(A - \lambda I)$.

Lembramos que se dois valores próprios λ, μ são distintos, então os vectores próprios associados a λ são independentes dos que estão associados a μ. Basta reparar que se $0 \neq v \in S_\lambda \cap S_\mu$, então $\lambda v = Av = \mu v \Rightarrow (\lambda - \mu)v = 0 \Rightarrow \lambda = \mu$.

Apenas nos interessa considerar o caso em que o espaço vectorial E tem dimensão finita N, que podemos identificar a um certo $\mathbb{R}^N$. No caso de operadores em dimensão infinita, o processo habitual é aproximar o operador linear por uma matriz (operador linear de dimensão finita) e aí determinar os valores próprios. Ou seja, 'formalmente' consideramos $A_n \to A$, e ao determinar $\lambda_n : A_n v_n = \lambda_n v_n$, obtemos uma sucessão tal que $\lambda_n \to \lambda$. *Note-se que isto é apenas possível quando o problema é regular e está demonstrada a dependência contínua.*

Começamos por rever algumas propriedades algébricas dos valores próprios em dimensão finita.

Como $S_\lambda = Ker(A - \lambda I) \neq \{0\}$, λ é valor próprio de A se e só se

$$p_A(\lambda) = \det(\lambda I - A) = 0,$$

o que define uma equação polinomial. Encontrando as raízes desta equação podemos obter a decomposição

$$p_A(\lambda) = (\lambda - \lambda_1)...(\lambda - \lambda_N)$$

em que $\lambda_1, ..., \lambda_N$ são os valores próprios de A. Podemos ter raízes múltiplas nessa equação e, nesse caso, dizemos que λ é um valor próprio com *multiplicidade algébrica* p se λ for uma raiz com multiplicidade p. Distinguimos multiplicidade algébrica de *multiplicidade geométrica,* que determina a dimensão do subespaço próprio S_λ. A multiplicidade geométrica nem sempre coincide com algébrica, para ilustrar esse facto, podemos dar como exemplo a matriz

$$\begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix}$$

onde $\lambda = 1$ é um valor próprio de multiplicidade algébrica 2, raiz da equação $(\lambda - 1)^2 = 0$, mas que tem apenas multiplicidade geométrica 1, no caso de $\varepsilon \neq 0$, porque tem apenas um vector próprio independente, $v = (1, 0)$, e que no caso $\varepsilon = 0$ tem multiplicidade geométrica 2.

Sabemos que a multiplicidade geométrica é sempre menor ou igual que a algébrica. No entanto, enquanto que a soma das multiplicidades algébricas é sempre igual à dimensão da matriz N, a soma das multiplicidades geométricas pode variar muito com pequenas variações das entradas da matriz... basta ver o exemplo anterior!

• Uma propriedade importante do polinómio característico é o *teorema de Hamilton-Cayley*, que afirma

$$p_A(A) = 0,$$

ou seja, a potência A^N pode ser obtida pela combinação linear das potências de grau inferior $I, A, A^2, ..., A^{N-1}$.

• Outras propriedades importantes são aquelas que relacionam valores próprios de diferentes matrizes.

Proposição 5.1 *Se duas matrizes A, B são* semelhantes, *ou seja, se existe uma matriz P invertível*

$$B = P^{-1}AP$$

(P é a matriz mudança de base), então os polinómios característicos são iguais. Portanto, os valores próprios coincidem com a sua multiplicidade, e temos:
v é vector próprio (associado a um valor próprio λ) de B se e só se *Pv for vector próprio (associado ao mesmo valor próprio λ) de A.*

Demonstração: Basta reparar que

$$p_B(\lambda) = \det(\lambda I - B) = \det(\lambda P^{-1}P - P^{-1}AP) = \det(P^{-1}(\lambda I - A)P) = \det(\lambda I - A) = p_A(\lambda),$$

porque $\det(P^{-1}) = 1/\det(P)$. A segunda afirmação resulta de

$$Bv = P^{-1}APv = P^{-1}(\lambda Pv) = \lambda v. \ \square$$

Observação *(decomposição - formas de Schur e Jordan).*

Podemos mesmo obter uma decomposição em que os valores próprios são os elementos da diagonal de uma matriz. A decomposição na *forma normal de Schur* diz-nos que existe uma matriz unitária U tal que:

$$T = U^*AU$$

é uma matriz triangular superior, e portanto $p_A(\lambda) = (\lambda - t_{11})...(\lambda - t_{NN})$.

No caso de A se tratar de uma matriz hermitiana, podemos obter T diagonal, ou seja

$$U^*AU = diag(\lambda_1, ..., \lambda_N)$$

em que vectores próprios associados a $\lambda_1, ..., \lambda_N$ formam uma base ortonormada do espaço.

No caso mais geral, apenas podemos obter a decomposição na forma canónica de Jordan:

$$P^{-1}AP = \begin{bmatrix} J_{n_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{n_2}(\lambda_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{n_r}(\lambda_r) \end{bmatrix}; \ J_{n_i}(\lambda_i) = \begin{bmatrix} \lambda_i & 1 & 0\dots & 0 \\ 0 & \lambda_i & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & \lambda_i \end{bmatrix}_{n_i \times n_i}$$

em que n_i corresponde à multiplicidade algébrica do valor próprio λ_i. O caso que nos interessará especialmente é aquele em a matriz A é diagonalizável, ou seja em que os blocos J_{n_i} têm apenas um elemento.

Observação *(matrizes Hermitianas).*

No caso em que A é uma matriz hermitiana, ou seja $A = A^*$, os valores próprios são reais e a forma normal de Schur assegura que existe uma matriz unitária U tal que é possível a *decomposição espectral:*

$$Ax = UDU^*x = \lambda_1(u_1 \cdot x)\, u_1 + ... + \lambda_N(u_N \cdot x)\, u_N$$

em que $u_1, ..., u_N$ são vectores próprios (ortonormais entre si) associados aos valores próprios $\lambda_1, ..., \lambda_N$. A matriz U é uma matriz de mudança de base formada por esses vectores próprios, enquanto que a matriz D é a matriz diagonal com os respectivos valores próprios. Trata-se de um caso em que a matriz é diagonalizável.

Não é difícil verificar que neste caso

$$A^kx = \lambda_1^k(u_1 \cdot x)\, u_1 + ... + \lambda_N^k(u_N \cdot x)\, u_N$$

Ora, isto permite definir, a partir da expansão em série de Taylor, funções analíticas (inteiras) em que a variável é uma matriz!

Assim, se $f(x) = \alpha_0 + \alpha_1 x + ... + \alpha_n x^n + ...$ obtemos

$$f(A)x = (\alpha_0 I + \alpha_1 A + ... + \alpha_n A^n + ...)x = f(\lambda_1)(u_1 \cdot x)\, u_1 + ... + f(\lambda_N)(u_N \cdot x)\, u_N$$

o que permite, por exemplo, definir a exponencial de uma matriz, a partir dos seus valores próprios:

$$e^A x = e^{\lambda_1}(u_1 \cdot x)\, u_1 + ... + e^{\lambda_N}(u_N \cdot x)\, u_N$$

Esta representação tem especial importância ao resolvermos um sistema de equações diferenciais $u'(t) = Au(t)$, pois nesse caso $u(t) = e^{At}u(0)$.

Reparando que os vectores próprios definem operadores de projecção $P_i x = (u_i \cdot x)\, u_i$ (assumimos $||u_i|| = 1$), podemos também escrever A na forma:

$$Ax = (\lambda_1 P_1 + ... + \lambda_N P_N)x,$$

e daqui $f(A)x = (f(\lambda_1)P_1 + ... + f(\lambda_N)P_N)x$. ((Como informação, observamos que esta representação pode ser generalizada a casos particulares de operadores lineares em espaços de dimensão infinita (operadores auto-adjuntos compactos...))

Observação *(valores singulares).*

Até aqui apenas falámos de valores próprios, noção que se aplica a uma matriz quadrada. Podemos introduzir uma noção adaptada a matrizes não quadradas. Sendo $B \in \mathbb{C}^M \times \mathbb{C}^N$ uma matriz.$M \times N$ com valores complexos (N pode ser igual a M) iremos considerar a matriz quadrada obtida pelo produto da matriz adjunta $B^* = \bar{B}^\top$ por ela própria. Dessa forma obtemos uma matriz quadrada $A = B^*B$ de dimensão $N \times N$, que será hermitiana e semi-definida positiva. Os valores próprios de A serão positivos ou nulos e é através desses valores que definimos valores singulares:

• Dizemos que $\mu \geq 0$ é *valor singular* de B se μ^2 for valor próprio de $A = B^*B$.

(i) Note-se que os núcleos de A e B coincidem, ie. $Ker(A) = Ker(B)$, porque se $Bx = 0$ é óbvio que $Ax = B^*Bx = 0$; e reciprocamente, se $Ax = 0$ temos

$$x^*Ax = x^*B^*Bx = (Bx)^*Bx = ||Bx||_2^2 = 0,$$

o que implica $Bx = 0$.□

Repare-se que isto significa que o número de valores singulares positivos (contando com a multiplicidade geométrica) será igual à característica da matriz B, e como é claro, os restantes valores singulares serão nulos.

(ii) A norma euclidiana de uma matriz não quadrada é dada por $||B||_2 = \sqrt{\rho(B^*B)}$, e assim concluímos que a norma euclidiana de uma matriz será igual ao maior valor singular.

(iii) A principal propriedade, é a *decomposição em valores singulares* (o análogo da decomposição espectral), que garante a existência de N vectores ortonormais $u_1, u_2, ...u_N \in \mathbb{C}^N$, e de M vectores ortonormais $v_1, ..., v_M \in \mathbb{C}^M$:

$$Bx = \mu_1(u_1 \cdot x)v_1 + ... + \mu_r(u_r \cdot x)v_r,$$

em que r é a característica da matriz B (os μ_i restantes seriam nulos). Note-se que $Bu_k = \mu_k v_k$, e que $B^*v_k = \mu_k u_k$. Quando um sistema da forma $Bx = y$ tem solução (ou soluções), então

$$x = \frac{1}{\mu_1}(y \cdot v_1)u_1 + ... + \frac{1}{\mu_r}(y \cdot v_r)u_r.$$

• Os valores singulares têm especial interesse na aproximação de sistemas mal condicionados. Por exemplo, estão relacionados com o problema de aproximação de dados pelo método dos mínimos quadrados, que iremos abordar no último capítulo.(Para maior detalhe, consultar p.ex.[18].)

5.1.1 Valores próprios e o polinómio característico

Já vimos que sendo A uma matriz $N \times N$, encontrar os valores próprios de A é encontrar as raizes $\lambda \in \mathbf{C}$:

$$p_A(\lambda) = \det(\lambda I - A) = 0,$$

em que $p_A(\lambda)$ é o polinómio característico de grau N, e isto corresponde a resolver uma equação polinomial.

- Encarando o determinante como forma multilinear, temos

$$p_A(\lambda) = \det(\lambda I - A) = \det(\lambda e^{(1)} - a^{(1)}, ..., \lambda e^{(N)} - a^{(N)})$$

em que $a^{(k)}$ são as linhas da matriz A e $e^{(k)}$ as linhas da matriz identidade (i.e: o vector k da base canónica).

Ora, se desenvolvermos $\det(\lambda e^{(1)} - a^{(1)}, ..., \lambda e^{(N)} - a^{(N)})$, obtemos

$$p_A(\lambda) = \lambda^N \det(e^{(1)}, ..., e^{(N)}) + ... + (-1)^N \det(a^{(1)}, ..., a^{(N)})$$

e reparamos que no termo constante $\det(a^{(1)}, ..., a^{(N)}) = \det(A)$. Por outro lado, como

$$p_A(\lambda) = (-1)^N \lambda_1 ... \lambda_N + ... - (\lambda_1 + ... + \lambda_N)\lambda^{N-1} + \lambda^N,$$

isto implica imediatamente que os termos constantes têm que ser iguais, ou seja $\det(A) = \lambda_1 ... \lambda_N$. Da mesma forma, podemos obter

$$\lambda_1 + ... + \lambda_N = \det(a^{(1)}, e^{(2)}, ..., e^{(N)}) + ... + \det(e^{(1)}, ..., e^{(N-1)}, a^{(N)}) = a_{11} + ... + a_{NN} = \mathrm{tr}(A),$$

ou seja, a soma dos valores próprios é igual a $\mathrm{tr}(A)$, o traço da matriz A.

Portanto, podemos concluir as relações

$$\begin{gathered} \mathrm{tr}(A) = \lambda_1 + ... + \lambda_N, \\ \det(A) = \lambda_1 ... \lambda_N. \end{gathered}$$

- Escrevendo o polinómio característico na forma

$$p_A(\lambda) = \alpha_1 + \alpha_2 \lambda + ... + \alpha_N \lambda^{N-1} + \lambda^N,$$

acabamos de concluir que os valores dos coeficientes α_k são obtidos pelo cálculo de determinantes, o que significa que há uma dependência contínua dos α_k face aos valores dos elementos da matriz. Sendo os valores próprios as raízes $\lambda_1, ..., \lambda_N$, do polinómio característico, resta saber se há uma dependência contínua das raízes face à variação dos coeficientes, para concluir que os valores próprios dependem de forma contínua das entradas da matriz. De facto isso verifica-se[1]:

[1] Uma demonstração alternativa, usando o teorema de Rouché pode ser vista em [25].

Lema 5.1 *Seja* $p(x) = \alpha_1 + \alpha_2 x + ... + \alpha_N x^{N-1} + x^N$, *com* $\alpha = (\alpha_1, ..., \alpha_N) \in \mathbb{C}^N$, *e seja* $z = (z_1, ..., z_N) \in \mathbb{C}^N$ *um vector que tem as* N *raízes de* p. *Consideremos agora uma perturbação do polinómio,* $\tilde{p}$, *com coeficientes em* $\tilde{\alpha}$ *e raízes em* $\tilde{z}$ *(ordenadas convenientemente, de forma a que* $\tilde{z}_k$ *seja a componente mais próxima de* z_k*). Então,*

$$\tilde{\alpha} \to \alpha \Rightarrow \tilde{z} \to z,$$

ou seja, há uma dependência contínua das raízes face aos coeficientes.

Demonstração:
Usando o teorema fundamental da álgebra,

$$p(x) = (x - z_1)...(x - z_N) = \alpha_1 + \alpha_2 x + ... + \alpha_N x^{N-1} + x^N$$

e a igualdade entre os polinómios permite escrever um sistema com N equações relacionando cada α_k como função contínua dos valores $z_1, ..., z_N$. Por exemplo, $\alpha_1 = (-1)^N z_1 ... z_N$, ou $\alpha_N = -(z_1 + ... + z_N)$. Ou seja, com notação vectorial, podemos escrever

$$\alpha = \mathcal{P}(z),$$

em que $\mathcal{P}$ é uma função vectorial de $\mathbb{C}^N$ em $\mathbb{C}^N$,

$$\mathcal{P}(z_1, ..., z_N) = ((-1)^N z_1 ... z_N, \cdots\cdots, -(z_1 + ... + z_N)).$$

Esta função é claramente contínua e a menos de permutação na lista $(z_1, ..., z_N)$ também é injectiva. Aliás, considerando a relação de equivalência $w \triangleq z$ quando $\sigma(z) = w$, onde σ é uma permutação dos coeficientes de z, podemos definir a bijectividade de

$$\begin{array}{rccc} \mathcal{P}: & \mathbb{C}^N / \triangleq & \longrightarrow & \mathbb{C}^N \\ & \dot{z} & \longmapsto & \alpha \end{array}$$

pelo teorema fundamental da álgebra. Como se trata de uma aplicação contínua e bijectiva a sua inversa é também contínua (i.e. trata-se de um homeomorfismo). Concluímos assim que com uma ordenação apropriada das raízes é possível obter o resultado. ■

Corolário 5.1 *Os valores próprios são funções contínuas dos elementos da matriz.*

5.2 Teorema de Gerschgorin

Podemos começar por retirar alguma informação acerca da localização dos valores próprios usando o teorema do Ponto Fixo. Com efeito, reparamos que se $\lambda \neq 0$, podemos escrever

$$Av = \lambda v \Leftrightarrow v = \frac{A}{\lambda} v$$

e se $||\frac{A}{\lambda}|| < 1$, temos uma contracção, logo a única solução será $v = 0$. Assim, para termos soluções não nulas, e consequentemente valores próprios, é necessário que $\lambda \leq ||A||$ (o caso $\lambda = 0$ é trivial). Isto reflecte a propriedade que já tinhamos visto acerca do raio espectral

$$\rho(A) \leq ||A||.$$

No caso de A ser uma matriz hermitiana é também possível obter uma minoração de forma simples,

$$x^*.Ax = x^*U^*DUx \leq x^*U^*(\lambda_{\max}I)Ux = \lambda_{\max}x^*.x = \lambda_{\max}||x||_2^2$$

o que significa que

$$\rho(A) \geq \max_{||u||_2=1} ||u^*Au||_2$$

Exercício 5.1 *(Quociente de Rayleigh). Mostre que se A for hermitiana então o maior valor próprio verifica*

$$\lambda_{\max} = \max_{x\neq 0} \frac{x^*Ax}{x^*x}.$$

No entanto, estes resultados podem ser melhorados. O próximo teorema permite obter informações *a priori,* mais concretas, acerca da localização dos valores próprios, através dos elementos da matriz.

Teorema 5.1 *(Gerschgorin).*
a) *Um valor próprio λ de uma matriz A verifica uma das seguintes desigualdades:*

$$|a_{kk} - \lambda| \leq \sum_{j=1, j\neq i}^{N} |a_{kj}| = r_k, \ (k = 1, .., N)$$

o que significa que os valores próprios pertencem a bolas fechadas com centro na diagonal e raio r_k, ou seja, $\lambda \in \bigcup_{k=1}^{N} \bar{B}(a_{kk}, r_k)$.
b) *Para além disso, se a reunião de m bolas forma uma componente conexa, haverá exactamente m valores próprios nessa componente (consideramos $m \geq 1$).*
c) *O mesmo argumento é válido se considerarmos linhas ao invés de colunas!*

Demonstração:
a) Um vector próprio v associado ao valor próprio λ verifica

$$[Av]_i = \sum_{j=1}^{N} a_{ij}v_j = \lambda v_i \Leftrightarrow \sum_{j=1, j\neq i}^{N} a_{ij}v_j + a_{ii}v_i = \lambda v_i$$

e daqui obtemos

$$\sum_{j=1, j\neq i}^{N} a_{ij}v_j = (\lambda - a_{ii})v_i$$

e portanto

$$|\lambda - a_{ii}|\,|v_i| \leq \sum_{j=1, j\neq i}^{N} |a_{ij}|\,|v_j|.$$

Considerando agora o índice k para o qual $|v_k| = \max_{i=1,\ldots,N} |v_i| = ||v||_\infty$ obtemos

$$|\lambda - a_{kk}|\,||v||_\infty \leq \sum_{j=1, j\neq i}^{N} |a_{kj}|\,|v_j| \leq \sum_{j=1, j\neq i}^{N} |a_{kj}|\,||v||_\infty$$

e assim, dividindo por $||v||_\infty \neq 0$ (porque é um valor próprio), obtemos o resultado.

b) Para mostrar a segunda parte, usamos argumentos analíticos. Consideramos um "*segmento* formado por matrizes"

$$A_t = D + t(A - D),\ (t \in [0, 1]),$$

que começa na matriz $D = diag(a_{11}, ..., a_{NN})$, quando $t = 0$, e termina em A, quando $t = 1$.

Essas matrizes A_t têm valores próprios associados $\Lambda_1(t), ..., \Lambda_N(t)$ que vão definir linhas contínuas (caminhos) no plano complexo. As matrizes A_t têm a mesma diagonal que a matriz A, e as outras entradas estão multiplicadas por t. Temos $\Lambda_i(0) = a_{kk}$, e também $\Lambda_i(1) = \lambda_i$.

Pelo que vimos em a), concluímos que os valores próprios $\lambda_i(t)$ pertencem à reunião das bolas, $\bigcup_k \bar{B}(a_{kk}, t\, r_k) \subseteq \bigcup_k \bar{B}(a_{kk},\ r_k)$.

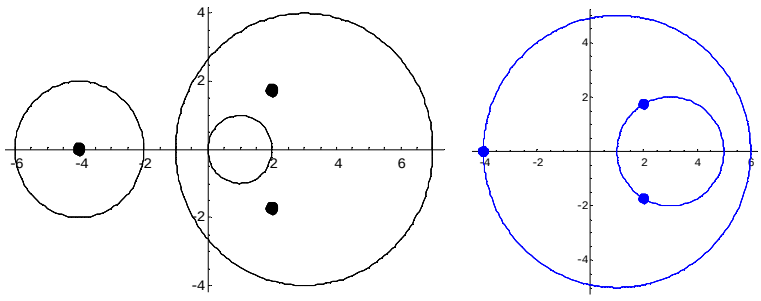
Pelo corolário do lema anterior as funções $\Lambda_k : [0, 1] \to \mathbf{C}$ são contínuas, consequentemente transformam conexos em conexos, logo $\Lambda_k([0, 1])$ é conexo, e por outro lado sabemos que tem que pertencer a $\bigcup_k \bar{B}(a_{kk},\ r_k)$. Isto implica que tem que pertencer a uma componente conexa dessa reunião. Assim, $\Lambda_k(1) = \lambda_k$ pertencem exactamente à componente conexa que contém $\Lambda_k(0) = a_{kk}$ e o resultado está provado.

c) Basta reparar que os valores próprios de A^T coincidem com os de A, porque $\det(A - \lambda I) = \det((A - \lambda I)^T) = \det(A^T - \lambda I)$.■

Observação: Quando a matriz é real, o polinómio característico tem coeficientes reais. Assim, nesse caso, os valores próprios complexos aparecem como pares conjugados. Logo, caso se conclua que há apenas um valor próprio numa componente conexa, ele terá que obrigatoriamente ser real. Isso acontece frequentemente quando a componente conexa é uma bola disjunta das restantes.

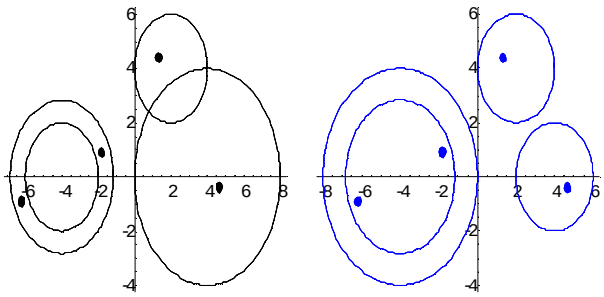
Exemplo 5.1 *Consideremos um primeiro caso em que a matriz é*

$$A = \begin{bmatrix} -4 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 4 & 3 \end{bmatrix}$$



Na primeira figura são representadas as três bolas $\bar{B}(-4,2), \bar{B}(1,1), \bar{B}(3,4)$, *bem como a localização dos valores próprios, no plano complexo. Estas três bolas são as que se obtêm fazendo uma análise por linhas com o Teorema de Gerschgorin. Como a primeira bola é disjunta das outras duas conclui-se que há um valor próprio em* $\bar{B}(-4,2)$ *(que é obrigatoriamente real, pela observação anterior) e dois valores próprios na reunião das outras duas,* $\bar{B}(1,1) \cup \bar{B}(3,4) = \bar{B}(3,4)$, *o que é confirmado na figura. Repare-se que na bola* $\bar{B}(1,1)$ *não há nenhum valor próprio.*
Na segunda figura é feita uma análise por colunas. Nesse caso, obtemos $\bar{B}(-4,0), \bar{B}(1,5), \bar{B}(3,2)$. *No entanto, reparamos que* $\bar{B}(4,0)$ *não é mais que o ponto* $z = -4$, *que é obviamente um valor próprio, pois basta considerar* $v = (1,0,0)$, *para termos* $Av = -4v$. *Isso acontece sempre nos casos em a matriz tem uma coluna (ou linha) em que o elemento não nulo está na diagonal. Assim, a análise por colunas permite concluir que há um valor próprio* $\lambda_1 = -4$ *e que os dois restantes,* λ_2, λ_3 *estão em* $B(1,5)$. *Juntando esta informação com a obtida por linhas, podemos concluir-se que* $\lambda_1 = -4$, *e que* $\lambda_2, \lambda_3 \in \bar{B}(3,4)$, *o que é confirmado nas figuras. Note-se que, neste caso, uma análise através da regra de Laplace permitiria resultados imediatos.*
Consideremos agora um outro exemplo em que a matriz é

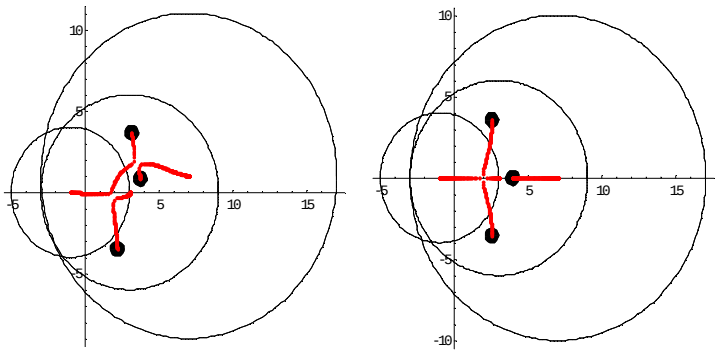
$$A = \begin{bmatrix} -4 & 2 & 0 & 0 \\ 2+2i & -4 & 0 & 0 \\ 0 & 0 & 2+4i & -2 \\ 0 & 2 & 2i & 4 \end{bmatrix}$$



Neste caso, fazendo uma análise por linhas, obtemos as bolas $B_1 = \bar{B}(-4,2), B_2 = \bar{B}(-4,2\sqrt{2}), B_3 = \bar{B}(2+4i,2), B_4 = \bar{B}(4,4)$. *A reunião tem duas componentes conexas, uma formada por* $B_1 \cup B_2 = B_2$, *que é disjunta da outra formada por* $B_3 \cup B_4$. *Pelo teorema concluímos que há dois valores próprios em cada uma destas componentes. Façamos agora uma análise por colunas. Nesse caso temos* $B'_1 = \bar{B}(-4,2\sqrt{2}), B'_2 = \bar{B}(-4,4), B'_3 = \bar{B}(2+4i,2), B'_4 = \bar{B}(4,-2)$. *Aqui há três componentes conexas, uma formada pela reunião* $B'_1 \cup B'_2 = B'_2$ *e duas outras formadas pelas bolas* B'_3 *e* B'_4. *Podemos assim concluir que há um valor próprio* $\lambda_3 \in B'_3$, *outro* $\lambda_4 \in B'_4$ *e que os outros dois* $\lambda_1, \lambda_2 \in \bar{B}(-4,4) = B'_1 \cup B'_2$. *Intersectando esta informação com a informação obtida por linhas, podemos mesmo concluir que* $\lambda_1, \lambda_2 \in \bar{B}(-4,2\sqrt{2})$.
Como curiosidade, apresentamos o gráfico da evolução dos valores próprios da matriz D *para a matriz* A, *através do segmento de matrizes formadas por* $A_t = D + t(A - D)$, *que*

foi utilizado na demonstração do teorema. Consideremos a matriz

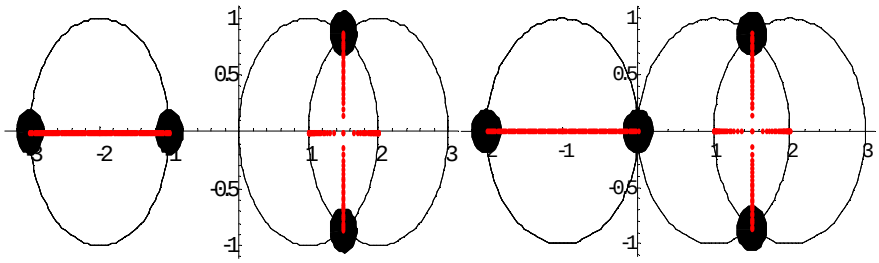
$$A = \begin{bmatrix} -1 & 2 & -2 \\ -4 & 3 & -2 \\ 5 & 5 & 7+\alpha i \end{bmatrix}$$



Podemos ver, no primeiro gráfico, para $\alpha = 1$, a evolução da posição dos valores próprios desde a diagonal D, que tem valores próprios $\Lambda_1(0) = -1, \Lambda_2(0) = 3, \Lambda_3(0) = 7 + i$, a que correspondem os pontos $x_1 = (-1, 0), x_2 = (3, 0), x_3 = (7, 1)$, até aos valores próprios da matriz final A. Repare-se na evolução do valor próprio Λ_1, que começando no centro x_1 acaba por sair da bola $\bar{B}(-1, 4)$. Isto retrata bem que pode haver bolas (definidas pela análise de linhas ou colunas) onde não há valores próprios, apenas podemos garantir a existência de valores próprios na componente conexa. No segundo gráfico mostramos a mesma evolução, mas para $\alpha = 0$. Repare-se que para um certo t os valores das trajectórias de Λ_1 e Λ_2 coincidem, o que corresponde a um valor próprio com multiplicidade 2. A partir desse t as trajectórias tomam sempre valores complexos conjugados, como é característico das matrizes reais; note-se que é indiferente dizer que a trajectória de Λ_1 é continuada para o valor próprio com parte imaginária positiva ou negativa, o que interessa é que a trajectória é contínua.

Terminamos com um exemplo em que os valores próprios estão na fronteira das bolas,

$$A = \begin{bmatrix} -\alpha & 1 & 0 & 0 \\ 1 & -\alpha & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$



Quer seja feita uma análise por linhas ou colunas o resultado é o mesmo. Se $\alpha = 2$ (primeira figura) conclui-se que deverá haver dois valores próprios na bola $\bar{B}(-2, 1)$ e dois valores próprios em $\bar{B}(1, 1) \cup \bar{B}(2, 1)$. Neste caso, como se pode observar, os valores próprios estão exactamente sobre a fronteira das componentes conexas, o que ilustra que a estimativa não poderia dar raios inferiores. Se $\alpha = 1$ (segunda figura) temos uma situação em que há uma translação da situação anterior para uma situação de contacto num único ponto. A reunião das bolas fechadas tem apenas uma componente conexa, que é ela própria. No entanto, nesta situação, em que a reunião das bolas abertas tem duas componentes conexas (o que significa, na prática, que há apenas intersecção num ponto), seguindo a demonstração do teorema, podemos concluir que a trajectória de um valor próprio $\Lambda_1(t)$ pertenceria sempre à bola $\bar{B}(-2, t)$, para $t < 1$, e portanto (sendo contínua) no instante $t = 1$ apenas poderia

estar na fronteira. É o que se passa neste caso. Da mesma forma, a conclusão do teorema mantém-se válida num caso geral, quando há intersecção num único ponto. Basta ver o que se passa nesse ponto, analisando se se trata ou não de um valor próprio (e qual a sua multiplicidade) e concluir para as componentes.

5.3 Método das Potências

Estando interessados em encontrar os valores próprios de uma matriz podemos pensar imediatamente num processo – encontrar as raízes do polinómio característico. Para esse efeito podemos usar qualquer método que vimos, como sejam os métodos de Bernoulli, de Newton, da Secante, ou de Steffensen. O primeiro apenas nos dá a maior raiz real, o segundo necessita do cálculo da derivada de um determinante, e os outros dois necessitam do cálculo em cada iterada de um determinante, o que é bastante moroso, para além de serem precisas boas aproximações iniciais...

Poderíamos ainda simplificar o processo determinando exactamente o polinómio através de interpolação polinomial usando apenas o cálculo de $N+1$ determinantes, o que reduziria o número de cálculos... mas mesmo este processo pode ser demasiado moroso.

Vamos começar por ver um processo extremamente simples, o *método das potências,* de *von Mises* (1929), que no entanto funciona apenas em circunstâncias particulares! É o método mais simples e pode ser encarado como um método de ponto fixo, em que se procura um vector próprio u de norma 1 (associado a um valor próprio $\lambda \neq 0$) no conjunto $S = \{x \in \mathbb{R}^N : ||x|| = 1\}$.

Escrevendo

$$Au = \lambda u \Leftrightarrow u = \frac{Au}{\lambda},$$

e reparando que $||Au|| = ||\lambda u|| = |\lambda|$, obtemos

$$u = \frac{|\lambda|}{\lambda} \frac{Au}{||Au||}.$$

O método iterativo poderia ficar

$$u^{(n+1)} = \frac{|\lambda|}{\lambda} \frac{Au^{(n)}}{||Au^{(n)}||},$$

mas isso implicava um conhecimento *a priori* do argumento $\theta_\lambda \in [0, 2\pi[$ do valor próprio, caso λ fosse um número complexo, pois $\frac{|\lambda|}{\lambda} = e^{-\theta_\lambda i}$.

No entanto, no caso de se tratar de um valor próprio real $\frac{|\lambda|}{\lambda} = \pm 1$, e a situação é mais fácil de resolver... sob certas condições. Devemos começar por reparar que, havendo sempre mais que um valor próprio, a convergência de uma tal sucessão não estaria *a priori* bem determinada. É preciso impor restrições.

• Admitiremos assim que a matriz é diagonalizável e que um dos valores próprios é dominante, ou seja,

$$|\lambda_1| > \max_{i=2,\ldots,N} |\lambda_i|,$$

e também que esse valor próprio dominante, λ_1, é real.

Observação: A condição de ser diagonalizável pode ser verificada imediatamente se a matriz for hermitiana. Nesse caso, a matriz tem valores próprios reais e basta mostrar que há um valor próprio maior que os restantes. Isso pode ser provado pelo teorema de Gerschgorin, mas, na prática, esta situação é quase sempre verificada, a menos que haja valores próprios com multiplicidade superior a 1, ou em que há um outro valor próprio dominante simétrico.

• Verificadas estas condições, podemos estabelecer o *método das potências*[2]**:**

$$\begin{cases} u^{(0)} : ||u^{(0)}|| = 1, \\ u^{(n+1)} = \sigma_n \frac{Au^{(n)}}{||Au^{(n)}||}, \end{cases}$$

em que $\sigma_n = \pm 1$ é o sinal da componente com maior módulo do vector $Au^{(n)}$. Todas as iteradas são vectores unitários para a norma considerada.

Habitualmente, considera-se a normalização usando a norma do máximo (também é frequente usar a norma euclidiana), que será a adoptada no que se segue.

A iterada inicial $u^{(0)}$ é normalmente um valor aleatório, de forma a que na base ortonormada $v_1, ..., v_N$ formada pelos vectores próprios tenha componente não nula relativamente ao vector próprio associado ao valor próprio dominante. Ou seja, exigimos que

$$u^{(0)} = \alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N$$

com $\alpha \neq 0$.

Proposição 5.2 *Seja A uma matriz diagonalizável (em particular, hermitiana) com um valor próprio dominante λ_1 real: $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_N|$. Se a iterada inicial $u^{(0)}$ tiver a componente $\alpha \neq 0$, o método das potências converge para v_1, e uma aproximação para o valor próprio dominante é*

$$\lambda^{(n)} = \frac{[Au^{(n)}]_i}{u_i^{(n)}}$$

[2]Uma outra possibilidade é considerar simplesmente

$$\begin{cases} x^{(0)} \in \mathbf{R}^d, \\ x^{(n+1)} = Ax^{(n)}, \end{cases}$$

e apenas normalizar no final dividindo por $||Ax^{(n)}||$ já que a divisão sucessiva por $||Ax^{(n)}||$ tem apenas como objectivo evitar a divergência, mantendo sempre o vector com norma 1.

É aliás fácil verificar que se μ_i são escalares,

$$\mu_n A(...(\mu_1 A(\mu_0 x^{(0)})...) = \mu_n ... \mu_0 A^n x^{(0)}$$

e portanto a normalização no final leva ao mesmo resultado. Como podemos ver, se $u^{(n)} = \frac{x^{(n)}}{||x^{(n)}||}$,

$$\frac{Au^{(n)}}{||Au^{(n)}||} = \frac{||x^{(n)}||}{||Ax^{(n)}||} A(\frac{x^{(n)}}{||x^{(n)}||}) = \frac{Ax^{(n)}}{||Ax^{(n)}||}.$$

No entanto, se não for efectuada a normalização, as iteradas podem tomar valores que crescem muito rapidamente e surgem problemas de cálculo e precisão.

para qualquer índice i (desde que $u_i^{(n)} \neq 0$*), sendo normalmente escolhido o índice com componente igual a 1. Estabelecemos, também, a estimativa de erro*[3]

$$||v_1 - u^{(n)}||_\infty \leq C \left|\frac{\lambda_2}{\lambda_1}\right|^n,$$

(onde a constante $C > 0$ *depende directamente de* $\frac{1}{|\alpha|}$*) e também*

$$|\lambda_1 - \lambda^{(n)}| \leq C' \left|\frac{\lambda_2}{\lambda_1}\right|^n.$$

Demonstração:[4]

Seja $v_1, ..., v_N$ a base de valores próprios unitária (i.e. $||v_k||_\infty = 1$), associada aos valores próprios $\lambda_1, ..., \lambda_N$.

Como supomos que $u^{(0)} = \alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N$, com $\alpha \neq 0$, vamos considerar o subconjunto fechado de $\mathbb{R}^N$

$$S_\alpha = \{\alpha v_1 + \alpha_2 v_2 + ... + \alpha_N v_N : \alpha_2, ..., \alpha_N \in \mathbb{R}\},$$

que é um subespaço afim. Para qualquer $x \in S_\alpha$, podemos considerar a decomposição $x = \alpha v_1 + \tilde{x}$.

Como podemos escrever $x = x'_1 v_1 + x'_2 v_2 + ... + x'_N v_N$, vamos considerar a norma em $\mathbb{R}^N$

$$||x||_1^* = |x'_1| + ... + |x'_N|,$$

que é equivalente a qualquer outra norma de $\mathbb{R}^N$. Por exemplo, temos $c_1||.||_\infty \leq ||.||_1^* \leq c_2||.||_\infty$, com $c_1 = 1, c_2 = N||P||_\infty$, em que P é a matriz de mudança para a base canónica, $Pv_k = e_k$, porque

$$\begin{aligned} ||x||_\infty &= ||x'_1 v_1 + ...x'_N v_N||_\infty \leq |x'_1| + ... + |x'_N| = ||x||_1^* \\ ||x||_1^* &= |x'_1| + ... + |x'_N| \leq N \max_{k=1,...,N} |x'_k| = N||Px||_\infty \leq N||P||_\infty ||x||_\infty, \end{aligned}$$

já que $x'_1 e_1 + ... + x'_N e_N = P(x'_1 v_1 + ... + x'_N v_N)$, e portanto $x'_k = [Px]_k$.

Note-se ainda que $||v_k||_1^* = 1$.

• Define-se a aplicação $T : S_\alpha \to S_\alpha$

$$Tx = \frac{Ax}{\lambda_1}$$

(note-se que $\lambda_1 \neq 0$, senão todos os valores próprios seriam nulos).

[3]Esta estimativa de erro é válida em qualquer norma, já que como todas as normas são equivalentes,

$$||v_1 - u^{(n)}|| \leq c_2||v_1 - u^{(n)}||_\infty,$$

e trata-se apenas de ajustar a constante.

[4]A demonstração clássica é eventualmente bastante mais simples (e.g. [1]), mas fornece menos informação.

Note-se que $T(S_\alpha) = S_\alpha$, pois quando $x \in S_\alpha$ temos $x = \alpha v_1 + x'_2 v_2 + ... + x'_N v_N = \alpha v_1 + \tilde{x}$, e é fácil ver que

$$Tx = \frac{A}{\lambda_1}(\alpha v_1 + \tilde{x}) = \frac{1}{\lambda_1}(\alpha A v_1 + A\tilde{x}) = \alpha \frac{\lambda_1}{\lambda_1} v_1 + \frac{1}{\lambda_1} A\tilde{x} \in S_\alpha.$$

porque $A\tilde{x} = A(x'_2 v_2 + ... + x'_N v_N) = x'_2 \lambda_2 v_2 + ... + x'_N \lambda_N v_N$.

• Vejamos agora que se trata de uma contracção:

$$\begin{aligned} ||Tx - Ty||_1^* = ||\tfrac{1}{\lambda_1} A\tilde{x} - \tfrac{1}{\lambda_1} A\tilde{y}||_\infty = \tfrac{1}{|\lambda_1|} ||(x'_2 - y'_2)\lambda_2 v_2 + ... + (x'_N - y'_N)\lambda_N v_N||_1^* = \\ = \left|\tfrac{\lambda_2}{\lambda_1}\right| ||(x'_2 - y'_2)v_2 + (x'_3 - y'_3)(\tfrac{\lambda_3}{\lambda_2})v_3 ... + (x'_N - y'_N)(\tfrac{\lambda_N}{\lambda_2})v_N||_1^* = \\ = \left|\tfrac{\lambda_2}{\lambda_1}\right| \left(|x'_2 - y'_2| + |x'_3 - y'_3|\,|\tfrac{\lambda_3}{\lambda_2}| + ... + |x'_N - y'_N|\,|\tfrac{\lambda_N}{\lambda_2}|\right) \tfrac{1}{N} \leq \left|\tfrac{\lambda_2}{\lambda_1}\right| \; ||x - y||_1^*. \end{aligned}$$

Portanto, pelo Teorema do Ponto Fixo existe um único $z \in S_\alpha : Tz = z$, esse valor é $z = \alpha v_1$ e temos

$$||z - T^n u^{(0)}||_1^* \leq \left|\frac{\lambda_2}{\lambda_1}\right|^n ||z - u^{(0)}||_1^*.$$

Como a norma $||.||_1^*$ é equivalente a $||.||_\infty$, temos

$$||z - T^n u^{(0)}||_\infty \leq \frac{c_2}{c_1} \left|\frac{\lambda_2}{\lambda_1}\right|^n ||z - u^{(0)}||_\infty,$$

em que $c_2 \geq c_1 > 0$ são as constantes que determinam a equivalência entre as normas.

Por outro lado, temos $||\frac{\mathbf{a}}{||\mathbf{a}||} - \frac{\mathbf{b}}{||\mathbf{b}||}|| \leq 2||\frac{\mathbf{a}}{||\mathbf{a}||} - \beta\mathbf{b}||, \forall \beta \geq 0$ (exercício[5]). Aplicando esta desigualdade com $\mathbf{a} = z, \mathbf{b} = T^n u^{(0)}, \beta = \frac{1}{||z||_\infty} = \frac{1}{|\alpha|}$,obtemos

$$||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty}||_\infty \leq 2||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||z||_\infty}||_\infty,$$

logo,

$$||\frac{z}{||z||_\infty} - \frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty}||_\infty \leq \frac{2}{|\alpha|}\frac{c_2}{c_1} \left|\frac{\lambda_2}{\lambda_1}\right|^n ||z - u^{(0)}||_\infty.$$

[5]Quando $||a|| = 1$, temos a desigualdade

$$||a - \frac{b}{||b||}|| \leq 2||a - b||,$$

porque

$$||a - \frac{b}{||b||}|| \leq ||a - b|| + ||b - \frac{b}{||b||}||$$

e como

$$||b - \frac{b}{||b||}|| = \frac{1}{||b||}||b\,(||b|| - 1)\,|| = |\,||b|| - 1|\,,$$

quando $||a|| = 1$ temos

$$||a - \frac{b}{||b||}|| \leq ||a - b|| + |\,||b|| - ||a||\,| \leq 2||a - b||.$$

Basta agora considerar $a = \frac{\mathbf{a}}{||\mathbf{a}||}$ e $b = \beta\mathbf{b}$.

Basta agora reparar que $T^n u^{(0)} = \frac{A^n u^{(0)}}{\lambda_1^n}$, e portanto

$$\frac{T^n u^{(0)}}{||T^n u^{(0)}||_\infty} = \frac{A^n u^{(0)}}{\lambda_1^n} || \frac{\lambda_1^n}{A^n u^{(0)}} ||_\infty = \left(\frac{|\lambda_1|}{\lambda_1} \right)^n \frac{A^n u^{(0)}}{||A^n u^{(0)}||_\infty} = (\pm 1)^n \sigma_1 ... \sigma_n \frac{A u^{(n)}}{||A u^{(n)}||_\infty} = \pm u^{(n+1)},$$

e por outro lado $\frac{z}{||z||_\infty} = \frac{\alpha}{|\alpha|} v_1 = \pm v_1$ (os sinais coincidem devido à construção), pelo que

$$||v_1 - u^{(n+1)}||_\infty \leq 2 \frac{c_2}{c_1} \left| \frac{\lambda_2}{\lambda_1} \right|^n ||v_1 - \frac{u^{(0)}}{|\alpha|}||_\infty.$$

Como $||v_1 - \frac{u^{(0)}}{|\alpha|}||_\infty = ||\frac{\alpha_2}{|\alpha|} v_2 + ... + \frac{\alpha_N}{|\alpha|} v_N||_\infty \leq \frac{1}{|\alpha|} ||u^{(0)}||_1^*$, e como $||u^{(0)}||_1^* \leq c_2 ||u^{(0)}||_\infty = c_2$, obtemos

$$||v_1 - u^{(n+1)}||_\infty \leq \frac{K}{|\alpha|} \left| \frac{\lambda_2}{\lambda_1} \right|^n$$

em que a constante $K = \frac{2c_2^2}{c_1}$ só poderia ser calculada se fossem conhecidos os valores e vectores próprios. Repare-se que se $\alpha \to 0$ o majorante tende para infinito[6], o que coloca em evidência que a componente segundo o vector próprio dominante não pode ser nula.

Finalmente, como $\lambda^{(n)} = [A u^{(n)}]_j$, em que j é o índice correspondente a $u_j^{(n)} = 1$, e temos $\lambda_1 = [A v_1]_j$, obtemos

$$|\lambda_1 - \lambda^{(n)}| = |[A v_1]_j - [A u^{(n)}]_j| \leq ||A v_1 - A u^{(n)}||_\infty \leq ||A||_\infty ||v_1 - u^{(n)}||_\infty. \blacksquare$$

Observação 1: No caso de ser utilizada a norma $||.||_2$ para a normalização, a estimativa obtida pode ser explícita no caso de matrizes simétricas, já que a matriz mudança de base P será unitária.

Observação 2: A restrição à escolha do vector inicial $u^{(0)}$ não é significativa. Dificilmente acontece a situação improvável de ser exactamente um vector cuja componente segundo v_1 fosse nula. Se usarmos números com todos os decimais (p. ex. gerados aleatoriamente), é praticamente uma situação inexistente, já que os próprios erros de arredondamento fazem aparecer uma pequena componente.

No entanto, em casos mais simples, quando usamos números inteiros, podemos cair facilmente em situações em que isso acontece. Apresentamos um exemplo simples. Consideremos a matriz

$$M = \begin{bmatrix} 5 & -7 & 7 \\ 6 & -9 & 8 \\ 6 & -7 & 6 \end{bmatrix}$$

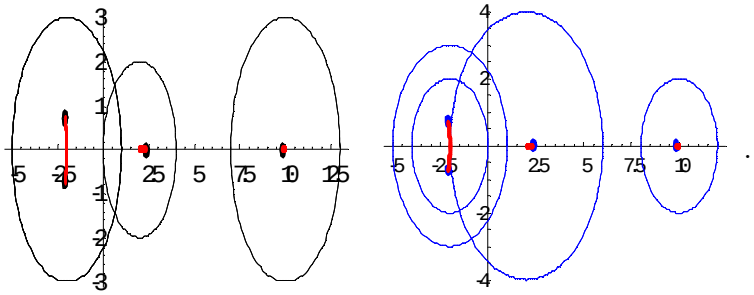
Se começarmos com $u^{(0)} = (1, 1, 0)$, obtemos $u^{(n)} \to v = (0.5, 1, 0.5)$. No entanto, o limite obtido não é o vector próprio associado ao valor próprio dominante! A matriz tem valores próprios $\lambda_1 = 5, \lambda_2 = -2, \lambda_3 = -1$, e é fácil ver que $Mv = -2v$, portanto v é o vector próprio associado ao valor próprio λ_2 e não a λ_1. A razão é simples, como os vectores próprios são $v_1 = (1, 1, 1), v_2 = (0.5, 1, 0.5), v_3 = (0, 1, 1)$, temos $u^{(0)} = 2v_2 - v_3$, ou seja,

[6] Isto não significa que não é possível majorar $||v_1 - u^{(n+1)}||_\infty$, pois é óbvio que será sempre menor que 2. Apenas significa que não há convergência para zero.

a componente segundo v_1 é $\alpha = 0$. Uma simples perturbação, usando $u^{(0)} = (1, 1, \varepsilon)$ com $\varepsilon \neq 0$, mesmo muito pequeno, já será suficiente para que o método convirja para v_1.

Exemplo 5.2 *Consideremos a matriz:*

$$A = \begin{bmatrix} -2 & 1 & 0 & 2 \\ -1 & 10 & -1 & -1 \\ -1 & 1 & -2 & 1 \\ 1 & 0 & -1 & 2 \end{bmatrix}$$



Pelo Teorema de Gerschgorin, aplicado a linhas, concluímos que os valores próprios devem estar na reunião das bolas

$$B_1 = \bar{B}(-2, 3), B_2 = \bar{B}(10, 3), B_3 = \bar{B}(-2, 3), B_4 = B(2, 2).$$

Imediatamente vemos que irá haver duas componentes conexas, uma que será B_2, onde haverá apenas um valor próprio real, e a outra componente que será a reunião das três restantes bolas, o que se resume a $B_1 \cup B_4$, onde haverá três valores próprios (ver a primeira figura). Feita uma análise por colunas, obtemos as bolas

$$B_1' = \bar{B}(-2, 3), B_2' = \bar{B}(10, 2), B_3' = \bar{B}(-2, 2), B_4' = B(2, 4),$$

e concluímos que há apenas um valor próprio real em B_2' e três valores próprios em $B_1' \cup B_4'$. Intersectando a informação, concluímos que há um valor próprio real $\lambda_1 \in [8, 12]$ e três valores próprios $\lambda_2, \lambda_3, \lambda_4 \in \bar{B}(-2, 3) \cup \bar{B}(2, 2)$, onde este último domínio é obtido pela intersecção de $B_1 \cup B_4$ com $B_1' \cup B_4'$.
Podemos concluir que o valor próprio real λ_1 é um valor próprio dominante, porque $|\lambda_1| \geq 8$ e $|\lambda_2|, |\lambda_3|, |\lambda_4| \leq 5$.
Admitindo que a matriz é diagonalizável, estamos nas condições de convergência do método das potências. Partimos do vector inicial $u^{(0)} = (0, 1, 0, 0)$, escolha que foi orientada pelo vector próprio dominante associado à matriz diagonal. Obtemos

$$u^{(1)} = \sigma_0 \frac{Au^{(0)}}{||Au^{(0)}||_\infty} = + \frac{(1, 10, 1, 0)}{||(1, 10, 1, 0)||_\infty} = (\frac{1}{10}, 1, \frac{1}{10}, 0)$$

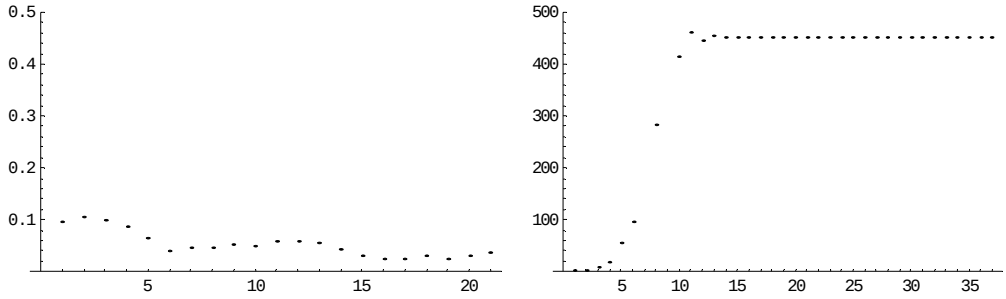
e sucessivamente $u^{(2)} = (0.0816, 1\ , 0.0714, 0), u^{(3)} = (0.0846, 1\ , 0.077, 0.0009)$. Em $u^{(3)}$ todos os dígitos apresentados já são correctos. Calculando $Au^{(3)} = (0.832, 9.835, 0.758, 0.0082)$, obtemos a aproximação $\lambda^{(3)} = 9.835$, que é dada pelo valor da segunda componente (pois é nessa que $u^{(3)}$ tem valor unitário... poderíamos obter outras aproximações dividindo a componente $Au_k^{(3)}$ por $u_k^{(3)}$, mas esta é a mais simples de obter). O valor exacto do valor próprio dominante é $\lambda_1 = 9.83703$. Como $\lambda_2 = 2.34741$, temos uma convergência linear com o factor $|\frac{\lambda_2}{\lambda_1}| = 0.238$, o que nos permitiria escrever

$$||v_1 - u^{(n)}||_\infty \leq 0.238^n C.$$

Como avaliar a constante C?
– Conhecendo o valor exacto do vector próprio, podemos avaliar os erros, e admitindo que

$$||v_1 - u^{(n)}||_\infty = |\frac{\lambda_2}{\lambda_1}|^n C_n,$$

colocamos num gráfico os valores de C_n obtidos (figura à esquerda),



e podemos concluir que os valores de C_n são inferiores a 0.1, *aproximando-se de* 0.03.
((Na figura da direita colocamos em evidência a dependência da constante C do valor α, componente segundo o vector próprio dominante. É considerada a matriz M usada na observação anterior e o vector inicial $u^{(0)} = (1, 1, \frac{1}{453})$. Com este vector inicial temos $\alpha = \frac{1}{453} \neq 0$, e o método converge para o vector próprio dominante. Como podemos ver no gráfico os valores de C_n tendem para um valor próximo de 450, *e para outros α o valor da constante seria próximo de $\frac{1}{\alpha}$. Como previsto na estimativa de erro, a constante depende directamente de $\frac{1}{\alpha}$, e isso é aqui verificado))*
– Como a priori não conhecemos os valores exactos, podemos no entanto obter informações avaliando o comportamento de $||u^{(n+1)} - u^{(n)}||_\infty$ para n suficientemente grande. Notamos que

$$||u^{(n+1)} - u^{(n)}||_\infty \leq ||u^{(n+1)} - v_1||_\infty + ||v_1 - u^{(n)}||_\infty,$$

e então

$$||u^{(n+1)} - u^{(n)}||_\infty \leq C_{n+1}|\frac{\lambda_2}{\lambda_1}|^{n+1} + C_n|\frac{\lambda_2}{\lambda_1}|^n.$$

Admitindo que $C_n \leq C$, então

$$||u^{(n+1)} - u^{(n)}||_\infty \leq 2C|\frac{\lambda_2}{\lambda_1}|^n.$$

Assim, é usual considerar a razão

$$K_n = \frac{||u^{(n+1)} - u^{(n)}||_\infty}{||u^{(n)} - u^{(n-1)}||_\infty} \approx \frac{2C|\frac{\lambda_2}{\lambda_1}|^n}{2C|\frac{\lambda_2}{\lambda_1}|^{n-1}} = |\frac{\lambda_2}{\lambda_1}|,$$

o que permite não apenas ter informação acerca da rapidez de convergência, mas também avaliar $|\lambda_2|$, já que $|\lambda_2| \sim K_n|\lambda_1|$. Com efeito, partindo dos valores $u^{(1)}, u^{(2)}, u^{(3)}$ calculados, poderíamos obter

$$K_2 = \frac{||u^{(3)} - u^{(2)}||_\infty}{||u^{(2)} - u^{(1)}||_\infty} = \frac{0.007328}{0.02857} = 0.2565,$$

o que não difere muito do valor 0.2386. *Como tinhamos obtido $\lambda^{(3)} = 9.835$, retiramos $|\lambda_2| \sim 0.2386 \times 9.835 = 2.3466$, o que é uma aproximação muito razoável, já que $\lambda_2 =$* 2.34741.

5.4 Método das iterações inversas

Este método é semelhante ao método das potências, mas baseia-se num conhecimento prévio da localização dos valores próprios. Continuamos a assumir uma diagonalização com os valores próprios que consideraremos reais (normalmente trabalharemos com matrizes hermiteanas). O método das potências apenas permitia aproximar o valor próprio dominante. Aqui consideramos qualquer um, mas precisamos de um conhecimento *a priori* sobre ele, que pode advir do Teorema de Gerschgorin ou de uma aproximação pelo método das potências.

Assim, é suposto termos λ como aproximação do valor próprio λ_m que pretendemos calcular. Logo,

$$Av = \lambda_m v \Leftrightarrow (A - \lambda I)v = (\lambda_m - \lambda)v \Leftrightarrow \frac{v}{\lambda_m - \lambda} = (A - \lambda I)^{-1} v,$$

e portanto se v é valor próprio de A, também é de $(A - \lambda I)^{-1}$. No entanto, os valores próprios são diferentes, λ_m é valor próprio de A e $\mu_m = \frac{1}{\lambda_m - \lambda}$ é valor próprio de $(A - \lambda I)^{-1}$ para o mesmo vector próprio!

A partir de uma iterada inicial $x^{(0)}$ (... com componente não nula no vector próprio), obtemos o *método das iterações inversas* (ou método de deflação de Wielandt)

$$x^{(n+1)} = \sigma_n \frac{(A - \lambda I)^{-1} x^{(n)}}{||(A - \lambda I)^{-1} x^{(n)}||_\infty}$$

em que σ_n é o sinal da componente de maior módulo de $(A - \lambda I)^{-1} x^{(n)}$. Reparamos, mais uma vez, tratar-se uma iteração do ponto fixo, pois como vimos,

$$Av = \lambda_m v \Leftrightarrow \frac{v}{\lambda_m - \lambda} = (A - \lambda I)^{-1} v,$$

e daqui obtemos $||(A - \lambda I)^{-1} v|| = \frac{||v||}{|\lambda_m - \lambda|}$. Assim:

$$\frac{v}{||v||} = \frac{\lambda_m - \lambda}{|\lambda_m - \lambda|} \frac{(A - \lambda I)^{-1} v}{||(A - \lambda I)^{-1} v||}$$

e mais uma vez substituimos $\frac{\lambda_m - \lambda}{|\lambda_m - \lambda|}$ pelo sinal da componente que determina o módulo (que designamos por σ).

A maneira para calcular de calcular as sucessivas iteradas baseia-se numa única factorização

$$A - \lambda I = LU,$$

seguida de sucessivas resoluções de sistemas (para cada n) :

$$LUw = x^{(n)} \Leftrightarrow \left\{ \begin{array}{c} Ly = x^{(n)} \\ Uw = y \end{array} \right.$$

o valor w é $(A - \lambda I)^{-1} x^{(n)}$. Assim, obtemos $x^{(n+1)} = \sigma_n w / ||w||_\infty$.

Reparamos que o método das iterações inversas dá-nos uma aproximação do vector próprio v, para calcularmos uma aproximação do valor próprio devemos fazer:

$$\lambda_m \sim \frac{[Ax^{(n)}]_i}{[x^{(n)}]_i}$$

De forma semelhante ao que conseguimos no método das potências podemos mostrar a convergência deste método desde que

$$L = \frac{|\lambda_m - \lambda|}{\min_{i \neq m} |\lambda_i - \lambda|} < 1.$$

Este resultado pode ser facilmente verificado se repararmos que isto corresponde a considerar

$$\max_{i \neq m} \frac{1}{|\lambda_i - \lambda|} < \frac{1}{|\lambda_m - \lambda|}$$

o que significa que $\frac{1}{|\lambda_m - \lambda|}$ é valor próprio dominante de $(A - \lambda I)^{-1}$. Depois basta aplicar o resultado obtido para o método das potências.

Exemplo 5.3 *Consideremos a matriz*

$$A = \begin{bmatrix} -15 & 0 & 1 & 1 \\ 2 & 10 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Pretendemos aproximar o valor próprio que se encontra no intervalo $[8, 12]$ *e escolhemos* $\lambda = 9$. *Ficamos com a factorização*

$$A - \lambda I = \begin{bmatrix} -24 & 0 & 1 & 1 \\ 2 & 1 & 0 & 0 \\ 1 & 1 & -8 & 1 \\ 1 & 1 & 1 & -8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/12 & 1 & 0 & 0 \\ -1/24 & 1 & 1 & 0 \\ -1/24 & 1 & \frac{-23}{193} & 1 \end{bmatrix} \begin{bmatrix} -24 & 0 & 1 & 1 \\ 0 & 1 & 1/12 & 1/12 \\ 0 & 0 & -\frac{193}{24} & \frac{23}{24} \\ 0 & 0 & 0 & -\frac{1530}{193} \end{bmatrix} = LU.$$

que iremos utilizar para calcular $(A - \lambda I)^{-1}$.
Escolhendo como iterada inicial $u^{(0)} = (0, 1, 0, 0)$. *A escolha deve-se às mesmas razões que as justificadas no exemplo anterior, atendendo a que agora procuramos o valor próprio que está na bola com centro em 10.*
Resolvendo $LUw = u^{(0)}$, *obtemos* $(0.0117, 0.9765, 0.1412, 0.1412)$, *portanto*

$$\begin{aligned} u^{(1)} &= \sigma_0 \frac{(A - \lambda I)^{-1} u^{(0)}}{||(A - \lambda I)^{-1} u^{(0)}||_\infty} = \frac{(0.0117, 0.9765, 0.1412, 0.1412)}{||(0.0117, 0.9765, 0.1412, 0.1412)||_\infty} \\ &= (0.0120482, 1., 0.144578, 0.144578) \end{aligned}$$

neste caso $Au^{(1)} = (0.1084, 10.024, 1.301, 1.301)$ *e portanto* $\lambda^{(1)} = 10.024...$
Continuando com as iterações, $u^{(2)} = (0.00975434, 1., 0.123194, 0.123194)$, *portanto* $\lambda^{(2)} = 10.0195...$, $\lambda^{(3)} = 10.0202$ *e aproximaríamos rapidamente o valor correcto* $\lambda = 10.0201$.

5.5 Métodos de Factorização

Um dos métodos mais utilizados para a determinação de valores próprios é o método QR, de Francis, que foi precedido por um outro método semelhante, devido a Rutishauser – o método LR, apresentado no final dos anos 50. A ideia principal destes métodos consiste em efectuar uma factorização da matriz num produto de matrizes mais simples, trocar a ordem do produto e obter uma nova matriz a que será aplicado o mesmo esquema!

Estes métodos baseiam-se na semelhança entre matrizes, pois escrevendo

$$B = P^{-1}AP,$$

a matriz A tem os mesmos valores próprios que B. Portanto, a ideia consiste em efectuar a iteração

$$A_{n+1} = P_n^{-1}A_nP_n,$$

começando com $A_0 = A$, se no limite tivermos uma matriz cujo cálculo dos valores próprios é simples (por exemplo, uma matriz triangular) então o problema fica simplificado, ou resolvido.

Alternativamente, estes métodos podem ser encarados como resultantes de uma factorização das matrizes. Assim, se for possível efectuar uma factorização do tipo $A_n = X_nY_n$, em que X_n é invertível, bastará considerar $A_{n+1} = Y_nX_n$ para termos

$$A_{n+1} = X_n^{-1}A_nX_n,$$

porque $Y_n = X_n^{-1}A_n$.

5.5.1 Método LR

No caso do método LR, de *Rutishauser*, efectuamos uma factorização $A = LU$ que por tradição é designada LR (left-right ao invés de lower-upper). Assim, começando com $A_0 = A$, e tendo obtido

$$A_n = L_nU_n$$

definimos a nova iterada como sendo

$$A_{n+1} = U_nL_n,$$

o que também significa que consideramos a iteração $A_{n+1} = L_n^{-1}A_nL_n$. Reparamos que a matriz A_{n+1} é semelhante a A_n e por isso os valores próprios são os mesmos, subsequentemente os mesmos que os de $A_0 = A$.

Se o método convergir, é suposto que a sucessão de matrizes A_n tenda para uma matriz triangular superior, cuja diagonal irá conter os valores próprios. No entanto, não é fácil obter condições de convergência para este método, podendo ser bastante instável. Sabe-se (cf. [24]) que se A for simétrica e definida positiva há convergência.

5.5.2 Método QR

• O método QR, de *Francis,* é baseado numa factorização menos conhecida

$$A = QR$$

em que Q é uma matriz unitária (ou seja, $QQ^* = Q^*Q = I$) e R uma matriz triangular superior.

Proposição 5.3 *A factorização $A = QR$ é única, a menos de produto por uma matriz diagonal, cujas entradas têm módulo 1.*

Demonstração:

Supondo que $A = Q_1R_1 = Q_2R_2$, então $R_1R_2^{-1} = Q_1^*Q_2$, o que significa que a matriz triangular superior $R_1R_2^{-1}$ seria uma matriz ortogonal (porque $Q_1^*Q_2$ é). No entanto, as únicas matrizes nestas condições são matrizes diagonais, logo $R_1R_2^{-1} = D$, ou seja $R_1 = DR_2$ e $Q_1^*Q_2 = D$, ou seja $Q_2 = Q_1D$. Verifica-se que essa diagonal verifica $DD^* = Q_1^*Q_2Q_2^*Q_1 = I$, ou seja $|d_{ii}| = 1$. ■

• *Construção da factorização QR através de matrizes de Householder.*

Uma matriz de Householder é uma matriz do tipo

$$H = I - 2vv^*$$

em que $v : ||v||_2 = 1$, ou seja $v^*v = 1$ (note-se que v^*v é uma matriz 1×1, identificada com um número, mas vv^* já é uma matriz $N \times N$).

As matrizes de Householder são unitárias porque $HH^* = H^*H = (I-2vv^*)(I-2vv^*) = I - 4vv^* + 4vv^*vv^* = I$.

Podemos considerar vectores $v^{(k)} = (0, .., 0, v_k, ..., v_N)$, que irão definir matrizes de Householder H_k. É possível efectuar a decomposição QR :

$$R = H_{N-1}...H_1A,$$
$$Q^* = H_{N-1}...H_1,$$

já que é fácil verificar que $QR = H_1^*...H_{N-1}^*H_{N-1}...H_1A = A$, faltando apenas ver que $H_{N-1}...H_1A$ é triangular superior calculando $v^{(k)}$ (cf.[1]).

• O *método QR* consiste em começar com $A_0 = A$, e tendo factorizado

$$A_n = Q_nR_n,$$

definir uma nova iterada

$$A_{n+1} = R_nQ_n,$$

ou seja, $A_{n+1} = Q_n^*A_nQ_n$ que é uma matriz semelhante a A_n.

Teorema 5.2 *(Francis) Se a matriz A for invertível e os seus valores próprios tiverem módulos diferentes, $|\lambda_1| > ... > |\lambda_N| > 0$, a matriz é diagonalizável, ou seja, $A = P^{-1}DP$. Se P admitir uma factorização $P = LU$, a sucessão de matrizes (A_n) converge para uma matriz triangular superior cujos valores da diagonal serão os valores próprios de A. Os vectores próprios associados encontram-se na matriz unitária Q.*

Demonstração: Ver, por exemplo, [7]. ■

No caso mais geral, pode convergir para uma matriz quase triangular (por blocos), cujos valores próprios são razovelmente fáceis de calcular. O método tem ainda normalmente a particularidade de apresentar os valores próprios ordenados, estando na primeira linha o maior e na última o mais pequeno. A rapidez de convergência para zero dos elementos não diagonais depende da relação $|\frac{\lambda_k}{\lambda_{k+1}}|$, o que pode constituir um obstáculo à rapidez do método, quando alguns valores próprios têm módulos semelhantes. Por isso é usada uma técnica de aceleração de convergência que veremos mais à frente.

Exemplo 5.4 *Consideramos a factorização QR da matriz A,*

$$A = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 3 & 2 \\ 2 & -2 & -4 \end{bmatrix} = \overbrace{\begin{bmatrix} \sqrt{2/3} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{2} & -1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & -1/\sqrt{3} \end{bmatrix}}^{Q} \overbrace{\begin{bmatrix} 2\sqrt{6} & \sqrt{3/2} & 0 \\ 0 & 5/\sqrt{2} & 3\sqrt{2} \\ 0 & 0 & \sqrt{3} \end{bmatrix}}^{R},$$

não especificando os cálculos inerentes... A partir deste ponto calculamos $A_1 = RQ$, e obtemos uma nova matriz cuja diagonal é $\{4.5, -0.5, -1\}$, estes são os primeiros valores que aproximam os valores próprios de A. Voltamos a efectuar a decomposição $A_1 = Q_1R_1$... que por razões óbvias não será aqui colocada. Calculando $A_2 = R_1Q_1$, obtemos na diagonal os valores $\{4.52, -3.52, 2\}$. Procedendo de forma semelhante nas iteradas seguintes, obtemos ao fim de 7 iterações, na diagonal de A_7, os valores $\{5.07..., -3.69..., 1.62...\}$, que não estão muito longe dos valores próprios correctos $\{5, -3.64..., 1.64...\}$. A matriz A_7 já é próxima de uma matriz triangular superior,

$$A_7 = \begin{bmatrix} 5.07467 & -1.89079 & 0.853412 \\ 0.343404 & -3.69541 & -3.5415 \\ 0.001455 & -0.0368519 & 1.62073 \end{bmatrix}.$$

O valor absoluto do maior elemento da subdiagonal determina, normalmente, um majorante do erro da aproximação.

Observação *(método de Jacobi).*

Outra possibilidade de obter a factorização QR é usar matrizes de rotações no plano ao invés de matrizes de Householder, ideia que também é usada no método de Jacobi. O método de Jacobi é válido para matrizes reais simétricas e baseia-se na utilização de

matrizes de rotação

$$U = \begin{bmatrix} 1 & 0 & & \cdots & 0 \\ 0 & \cos(\theta) & -\sin(\theta) & & \vdots \\ & \sin(\theta) & \cos(\theta) & & \\ \vdots & & & 1 & 0 \\ & & & \ddots & \\ 0 & \cdots & & 0 & 1 \end{bmatrix},$$

mas não falaremos dele aqui (ver, por exemplo, [18]).

Ainda uma outra possibilidade para a efectuar a factorização QR é considerar o processo de ortonormalização de Gram-Schmidt, que no entanto é instável numericamente, devido ao cancelamento subtractivo.

5.5.3 Método QR com deslocamento

Os métodos de factorização são computacionalmente dispendiosos em termos de tempo (aproximadamente $\frac{2}{3}N^3$ operações por iteração) e como já referimos a sua convergência pode ser lenta. Uma possibilidade para acelerar a convergência destes métodos é utilizar uma técnica de deslocamento (ou *shift*), reparando que considerando

$$\tilde{A} = A - \alpha I$$

se uma matriz B for semelhante a $\tilde{A}$ então $\tilde{B} = B + \alpha I$ será semelhante a A, porque

$$B = P^{-1}\tilde{A}P = P^{-1}(A - \alpha I)P = P^{-1}AP - \alpha I.$$

Assim, para o método QR, podemos estruturar os passos usando um deslocamento α_n diferente, em cada passo, de forma a que efectuamos primeiro a decomposição QR da matriz $A_n - \alpha_n I$ e depois trocamos a ordem somando $\alpha_n I$. Ou seja,

$$\begin{aligned} A_n - \alpha_n I &= Q_n R_n, \\ A_{n+1} &= R_n Q_n + \alpha_n I, \end{aligned}$$

ficando com

$$A_{n+1} = Q_n^*(A_n - \alpha_n I)Q_n + \alpha_n I = Q_n^* A_n Q_n,$$

e desta forma, A_{n+1} continua a ser uma matriz semelhante a A_n. A escolha do deslocamento α_n é discutida em [29] e uma das possibilidades é considerar α_n como sendo o elemento de menor módulo da diagonal (normalmente o último).

Observação 1: Apesar de ser o método mais utilizado para o cálculo de valores próprios, o método QR com *shift* tem resistido à demonstração da sua convergência no caso mais geral (cf.[7]).

Observação 2: O *Mathematica* tem implementadas as rotinas Eigenvalues e Eigenvectors, que permitem o cálculo de valores e vectores próprios de matrizes, usando um método QR com *shift.* A factorização QR pode ser obtida usando a rotina QRDecomposition.(o resultado é uma lista com a transposta da matriz Q e com a matriz R).

5.6 Condicionamento do cálculo de valores próprios

Apresentamos agora um resultado relativo ao condicionamento do cálculo de valores próprios.

Teorema 5.3 *(Bauer-Fike). Seja A uma matriz hermitiana. No caso de $\tilde{A}$ ser uma aproximação (hermitiana) de A, temos o resultado*

$$\forall j \, \exists i : |\lambda_i - \tilde{\lambda}_j| \leq ||A - \tilde{A}||_2 \tag{5.1}$$

em que λ_i são os valores próprios de A e $\tilde{\lambda}_j$ os de $\tilde{A}$.

No caso mais geral, em que há a matriz tem forma canónica de Jordan diagonal, $A = P^{-1}DP$ (com $D = diag(\lambda_1, ..., \lambda_N)$), temos

$$\forall j \, \exists i : |\lambda_i - \tilde{\lambda}_j| \leq cond_\infty(P)||A - \tilde{A}||_\infty. \tag{5.2}$$

(o que também é válido para algumas outras normas, como $||.||_1, ||.||_2$).

Demonstração:

i) Começamos por ver que o resultado sai facilmente para a norma $||.||_\infty$ (ou mesmo para $||.||_1$).

Seja $B = P(A - \tilde{A})P^{-1}$, temos $B = D - C$ em que $C = P\tilde{A}P^{-1}$ tem os valores próprios de $\tilde{A}$. Pelo teorema de Gerschgorin, aplicado a $C = D - B$, sabemos que dado um valor próprio $\tilde{\lambda}_j$ de C existe uma linha i :

$$|\lambda_i - b_{ii} - \tilde{\lambda}_j| \leq \sum_{k \neq i} |b_{ik}|,$$

e portanto

$$|\lambda_i - \tilde{\lambda}_j| \leq \sum_k |b_{ik}| \leq ||B||_\infty \leq ||P||_\infty ||A - \tilde{A}||_\infty ||P^{-1}||_\infty.$$

ii) Para mostrar que é válido para a norma $||.||_2$, vemos que

$$\min_{i=1,...,N} |\lambda_i - \tilde{\lambda}| \leq \text{cond}_2(P)||A - \tilde{A}||_2,$$

para qualquer valor próprio $\tilde{\lambda}$, e a partir daqui podemos aplicar de novo o teorema de Gerschgorin para concluir o teorema.

Suponhamos que $\tilde{\lambda} \neq \lambda_i$ para qualquer i (senão seria trivial, pois o mínimo seria zero) e seja $\tilde{v}$ um vector próprio de $\tilde{A}$.

Como $\tilde{A}\,\tilde{v} = \tilde{\lambda}\tilde{v}$,

$$(\tilde{\lambda}I - A)\tilde{v} = (\tilde{A} - A)\tilde{v} \tag{5.3}$$

e substituindo A, temos $(\tilde{\lambda}I - A)\tilde{v} = (\tilde{\lambda}I - P^{-1}DP)\tilde{v} = P^{-1}(\tilde{\lambda}I - D)P\tilde{v}$ o que implica, por (5.3), que

$$(\tilde{\lambda}I - D)P\tilde{v} = P(\tilde{A} - A)P^{-1}P\tilde{v}.$$

Como $\tilde{\lambda} \neq \lambda_i$, a matriz diagonal $\tilde{\lambda}I - D$ tem inversa, e obtemos

$$P\tilde{v} = (\tilde{\lambda}I - D)^{-1}P(\tilde{A} - A)P^{-1}P\tilde{v}.$$

Notando que $||(\tilde{\lambda}I - D)^{-1}||_2 = \rho((\tilde{\lambda}I - D)^{-1}) = \frac{1}{\min|\tilde{\lambda}-\lambda_i|}$ (o que também é válido para outras normas ditas 'monótonas'), temos

$$||P\tilde{v}||_2 \leq \frac{1}{\min|\tilde{\lambda} - \lambda_i|} ||P(\tilde{A} - A)P^{-1}||_2 \, ||P\tilde{v}||_2$$

o que origina

$$\min_{i=1,\ldots,N} |\lambda_i - \tilde{\lambda}| \leq ||P||_2||P^{-1}||_2||A - \tilde{A}||_2.$$

No caso de matrizes hermitianas, basta referir que pela decomposição na forma normal de Schur podemos encontrar matrizes P unitárias tal que $A = P^*DP$, pelo que $||P||_2 = ||P^*||_2 = 1$. ■

Observação:
A propriedade que provámos traduz também o bom condicionamento no cálculo de valores próprios para as matrizes hermitianas. Para outro tipo de matrizes, o cálculo dos valores próprios poderá ser um problema mal condicionado, dependendo do número de condição da matriz P.

Como a estimativa do número de condição de P não é normalmente possível (se P fosse conhecido também seriam os seus valores próprios), apenas temos a informação da possibilidade de ocorrerem problemas de condicionamento no cálculo dos valores próprios.

5.7 Cálculo de raízes polinomiais

Terminamos este capítulo referindo que um excelente processo de obter resultados acerca das raízes de polinómios é a utilização da noção de matriz companheira de um polinómio.

Definição 5.1 *Dizemos que $\mathcal{C}$ é a matriz companheira do polinómio $p(x) = a_0 + a_1x + \ldots + a_{N-1}x^{N-1} + x^N$, se*

$$\mathcal{C} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{N-2} & -a_{N-1} \end{bmatrix}$$

notando que o polinómio característico de $\mathcal{C}$ é exactamente p.

Esta noção pode ser aplicada para a localização das raízes de polinómios através do teorema de Gerschgorin (ver exercício 2, no final do capítulo) ou mesmo para aproximá-las usando um qualquer método de valores próprios, já que identificar os valores próprios de

C é equivalente a determinar as raízes de p. Deste facto retiramos que a *determinação de valores próprios é um problema teoricamente equivalente à resolução de equações algébricas.*

Exemplo 5.5 *Tomemos como exemplo o método das potências aplicado a $\mathcal{C}$. Executar a iteração*

$$x^{(n+1)} = \mathcal{C}\, x^{(n)}$$

é equivalente a considerar

$$\begin{cases} x_i^{(n+1)} = x_{i+1}^{(n)} & \text{se } i = 1, ..., N-1, \\ x_N^{(n+1)} = -a_0 x_1^{(n)} - ... - a_{N-1} x_N^{(n)} & \text{caso } i = N. \end{cases}$$

Reparamos assim que $x_1^{(n)} = x_2^{(n-1)} = ... = x_N^{(n-N+1)}$, $x_2^{(n)} = ... = x_N^{(n-N+2)}$, etc... de um modo geral, $x_i^{(n)} = x_N^{(n-N+i)}$, o que corresponde a substituir valores na iterada n por valores em iteradas anteriores.

Ora, designando $y_k = x_N^{(k-N+1)}$, obtemos $x_i^{(n)} = y_{n+i-1}$, pelo que o sistema anterior reduz-se à equação às diferenças

$$y_{n+N} = -a_0 y_n - ... - a_{N-1} y_{n+N-1}.$$

A mesma equação às diferenças que encontrámos no método de Bernoulli.

Para concluirmos que o método de Bernoulli aparece como um caso particular do método das potências*, reparamos que no caso do método das potências consideramos como aproximação do valor próprio dominante*[7]*:*

$$\lambda^{(n)} = \frac{[\mathcal{C}\, x^{(n)}]_1}{x_1^{(n)}} = \frac{x_1^{(n+1)}}{x_1^{(n)}} = \frac{y_{n+1}}{y_n},$$

ou seja, a mesma aproximação que consideramos no método de Bernoulli para a raiz dominante!

Outros métodos para valores próprios levam a outras aproximações, não havendo necessariamente um método específico para polinómios que lhes corresponda, como neste caso aconteceu com o método de Bernoulli.

[7] Ver também a nota de rodapé anterior, considerando o método das potências sem a normalização sucessiva!

Consideramos aqui a primeira componente, mas para qualquer componente j obteríamos

$$\lambda^{(n)} = \frac{x_j^{(n+1)}}{x_j^{(n)}} = \frac{y_{n+j}}{y_{n+j-1}}$$

o que corresponde ao mesmo resultado.

Observação:
Como curiosidade reparamos que a matriz inversa da matriz companheira é

$$\mathcal{C}^{-1} = \begin{bmatrix} -\frac{a_1}{a_0} & -\frac{a_1}{a_0} & \cdots & -\frac{a_{N-1}}{a_0} & -\frac{1}{a_0} \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

que tem associada como polinómio característico $q(y) = \frac{1}{a_0} + \frac{a_{N-1}}{a_0}y + ... + \frac{a_1}{a_0}y^{N-1} + y^N$ cujas raízes são as inversas de $p(x)$, como vimos num exercício do Capítulo 2 (basta tomar $y = 1/x$). Isto é perfeitamente natural, já que é claro que os valores próprios da matriz inversa são os inversos da original.

5.8 Exercícios

1. (Método de Krylov) Considere o seguinte método, baseado na aplicação do teorema de Hamilton-Cayley, para encontrar o polinómio característico de uma matriz A de dimensão N :

- Calcular A^k, para $k = 2, ..., N$
- Determinar os coeficientes α_i tais que $\alpha_0 I + \alpha_1 A + ... + \alpha_{N-1}A^{N-1} + A^N = 0$.

a) Indique uma estimativa do número de operações $(*, /)$ necessárias a esse cálculo.

b) Use este método para determinar a equação característica de uma matriz 2×2.

c) Ao invés de calcular A^k, considere um vector inicial $x^{(0)}$ e defina $x^{(k)} = Ax^{(k-1)}$. Apresente um processo equivalente para determinar o polinómio característico. Comente quanto ao número de operações e quanto à solubilidade do sistema.

2. Considere a matriz companheira do polinómio com coeficientes reais $p(x) = a_0 + a_1x + ... + a_{n-1}x^{n-1} + x^n$.

a) Mostre que se $|a_{n-1}| > 1 + M$, com $M = \max\{|a_0|, |a_1| + 1, ..., |a_{n-2}| + 1\}$ então existe uma e uma só raiz real dominante em $[-a_{n-1} - 1, -a_{n-1} + 1]$, e que as restantes se encontram na bola $\{|z| \leq M\}$.

b) Considere $p(x) = 2 - 6x^2 + 4x^3 - 16x^4 + 2x^5$.

Localize as raízes dominante num intervalo de comprimento 2 e as restantes numa bola de raio 1.

Determine aproximadamente a raiz dominante usando duas iterações do método das potências.

3. Seja A uma matriz real $N \times N$, que verifica:

$$|a_{ii} - a_{jj}| > r_i + r_j, \forall i, j = 1, ..., N \ (i \neq j)$$

em que

$$r_k = -|a_{kk}| + \sum_{j=1}^{N} |a_{kj}|$$

Mostre que os valores próprios da matriz são reais.

4. Considere uma matriz $A \in \mathbb{C}^N \times \mathbb{C}^N$ e várias sucessões $\mu^{(k)} \in l^1$. Supondo que

$$|a_{ii}| > ||\mu^{(i)}||_1 \ \ \forall i = 1, \ldots, N$$

$$|a_{ij}| \leq |\mu_j^{(i)}| \ \ \forall i, j = 1, \ldots, N, \ \ (i \neq j)$$

a) Mostre que a matriz A é invertível.

b) Mostre que se A for hermitiana e tiver a diagonal positiva, então é definida positiva e o raio espectral verifica

$$\rho(A) \leq \max_{i=1,\ldots,N} \left(|a_{ii}| + ||\mu^{(i)}||_1 \right).$$

c) Mostre que é possível resolver o sistema $Ax = b$, para qualquer $b \in I\!R^N$, usando o método de Jacobi, e que se verifica:

$$||x - x^{(n)}||_\infty \leq \frac{L^n}{1-L} \frac{||b||_\infty}{K}$$

considerando $x^{(0)} = 0$, com

$$L = 1 - \min_{i=1,\ldots,n} \left(\frac{|\mu_i^{(i)}|}{||\mu^{(i)}||_1} \right), \ \text{ e com } K = \min_{i=1,\ldots,n} ||\mu^{(i)}||_1.$$

5. Considere a matriz

$$A = \begin{bmatrix} 6 & 0 & 1 \\ 2 & 1 & 0 \\ 2 & -1 & -1 \end{bmatrix}.$$

a) Aplicando o T. de Gerschgorin determine um domínio em $\mathbb{C}$ onde se encontram os valores próprios de A.

b) Conclua que existe um valor próprio dominante para A, e determine uma aproximação utilizando o método das potências.

c) Diga qual o raio espectral da matriz $A/10$? O que pode concluir acerca da convergência do seguinte método:

6. Considere a matriz

$$\begin{bmatrix} -1+i & 1 & 1 \\ 1 & -1-i & 1 \\ 1 & 0 & 3+4i \end{bmatrix}$$

a) Indique um domínio do plano complexo onde se situam os valores próprios.

b) Determine um majorante para o módulo do determinante da matriz.

c) Entre que valores se pode situar o raio espectral da matriz? A matriz é invertível?

7. Considere a matriz

$$\begin{bmatrix} 8 & 1 & -1 \\ 1 & -3 & 1 \\ 0 & 1/2 & 1 \end{bmatrix}$$

a) Justifique que todos os valores próprios da matriz são reais, e indique intervalos que os contenham.

b) Verifique que a matriz possui um valor próprio dominante e aproxime-o considerando três iteradas do método das potências, usando como vector inicial $v^{(0)} = (1, 0, 0)$.

8. Considere a matriz

$$A = \begin{bmatrix} 10 & 3 - 2\cos(b) & \cos(b) \\ 1 & 25 & 5\sin(a) \\ 1 & 5\sin(a) + \sin(b) & 50 \end{bmatrix}$$

a) Localize os valores próprios de A usando o teorema de Gerschgorin.

b) Indique os valores de b para os quais podemos obter uma decomposição $A = LL^T$, em que L é uma matriz triangular inferior real.

c) Para que valores de $h \in \mathbb{R}^3$ é possível utilizar o método de Jacobi para resolver um sistema $Ax = h$? Indique uma estimativa de erro para $||e^{(n)}||_\infty$ em função de $||h||_\infty$, sabendo que $x^{(0)} = 0$.

9. Considere uma matriz $A \in \mathbb{C}^N \times \mathbb{C}^N$ e várias sucessões $\mu^{(k)} \in l^1$. Supondo que

$$|a_{ii}| > ||\mu^{(i)}||_1 \ \ \forall i = 1, \ldots, N$$

$$|a_{ij}| \leq |\mu_j^{(i)}| \ \ \forall i, j = 1, \ldots, N, \ \ (i \neq j)$$

a) Mostre que a matriz A é invertível.

b) Mostre que se A for hermitiana e tiver a diagonal positiva, então é definida positiva e o raio espectral verifica

$$\rho(A) \leq \max_{i=1,\ldots,N} \left(|a_{ii}| + ||\mu^{(i)}||_1 \right).$$

c) Mostre que é possível resolver o sistema $Ax = b$, para qualquer $b \in \mathbb{R}^N$, usando o método de Jacobi, e que se verifica:

$$||x - x^{(n)}||_\infty \leq \frac{L^n}{1 - L} \frac{||b||_\infty}{K}$$

considerando $x^{(0)} = 0$, com

$$L = 1 - \min_{i=1,\ldots,n} \left(\frac{|\mu_i^{(i)}|}{||\mu^{(i)}||_1} \right), \ \text{ e com } K = \min_{i=1,\ldots,n} ||\mu^{(i)}||_1.$$

10. Suponha que obteve

$$A = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 0 & -1 \end{bmatrix}.$$

a) Calcule a primeira iteração pelo método QR.

b) Compare as aproximações dos valores próprios com os valores próprios correctos.

11. Considere

$$A = \begin{bmatrix} a\cos(\theta) & -a\sin(\theta) \\ a\sin(\theta) & a\cos(\theta) \end{bmatrix}.$$

Qual a factorização QR de A? Como se processa a iteração do método QR neste caso? Calcule os valores próprios de A e justifique.

12. Mostre que se os elementos de uma matriz forem números racionais então os valores próprios dessa matriz não podem ser números transcendentes.

Capítulo 6

Introdução a métodos de optimização

Um dos problemas correntes em aplicações é encontrar valores que minimizem[1] determinadas quantidades. O cálculo em $\mathbb{R}$ mostra que se uma função for C^1, numa vizinhança do ponto de mínimo, isso implica que a derivada se anule. Isto sugere que se procure o ponto de mínimo como um zero da derivada... mas esta ideia nem sempre admite generalizações nem facilidades práticas. A apresentação que aqui faremos é bastante superficial e de carácter introdutório, para uma consulta mais aprofundada existem bastantes referências, por exemplo [7], [22].

Começamos por observar que os métodos de minimização podem ser utilizados para resolver equações $f(x) = 0$, bastando considerar a minimização de $||f(x)||$, ou mais frequentemente $||f(x)||_2^2$, já que nesse caso os pontos de mínimo irão coincidir com as raízes!

Iremos começar por falar de um método de minimização simples, o método dos mínimos quadrados, que também pode ser encarado como um método para a aproximação de uma função, ou de um conjunto de pontos por uma função. Devido a esse facto, e ao facto ser substancialmente diferente dos restantes, muitas vezes não é enquadrado no contexto da optimização.

6.1 Método dos mínimos quadrados

A minimização através do método dos mínimos quadrados, que pode ser visto como um caso particular de encontrar a distância mínima a um conjunto convexo abstracto. No entanto, no caso que iremos abordar, esse conjunto convexo é simplesmente um subespaço vectorial de dimensão finita. Um caso particular de método dos mínimos quadrados é a regressão linear (e é o mais simples), em que a partir de vários dados numéricos (normalmente obtidos experimentalmente) se pretende obter a recta que melhor se aproxima desse conjunto de pontos. Assim, tenta-se obter uma relação linear entre os diferentes valores. No entanto, nem sempre é conveniente obter uma aproximação desses valores por rectas, podemos pensar noutro tipo de funções... polinómios de grau superior, ou várias outras... Também pode ser

[1] Ou que maximizem... como é óbvio, tratam-se de problemas equivalentes!

necessário aproximar uma função, e não apenas um conjunto de pontos, por outras funções mais simples (ou de características convenientes para o problema que queremos resolver).

Vamos resolver estas questões de forma simples – através do método dos mínimos quadrados. Podemos encarar o método dos mínimos quadrados como um método com restrições, porque o ponto de mínimo que pretendemos determinar está num subespaço (e vai ser a projecção ortogonal da função dada, no subespaço formado pela base de "funções aproximadoras"). No entanto, reparamos que se considerarmos o subespaço como o próprio espaço, já não temos restrições... é isso que faremos!

• Tudo se vai resumir a minimizar uma distância num espaço de Banach em que está definido um produto interno, ou seja um espaço de Hilbert. Neste caso essa distância é dada através do produto interno

$$d(x,y) = ||x-y|| = \sqrt{\langle x-y,\, x-y\rangle}.$$

O objectivo é minimizar a distância de um ponto w a um subespaço vectorial S, ou o que é equivalente, minimizar o seu quadrado $||x-w||^2$, para $x \in S$. Podemos fazer isso determinando quando a derivada (de Fréchet) se anula.

Proposição 6.1 *Seja H um espaço de Hilbert e $w \in H$. A derivada de Fréchet de $J(x) = \frac{1}{2}||x-w||^2$ é*

$$J'_x h = \langle x-w,\, h\rangle .$$

Demonstração:
Note-se que $J : S \to \mathbb{R}$, e temos

$$\begin{aligned} J(x+h) - J(x) &= \frac{1}{2}\langle x+h-w, x+h-w\rangle - \frac{1}{2}\langle x-w,\, x-w\rangle \\ &= \frac{1}{2}\langle h, x+h-w\rangle + \frac{1}{2}\langle x+h-w,\, h\rangle \\ &= \langle x-w,\, h\rangle + ||h||^2. \end{aligned}$$

Portanto a derivada de Fréchet de J em x é o operador linear $h \mapsto \langle x-w,\, h\rangle$. ■

Proposição 6.2 *x é ponto de mínimo da função J (definida na proposição anterior) se e só se*

$$J'_x = 0,$$

ou seja, se $\langle x-w,\, h\rangle = 0$, para qualquer $h \in H$.

Demonstração:
Se $J'_x = 0$ então

$$J(x+h) - J(x) = J'_x h + ||h||^2 = ||h||^2 > 0,\ \forall h \neq 0,$$

e portanto x é ponto de mínimo estrito. Reciprocamente, se

$$J(x+h) - J(x) > 0,\ \ \forall h \neq 0,$$

então $J'_x h + ||h||^2 > 0$, e para $h = \varepsilon v$, com $||v|| = 1, \varepsilon > 0$, temos

$$0 < J'_x v + \varepsilon.$$

Fazendo $\varepsilon \to 0$, ficamos com $0 \leq J'_x v$, para qualquer $||v|| = 1$, o que significa que será válido para $-v$, logo $0 \leq J'_x(-v) = -J'_x v$. Conclui-se assim que $J'_x v = 0$, para qualquer $||v|| = 1$, e portanto $J'_x h = ||h|| J'_x(\frac{h}{||h||}) = 0$, para qualquer $h \neq 0$ (para $h = 0$, é trivial). ■

• Portanto, dado $w \notin S$ o problema a resolver passa a ser determinar $x \in S$:

$$\langle x - w,\, h \rangle = 0, \forall h \in S,$$

o que corresponde a encontrar a *projecção ortogonal* de w a S.

Como S é um subespaço, basta verificar para cada elemento da base desse subespaço $\{v_1, ..., v_N, ...\}$:

$$\langle x,\, v_i \rangle = \langle w,\, v_i \rangle\,, \quad \text{para } i = 1, ..., N, ...$$

Por outro lado escrevendo $x = \sum_{j \geq 1} x_j v_j$, obtemos

$$\sum_{j \geq 1} x_j\, \langle v_j,\, v_i \rangle = \langle w,\, v_i \rangle\,, \quad \text{para } i = 1, ..., N, ...$$

No caso de estarmos num espaço com dimensão finita N, corresponde a um sistema $\mathbf{V}x = \mathbf{w}$:

$$\begin{bmatrix} < v_1, v_1 > & \dots & < v_1, v_N > \\ \vdots & \ddots & \vdots \\ < v_N, v_1 > & \dots & < v_N, v_N > \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} < w, v_1 > \\ \vdots \\ < w, v_1 > \end{bmatrix}$$

denominado *sistema normal*.

Por construção, a solução x do sistema $\mathbf{V}x = \mathbf{w}$, será o ponto que minimiza $J(y) = \frac{1}{2}||y - w||^2$.

Proposição 6.3 *A matriz do sistema normal* $\mathbf{V}$ *é simétrica e definida positiva (supondo que* $v_1, ..., v_N$ *são linearmente independentes).*

Demonstração:

A simetria resulta do produto interno (no caso de números complexos a matriz será hermitiana). Para verificar que é definida positiva, consideramos

$$(\mathbf{V}x)_i = \sum_{j=1}^{N} \langle v_j,\, v_i \rangle\, x_j = \left\langle v_i, \sum_{j=1}^{N} x_j v_j \right\rangle = \langle v_i, x \rangle\,,$$

e portanto

$$x^\top \mathbf{V} x = \sum_{i=1}^{N} x_i\, \langle v_i, x \rangle = \langle x, x \rangle = ||x||^2 > 0, \text{ se } x \neq 0. \ ■$$

Observação:

(i) Reparamos que se a base for ortogonal obtemos uma matriz diagonal e a solução é explicitamente dada pelas projecções sobre cada vector da base, i.e:

$$x_k = P_{v_k} w = \frac{\langle w, v_k \rangle}{||v_k||^2}.$$

Um exemplo em que isso acontece é quando se consideram polinómios ortogonais.

(ii) Repare-se que, como a matriz $\mathbf{V}$ do sistema normal é simétrica e definida positiva ela define um produto interno.

(iii) A resolução do sistema normal $\mathbf{V}x = \mathbf{w}$ será equivalente a minimizar $J(x) = \frac{1}{2}x^\top \mathbf{V}x - x^\top \mathbf{w}$, e por isso os métodos de minimização do tipo gradiente ou gradiente conjugado, tal como o método de Cholesky, são adequados para a resolução do sistema normal.

6.1.1 Aproximação no caso discreto

No caso de uma aproximação em que se pretende determinar a função da forma $x = x_1 v_1 + ... + x_N v_N$ que melhor aproxima os valores $w(t_i)$ obtidos no conjunto de pontos $t_1, ..., t_M$, consideramos o produto interno em $\mathbb{R}^M$

$$< u, v >= \sum_{k=1}^{M} u(t_k) v(t_k).$$

que está associado à distância que nos interessa minimizar: $||x-w||_2^2 = \sum_{k=1}^{M} (x(t_k) - w(t_k))^2$.

No caso discreto, a matriz do sistema normal pode ser escrita como o produto de uma matriz não quadrada pela sua transposta $\mathbf{S}$. A matriz $\mathbf{S}$ é constituída pelos valores das N funções base nos M pontos, ou seja,

$$\mathbf{S} = \begin{bmatrix} v_1(t_1) & \dots & v_N(t_1) \\ \vdots & \ddots & \vdots \\ v_1(t_M) & \dots & v_N(t_M) \end{bmatrix}_{M \times N}, \text{ e temos } \mathbf{V} = \mathbf{S}^\top \mathbf{S}.$$

Nota: No caso complexo, $\mathbf{V} = \mathbf{S}^* \mathbf{S}$.

É também claro que os valores singulares da matriz $\mathbf{S}$ são as raízes quadradas dos valores próprios de $\mathbf{V}$.

Exercício 6.1 *Considere $w(t_i)$ valores obtidos para os pontos $t_1, ..., t_n$. Mostre que se pretendermos uma aproximação polinomial com a base $v_k(t) = t^k$, obtemos $v = \alpha_0 + \alpha_1 t + ... + \alpha_p t^p$, em que*

$$\begin{bmatrix} n+1 & \dots & \sum_{i=1}^n t_i^p \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n t_i^p & \dots & \sum_{i=1}^n t_i^{2p} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n w(t_i) \\ \vdots \\ \sum_{i=1}^n t_i^p w(t_i) \end{bmatrix}$$

e obtenha as fórmulas para a regressão linear (o caso $p = 1$).

Observação: Para o caso discreto, o *Mathematica* tem implementada a rotina `Fit`, que permite introduzir uma lista com os pontos a aproximar e uma outra lista com as funções base consideradas.

6.1.2 Aproximação no caso contínuo

No caso de uma aproximação em que se pretende determinar a função da forma $x = x_1 v_1 + ... + x_N v_N$ que melhor aproxima uma função $w(t)$ definida num intervalo $[a, b]$,consideramos o produto interno em $L^2([a, b])$

$$< u, v >_{L^2(a,b)} = \int_a^b u(t)v(t)\, dt.$$

que está associado à distância que nos interessa minimizar: $||x - w||^2_{L^2(a,b)} = \int_a^b |x(t) - w(t)|^2\, dt$.

Exercício 6.2 *Considere uma função w definida no intervalo $[0, 1]$. Mostre que se pretendermos uma aproximação polinomial com a base $v_k(t) = t^k$, obtemos $v = \alpha_0 + \alpha_1 t + ... + \alpha_p t^p$, em que*

$$\begin{bmatrix} 1 & \cdots & \frac{1}{p+1} \\ \vdots & \ddots & \vdots \\ \frac{1}{p+1} & \cdots & \frac{1}{2p+1} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} \int_0^1 w(t)\, dt \\ \vdots \\ \int_0^1 t^p w(t)\, dt \end{bmatrix}$$

e calcule o número de condição da matriz (chamada matriz de Hilbert) para $p = 1$ e $p = 2$.
Nota: *Esta matriz é torna-se bastante mal condicionada, mesmo para valores de p não muito grandes.*

6.1.3 Dependência não linear nos coeficientes

Supondo que queremos aproximar uma função $f(x)$ por uma função do tipo

$$g(x) = ae^{bx},$$

reparamos que não há dependência linear nos coeficientes, pelo que há normalmente duas soluções.

(i) Solução exacta.

Calcula-se o gradiente de

$$J(a, b) = ||f(x) - ae^{bx}||^2$$

e procuram-se os pontos críticos, tais que $\nabla J(a, b) = 0$, o que irá levar a um sistema não linear em a e b.

Neste caso,

$$J(a, b) = \left\langle f(x) - ae^{bx}, f(x) - ae^{bx} \right\rangle = ||f(x)||^2 - 2a \left\langle e^{bx}, f(x) \right\rangle + a^2 \left\langle e^{bx}, e^{bx} \right\rangle$$

e portanto,

$$\nabla J(a,b) = \begin{bmatrix} -2\left\langle e^{bx}, f(x)\right\rangle + 2a\left\langle e^{bx}, e^{bx}\right\rangle \\ -2a\left\langle xe^{bx}, f(x)\right\rangle + 2a^2\left\langle xe^{bx}, e^{bx}\right\rangle \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

leva-nos a um sistema de duas equações a duas incógnitas. Neste caso simples, retiramos

$$a = \frac{\left\langle e^{bx}, f(x)\right\rangle}{\left\langle e^{bx}, e^{bx}\right\rangle}, \text{ e se } a \neq 0,$$
$$\left\langle xe^{bx}, f(x)\right\rangle \left\langle e^{bx}, e^{bx}\right\rangle = \left\langle e^{bx}, f(x)\right\rangle \left\langle xe^{bx}, e^{bx}\right\rangle.$$

A segunda equação é uma equação não linear em b que deverá ser resolvida.

(ii) Solução aproximada.

Como queremos $f(x) \approx g(x)$, efectua-se a transformação $\log(f(x)) \approx \log(g(x)) = \log a + bx$.

Desta forma podemos pensar em fazer a aproximação dos pontos $\log(f(x))$, que são facilmente calculáveis, através de uma regressão linear com $G(x) = A + Bx$. Tendo obtido os valores de A e B, substituímos $a = e^A, b = B$.

É claro que a solução aproximada pode diferir razoavelmente dos valores correctos, basta reparar que neste caso,

$$\log f(x) = \log a + bx + E(x) \Rightarrow f(x) = e^{E(x)} a e^{bx},$$

ou seja, o erro $E(x)$ irá aparecer como coeficiente multiplicativo e exponencial. Caso os valores de $E(x)$ sejam pequenos, teremos $e^{E(x)}$ próximo de 1, o que torna a aproximação aceitável (especialmente quando não é garantida a exactidão dos próprios valores de f, como acontece em ciências experimentais).

• Este tipo de técnica pode ser aplicada com outro tipo de funções, quando uma simples transformação permita passar de uma dependência não linear para uma dependência linear... com as devidas precauções.

6.2 Minimização sem restrições

Iremos agora abordar superficialmente alguns métodos de optimização que se caracterizam por procurar o mínimo de uma função sem impor restrições a que esse mínimo esteja especificamente num conjunto, ou o que pode ser equivalente, que verifique uma determinada propriedade. Não será assim abordado o assunto da programação linear e não linear, a minimização com restrições, que no caso não linear se baseia parcialmente na generalização da teoria dos multiplicadores de Lagrange (relações de Kuhn-Tucker) e na ideia de aproximar problemas de minimização com restrições por uma sucessão de problemas sem restrições (método de Uzawa)... Nem tão pouco iremos introduzir a parte de programação linear, em que o método do simplex é um paradigma de simplicidade e eficácia. O assunto da optimização é um vasto campo, e aqui apenas abordaremos alguns dos métodos mais conhecidos na minimização sem restrições.

Começamos por relembrar as noções de pontos de mínimo para uma função $f : X \subseteq \mathbb{R}^N \to \mathbb{R}$:

– Dizemos que $x \in X$ é um *ponto de mínimo absoluto estrito* de f em X, se $f(y) > f(x), \forall y \in X$, com $y \neq x$.

– Dizemos que $x \in X$ é um *ponto de mínimo relativo estrito* de f, se $f(y) > f(x), \forall y \in V_x$, com $y \neq x$, onde V_x é uma vizinhança de x.

Definição 6.1 *Seja $f : \mathbb{R}^N \to \mathbb{R}$ de classe C^1. Dizemos que x é um ponto crítico de f se $\nabla f(x) = 0$.*

Relembramos os seguintes teoremas da análise matemática.

Teorema 6.1 *(condição suficiente). Se $f \in C^2$ e a matriz hessiana $\boldsymbol{\nabla}^2 f$ for definida positiva num ponto crítico x, então x é um ponto de mínimo relativo estrito de f (se for definida negativa será um ponto de máximo relativo estrito).*

Demonstração. Resulta do desenvolvimento em série de Taylor,

$$f(x+h) = f(x) + \nabla f(x)h + \frac{1}{2}h^\top \boldsymbol{\nabla}^2 f(x+\theta h)h, \text{ com } \theta \in]0,1[,$$

já que nesse caso ficamos com $f(x+h) - f(x) = h^\top \boldsymbol{\nabla}^2 f(x+\theta h)h > 0$ se $h \neq 0$ for suficientemente pequeno.□

Teorema 6.2 *(condição necessária). Se $f \in C^1$ e x for um ponto de mínimo ou máximo relativo, então $\nabla f(x) = 0$. Se $f \in C^2$, então a matriz hessiana de $\boldsymbol{\nabla}^2 f(x)$ é semidefinida positiva ou semidefinida negativa, consoante x seja um ponto de máximo ou mínimo relativo.*

Demonstração. Resulta também do desenvolvimento em série de Taylor,

$$f(x+h) = f(x) + \nabla f(x)h + o(||h||).$$

Como $f(x+h) - f(x) \geq 0$, para qualquer $h \neq 0$, temos $\nabla f(x)\frac{h}{||h||} \geq o(1)$, o que no limite significa

$$\nabla f(x)v \geq 0, \ \forall v : ||v|| = 1,$$

desigualdade que aplicada a $-v$ dá $\nabla f(x)v \leq 0$, e portanto

$$\nabla f(x)v = 0, \ \forall v : ||v|| = 1.$$

Particularizando para $v = \mathbf{e}^{(k)}$ (vectores da base canónica), obtemos $\nabla f(x) = 0$. O outro resultado é semelhante, considerando o desenvolvimento

$$f(x+h) = f(x) + \frac{1}{2}h^\top \boldsymbol{\nabla}^2 f(x)h + o(||h||^2),$$

pois já mostrámos que $\nabla f(x) = 0$. □

Definição 6.2 *Uma função diz-se convexa em D se para quaisquer $x, y \in D$ verificar*

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \forall \theta \in [0,1]$$

e estritamente convexa se a desigualdade for estrita (excepto para $x = y$).

• Uma função será convexa se $f(y) \geq f(x) + \nabla f(x)(y-x)$, caso $f \in C^1(D)$, e se $f \in C^2(D)$ será convexa quando a matriz hessiana for semidefinida positiva (estritamente convexa se for definida positiva).

Um processo de determinar mínimos relativos será, portanto, encontrar $x : \nabla f(x) = 0$, o que corresponde a resolver um sistema não linear com N equações (as derivadas parciais $\partial_1 f = 0, ..., \partial_N f = 0$) e N incógnitas, $x_1, ..., x_N$, assegurando que a matriz hessiana é definida positiva.

6.2.1 Métodos de descida

Começamos por apresentar um tipo de métodos, designados como *métodos de descida,* em que o processo de encontrar um ponto de mínimo consiste em associar à iteração uma função de descida, de forma a assegurar que através do processo iterativo se esteja mais próximo de um mínimo. Genericamente podemos descrever os métodos de descida como métodos iterativos (semelhantes a métodos de ponto fixo), que podem ser apresentados na forma geral[2]:

$$x_{n+1} = A(x_n),$$

mas que estão sujeitos a um critério de atribuição definido por uma função de descida Z.

A *função de descida* para A, é uma função contínua num conjunto X com valores reais que deve verificar $Z(A(x)) < Z(x)$, para todo o $x \neq z, x \in X$, onde z são os pontos de mínimo (poderá ser apenas um, ou vários).

Normalmente a função Z poderá ser a própria função f, e nesse caso os pontos de mínimo de Z coincidem com os pontos de mínimo de f. Outra possibilidade é escolher $Z = |\nabla f|$, e nesse caso poderá haver pontos de mínimo para Z que não o sejam para f (pois mesmo que o gradiente seja nulo não implicará que se trate de um ponto de mínimo de f).

Teorema 6.3 *(convergência global). Se A for uma função contínua e X for um conjunto compacto então a sucessão (x_n), dada por um método de descida, tem subsucessões convergentes que convergem para pontos de mínimo.*

Demonstração:

Como admitimos que X é compacto então é possível retirar subsucessões convergentes. Consideremos (y_n) uma dessas subsucessões convergentes, temos $y_n \to y$, e suponhamos por absurdo que y não é ponto de mínimo, logo $Z(A(y)) > Z(y)$. Como A é contínua,

$$y_n = A(y_{n-1}) \to A(y),$$

[2]Seguimos [22], com algumas simplificações. Grande parte da teoria geral é devida a Zangwill (final dos anos 60).

ou seja $A(y) = y$, e portanto $Z(A(y)) = Z(y)$,.contradição que implica que y seja ponto de mínimo.■

• No caso em que há apenas um ponto de mínimo, a sucessão definida pelo método terá que convergir para ele.

Observação: Quando se trata de minimizar uma função num intervalo de $\mathbb{R}$ existe um método de certa forma análogo ao método da bissecção, designado *método da secção dourada* (porque a divisão do intervalo é feita de acordo com a sucessão de Fibonacci). Analisando a função nesses pontos, de forma a escolher intervalos em que o valor da função seja mais pequeno (usamos $Z = f$ como função de descida), obtemos uma convergência para um ponto de mínimo. Quando há apenas um mínimo nesse intervalo, o método permite localizá-lo com uma convergência linear em que o factor assimptótico é o número de ouro $\frac{\sqrt{5}-1}{2}$.

6.2.2 Método do gradiente

Começamos por apresentar um método de descida bastante conhecido, o método do gradiente, relacionando-o com a pesquisa de um ponto crítico x, onde $\nabla f(x) = 0$. Considerando ω não nulo, podemos estabelecer a equivalência,

$$\nabla f(x) = 0 \Leftrightarrow x = x - \omega \nabla f(x).$$

Esta equivalência, semelhante à deduzida para o método do ponto fixo com relaxação, sugere a implementação do método iterativo

$$x^{(n+1)} = x^{(n)} - \omega \nabla f(x^{(n)}).$$

Podemos reparar que $-\nabla f(x^{(n)})$ será uma direcção que determina a descida de f no ponto $x^{(n)}$. Basta reparar que, admitindo que $f \in C^2$, e sendo $d = -\nabla f(x)$, pela fórmula de Taylor tem-se

$$f(x + \omega d) = f(x) + \omega \nabla f(x) d + o(1),$$

ou seja,

$$f(x + \omega d) = f(x) - \omega |\nabla f(x)|^2 + o(1).$$

Isto significa que, para $\omega > 0$ suficientemente pequeno, o termo $o(1)$ é negligenciável, e teremos

$$f(x + \omega d) < f(x).$$

Substituindo os valores, ie. x por $x^{(n)}$ e d por $-\nabla f(x^{(n)})$, concluímos que nessas condições,

$$f(x^{(n+1)}) < f(x^{(n)}),$$

ou seja, a função de descida é $Z = f$, e o processo iterativo

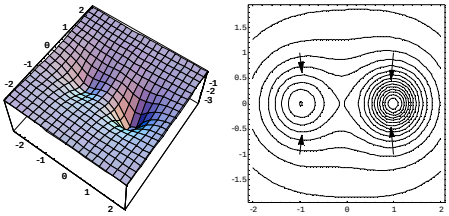
$$x^{(n+1)} = x^{(n)} - \omega_n \nabla f(x^{(n)}),$$

constitui um método de descida, desde que os $\omega_n > 0$ sejam suficientemente pequenos. A convergência é assegurada pelo teorema de convergência global. Acabamos de apresentar uma variante a um método bastante conhecido, denominado *método do gradiente* (ou do *declive máximo...* em inglês, *steepest descent*). As diferenças entre as variantes residem na escolha do valor ω_n, já que no caso do método do gradiente clássico esse valor ω_n é bem determinado em cada passo. Antes de entrarmos no assunto da escolha de ω_n, apresentamos um exemplo.

Exemplo 6.1 *Consideramos a função,*

$$f(x,y) = \frac{-1}{(x+1)^2+y^2+0.5} + \frac{-1}{(x-1)^2+y^2+0.25}.$$

Vemos facilmente que esta função tem dois pontos de mínimo relativo, um deles próximo de $(x,y) = (-1,0)$, *onde a função atinge o valor* $f(-1,0) = -2 - \frac{4}{17}$, *e um outro próximo de* $(x,y) = (1,0)$, *onde a função atinge o valor* $f(1,0) = -4 - \frac{2}{9}$, *que se trata mesmo do mínimo global (ver figura em baixo, à esquerda). Considerando quatro pontos iniciais distintos,* $(x_0,y_0) = (\pm 1, \pm 1)$, *verificamos que as direcções definidas pelo vector gradiente (nesses 4 pontos) podem tanto apontar para um dos mínimos relativos como para o outro (figura em baixo, à direita, onde a função é representada pelas curvas de nível). A convergência para um outro mínimo relativo irá depender do valor inicial considerado. Assim, para os pontos iniciais* $(-1, \pm 1)$ *haverá, a priori, convergência para o ponto de mínimo relativo próximo de* $(-1,0)$, *e para os pontos iniciais* $(1, \pm 1)$ *haverá convergência para o ponto de mínimo relativo próximo de* $(1,0)$, *que também é ponto de mínimo global*[3].



• **Escolha do valor** ω_n.

No *método do declive máximo,* o valor ω_n é o ponto de mínimo de $f(x^{(n)} + \omega d^{(n)})$ enquanto função real de ω, o que implica resolver a equação em $\mathbb{R}$,

$$\phi'(\omega) = 0, \text{ em que } \phi(\omega) = f\left(x^{(n)} - \omega \nabla f(x^{(n)})\right).$$

Para resolver a equação não linear $\phi'(\omega) = 0$ podemos utilizar qualquer um dos métodos que estudámos para equações não lineares com uma variável.

[3]É importante notar que se é possível desenvolver métodos para encontrar mínimos relativos, já não o será para encontrar os mínimos globais, pelo menos no caso geral. Apenas com um razoável conhecimento da função, por exemplo, no que diz respeito à sua convexidade, ou ao número máximo de mínimos relativos, será possível obter o ponto em que se atinge um mínimo global.

Exemplo 6.2 *Retomando o exemplo anterior, começamos com* $(x_0, y_0) = (1, 1)$, *e temos* $\nabla f(1,1) = (0.132, 1.346) = -d^{(0)}$. *Assim, a primeira iterada seria definida pela resolução de*

$$\frac{d}{d\omega} f\left((1,1) - \omega(0.132, 1.346)\right) = 0,$$

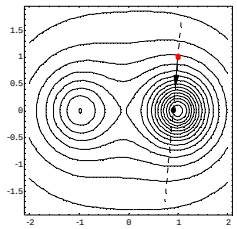
ou seja,

$$\begin{aligned} 0 &= \frac{d}{d\omega} f\left(1 - 0.132\omega, 1 - 1.346\omega\right) \\ &= \frac{d}{d\omega}\left(\frac{-1}{(2-0.132\omega)^2 + (1-1.346\omega)^2 + 0.5} + \frac{-1}{(-0.132\omega)^2 + (1-1.346\omega)^2 + 0.25}\right) \\ &\Leftrightarrow \quad (0.736 - \omega)(1.67 - 0.763\omega + \omega^2)(2.87 - 2.54\omega + \omega^2) = 0. \end{aligned}$$

A equação não linear é aqui apresentada numa forma factorizada, mas normalmente haveria que encontrar as suas raízes através de um método numérico. Neste caso, há apenas uma raiz real que é $\omega_0 = 0.736$ *e será esse o ponto de mínimo, segundo a direcção dada pelo gradiente, que define a próxima iteração, pois ficamos com*

$$x^{(1)} = x^{(0)} - \omega_0 \nabla f(x^{(0)}) = (1,1) - 0.736(0.132, 1.346) = (0.903, 0.009).$$

Este valor é já uma aproximação razoável do valor correcto e, neste caso, em poucas iterações obteríamos um valor muito próximo do valor correcto. Ilustramos esta situação na figura seguinte, onde está representada a iterada a inicial, o vector e a direcção definidas pelo gradiente. É ao longo da linha a tracejado que deve ser procurado o ponto de mínimo, é esse o valor $x^{(1)}$ *que está representado como um ponto negro, já muito próximo do ponto de mínimo, que é* $(0.99376, 0)$.



Observação 1 *(escolha aproximada do valor* ω_n*).*

Normalmente encontrar ω_n que seja o zero da derivada pode envolver demasiados cálculos, pelo que é comum aproximar a função

$$\phi(\omega) = f\left(x^{(n)} - \omega \nabla f(x^{(n)})\right)$$

por um polinómio do segundo grau $p(\omega) = \alpha_0 + \alpha_1\omega + \alpha_2\omega^2$. Esta aproximação *pressupõe* que f seja *convexa*, mas poderá ser adaptada a outros casos. A aproximação da função ϕ através de uma parábola pode ser efectuada com interpolação em três pontos w_1, w_2, w_3, positivos, escolhidos próximo de 0, o que corresponde a resolver três equações

$$p(w_1) = \phi(w_1),\ p(w_2) = \phi(w_2),\ p(w_3) = \phi(w_3),$$

para determinar as incógnitas α_k. Podemos colocar o problema em termos de sistema,

$$\begin{bmatrix} 1 & w_1 & w_1^2 \\ 1 & w_2 & w_2^2 \\ 1 & w_3 & w_3^2 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \phi(w_1) \\ \phi(w_2) \\ \phi(w_3) \end{bmatrix},$$

mas notamos que a solução deste sistema é obtida de forma simples usando uma fórmula de interpolação,

$$p(\omega) = \phi(w_1) + \phi[w_1, w_2](\omega - w_1) + \phi[w_1, w_2, w_3](\omega - w_1)(\omega - w_2),$$

em que $\phi[w_1, w_2] = \frac{\phi(w_1) - \phi(w_2)}{w_1 - w_2}$, e $\phi[w_1, w_2] = \frac{\phi[w_1, w_2] - \phi[w_2, w_3]}{w_1 - w_3}$.

Tendo obtido estes valores, calculamos o ponto de mínimo de p, que será $\tilde{\omega}$ tal que $p'(\tilde{\omega}) = 0$, ou seja,

$$0 = \phi[w_1, w_2] + \phi[w_1, w_2, w_3](2\tilde{\omega} - w_1 - w_2),$$

e portanto,

$$\tilde{\omega} = \frac{w_1 + w_2}{2} - \frac{\phi[w_1, w_2]}{2\phi[w_1, w_2, w_3]}.$$

Podemos usar este valor como aproximação de ω_n.

No entanto, como referimos, esta aproximação pressupõe que a função seja convexa, já que doutra forma, ao calcular o zero da derivada do polinómio, poderemos estar a encontrar um máximo ao invés de um mínimo... veremos mais concretamente o método e este problema no exemplo seguinte.

Exemplo 6.3 *Retomamos ainda os dados do exemplo anterior, com* $(x_0, y_0) = (1, 1)$. *Notamos que a função não é sempre convexa, por isso podem surgir problemas com a aproximação. Com efeito, considerando inicialmente* $w_1 = 0, w_2 = 0.1, w_3 = 0.2$, *obtemos*

$$\tilde{\omega} = \frac{w_1 + w_2}{2} - \frac{\phi[w_1, w_2]}{2\phi[w_1, w_2, w_3]} = 0.05 - \frac{-5.596}{2 \times (-15.21)} = -0.133.$$

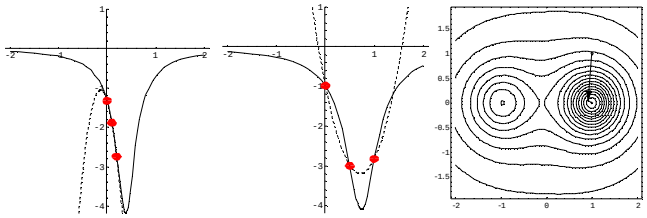
Este valor é anómalo, porque é negativo e porque reparamos que se trata de uma parábola concâva, já que o coeficiente $\phi[w_1, w_2, w_3]$ *que determina a orientação da parábola é negativo (ver figura em baixo, à esquerda). Ainda que neste caso haja problemas, eles podem ser contornados pela escolha de valores diferentes para* w_1, w_2, w_3 *até que o coeficiente* $\phi[w_1, w_2, w_3]$ *seja positivo, e assim o extremo será um mínimo. Outra hipótese é considerar* $\tilde{\omega} = w_k$ *como o valor mais baixo de* $\phi(w_k)$, *o que neste caso acontece para* w_3.
Optamos pela primeira hipótese, escolhendo novos valores, $w_1 = 0, w_2 = 0.5, w_3 = 1.0$, *e obtemos agora*

$$\tilde{\omega}_0 = \frac{w_1 + w_2}{2} - \frac{\phi[w_1, w_2]}{2\phi[w_1, w_2, w_3]} = 0.25 - \frac{-4.03}{2 \times 4.38} = 0.7107...$$

Portanto, como a parábola já é convexa (ver figura em baixo, ao centro), já podemos tomar como aproximação do ponto de mínimo,

$$x^{(1)} = x^{(0)} - \tilde{\omega}_0 \nabla f(x^{(0)}) = (1, 1) - 0.711(0.132, 1.346) = (0.906, 0.0403).$$

Repetindo o procedimento, notamos que os valores de w_1, w_2, w_3 *devem ser reajustados, e mais pequenos. Para* $w_1 = 0, w_2 = 0.1, w_3 = 0.2$, *obteríamos* $x^{(2)} = (0.9947, -0.0013), ...$ *já muito próximo do valor correcto* $(0.99376..., 0)$.



Observação 1. Outra possibilidade, para funções regulares, consiste em considerar a aproximação usando derivadas, por exemplo, efectuando uma aproximação de ϕ pelo desenvolvimento em série de MacLaurin, $\phi(\omega) = \phi(0) + \phi'(0)\omega + \frac{1}{2}\phi''(0)\omega^2 + o(1)$, mas envolve um cálculo das derivadas de ordem superior de f.

Observação 2 *(ordem de convergência).*
É possível ver (cf. [22]) que o método do gradiente quando aplicado à minimização de formas quadráticas tem convergência linear, e que o factor assimptótico de convergência é dado por $\left(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right)^2$ onde $\lambda_{\max}, \lambda_{\min}$ são o maior e o menor valor próprio da matriz simétrica e definida positiva A (referida na observação anterior). No caso geral, esse mesmo resultado poderá ser obtido para $A = \boldsymbol{\nabla}^2 f(z)$.

6.2.3 Aplicação à resolução de sistemas

Os métodos de optimização podem também aplicar-se à resolução de sistemas (e os próprios métodos iterativos podem ser vistos como um caso de métodos de optimização – é o que acontece com o método de Gauss-Seidel, que resulta da aplicação de um método de relaxação de optimização a um sistema linear (cf.[7]).

Seja A uma matriz hermitiana e definida positiva. Consideramos o sistema

$$Ax = b,$$

e a equivalência

$$Ax = b \Leftrightarrow J(x) = \min_y J(y), \text{ em que } J(y) = \frac{1}{2}y^*Ay - y^*b,$$

que podemos estabelecer, de forma semelhante ao que fizémos no caso dos mínimos quadrados *(exercício),*.notando que

$$\begin{aligned} J(y+h) - J(y) &= \frac{1}{2}(y+h)^*A(y+h) - (y+h)^*b - \frac{1}{2}y^*Ay + y^*b \\ &= \frac{1}{2}y^*Ah + \frac{1}{2}h^*A(y+h) - h^*b = \\ &= \frac{1}{2}h^*A^*y + \frac{1}{2}h^*Ay - h^*b + \frac{1}{2}h^*Ah = h^*(Ay-b) + O(||h||^2) \end{aligned}$$

(porque A é hermitiana), e portanto $\nabla J(y)h = h^*(Ay - b)$, ou seja

$$\nabla J(y) = Ay - b.$$

A equivalência sugere a aplicação do método do gradiente à função $J(x)$, e temos

$$x^{(n+1)} = x^{(n)} - \omega_n \nabla J(x^{(n)})$$

em que ω_n será o mínimo de $J(x^{(n)} - \omega d^{(n)})$, sendo $d^{(n)} = Ax^{(n)} - b$.

Podemos calcular explicitamente esse mínimo, pois

$$\begin{aligned} J(x^{(n)} - \omega d^{(n)}) &= \frac{1}{2}(x^{(n)} - \omega d^{(n)})^* A(x^{(n)} - \omega d^{(n)}) - (x^{(n)} - \omega d^{(n)})^* b = \\ &= \frac{1}{2}\omega^2 (d^{(n)})^* A d^{(n)} - \omega\,(d^{(n)})^* A x^{(n)} + \omega (d^{(n)})^* b + C, \end{aligned}$$

notando que A é hermitiana (assim $\frac{1}{2}(x^{(n)})^* A d^{(n)} = \frac{1}{2}(d^{(n)})^* A x^{(n)}$), e em que $C = (x^{(n)})^* A x^{(n)} - (x^{(n)})^* b$ não depende de ω.

Portanto ao calcular $\frac{\mathrm{d}}{\mathrm{d}\omega} J(x^{(n)} - \omega d^{(n)})$, ficamos com

$$\frac{\mathrm{d}}{\mathrm{d}\omega} J(x^{(n)} - \omega d^{(n)}) = \omega (d^{(n)})^* A d^{(n)} - (d^{(n)})^* A x^{(n)} + (d^{(n)})^* b,$$

e o mínimo será atingido quando a derivada se anular, o que acontece quando

$$\omega (d^{(n)})^* A d^{(n)} = (d^{(n)})^* A x^{(n)} - (d^{(n)})^* b,$$

ou seja,

$$\omega = \frac{(d^{(n)})^*(Ax^{(n)} - b)}{(d^{(n)})^* A d^{(n)}} = \frac{||d^{(n)}||^2}{(d^{(n)})^* A d^{(n)}}.$$

• Resumindo, o *método do gradiente* (aplicado a sistemas de equações) fica

$$x^{(n+1)} = x^{(n)} + \omega_n d^{(n)}, \quad \text{com} \quad \begin{cases} d^{(n)} = Ax^{(n)} - b, \\ \omega_n = \frac{||d^{(n)}||^2}{(d^{(n)})^* A d^{(n)}}. \end{cases}$$

Observação:

(i) Convém notar que esta escolha de ω é também adoptada para sistemas não lineares escritos sob a forma $A(x) = b(x)$.

(ii) Se a matriz não estiver nas condições referidas, consideramos a matriz A^*A e resolvemos o sistema equivalente

$$A^*Ax = A^*b.$$

6.2.4 Método do gradiente conjugado

No princípio dos anos 50 apareceu um outro método, devido a *Hestenes e Stiefel*, designado por *método das direcções conjugadas*, que permitia melhorar a dedução que acabamos de ver para sistemas lineares. Não seguia a ideia da descida do método do gradiente, procurando

antes direcções ortogonais face à matriz, ditas A-conjugadas. A aplicação desse método à minimização de quaisquer funções, é designada por *método do gradiente conjugado.*

Recuperando a dedução anterior, notamos que

$$x^{(n)} = x^{(n-1)} + \omega_{n-1} d^{(n-1)} = x^{(0)} + \omega_{n-1} d^{(n-1)} + ... + \omega_0 d^{(0)}$$

e portanto se as direcções $d^{(k)}$ e os coeficientes ω_k fossem escolhidos apropriadamente, obter $x^{(n)} = x$, corresponderia a encontrar essas direcções e coeficientes tais que

$$x - x^{(0)} = \omega_{n-1} d^{(n-1)} + ... + \omega_0 d^{(0)}.$$

Se os vectores $d^{(0)}, ..., d^{(N-1)}$ constituirem uma base de $\mathbb{R}^N$, isso significa que seria possível ao fim de N iterações atingir a solução exacta, caso fosse possível escolhê-los apropriadamente... sem conhecer a solução!. No caso de formas quadráticas associadas a sistemas lineares, isso é possível!

Com efeito, começamos por reparar que A, matriz hermitiana e definida positiva, define um produto interno

$$\langle u, v \rangle_A = u^* A v,$$

e podemos escolher $d^{(0)}, ..., d^{(N-1)}$ de forma a constituirem uma base ortogonal para esse produto interno. Nesse caso, os valores $\omega_0, ..., \omega_N$ serão as projecções segundo essa base, usando o produto interno definido por $\langle ., . \rangle_A$, tendo-se,

$$\omega_k = \frac{\left\langle x - x^{(0)}, d^{(k)} \right\rangle_A}{\langle d^{(k)}, d^{(k)} \rangle_A} = \frac{(d^{(k)})^* A (x - x^{(0)})}{(d^{(k)})^* A d^{(k)}} = \frac{(d^{(k)})^* (b - A x^{(0)})}{(d^{(k)})^* A d^{(k)}}.$$

Resta saber como encontrar as direcções $d^{(k)}$.

Dado $x^{(0)}$, consideramos $d^{(0)} = b - A x^{(0)}$, o que permite definir $x^{(1)} = x^{(0)} + \omega_0 d^{(0)}$. Ao resto, $r^{(1)} = b - A x^{(1)}$, aplicamos o processo de ortogonalização de Gram-Schmidt (com o produto interno associado a A), para definir uma direcção $d^{(1)}$ que seja A-ortogonal a $d^{(0)}$, dizendo-se que são direcções A-conjugadas (o nome do método surge daqui), ou seja,

$$d^{(1)} = r^{(1)} - \frac{\left\langle r^{(1)}, d^{(0)} \right\rangle_A}{\langle d^{(0)}, d^{(0)} \rangle_A} \, d^{(0)} = r^{(1)} - \frac{(d^{(0)})^* A r^{(1)}}{(d^{(0)})^* A d^{(0)}} \, d^{(0)}.$$

O mesmo processo é aplicado nos passos seguintes.

• Para $r^{(k)} = b - A x^{(k)}$, definimos

$$d^{(k)} = r^{(k)} - \frac{(d^{(k-1)})^* A r^{(k)}}{(d^{(k-1)})^* A d^{(k-1)}} \, d^{(k-1)},$$

e tal como deduzido no método do gradiente, o valor ω_k é dado por

$$\omega_k = \frac{(d^{(k)})^* (b - A x^{(k)})}{(d^{(k)})^* A d^{(k)}} = \frac{(d^{(k)})^* r^{(k)}}{(d^{(k)})^* A d^{(k)}},$$

tendo-se

$$x^{(k+1)} = x^{(k)} + \omega_k d^{(k)}$$

No final, podemos escrever

$$x = x^{(0)} + \frac{(d^{(0)})^* r^{(0)}}{(d^{(0)})^* A d^{(0)}} d^{(0)} + ... + \frac{(d^{(N-1)})^* r^{(N-1)}}{(d^{(N-1)})^* A d^{(N-1)}} d^{(N-1)},$$

ou seja, a solução é atingida ao fim de N iterações, *no caso de sistemas lineares.*

• Estas expressões podem ser simplificadas, em termos de cálculo, usando a ortogonalidade, pois

$$(d^{(k)})^* r^{(k)} = (r^{(k)})^* r^{(k)} - \frac{(d^{(k-1)})^* A r^{(k)}}{(d^{(k-1)})^* A d^{(k-1)}} \; (d^{(k-1)})^* r^{(k)} = (r^{(k)})^* r^{(k)},$$

porque $(d^{(k-1)})^* r^{(k)} = 0$. Por outro lado, como $Ax^{(k)} = Ax^{(k-1)} + \omega_{k-1} A d^{(k-1)}$, obtemos

$$A d^{(k-1)} = \frac{1}{\omega_{k-1}} (r^{(k)} - r^{(k-1)}),$$

e como $(r^{(k)})^* r^{(k-1)} = 0$,

$$(d^{(k-1)})^* A r^{(k)} = (r^{(k)})^* A d^{(k-1)} = \frac{1}{\omega_{k-1}} (r^{(k)})^* r^{(k)} - \frac{1}{\omega_{k-1}} (r^{(k)})^* r^{(k-1)} = \frac{1}{\omega_{k-1}} (r^{(k)})^* r^{(k)}$$

e também

$$(d^{(k-1)})^* A d^{(k-1)} = \frac{1}{\omega_{k-1}} (d^{(k-1)})^* (r^{(k)} - r^{(k-1)}) = \frac{1}{\omega_{k-1}} (d^{(k-1)})^* r^{(k-1)} = \frac{1}{\omega_{k-1}} (r^{(k-1)})^* r^{(k-1)}.$$

• Resumindo, o *método das direcções conjugadas* (ou do gradiente conjugado aplicado a sistemas) fica,

$$r^{(k)} = b - Ax^{(k)},$$

$$\begin{cases} d^{(0)} = r^{(0)} \\ d^{(k)} = r^{(k)} - \frac{(r^{(k)})^* r^{(k)}}{(r^{(k-1)})^* r^{(k-1)}} \; d^{(k-1)} \end{cases}$$

e

$$x^{(k+1)} = x^{(k)} + \omega_k d^{(k)}$$

com

$$\omega_k = \frac{(r^{(k)})^* r^{(k)}}{(d^{(k)})^* A d^{(k)}}.$$

Se no caso dos sistemas lineares o método do gradiente conjugado atinge a solução *exacta* ao fim de um número de iterações menor ou igual que a dimensão da matriz, a sua extensão para o caso geral, não verifica essa propriedade, como é natural.

• Seguindo as mesmas etapas, o *método do gradiente conjugado,* para quaisquer funções regulares, resume-se a

$$\begin{cases} d^{(0)} = \nabla f(x^{(0)}), \\ d^{(k)} = \nabla f(x^{(k)}) - \frac{||\nabla f(x^{(k)})||^2}{||\nabla f(x^{(k-1)})||^2} \; d^{(k-1)}. \end{cases}$$

e

$$x^{(k+1)} = x^{(k)} + \omega_k d^{(k)}$$

em que ω_k resulta de minimizar $f(x^{(k)} + \omega d^{(k)})$ enquanto função de ω, tal como no método do gradiente.

Observação:

(i) A generalização do método do gradiente conjugado é consistente com o caso particular de sistemas, pois o gradiente da forma quadrática verifica $\nabla J(x^{(k)}) = r^{(k)}$. Como acontecia no caso do método do gradiente, não há fórmula explícita para ω_k, sendo ainda necessário efectuar a minimização no parâmetro ω. Essa minimização pode também ser feita de forma aproximada usando a interpolação por uma parábola, tal como no método do gradiente. A justificação da generalização do método para funções regulares resulta da aproximação da função próximo do ponto de mínimo por uma forma quadrática definida pela matriz hessiana.

(ii) No que diz respeito à resolução de sistemas lineares, o método do gradiente conjugado aplicado a matrizes hermitianas, definidas positivas, consome aproximadamente N^3 operações $(\times, \backslash)$ e é apenas preferível ao método de Cholesky no caso de matrizes suficientemente esparsas.

6.2.5 Método de Newton

Para obtermos um método com convergência mais rápida, quadrática, quando a função é regular ($f \in C^2$) podemos considerar o método de Newton, definindo-o de forma semelhante a um método de descida com

$$\omega_n d^{(n)} = -[\boldsymbol{\nabla}^2 f]^{-1}(x^{(n)})\ \nabla f(x^{(n)}).$$

Ou seja, resolvemos sucessivamente

$$\boldsymbol{\nabla}^2 f(x^{(n)})(x^{(n+1)} - x^{(n)}) = -\nabla f(x^{(n)}).$$

• Repare-se que vimos que uma condição suficiente para que a função fosse mínimo relativo, seria que a matriz hessiana fosse definida positiva. Como assumimos que a função é regular podemos assumir que geralmente, próximo do ponto de mínimo, a hessiana $\boldsymbol{\nabla}^2 f$ será definida positiva, o que nos garante que o método está bem definido.

Note-se ainda que o método envolve um elevado número de operações relativo ao cálculo da matriz hessiana e do sistema linear.

• *Método de Levenberg-Marquardt.*

Como, longe do ponto de mínimo, a matriz hessiana não será habitualmente definida positiva, podendo mesmo não existir, então uma possibilidade é considerar métodos de Newton modificados. Uma alternativa bastante utilizada, consiste em fazer uma perturbação na matriz hessiana, adicionando um $\varepsilon\mathbf{I}$, ou seja, uma pequena perturbação na diagonal da matriz. Ficamos assim com

$$\omega_n d^{(n)} = -(\varepsilon_n \mathbf{I} + [\boldsymbol{\nabla}^2 f])^{-1}(x^{(n)})\ \nabla f(x^{(n)}),$$

o que corresponderá a resolver o sistema

$$(\varepsilon_n \mathbf{I} + \boldsymbol{\nabla}^2 f)(x^{(n)})(x^{(n+1)} - x^{(n)}) = -\nabla f(x^{(n)}).$$

Os valores $\varepsilon_n > 0$ devem ser pequenos, mas escolhidos de forma a que a matriz $\varepsilon_n \mathbf{I} + [\boldsymbol{\nabla}^2 f]$ seja definida positiva. Isso consegue-se sempre porque para $\varepsilon > 0$ suficientemente grande teríamos uma matriz simétrica com a diagonal estritamente dominante e positiva. Este método, ou as suas variantes, são por vezes designados *métodos de Levenberg-Marquardt*. Note-se que no próprio processo de resolução do sistema, poderá ser avaliado se a matriz é ou não definida positiva. Isso será particularmente simples com o método de Cholesky, já que as iterações seriam interrompidas pela existência de um pivot negativo, sendo necessário incrementar o valor de ε.

Observação: Há ainda todo um conjunto de métodos derivados do método de Newton, com os mais variados objectivos, que incluem evitar a resolução sistemática do sistema, ou o cálculo da matriz hessiana. Esses métodos são designados genericamente por *métodos quasi-Newton.*

• O *Mathematica* tem implementada a rotina `FindMinimum`, em que basta introduzir a função a minimizar e uma iterada inicial.

6.2.6 Método das coordenadas

Um dos métodos mais simples consiste em considerar a avaliação da função, coordenada a coordenada. Ou seja, dado um ponto inicial $(y_1, ..., y_N)$ iremos corrigir as suas componentes procurando primeiro x_1 tal que

$$f(x_1, y_2, ..., y_N) = \min_{y_1} f(y_1, y_2, ..., y_N),$$

depois x_2 tal que

$$f(x_1, x_2, y_3, ..., y_N) = \min_{y_2} f(x_1, y_2, y_3, ..., y_N),$$

até encontrarmos x_N tal que

$$f(x_1, ..., x_N) = \min_{y_N} f(x_1, ..., x_{N-1}, y_N).$$

Isto permite reduzir o problema a várias dimensões, a vários problemas unidimensionais. Eventualmente poderá acontecer a convergência para um mínimo relativo, mas no caso mais geral isso pode não acontecer. Para complementar esta técnica, é habitualmente feita uma procura no sentido inverso. Ou seja, tendo encontrado o ponto $(x_1, ..., x_N)$ podemos repetir o procedimento começando pela coordenada N.

No caso em que há apenas um mínimo segundo cada direcção, podemos mostrar a convergência do método usando o teorema anterior, para isso basta considerar como função de descida a própria função f.

As direcções de descida aqui consideradas são os vectores da base canónica, ou seja, temos $d^{(k)} = \pm \mathbf{e}^{(k)}$, e o valor ω_k será $x_k - y_k$, de acordo com as notações usadas em cima.

Observação: Uma técnica, não fiável, mas que poderá evitar um cálculo exaustivo do mínimo segundo cada direcção consiste em efectuar uma pequena perturbação nas coordenadas. A ideia é fixar um parâmetro $\varepsilon > 0$ e avaliar o menor dos valores entre $f(x_1, .., x_N)$ e

$$f(x_1 \pm \varepsilon, x_2, ..., x_N), ..., \ f(x_1, ..., x_{N-1}, x_N \pm \varepsilon).$$

Repete-se sucessivamente este procedimento, até que o valor de $f(x_1, .., x_N)$ seja o mais pequeno. Nessa altura, deverá diminuir o valor de ε. Voltamos a insistir que esta técnica pode não levar à obtenção de um mínimo, mas poderá ser um processo para inicializar um outro método mais eficaz.

6.3 Exercícios

1. Considere uma função f que toma os seguintes valores:

x	-2	-1	0	1	2
f(x)	-10	-4	-2	1	5

e que tem um único zero $z \in [0, 1]$. Pretende-se aproximar esta função por uma função do tipo $g(x) = a + bx + cx^3$, no sentido dos mínimos quadrados.

a) Mostre que os valores a, b, c verificam:

$$\begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 34 \\ 0 & 34 & 130 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -10 \\ 35 \\ 125 \end{bmatrix}$$

b) Seja $a = -2, \ b = 25/12, \ c = 5/12$ a solução do sistema da alínea anterior. Determine uma aproximação da raiz w do polinómio g em $[0, 1]$ calculando x_2 pelo método de Newton com $x_0 = 1$ (justifique a convergência).

c) Sabendo que $|z - w| < 0.01$, determine um majorante para o erro absoluto que cometemos se utilizarmos o valor x_2 obtido em b) para aproximar z.

2. Pretende-se aproximar a tabela de pontos dada no exercício anterior por uma função do tipo

$$g(x) = ae^{bx}.$$

a) Efectue uma transformação nos pontos de forma a reduzir a aproximação a um problema de mínimos quadrados (linear).

b) Calcule os valores a e b após a resolução do sistema normal e avalie a diferença na norma euclidiana.

c) Efectue a minimização através da derivação em ordem a a e b e compare com os resultados obtidos em b).

3. Pretende-se minimizar a função

$$f(x, y) = x^2 + y^2 + x + y - \sin(xy)/2$$

no conjunto $X = [-1, 1] \times [-1, 1]$

a) Mostre que existe um e um só ponto crítico no conjunto X.

Sugestão: Escreva a equação que permite obter os pontos críticos, e aplique o teorema do ponto fixo.

b) Prove que esse ponto crítico é o mínimo da função em X.

c) Usando o método do ponto fixo, determine uma aproximação para esse mínimo, com um erro absoluto nas componentes inferior a 0.01.

d) Aproxime esse mínimo usando duas iteradas do método do gradiente com $x^{(0)} = 0$.

4. Pretende-se encontrar o mínimo absoluto em $\mathbb{R}$ de

$$f(x) = 1 + |5x - \cos(2x) + 2\sin(x)|$$

a) Mostre que há um único ponto em $\mathbb{R}$ que é mínimo absoluto de f.

b) Determine aproximadamente o valor desse mínimo, de forma a que o erro absoluto seja inferior a 0.01, e determine exactamente o valor da função nesse mínimo.

5. Pretende-se encontrar a função da forma $g(x) = a\exp(x) + b\exp(-x)$ que melhor aproxima a função $f(x) = \exp(x/2)$ no intervalo $[-2, 2]$

a) Para determinar a e b utilize o método dos mínimos quadrados discreto, considerando os pontos $\{-2, -1, 0, 1, 2\}$.

(*Nota: Não esquecer de mostrar que as funções base são linearmente independentes, para esse conjunto de pontos*)

b) Para determinar a e b utilize agora o método dos mínimos quadrados contínuo, considerando todo o intervalo $[-2, 2]$.

c) Compare os valores obtidos nas alíneas anteriores e comente. Comente o condicionamento das matrizes obtidas nas alíneas anteriores

6. Considere uma função f que toma os seguintes valores:

x	-2	-1	0	1	2
f(x)	-10	-4	-2	1	5

e que tem um único zero $z \in [0, 1]$. Pretende-se aproximar esta função por uma função do tipo $g(x) = a + bx + cx^3$, no sentido dos mínimos quadrados.

a) Mostre que os valores a, b, c verificam:

$$\begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 34 \\ 0 & 34 & 130 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -10 \\ 35 \\ 125 \end{bmatrix}$$

b) Seja $a = -2,\ b = 25/12,\ c = 5/12$ a solução do sistema da alínea anterior. Determine uma aproximação da raiz w do polinómio g em $[0, 1]$ calculando x_2 pelo método de Newton com $x_0 = 1$ (justifique a convergência).

c) Sabendo que $|z - w| < 0.01$, determine um majorante para o erro absoluto que cometemos se utilizarmos o valor x_2 obtido em b) para aproximar z.

7. Considere a função $f : \mathbb{R}^3 \to \mathbb{R}$:

$$f(x) = e^{||x||_2^2} + x_1 + x_2 + x_3$$

a) Mostre que a função $h(t) = t/e^{t^2}$ verifica $h(t) < 0.5$. *Sugestão: Determine uma aproximação do valor do máximo positivo.*

b) Mostre que existe um e um só ponto crítico de f em $\mathbb{R}^3$ e que pode ser determinado usando o método iterativo:

$$x^{(n+1)} = -0.5[1 \quad 1 \quad 1]^T \exp(-||x^{(n)}||_2^2)$$

para qualquer $x^{(0)} \in \mathbb{R}^3$. *Sugestão: Usando a desigualdade de Cauchy-Schwarz, mostre que* $||x||_1 \leq \sqrt{3}||x||_2$ *e utilize a alínea a)*

8. Para aproximar o mínimo de $f : \mathbb{R}^3 \to \mathbb{R}$:

$$f(x) = e^{||x||_2^2} + x_1 + x_2 + x_3,$$

calcule duas iteradas, para $x^{(0)} = (-1, 1, 1)$, usando:

a) Método do gradiente sem aproximação.

b) Método do gradiente com aproximação quadrática.

c) Método do gradiente conjugado sem aproximação.

d) Método de Newton.

9. Considere $f : \mathbb{R}^2 \to \mathbb{R}$:

$$f(x, y) = \frac{1}{(x-1)^2 + y^2 + 0.1}.$$

a) Mostre que o ponto de mínimo de f é $(1, 0)$.

b) Calcule duas iteradas, para $x^{(0)} = (-1, -1)$, usando o método do gradiente sem e com aproximação.

c) Compare os valores obtidos com os valores exactos e comente.

10. Suponha que f é uma função em que a matriz hessiana é sempre definida positiva.

a) Mostre que se $\nabla f(z) = 0$, z é o único ponto de mínimo para f.

b) Suponha que sabe o valor de $f(z)$, mas não sabe z. Mostre em que condições o método do gradiente converge para esse ponto de mínimo z, considerando $\omega_n = |f(z) - f(x^{(n)})|$.

Capítulo 7

Anexos

7.1 Resultados Elementares de Análise

7.1.1 Funções de várias variáveis reais

No caso unidimensional, um teorema fundamental que pode ser visto como uma generalização do teorema do valor médio, é a expansão em série de Taylor com resto de Lagrange:

Teorema 7.1 *(Taylor) :Se* $f \in C^{p+1}(]a,b[)$, *então para quaisquer* $x, y \in]a,b[$, *existe* $\xi \in]a,b[$:

$$f(y) = f(x) + f'(x)(y-x) + ... + \frac{1}{p!}f^{(p)}(x)(y-x)^p + \frac{1}{(p+1)!}f^{(p+1)}(\xi)(y-x)^{p+1}.$$

Convém aqui observar que, no caso limite, se $f \in C^{\infty}(]a,b[)$, a série de Taylor

$$\sum_{p\geq 0} \frac{1}{p!}f^{(p)}(x)(y-x)^p,$$

pode não coincidir com a função. O exemplo clássico é a função

$$f(x) = \begin{cases} xe^{x^2} & \text{se } x > 0 \\ 0 & \text{se } x \leq 0, \end{cases}$$

já que embora $f \in C^{\infty}(\mathbb{R})$, f tem todas as derivadas nulas em zero, i.e. $f^{(p)}(0) = 0, \forall p$.

Para que a série de Taylor coincida com a função, é necessário exigir mais que a diferenciabilidade *ad infinitum*, é necessário exigir que a função seja analítica. Uma função analítica em I é aquela que admite uma representação em série de potências nesse intervalo, sendo no fundo uma generalização natural do conceito de polinómio, ou seja a generalização de um conceito algébrico. Isto reflecte bem a diferença entre conceitos algébricos e conceitos da análise. Quando trabalhamos com funções de variável complexa a situação é diferente, e pode-se provar que aí coincidem as noções de diferenciabilidade e analiticidade.

Definição 7.1 *Dizemos que* $F : X \subseteq \mathbb{R}^N \to \mathbb{R}^M$ *é de classe* C^1 *em* X *se as derivadas parciais* $\partial_i F_j = \frac{\partial F_j}{\partial x_i}$ *(para* $i = 1, ..., N,\ \ j = 1, ..., M$*) existirem e forem contínuas em* X*. Ao vector* $\nabla F_j = (\partial_1 F_j, ..., \partial_N F_j)$ *chamamos gradiente de* F_j*, e à matriz (cujas linhas são gradientes)*

$$\nabla F = \begin{bmatrix} \partial_1 F_1 & ... & \partial_N F_1 \\ & \ddots & \\ \partial_1 F_M & & \partial_N F_M \end{bmatrix}$$

chamamos matriz jacobiana *de* F*.*

De forma análoga, dizemos que F é de classe C^p em X se as derivadas parciais $\partial_i^\alpha F_j$ para $|\alpha| \leq p$, existirem e forem contínuas em X. Aqui α designa um multi-índice, ou seja $\alpha = (i_1, ..., i_N) \in \mathbb{N}^N$, e $|\alpha| = i_1 + ... + i_N$.

Seja $f : X \subseteq \mathbb{R}^N \to \mathbb{R}$. Interessa-nos ainda recordar a noção de matriz Hessiana de uma função f com valores reais

$$\boldsymbol{\nabla}^2 f = \begin{bmatrix} \partial_{11} f & ... & \partial_{1N} f \\ & \ddots & \\ \partial_{N1} f & & \partial_{NN} f \end{bmatrix},$$

que intervém na expansão em série de Taylor de ordem 2.

Teorema 7.2 *(expansão de Taylor de ordem 2) Seja* $f \in C^2(X)$ *uma função definida num convexo* X *com valores reais.*
Então, para quaisquer $x, x + h \in X$*, existe* $\theta \in]0, 1[$*:*

$$f(x + h) = f(x) + \nabla f(x)h + h^\top \boldsymbol{\nabla}^2 f(x + \theta h)h$$

É claro que a expansão em série de Taylor pode ser generalizada para ordens superiores. A maneira mais conveniente de o fazer é considerar formas diferenciais. No entanto, no âmbito do curso esta expansão será suficiente.

• Referimos também um resultado em que apenas se assegura a existência de ponto fixo, não se retirando informação acerca da unicidade nem da convergência, pois não se trata de um teorema construtivo. É apenas válido (nesta forma) em $\mathbb{R}^N$. A referência a este resultado é importante já que em algumas situações não é praticável a demonstração de existência de um ponto fixo de outra forma.

Teorema 7.3 **(***do Ponto Fixo de Brouwer***).** *Seja* $\bar{B}(0, 1)$ *uma bola unitária em* $\mathbb{R}^N$*. Uma função contínua* $G : \bar{B}(0, 1) \to \bar{B}(0, 1)$ *tem pelo menos um ponto fixo em* $.\bar{B}(0, 1)$*.*

Demonstração: Sai fora do âmbito do curso. Note-se que a demonstração não é construtiva e portanto não fornece um método numérico[1].■

[1]Apesar de Brouwer ser um acérrimo defensor do construtivismo, grande parte do seu trabalho mais conhecido usa argumentos não construtivistas, como é este o caso.

Teorema 7.4 *Seja X um conjunto fechado homeomorfo à bola $\bar{B}(0,1)$, e $G : X \to X$ uma aplicação contínua. Então existe pelo menos um ponto fixo de G em X.*

Demonstração: Basta usar a definição de homeomorfo: X é homeormorfo a Y se existir uma aplicação contínua $F : X \to Y$ bijectiva e com inversa contínua. Nesse caso consideramos $Y = \bar{B}(0,1)$, e temos $H = FGF^{-1} : \bar{B}(0,1) \to \bar{B}(0,1)$ que é contínua por hipótese, logo pelo Teorema de Brouwer, existe pelo menos um $z : Hz = z$, o que é equivalente a $GF^{-1}z = F^{-1}z$, logo $F^{-1}z$ é um ponto fixo de G. ■

Observação: Um *estrelado* é um conjunto $X = \{x \in \mathbb{R}^d : x = a + \lambda r(\hat{x})\hat{x}$, com $\lambda \in [0,1]$, $\hat{x} = \frac{x}{||x||} \in S^2\}$ (onde S^2 designa a esfera unitária $\{x \in \mathbb{R}^d : ||x|| = 1\}$), e r é uma função contínua em S^2). Todos os estrelados, e em particular os convexos, são homeomorfos à bola unitária, porque podemos definir a aplicação $f : \bar{B}(0,1) \to X$, por $f(x) = a + r(\hat{x})x$, que é uma aplicação bijectiva contínua de inversa $f^{-1}(y) = y - a$. O resultado aparece normalmente enunciado apenas para convexos.

• Um outro resultado que é referido no texto, permite concluir a existência de pontos de máximo e mínimo num conjunto compacto.

Teorema 7.5 *(Weierstrass). Uma função contínua $f : X \to \mathbb{R}$ definida num compacto*[2] *X de $\mathbb{R}^N$ admite um ponto de mínimo e um ponto de máximo em X.*

7.1.2 Funções de uma variável complexa

Definição 7.2 *Uma função $f(z) = u(z) + \mathbf{i}v(z)$ é diferenciável em $z = x + \mathbf{i}y \in \mathbb{C}$, se se verificarem as condições de Cauchy-Riemann*

$$\begin{cases} \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \\ \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \end{cases}$$

Estas condições podem ser resumidas na expressão $\frac{df}{d\bar{z}} = 0$, fazendo a mudança de variáveis $x = \frac{1}{2}(z + \bar{z})$, $y = \frac{1}{2\mathbf{i}}(z - \bar{z})$. Com efeito,

$$\frac{df}{d\bar{z}} = \frac{1}{2}(\frac{\partial u}{\partial x} + \mathbf{i}\frac{\partial u}{\partial y}) + \frac{1}{2}(\mathbf{i}\frac{\partial v}{\partial x} - \frac{\partial v}{\partial y})$$

e a condição $\frac{df}{d\bar{z}} = 0$ é equivalente a $\frac{1}{2}(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}) = 0, \frac{1}{2}\mathbf{i}(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}) = 0$, ou seja, às condições de Cauchy-Riemann. Um polinómio verifica trivialmente as condições de Cauchy-Riemann em $\mathbb{C}$, visto $\frac{dz^k}{d\bar{z}} = 0$, no entanto, a função $f(z) = \bar{z}$ não é diferenciável, pois $\frac{d\bar{z}}{d\bar{z}} = 1 \neq 0$, o

[2]Em espaços vectoriais de dimensão *finita* os conjuntos compactos são limitados e fechados. Deve notar-se que em espaços de dimensão infinita isto não é verdade, já que por exemplo, em l^∞ a bola $\bar{B}(0,1)$ sendo limitada e fechada não é compacta. Basta pensar que da sucessão definida pelos elementos da base não é possível extrair uma subsucessão convergente!

mesmo acontecendo com $f(z) = Re(z)$, pois $Re(z) = \frac{1}{2}(z+\bar{z})$, e assim $\frac{d}{d\bar{z}}[\frac{1}{2}(z+\bar{z})] = \frac{1}{2} \neq 0$. Por outro lado, analogamente,

$$\frac{df}{dz} = \frac{1}{2}(\frac{\partial u}{\partial x} - \mathbf{i}\frac{\partial u}{\partial y}) + \frac{1}{2}(\mathbf{i}\frac{\partial v}{\partial x} + \frac{\partial v}{\partial y}),$$

e quando a função é diferenciável, as condições de Cauchy-Riemann dão $\frac{df}{dz} = \frac{1}{2}(\frac{\partial u}{\partial x} + \mathbf{i}\frac{\partial v}{\partial x}) + \frac{1}{2}(\mathbf{i}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial x})$, ou seja, a derivada fica $f'(z) = \frac{\partial f}{\partial x}(z)$, ou equivalentemente $f'(z) = -\mathbf{i}\frac{\partial f}{\partial y}(z)$.

Entende-se por *função analítica* num aberto X, uma função que admite expansão em série de potências em torno de qualquer $x \in X$ que é válida numa vizinhança desse ponto. Há que distinguir a analiticidade em $\mathbb{C}$ de duas maneiras, já que a identificação topológica entre $\mathbb{C}$ e $\mathbb{R}^2$ pode prestar-se a confusões. Assim, uma função analítica em $\mathbb{R}^2$, em que as potências são potências de cada uma das variáveis, pode não ser analítica em $\mathbb{C}$, já que nesse caso as potências resultam apenas de multiplicação complexa com uma variável! O recíproco é verdadeiro.

Teorema 7.6 *As funções de variável complexa, diferenciáveis num aberto $X \subseteq \mathbb{C}$, são analíticas em X.*

A soma, o produto e a composição de funções analíticas origina ainda funções analíticas. A divisão origina também uma função analítica excepto nos pontos em que o denominador se anula. Uma outra designação para uma função complexa analítica é a de *função holomorfa*.

Caso uma função seja holomorfa em todo o $\mathbb{C}$, dizemos também que se trata de uma função *inteira*. Os polinómios são funções inteiras e, para além disso, as funções habituais

$$e^z = e^x e^{\mathbf{i}y} = e^x(\cos(y) + \mathbf{i}\sin(y)),\ \sin(z) = \frac{e^{\mathbf{i}z} - e^{-\mathbf{i}z}}{2\mathbf{i}},$$
$$\cos(z) = \frac{e^{\mathbf{i}z} + e^{-\mathbf{i}z}}{2},\ \sinh(z) = \frac{e^z - e^{-z}}{2},\ \cosh(z) = \frac{e^z + e^{-z}}{2}$$

são também funções inteiras.

Especial cuidado deve-se ter com a potenciação em geral. Isto deve-se ao logaritmo estar apenas bem definido em certos *ramos*. Com efeito, sendo

$$z^w = e^{w\log(z)},$$

a potenciação fica dependente da boa definição do logaritmo. Ora como o logaritmo é a função inversa da exponencial, verificamos imediatamente que a exponencial não é uma função injectiva em $\mathbb{C}$, pois $e^{x+\mathbf{i}y} = e^{x+\mathbf{i}(y+2k\pi)}$. Isto deve-se ao facto de senos e co-senos não serem funções injectivas em $\mathbb{R}$. No entanto, se nos restringirmos a um intervalo em particular, como $]-\pi, \pi]$ ou $[0, 2\pi[$, já podemos assegurar injectividade. São este tipo de intervalos que definem os *ramos*.

Outro aspecto especial aparece nas divisões. Quando um denominador se anula podemos ter três tipos de singularidades a saber, removíveis, isoladas ou essenciais. Com efeito, devemos considerar um teorema que surge como uma generalização da expansão em série de Taylor.

Teorema 7.7 *(Laurent). Se f é analítica na coroa $A = \{z : r_1 < |z - z_0| < r_2\}$, então para qualquer $z \in A$,*

$$f(z) = \sum_{k \in \mathbf{Z}} a_k (z - z_0)^k.$$

Os coeficientes a_k são dados por

$$a_k = \frac{1}{2\pi i} \int_{|z-z_0|=r} \frac{f(z)}{(z - z_0)^{k+1}}\, dz,$$

para qualquer r entre r_1 e r_2.

Se $a_k = 0$, para $k < 0$, a função é analítica. Se existir $k < 0$ tal que $a_k \neq 0$, então f tem uma singularidade em z_0, que se designa *essencial* se houver uma infinidade de $k < 0$ para os quais $a_k \neq 0$, e que se designa *pólo* no caso contrário.

Teorema 7.8 *(Liouville). As únicas funções inteiras limitadas são as constantes.*

Podemos ainda enunciar um resultado útil, que permite boas estimativas para a localização de raízes, e cuja verificação das condições é simples.

Teorema 7.9 *(Rouché) Sejam f e ϕ funções analíticas num domínio D tais que*

$$|f(x) - \phi(x)| \leq |\phi(x)|,\ \forall x \in \partial D$$

e ϕ não se anula em ∂D. Então f e ϕ têm o mesmo número de zeros em D. □

Exemplo: Consideramos $D = B(0, \frac{3}{2})$, $f(x) = x^4 + x^3 + 1$, e $\phi(x) = x^4$ (que tem quatro zeros em D e não se anula na fronteira). Como

$$|f(x) - \phi(x)| = |x^3 + 1| \leq |x|^3 + 1,$$

quando $|x| = \frac{3}{2}$ obtemos $\frac{27}{8} + 1 \leq \frac{81}{16}$, e concluímos que há 4 zeros de f em D.

7.2 Equações às Diferenças

Entendemos por equação às diferenças de ordem $p+1$, uma equação cuja incógnita é uma sucessão (x_n) :

$$x_{n+p+1} + a_p x_{n+p} + \ldots + a_0 x_n = b_n, \ (n \in \mathbb{N}_0) \tag{7.1}$$

e em que, no caso mais simples, os coeficientes $a_0, \ldots, a_k$ são constantes reais[3]. Este caso – o de equações lineares com coeficientes constantes – será o único caso que iremos tratar aqui.

Diremos que a equação às diferenças (7.1) é *homogénea* se $b_n \equiv 0$. Os valores da sucessão ficam determinados de forma única a partir de valores iniciais atribuídos a $x_0, ..., x_p$ através da recorrência

$$x_{n+p+1} = b_n - a_p x_{n+p} - \ldots - a_0 x_n,$$

ou seja, baseando-nos nesta construtividade é imediata a prova de existência e unicidade de solução de uma equação às diferenças (7.1).

Proposição 7.1 *Dados os valores iniciais $x_0, ..., x_p$ existe uma e uma só sucessão que é solução da equação às diferenças (7.1).*

Exemplo 7.1 *Como exemplo, podemos considerar a sucessão de Fibonacci, que é solução da equação às diferenças homogénea*

$$x_{n+2} - x_{n+1} - x_n = 0 \tag{7.2}$$

com condições iniciais $x_0 = 0$, $x_1 = 1$. Através de $x_{n+2} = x_{n+1} + x_n$ podemos determinar recursivamente

$$x_2 = 1, \ x_3 = 2, \ x_4 = 3, \ x_5 = 5, \ x_6 = 8, \ x_7 = 13, \ \ etc...$$

[3]Nesta matéria iremos encontrar uma semelhança com a teoria das equações diferenciais ordinárias, que podemos perceber melhor se encararmos estas equações às diferenças como resultado final de uma discretização. Assim, pensemos em discretizar uma equação diferencial

$$x''(t) + ax'(t) + bx(t) = 0$$

num intervalo $[a, b]$, usando N pontos igualmente espaçados $t_n = a + nh$, com $h = \frac{b-a}{N}$.

Ao considerar uma aproximação das derivadas na seguinte forma

$$x'(t_n) \sim \frac{x(t_n + h) - x(t_n)}{h} = \frac{x_{n+1} - x_n}{h}$$

e

$$x''(t_n) \sim \frac{x_{n+1} - 2x_n + x_{n-1}}{h^2}$$

obtemos

$$\frac{x_{n+1} - 2x_n + x_{n-1}}{h^2} + a\frac{x_{n+1} - x_n}{h} + bx_n = 0$$

ou seja, uma equação às diferenças... (neste caso trata-se de uma aproximação por diferenças finitas, que será apenas estudada em detalhe em Análise Numérica II).

No entanto, estes valores também podem ser obtidos usando a fórmula

$$x_n = \frac{1}{\sqrt{5}}\left((\frac{1+\sqrt{5}}{2})^n - (\frac{1-\sqrt{5}}{2})^n\right). \tag{7.3}$$

A questão que se coloca é a de saber como é possível obter, a partir de uma equação suficiente geral como (7.1), uma solução simplificada tal como apresentámos (7.3) para a equação (7.2), mas com quaisquer condições iniciais.

Não especificando as condições iniciais, começamos por observar que qualquer solução da equação (7.1) é dada através da soma de uma solução particular com uma solução da equação homogénea.

Proposição 7.2 *Seja (x_n) uma solução da equação (7.1) para certas condições iniciais. Qualquer solução (y_n) de (7.1) para outras condições iniciais resulta da soma de (x_n) com uma certa solução da equação homogénea associada.*

Demonstração: Segundo as hipóteses consideradas, temos

$$x_{n+p+1} + a_p x_{n+p} + \ldots + a_0 x_n = b_n$$

e

$$y_{n+p+1} + a_p y_{n+p} + \ldots + a_0 y_n = b_n,$$

para $n \geq 0$. É óbvio que se subtrairmos as duas equações obtemos

$$(y_{n+p+1} - x_{n+p+1}) + a_p(y_{n+p} - x_{n+p}) + \ldots + a_0(y_n - x_n) = 0$$

e portanto (y_n) resulta da soma da solução particular (x_n) com $(y_n - x_n)$ que é solução da equação homogénea associada a (7.1). □

Proposição 7.3 *As soluções de uma equação às diferenças homogénea de ordem $p+1$ formam um subespaço vectorial do espaço das sucessões cuja dimensão é $p+1$.*

Demonstração: Basta reparar que se (x_n) e (y_n) são soluções da equação às diferenças homogénea, ou seja,

$$x_{n+p+1} + a_p x_{n+p} + \ldots + a_0 x_n = 0, \quad y_{n+p+1} + a_p y_{n+p} + \ldots + a_0 y_n = 0,$$

então para quaisquer α, $\beta \in \mathbb{R}$ a sucessão $(\alpha x_n + \beta y_n)$ também é solução. Fica assim provado que se trata de um subespaço vectorial $\mathbf{S}$. O facto de ter dimensão $p+1$ resulta de haver $p+1$ graus de liberdade que são determinados pelos $p+1$ valores iniciais $x_0, ..., x_p$. Com efeito, como as soluções são determinadas de forma única a partir desses valores iniciais, resulta que a cada vector $(x_0, ..., x_p) \in \mathbb{R}^{p+1}$ corresponde uma e uma só solução, estabelecendo-se uma transformação bijectiva entre $\mathbb{R}^{p+1}$ e o subespaço vectorial das soluções definido por

$$\begin{array}{rccc} T: & \mathbb{R}^{p+1} & \longrightarrow & \mathbf{S} \\ & (x_0, ..., x_p) & \longmapsto & (x_n) \end{array}$$

que é um isomorfismo, porque T também é linear (exercício). □

7.2.1 Soluções de uma Equação às Diferenças Homogénea

Comecemos por ver o caso trivial em que temos uma equação às diferenças homogénea de ordem 1:

$$x_{n+1} + ax_n = 0$$

a sua solução é obviamente

$$x_n = x_0(-a)^n.$$

Se considerarmos agora a equação às diferenças homogénea de ordem 2:

$$x_{n+2} + a_1 x_{n+1} + a_0 x_n = 0 \tag{7.4}$$

em que substituímos a variável, i.e: a sucessão (x_n), por uma sucessão particular

$$x_n = z^n$$

em que z é um qualquer número complexo, obtemos

$$z^n(z^2 + a_1 z + a_0) = 0.$$

Temos soluções não triviais para a equação homogénea se resolvermos a equação do segundo grau:

$$z^2 + a_1 z + a_0 = 0,$$

que sabemos ter duas raízes complexas, que designaremos por z_1 e z_2. Assim, se $z_1 \neq z_2$, as soluções da equação homogénea (7.4) serão as sucessões da forma z_1^n e z_2^n, ou combinações lineares delas:

$$x_n = c_1 z_1^n + c_2 z_2^n.$$

Se a multiplicidade da raiz for dupla, ou seja, $z_1 = z_2 = z$, então para além da sucessão z^n, necessitamos de uma outra sucessão podemos encontrar uma outra solução $x_n = nz^n$ (o que não é difícil de verificar, pois nesses casos $z = -a_1/2$), e assim a solução geral virá

$$x_n = c_1 z^n + c_2 n z^n.$$

De uma forma geral temos o seguinte resultado.

Teorema 7.10 *As soluções da equação às diferenças homogénea de ordem $p+1$:*

$$x_{n+p+1} + a_p x_{n+p} + \ldots + a_0 x_n = 0$$

são obtidas resolvendo a equação característica

$$z^{p+1} + a_p z^p + \ldots + a_1 z + a_0 = 0.$$

Se as $p+1$ raízes $z_0, z_1, \ldots, z_p$ desta equação polinomial forem distintas, uma solução geral (x_n) é dada pelas combinações lineares das soluções $(z_i)^n$

$$x_n = c_0(z_0)^n + c_1(z_1)^n + \ldots + c_p(z_p)^n.$$

Caso haja raízes múltiplas, se, por exemplo, z_j for uma raiz de multiplicidade m, então devemos considerar combinações lineares não apenas dos $(z_i)^n$, em que $i \neq j$, mas também de

$$z_j^n,\ nz_j^n, \ldots,\ n^m z_j^n.$$

Exemplo 7.2 *Consideremos a sucessão definida por* $x_0 = 0, x_1 = 0, x_2 = 1$,*e recursivamente por*

$$x_{n+3} + x_{n+2} - x_{n+1} - x_n = 0.$$

O objectivo é determinar uma expressão explícita que nos dê, por exemplo, o termo x_{378} *sem ter que calcular recursivamente os 375 termos anteriores!! A equação característica associada à equação às diferenças é*

$$z^3 + z^2 - z - 1 = 0$$

que tem como raízes 1 *(raiz simples) e* -1 *(raiz dupla), portanto a solução geral será*

$$x_n = c_0 + c_1(-1)^n + c_2 n(-1)^n.$$

Tendo considerado como valores iniciais, $x_0 = 0, x_1 = 0, x_2 = 1$, *podemos obter um sistema que dá a solução particular neste caso*

$$\left\{ \begin{array}{c} c_0 + c_1 = x_0 = 0 \\ c_0 - c_1 - c_2 = x_1 = 0 \\ c_0 + c_1 + 2c_2 = x_2 = 1 \end{array} \right. \quad \left\{ \begin{array}{c} c_0 + c_1 = 0 \\ c_0 - c_1 = 1/2 \\ c_2 = 1/2 \end{array} \right.$$

ou seja, $c_0 = 1/4, c_1 = -1/4, c_2 = 1/2,$ *e portanto*

$$x_n = \frac{1}{4}(1 - (-1)^n + 2n(-1)^n).$$

Agora é fácil estabelecer que $x_{378} = (1 - 1 + 2 \cdot 378)/4 = 189$.

7.2.2 Equações às diferenças não homogéneas

Consideremos agora o caso em que $b_n \neq 0$. Já vimos que basta encontrar uma solução particular de (7.1) e a solução geral da equação homogénea associada, para obtermos qualquer solução de (7.1) que será a soma das duas.

Para encontrar uma solução particular de (7.1), podemos recorrer a dois processos, o primeiro intuitivo, e que consiste em tentar descobrir uma possível solução através do aspecto do segundo membro, ou seja de b_n. Assim, por exemplo, se tivermos uma equação do tipo

$$x_{n+2} + ax_{n+1} + bx_n = A + Bn,$$

procuramos soluções na forma $x_n = \alpha + \beta n$, e obtemos

$$\alpha(1 + a + b) + \beta(2 + a) + \beta(1 + a + b)n = A + Bn$$

$$\left[\begin{array}{cc} 1 + a + b & 2 + a \\ 0 & 1 + a + b \end{array} \right] \left[\begin{array}{c} \alpha \\ \beta \end{array} \right] = \left[\begin{array}{c} A \\ B \end{array} \right]. \tag{7.5}$$

que só não possui solução se $a + b = -1$.

Exemplo 7.3 *1) Num caso concreto,*

$$x_{n+2} - x_{n+1} + x_n = 2n + 2,$$

facilmente podemos obter que $x_n = 2n$ *é uma solução particular, e portanto, a partir da solução da equação homogénea, a solução geral fica*

$$x_n = 2n + c_1(\frac{1 - i\sqrt{3}}{2})^n + c_2(\frac{1 + i\sqrt{3}}{2})^n.$$

Se exigirmos ainda $x_0 = 0, x_1 = 1$, *vamos obter* $c_1 = \frac{i}{\sqrt{3}}$, $c_2 = -\frac{i}{\sqrt{3}}$. *Apesar de estarem envolvidos valores complexos, o resultado final será sempre um número inteiro!*
2) Outros exemplos levam a outras escolhas, por exemplo, num outro caso:

$$x_{n+2} - x_{n+1} + x_n = 2^n,$$

escolhemos soluções do tipo $x_n = \alpha 2^n$ *e obtemos*

$$(2^2\alpha - 2\alpha + \alpha)2^n = 2^n,$$

o que nos dá $\alpha = 1/3$, *e portanto a solução geral será*

$$x_n = \frac{1}{3}(2^n) + c_1(\frac{1 - i\sqrt{3}}{2})^n + c_2(\frac{1 + i\sqrt{3}}{2})^n.$$

No entanto, este método baseia-se numa certa intuição à qual não podemos recorrer em casos mais complicados. Temos porém um método que é similar ao que é usado em equações diferenciais ordinárias – o método da variação de constantes.

7.2.3 Método da variação de constantes

Dado que a equação homogénea de ordem p tem p soluções distintas (linearmente independentes), associadas às raizes, chamemos a essas soluções

$$x^{(1)}, \ldots, x^{(p)}$$

que se tratam de sucessões. O facto de serem linearmente independentes está directamente relacionado com o facto da sucessão dos determinantes

$$w_n = \det \begin{bmatrix} x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(p)} \\ x_{n-1}^{(1)} & x_{n-1}^{(2)} & \cdots & x_{n-1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-p+1}^{(1)} & x_{n-p+1}^{(2)} & \cdots & x_{n-p+1}^{(p)} \end{bmatrix}$$

nunca ser nula. A esta sucessão de determinantes chamamos Wronskiano (em analogia ao caso das equações diferenciais ordinárias). O Wronskiano permite construir uma solução particular para uma sucessão (b_n) qualquer (e.g. [15])

$$x_n = \sum_{i=0}^{n} \frac{b_i}{w_i} \det \begin{bmatrix} x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(p)} \\ x_{i-1}^{(1)} & x_{i-1}^{(2)} & \cdots & x_{i-1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i-p+1}^{(1)} & x_{i-p+1}^{(2)} & \cdots & x_{i-p+1}^{(p)} \end{bmatrix}$$

(...estamos a supor aqui que as condições iniciais são dadas em $x_{-p+1}, ..., x_0$, para que o somatório comece em 0, senão deveriamos começar o somatório em $p-1$).

Exemplo 7.4 *Consideremos a equação não homogénea*

$$x_n - 2x_{n-1} + x_{n-2} = 1.$$

Neste caso, seríamos tentados a escolher $x_n = a+bn$ para descobrir uma solução particular, mas isso dá-nos a situação $a+b = -2+1 = -1$ em que o sistema (7.5) não tem solução. Através do método de variação de constantes podemos obter uma solução. Começando por reparar que a equação homogénea tem como soluções independentes

$$x_n^{(1)} = 1, \ e \ x_n^{(2)} = n,$$

pois a equação caracteristica tem 1 como raiz dupla, obtemos como Wronskiano

$$w_n = \det \begin{bmatrix} 1 & n \\ 1 & n-1 \end{bmatrix} = -1.$$

Assim, pelo método de variação de constantes temos

$$x_n = \sum_{i=0}^{n} \frac{1}{-1} \det \begin{bmatrix} 1 & n \\ 1 & i-1 \end{bmatrix} = -\sum_{i=0}^{n} (i-1-n) = \frac{(n+1)(n+2)}{2}$$

e a solução geral será então dada por

$$x_n = \frac{(n+1)(n+2)}{2} + C_1 + C_2 n.$$

Se $x_0 = 0$, $x_1 = 0$, as constantes seriam $C_1 = -1, C_2 = -2$.

7.2.4 Exercícios

1. Considere a sucessão definida por $x_{n+2} = \frac{9}{2}x_{n+1} - 2x_n$em que $x_0 = 2$ e $x_1 = 1$. Mostre que x_n converge para zero.

2. O processo de construção de uma estrutura com 10 colunas obedece à seguinte regra: – A altura a_k de uma coluna k deve ser igual à seguinte média ponderada de alturas de colunas mais próximas $a_k = \frac{1}{2}(a_{k-1} - a_{k+1}) + a_{k-2}$.

Tendo já sido construídas as duas primeiras colunas, com alturas $a_1 = 4, \ a_2 = 8$, bem como a última, com $a_{10} = 180$, determine qual a altura da coluna a_5.

3. Considere a sucessão $x_0 = 1$, $x_1 = \frac{4}{9}$, com $x_{n+1} = \frac{8104}{9}x_n - 400x_{n-1}$. Introduzindo o valor de x_1 numericamente, ao fim de algumas iterações obtemos resultados para a sucessão que nada têm a ver com a solução $x_n = (\frac{4}{9})^n$. Comente e justifique os resultados quanto à estabilidade numérica.

7.3 Teoria de Erros em Espaços Normados

Começamos por generalizar os conceitos elementares de erros para espaços normados.

7.3.1 Erro, Absoluto e Relativo

A generalização natural das noções de erro, é feita através da norma.

Definição 7.3 . *Seja E um espaço normado com norma $||.||$ e $\tilde{x} \in E$ um valor aproximado de $x \in E$. Definimos:*

- Erro : $e_x = x - \tilde{x}$
- Erro Absoluto : $||e_x|| = ||x - \tilde{x}||$
- Erro Relativo : Se $x \neq 0$, definimos $\delta_x = \frac{x-\tilde{x}}{||x||}$, e portanto $||\delta_x|| = \frac{||x-\tilde{x}||}{||x||}$.

É preciso ter em mente que mesmo em $\mathbb{R}^N$ podemos ter medições de erro diferentes, consoante a norma utilizada. Assim se o valor exacto for $(\pi, \sqrt{2}, 1)$ e tivermos obtido como valor aproximado $(3.14, 1.4, 1.01)$, o erro absoluto com a norma $||.||_\infty$ é igual a

$$||e||_\infty = \max\{|\pi - 3.14|, |\sqrt{2} - 1.5|, |1 - 1.001|\} = 0.0857864...$$

enquanto se considerarmos a norma $||.||_1$ já obtemos um valor diferente,

$$||e||_1 = |\pi - 3.14| + |\sqrt{2} - 1.5| + |1 - 1.001| = 0.0883791...$$

e o mesmo se passa com os erros relativos.

7.3.2 Propagação de Erros

Se tivermos um ponto $\tilde{x}$ que aproxima x, ao calcularmos a imagem por um operador Fréchet diferenciável A num conjunto X que contém os pontos $\tilde{x}, x$, vamos obter um valor aproximado $A\tilde{x}$ que será diferente do valor Ax. Para controlarmos o erro que se propaga ao aplicarmos esta função, usamos a própria definição,

$$e_{Ax} = Ax - A\tilde{x} = A'_x e_x + o(e_x)$$

quando $||e_x||$ tende para zero. Desta forma, desprezando o termo $o(e_x)$, podemos definir

$$\tilde{e}_A x = A'_x e_x$$

e para o erro relativo, quando $Ax \neq 0$, obtemos

$$||\tilde{\delta}_A x|| = \frac{||\tilde{e}_A x||}{||Ax||} = \frac{||A'_x e_x||}{||Ax||} \leq \frac{||x||\,||A'_x||}{||Ax||}||\delta_x||$$

Observação: Se A for um operador linear contínuo, bijectivo, com inversa contínua, então reencontramos o número de condição $||A^{-1}||\ ||A||$,

$$||\tilde{\delta}_A x|| \leq ||A^{-1}||\ ||A||\ ||\delta_x||,$$

porque, como já referimos a derivada de Fréchet de um operador linear é o próprio operador, e por outro lado,

$$||A^{-1}|| = \sup_{y\neq 0} \frac{||A^{-1}y||}{||y||} = \sup_{Ax\neq 0} \frac{||A^{-1}Ax||}{||Ax||} = \sup_{Ax\neq 0} \frac{||x||}{||Ax||}.$$

Exemplo 7.5 *Como exemplo, podemos tentar avaliar o comportamento de um integral* $\int_a^b f(x)dx$ *através da variação da função* f. *Definimos*

$$Af = \int_a^b f(x)dx,$$

notando que $A : C([a,b]) \to R$. *Como o integral é linear já vimos que* $A'_f = A$, *e como*

$$||A||_\infty = \sup_{f\neq 0} \frac{|Af|}{||f||_\infty} \leq \sup_{f\neq 0} \frac{|Af|}{||f||_\infty} = |b-a|,$$

temos

$$|\tilde{\delta}_A f| \leq \frac{||f||_\infty |b-a|}{|Af|} ||\delta_f||_\infty.$$

Assim, se $f(x) = x^2+1$, *e* $[a,b] = [-1,1]$, *temos* $|\tilde{\delta}_A| \leq \frac{2\cdot 2}{8/3}||\delta_f||_\infty = \frac{3}{2}||\delta_f||_\infty$. *Assim se aproximarmos* $f(x)$ *por* $\frac{3}{2}$ *nesse intervalo, estimamos que o erro relativo do integral não será maior que* $\frac{3}{8}$, *porque* $||\delta_f||_\infty = \frac{||x^2+1-3/2||_\infty}{||x^2+1||_\infty} \leq \frac{1/2}{2} = \frac{1}{4}$. *Com efeito,* $|\delta_A| = \frac{|8/3-3|}{8/3} = \frac{1}{8}$.

• Propagação de erros em espaços de dimensão finita

Na prática é este o caso que nos interessa, já que num algoritmo acabaremos sempre por trabalhar em espaços de dimensão finita.

Se pensarmos em $\mathbb{R}^N$, a derivada de Fréchet pode ser identificada à matriz Jacobiana, e portanto para uma qualquer função $f : \mathbb{R}^N \to \mathbb{R}^M$, temos

$$\tilde{e}_f(x) = \nabla f(x) e_x,$$

o que é coerente com os resultados obtidos para operações, que podem ser vistas como funções $\phi : \mathbb{R}^2 \to \mathbb{R}$. Nesse caso a matriz Jacobiana não é mais que um vector com duas componentes, normalmente designado por gradiente, ou seja $\nabla f(x) = [\frac{\partial\phi}{\partial x}\ \frac{\partial\phi}{\partial y}]$ e temos

$$\tilde{e}_\phi(x) = [\frac{\partial\phi}{\partial x}\ \frac{\partial\phi}{\partial y}] \begin{bmatrix} e_x \\ e_y \end{bmatrix} = \frac{\partial\phi}{\partial x}(x,y)e_x + \frac{\partial\phi}{\partial y}(x,y)e_y.$$

Para o erro relativo podemos estabelecer resultados semelhantes,

$$\tilde{\delta}_{f_i}(x) = \frac{\nabla f_i(x) e_x}{f_i(x)},$$

relembrando que $e_x = (x_1\delta_{x_1}, ..., x_N\delta_{x_N})$, o que no caso de $\phi : \mathbb{R}^2 \to \mathbb{R}$ dá

$$\tilde{\delta}_\phi(x,y) = \frac{x\frac{\partial\phi}{\partial x}(x,y)}{\phi(x,y)}\delta_x + \frac{y\frac{\partial\phi}{\partial y}(x,y)}{\phi(x,y)}\delta_y$$

Outra possibilidade é estabelecer uma desigualdade em termos de normas,

$$||\tilde{\delta}_f(x)|| = \frac{||\tilde{e}_f(x)||}{||f(x)||} = \frac{||\nabla f(x) e_x||}{||f(x)||} \leq \frac{||\nabla f(x)||\,||x||}{||f(x)||}||\delta_x||.$$

Exemplo 7.6 *Assim, quando tivermos uma rotina que dependa de várias variáveis que podem estar afectadas de erro, por exemplo, ao calcularmos uma rotina*[4]

```
AproxInteg[Exp[-c x^2],{x,a,b}]
```

os valores de a, b, c podem vir afectados de um erro que pode condicionar o resultado. Encarando esta rotina como uma função regular $\phi : \mathbb{R}^3 \to \mathbb{R}$, bastará obter valores para $\frac{\partial\phi}{\partial a}$, $\frac{\partial\phi}{\partial b}$, $\frac{\partial\phi}{\partial c}$, para que possamos ter uma ideia dos erros relativos envolvidos. Mas como é claro, o cálculo destas derivadas parciais para além de nem sempre ser fácil, implicaria um conhecimento exacto da rotina AproxInteg(... neste caso concreto isso até seria fácil, porque a rotina em causa reflecte o cálculo aproximado de um integral paramétrico, cuja derivação seria elementar).
Para contornar na prática este tipo de problemas, podemos ser levados a recorrer a soluções menos correctas do ponto de vista teórico. Ou seja, podemos procurar uma aproximação numérica da derivada, em torno dos pontos a, b, c.
Num caso concreto, suponhamos que queremos ver qual o comportamento da rotina para valores de $a \sim -1, b \sim 1, c \sim 1$.
Calculando $\frac{1}{\epsilon}(\phi(-1+\epsilon, 1, 1) - \phi(-1, 1, 1))$, para aproximar $\frac{\partial\phi}{\partial a}$, usando $\epsilon = 0.001$ obtemos $\frac{\partial\phi}{\partial a} \sim -0.368247$, e de forma semelhante $\frac{\partial\phi}{\partial b} \sim 0.367512$, $\frac{\partial\phi}{\partial c} \sim -0.378844$. Por outro lado vemos que $\phi(-1, 1, 1) \sim 1.49365$. Assim, usando a norma do máximo $||J_\phi(-1, 1, 1)||_\infty \sim 0.378844$, e obtemos

$$|\delta_\phi(-1, 1, 1)| \leq \frac{0.378844}{1.49365}||e_{(-1,1,1)}||_\infty.$$

Se calcularmos agora $\phi(-1.03, 1.01, 1.01)$, como $||e_{(-1,1,1)}||_\infty = 0.03$, não devemos ter um erro relativo superior a 0.00763. Com efeito, se calcularmos o valor, vemos que dá 1.50407, que corresponderá a um erro relativo de 0.006976, dentro dos limites estimados.
Poderá haver algumas diferenças, que dependerão, entre outros factores, da proximidade dos coeficientes aproximados face aos exactos, ou ainda da norma escolhida, mas os princípios gerais mantêm-se presentes.

[4]O nome AproxInteg é apenas ilustrativo, só nos interessa saber que a rotina é regular (F-diferenciável) para a abordagem numérica, de índole experimental, que expômos neste exemplo. Com efeito, a rotina usada foi NIntegrate, que aproxima o valor de um integral num intervalo.

7.4 Operadores integrais e matrizes

As matrizes são os operadores lineares em dimensão finita e podem ser vistas, nalguns casos como discretizações de operadores integrais lineares. Sem querer entrar em detalhes acerca do assunto, que é suficientemente complicado para uma abordagem rigorosa, fora do âmbito introdutório do curso, iremos apresentar uma motivação para estudos posteriores.

Uma possibilidade para definir uma matriz consiste em considerar uma função $f : \mathbb{R}^2 \to \mathbb{R}$ vulgar, escolher alguns pontos $x_0, ..., x_N$ e colocar o resultado numa matriz. Obtemos assim uma matriz A, em que

$$A_{ij} = f(x_i, x_j).$$

Se no *Mathematica* fizermos literalmente ListPlot3D[A], o resultado é sobejamente conhecido para quem já alguma vez procurou ver o gráfico de uma função... mas talvez poucos terão sido tentados a avaliar o determinante ou os valores próprios dessa matriz. Poderá parecer que não faz muito sentido calcular o determinante, ou os valores próprios associados à função... mas tem! Nas figuras que se seguem vamos ilustrar algumas situações. Iremos escolher um intervalo $[a, b]$, definir os pontos $x_k = a + kh$, com $h = \frac{b-a}{N}$, e apresentar figuras para as matrizes $A_{ij} = h\, f(x_i, x_j)$... o facto de multiplicarmos por h será justificado mais à frente.

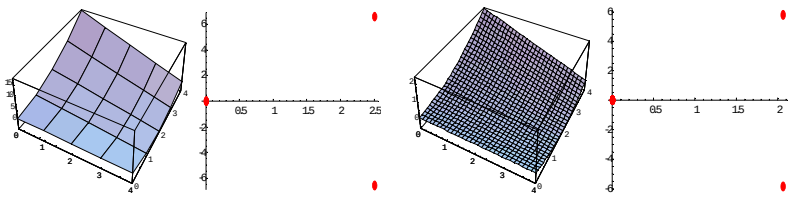
Começamos com $N = 4$, com o intervalo $[0, 4]$, considerando os cinco pontos x_k dados por $0, 1, 2, 3, 4$, e a função

$$f(x, y) = x^2 - xy + 1$$

isso dá-nos a matriz

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & -1 & -2 \\ 5 & 3 & 1 & -1 & -3 \\ 10 & 7 & 4 & 1 & -2 \\ 17 & 13 & 9 & 5 & 1 \end{bmatrix}$$

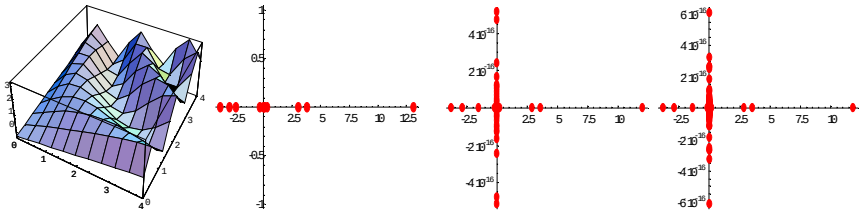
e dificilmente podemos dizer imediatamente se se trata ou não de uma matriz invertível. O gráfico obtido pelo ListPlot3D é apresentado em baixo, à esquerda. A segunda figura tem a localização dos valores próprios. Neste caso há três valores próprios nulos, e portanto a matriz não é invertível. Aumentamos N para 30... é claro que não apresentamos a matriz, mas como será de esperar, seria igualmente dificil avaliar se seria ou não invertível. Tal como para o valor de N anterior, colocamos o gráfico da função e os valores próprios obtidos (terceira e quarta figura, em baixo)..



O gráfico da função é semelhante, apenas o factor h fez diminuir a proporção, mas em nada influi na invertibilidade da matriz. Como podemos constatar, a matriz continuará a não ser invertível (terá 28 valores próprios que dão zero!), e o aspecto da distribuição

dos valores próprios é semelhante! Se aumentarmos a dimensão de N iremos reparar que acontecerá a mesma coisa! Será este um facto isolado, devido a alguma particularidade da função escolhida?

Vamos considerar um outro exemplo, em que $f(x,y) = 3\sin(xy) + x + y - 2$. Notamos que a matriz associada é simétrica, e portanto os valores próprios têm que ser reais. Apresentamos um gráfico semelhante para $N = 10$ (figura em baixo, à esquerda), e o gráfico com a localização do valores próprios, na figura seguinte. Nas restantes duas figuras, como é claro que o gráfico da função seria semelhante, colocamos apenas os valores próprios para $N = 50$ e $N = 200$, respectivamente.



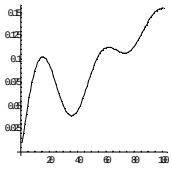
Há várias observações a fazer.

(i) Os dois últimos gráficos apresentam valores com parte imaginária, que são admissíveis, já que a ordem de grandeza dos valores é totalmente insignificante, e resulta apenas de erros de arredondamento numérico. Devemos considerá-los quase nulos.

(ii) Verificamos uma convergência para os mesmos valores.

(iii) À excepção de seis valores próprios, todos os restantes são quase nulos, independentemente do valor N. Mais uma vez, este estranho fenómeno ocorre!... E, podemos considerar que, na prática, a matriz A não será invertível!

(iv) Se procurarmos o vector próprio associado ao valor próprio mais elevado, ao colocarmos o vector em forma de gráfico, obtemos a figura seguinte.



Suficientemente estranho? Não será, se procurarmos compreender o que estamos a calcular... é claro que, como não iremos entrar em detalhes, apenas daremos uma justificação ligeira.

Multiplicando a matriz A por um vector v, obtemos o vector

$$[Av]_i = \sum_{k=0}^{N} h f(x_k, x_i) v_k,$$

e podemos encarar o valor das componentes como uma aproximação em x_i de

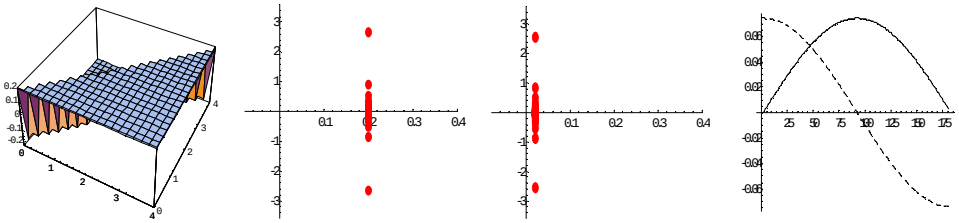
$$(\mathcal{A}v)(y) = \int_a^b f(x,y)v(x)\,dx,$$

já que a colocação dos pontos permite pensar numa aproximação do integral por somas de Riemann, considerando pequenos intervalos de comprimento h, e notando que v_k será $v(x_k)$. Portanto, quando N é grande, estaremos a avaliar uma aproximação do integral. Quanto menos irregular for a função, melhor será a aproximação. Isto explica, parcialmente, o ponto (ii). Repare-se que $\mathcal{A}$ é um operador linear.

Desde que a função f seja contínua, foi provado, no início do século XX (devido a Fredholm[5]), que esse operador integral é aquilo a que se designa por *operador compacto*, e os seus valores próprios λ, para os quais $\mathcal{A}v = \lambda v$ (com v não nula), verificam uma propriedade curiosa que caracteriza este tipo de operadores. Ou há apenas um número finito que não é nulo ou, no caso de haver um número infinito de valores não nulos, formam um ponto de acumulação em zero. A situação que retratámos nos exemplos apresentados foi a primeira, e isto explica parcialmente (iii). A explicação de (iv) resulta de associar o vector próprio à função própria v, que verifica $\mathcal{A}v = \lambda v$, resultado da discretização.

Como curiosidade, examinamos mais alguns casos.

O primeiro caso é o de uma matriz triangular superior, definida pela função descontínua $f(x, y) = -1$, se $x > y$, $f(x, y) = 1$, se $x \leq y$, representada na primeira figura, em baixo. Nas duas figuras seguintes, mostramos os valores próprios obtidos para $N = 20$ e $N = 200$. Neste caso, a convergência é mais lenta, mas os valores próprios irão ficar no eixo imaginário, formando um ponto de acumulação em zero. No gráfico da direita podemos ver a parte real (a tracejado), e a parte imaginária (a cheio), do maior vector próprio (com parte imaginária positiva). Independentemente da escala, que é irrelevante num vector próprio (pois poderá sempre ser multiplicado por qualquer valor), as funções apresentadas parecem-nos familiares.



Vejamos com maior atenção. No limite, em termos de integrais, teremos para uma função própria v,

$$\mathcal{A}v = \lambda v \Leftrightarrow \int_0^4 f(x,y)v(x)dx = \lambda v(y) \Leftrightarrow \int_0^y v(x)dx - \int_y^4 v(x)dx = \lambda v(y).$$

e se derivarmos obtemos $2v(y) = \lambda v'(y)$, ou seja, $v(y) = Ce^{2y/\lambda}$ (a constante é multiplicativa, logo não interfere num vector próprio). Para obtermos os valores de λ devemos voltar à igualdade anterior. Verificamos então que se $\lambda \neq 0$, a primitivação de v dá

$$\frac{\lambda}{2}(e^{2y/\lambda} - 1) - \frac{\lambda}{2}(e^{8/\lambda} - e^{2y/\lambda}) = \lambda e^{2y/\lambda} \Leftrightarrow \frac{\lambda}{2}(e^{8/\lambda} + 1) = 0$$

[5]A equação integral $\mathcal{A}v = g$ é denominada equação integral de Fredholm de primeira espécie, para contrastar com as equações do tipo $\mathcal{A}v + v = g(y)$, denominadas de segunda espécie... o que acontece neste caso aos valores próprios?

e portanto $\frac{8}{\lambda} = (2k+1)\pi i$, ou seja, os valores próprios serão

$$\pm\frac{8i}{\pi},\ \pm\frac{8i}{3\pi},\ \ldots\ ,\pm\frac{8i}{(2n+1)\pi},\ldots$$

sucessão que tem zero como ponto de acumulação. Relativamente a $\lambda = \frac{8}{\pi}$, vemos que a função própria será

$$v(x) = \exp(x\frac{\pi}{4}i) = \cos(x\frac{\pi}{4}) + i\sin(x\frac{\pi}{4}).$$

Neste caso reparamos que se trata da diferença entre dois operadores do tipo

$$(\mathcal{A}v)(y) = \int_a^x f(x,y)v(x)\,dx,$$

que são designados por *operadores de Volterra,* e correspondem no caso discreto a considerar matrizes triangulares. É sabido que operadores de Volterra são também operadores compactos (com espectro nulo) e aqui evidencia-se que a diferença entre eles é ainda um operador compacto (embora o espectro já não seja nulo).

Até aqui não apresentámos nenhum operador integral que não fosse compacto! Podemos entender isto, porque os elementos da matriz $A_{ij} = hf(x_i, x_j)$ tendem para zero quando $h \to 0$. Para que isso não aconteça, uma possibilidade é considerar funções que não sejam limitadas. Um caso de um operador integral não compacto, é o *operador de Abel*, por exemplo,

$$(\mathcal{A}v)(y) = \int_a^x \frac{v(x)}{\sqrt{x^2-y^2}}\,dx.$$

Outro caso de operador não compacto (nos complexos) é o *operador de Cauchy*,

$$(\mathcal{A}v)(z) = \frac{1}{\pi i}\int_\gamma \frac{v(x)}{x-z}\,dx,$$

onde mais uma vez é visível a singularidade.

Finalmente, apresentamos um outro tipo de operador não compacto, em que $f(x,y) = e^{-|x-y|}$ é uma função contínua, mas em que o intervalo *não é limitado,* trata-se do *operador de Wiener-Hopf*,

$$(\mathcal{A}v)(y) = \int_0^\infty e^{-|x-y|}v(x)\,dx.$$

Terminamos, referindo que o caso mais curioso é o do operador identidade, que não é compacto num espaço de dimensão infinita (o único valor próprio é 1... mas terá uma 'multiplicidade infinita contínua'). Sob o ponto de vista dos operadores integrais pode ser encarado como um integral formal

$$(\mathcal{I}v)(y) = \int_a^b \delta(x-y)v(x)\,dx = v(y),$$

em que a 'função' δ é designada delta de Dirac, e pode ser encarada intuitivamente como uma função que vale infinito em zero e zero nos restantes pontos. Sob o ponto de vista das matrizes será o equivalente a evitar que a matriz $A_{ij} = h\delta(x_i - x_j)$ tenda para zero quando $h \to 0$, exigindo que seja infinito quando $x_i = x_j$. Esta entidade, o delta de Dirac, é formalmente definida em contextos mais abstractos, na teoria da medida e na teoria das distribuições.

Capítulo 8

Exercícios resolvidos

8.1 Exercícios de avaliações anteriores

8.1.1 1a. Avaliação (97)

1. Considere o intervalo $I = [a, b] \subset \mathbb{R}$, e as funções $g, h \in C^1(I)$ tais que $g \circ h \neq h \circ g$. Sabemos que $g(I) \subseteq I, h(I) \subseteq I$, e que

$$|g'(x)| \leq L_1, \;\; |h'(x)| \leq L_2, \;\; \forall x \in I,$$

com $L_1 L_2 < 1$.

a) Se estabelecermos em I:

$$x = g(x) \Leftrightarrow x = h(x) \Leftrightarrow f(x) = 0,$$

mostre que existe uma única raiz $z \in I$ de $f(x) = 0$ e indique (justificando) três funções iteradoras distintas que assegurem uma convergência para essa raiz, qualquer que seja $x_0 \in I$.

b) Supondo que $a < 0$ e $b \geq 0$, mostre que o zero z da função f em I verifica:

$$|z| \leq \frac{\min\{|h(g(0))|, |g(h(0))|\}}{1 - L_1 L_2}$$

c) Suponha agora que os pontos fixos de g e h em I são diferentes. A equação $x = g(h(x))$ tem uma solução única em I? Justifique.

d) Mostre que as equações:

$$2x - \cos(\frac{0.5}{x^2 + 1}) = \sin(\frac{1}{x^2 + 1})$$

$$x = \frac{4}{(\cos(x/2) + \sin(x))^2 + 4}$$

têm uma única solução no intervalo $[0, 1]$ e indique funções iteradoras convergentes para a solução.

Sugestão: Usar funções g e h apropriadas e aplicar o anterior.

2. Considere uma função g contínua no intervalo $[0,1]$, diferenciável, que verifica:

$$g(0) = \frac{1}{2}, \qquad g'(x) = \frac{1}{2} x \cos^2(g(x))$$

Mostre que a função g tem um único ponto fixo no intervalo $[0,1]$.

3. Considere a matriz

$$A = \begin{bmatrix} 3 & \sin(a^2) & \cos(b) \\ \sin(a^2) & \exp(b/\pi) + 2 & \sin(a+b) \\ \cos(b) & \sin(a+b) & \frac{5}{2} \end{bmatrix}$$

a) Mostre que a matriz A é definida positiva para quaisquer $a, b \in \mathbb{R}$ e que o método de Gauss-Seidel converge quando aplicado a um sistema do tipo $Ax = v$ com $v \in \mathbb{R}^3$.

b) Localize intervalos para os valores próprios de A quando $b = 0$.

c) Mostre que a matriz tem um valor próprio dominante quando $b > \pi \log(5)$. Determine esse valor próprio, quando $a = 0,\ b = 7\pi$.

4. Seja E um espaço de Banach, e A um operador linear contínuo em E, isto é $A \in \mathcal{L}(E,E)$.

a) Mostre que se $||A||_{\mathcal{L}(E,E)} < \lambda$, então o operador $(\lambda I - A)^{-1}$ existe e pertence a $\mathcal{L}(E,E)$. Para além disso, mostre que se verifica:

$$(\lambda I - A)^{-1} = \sum_{k=0}^{\infty} \frac{A^k}{\lambda^{k+1}}, \text{ e que } ||(\lambda I - A)^{-1}||_{\mathcal{L}(E,E)} \leq \frac{1}{\lambda - ||A||_{\mathcal{L}(E,E)}}$$

Sugestão:

Use a equivalência $(\lambda I - A)X = I \Leftrightarrow X = (I + AX)/\lambda$, e lembre que:

$$||AB||_{\mathcal{L}(E,E)} \leq ||A||_{\mathcal{L}(E,E)} ||B||_{\mathcal{L}(E,E)}$$

b) Considere a matriz definida no exercício 3 com $b = 0, a \in \mathbb{R}$. Baseado na alínea anterior, mostre que a matriz $B = A - 6I$ é invertível e que:

$$||B^{-1}||_1 \leq 1$$

5. Considere uma função f que toma os seguintes valores:

x	-2	-1	0	1	2
f(x)	-10	-4	-2	1	5

e que tem um único zero $z \in [0,1]$. Pretende-se aproximar esta função por uma função do tipo $g(x) = a + bx + cx^3$, no sentido dos mínimos quadrados.

a) Mostre que os valores a, b, c verificam:

$$\begin{bmatrix} 5 & 0 & 0 \\ 0 & 10 & 34 \\ 0 & 34 & 130 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -10 \\ 35 \\ 125 \end{bmatrix}$$

b) Seja $a = -2,\ b = 25/12,\ c = 5/12$ a solução do sistema da alínea anterior. Determine uma aproximação da raiz w do polinómio g em $[0,1]$ calculando x_2 pelo método de Newton com $x_0 = 1$ (justifique a convergência).

c) Sabendo que $|z - w| < 0.01$, determine um majorante para o erro absoluto que cometemos se utilizarmos o valor x_2 obtido em b) para aproximar z.

8.1.2 2a. Avaliação (97)

1. Considere a seguinte fórmula (deduzida através da resolução de uma equação às diferenças) utilizada para calcular o valor de uma prestação p de um empréstimo D a uma taxa $T\%$ ao longo de N prestações:

$$p = tD\left(1 + \frac{1}{(1+t)^N - 1}\right) \tag{8.1}$$

com $t = T/1200$.

a) Aproximamos a fórmula anterior, considerando $(1+t)^N \sim e^x$ com $x = Nt$ e obtemos

$$\tilde{p} = x\frac{D}{N}\frac{e^x}{e^x - 1}. \tag{8.2}$$

Mostre que se $N\tilde{p} = 1.62D$, existe um e um só $x \in [1, 2]$ que verifica a equação (8.2).

b) Considerando $D = 10000$, $N = 180$, $p = 90$, determine aproximadamente o valor de $t > 0$ que verifica a equação (8.1), usando como critério de paragem $|t_{n+1} - t_n| < 0.001$. Conclua, indicando o valor da taxa T. *(Escolha um método e uma iterada inicial conveniente) (Isto corresponde a calcular a taxa a negociar para um empréstimo de 10 mil contos com uma prestação mensal de 90 contos durante 15 anos.)*

2. Considere $f(x) = 0 \Leftrightarrow x = g(x)$ uma equação em $\mathbb{R}$ que tem pelo menos duas raízes z_1 e z_2 consecutivas (ou seja, não existe nenhuma outra raiz entre elas).

a) Mostre que se $g \in C^1(\mathbb{R})$ e $|g'(z_1)| < 1$ então $g'(z_2) \geq 1$.

b) Suponha que $z_2 \in I = [a, b]$, que $|g'(x)| > 1, \forall x \in I$ e que $I \subseteq g(I)$. Mostre que o método $x_{n+1} = g^{-1}(x_n)$ converge para z_2 qualquer que seja $x_0 \in I$.

c) Seja $f \in C^p(\mathbb{R})$, tal que a raiz z_2 tem multiplicidade $p \geq 1$, e seja g tal que $g'(z_2) > 1$. Indique uma função iteradora que assegure uma convergência local linear para z_2, e uma outra que assegure convergência quadrática, para cada caso de p.

3. Considere a matriz

$$A = \begin{bmatrix} 10 & 3 - 2\cos(b) & \cos(b) \\ 1 & 25 & 5\sin(a) \\ 1 & 5\sin(a) + \sin(b) & 50 \end{bmatrix}$$

a) Localize os valores próprios de A usando o teorema de Gershgorin.

b) Indique os valores de b para os quais podemos obter uma decomposição $A = LL^T$, em que L é uma matriz triangular inferior real.

c) Para que valores de $h \in \mathbb{R}^3$ é possível utilizar o método de Jacobi para resolver um sistema $Ax = h$? Indique uma estimativa de erro para $||e^{(n)}||_\infty$ em função de $||h||_\infty$, sabendo que $x^{(0)} = 0$.

d) Seja A uma matriz real $N \times N$, que verifica:

$$|a_{ii} - a_{jj}| > r_i + r_j, \forall i, j = 1, ..., N \ (i \neq j)$$

em que

$$r_k = -|a_{kk}| + \sum_{j=1}^{N} |a_{kj}|$$

Mostre que os valores próprios da matriz são reais.

4. Considere o operador A em $E = C[0, a]$, com $a < 1/2$ definido por:

$$(Af)(x) = \int_0^x (f(t))^2 dt$$

e sabemos que a sua derivada de Fréchet é:

$$(A'_f)h(x) = 2\int_0^x h(t)f(t)dt$$

Pretende-se resolver a equação em E:

$$f(x) = \int_0^x (f(t))^2 dt - \phi(x) \tag{8.3}$$

em que ϕ é conhecido e verifica $||\phi||_\infty \leq a$.

a) Mostre que existe uma e uma só solução de (8.3) em $X = \{f \in C[0, a] : ||f||_\infty \leq 1\}$ e indique uma sucessão que converge para a solução.

b) Mostre que a solução de (8.3) verifica

$$||f||_\infty \leq \frac{a}{1 - 2a}$$

5. Pretende-se aproximar a função $f(x)$ no intervalo $[-1, 1]$ através de funções do tipo $g(x) = \alpha + \beta f'(x)$, no sentido dos mínimos quadrados. Mostre que os valores α e β ficam perfeitamente determinados pelos quatro valores

$$\int_{-1}^{1} |f'(x)|^2 dx, \quad \int_{-1}^{1} f(x)dx, \quad f(1), \; f(0),$$

e comente os resultados quando a função f é par ou impar.

8.1.3 Teste (98)

1. Suponha que temos $g([1, 2]) \subset [1, 2]$ e $|g'(x)| \leq \frac{1}{4}$ em $[1, 2]$.

Pretende-se calcular $y = g^{20}(a) = g \circ ... \circ g(a)$, mas o valor de $a \in]1.1, 1.9[$ é dado com um erro relativo $|\delta_a| \sim 10^{-3}$. Apresente, justificando com cálculos, uma estimativa para o erro relativo de y. Comente os resultados relativamente ao condicionamento.

2.a) Seja $f \in C^1(\mathbb{R})$, tal que $f(0) = 0$ e $|f'(x)| \leq 1, \forall x \in \mathbb{R}$. Mostre que para $x \geq 0$ temos

$$2f(x) - \cos(f(\frac{x}{2})) \leq 3x - 1.$$

Sugestão: Usar o teorema do ponto fixo em $\mathbb{R}$.

b) Considere a equação

$$|2\sin(x) - \cos(\sin(\frac{x}{2})) - 3x + 1| + 2\sin(x) = 4x - 2.$$

Mostre que há apenas uma raiz positiva e determine uma aproximação tal que o erro absoluto seja inferior a 10^{-2}.

3. Considere a sucessão definida por

$$x_{n+2} = \frac{9}{2}x_{n+1} - 2x_n$$

em que $x_0 = 2$ e $x_1 = 1$. Mostre que x_n converge para zero.

4.a) Seja E um espaço de Banach e X um conjunto não vazio, fechado. Mostre que um operador contínuo $A : E \to E$ que verifica $X \subseteq A(X)$ (pode assumir que $A(X)$ é fechado) e

$$||Ax - Ay|| \geq L||x - y||, \forall x, y \in X$$

para um certo $L > 1$, tem um e um só ponto fixo $z \in X$, e que o método $x_n = Ax_{n+1}$ converge para esse ponto fixo, qualquer que seja $x_0 \in X$.

b) Seja A uma contracção num aberto não vazio X, onde se sabe que existe um ponto fixo z de A.

Mostre que existe um subconjunto de X onde são verificadas as condições do Teorema do Ponto Fixo de Banach, e conclua que pode assegurar convergência local do método do ponto fixo.

8.1.4 1a. Avaliação (98)

1. Considere a equação em $\mathbb{R}$:

$$\sqrt{|\sin(x) + 3x| + \sin(x)} - 1 = 0.$$

a) Mostre que a equação tem duas raízes, uma positiva e uma negativa.

b) Determine uma aproximação da raiz positiva pelo método de Newton, mostrando a convergência num intervalo apropriado, de forma a garantir um erro relativo inferior a 10^{-3}.

c) Ao calcular a sucessão definida por $x_0 = 0,\ x_{n+1} = \frac{1}{3} - \frac{2}{3}\sin(x_n)$, aproximou-se o cálculo do seno por $\sin(x) \sim x - \frac{1}{6}x^3$. Considerando que essa aproximação implica um erro absoluto inferior a $\frac{|x|^5}{120}$, determine um majorante do erro $|x - y_n|$ em que x é o limite de (x_n), e em que y_n é dado por

$$y_0 = 0,\ y_{n+1} = \frac{1}{3} - \frac{2}{3}y_n + \frac{1}{9}y_n^3..$$

d) Localize as raízes de $0 = \frac{1}{3} - \frac{5}{3}y + \frac{1}{9}y^3$, começando por indicar a coroa circular que as contém. Aplique o método de Bernoulli efectuando 5 iterações, justificando a convergência..

2. Sabemos que podemos aproximar $\sqrt{a}$, para qualquer $a \in \mathbb{N}$, usando o método do ponto fixo com $x_0 = 1$ e com a função iteradora $g(x) = \frac{x^2+a}{2x}$, que se trata de uma função racional (quociente de dois polinómios de coeficientes inteiros), $g : \mathbb{Q} \to \mathbb{Q}$. Da mesma forma, é possível encontrar uma função iteradora racional, tal que o método do ponto fixo convirja para o número π? Em caso afirmativo apresente-a, em caso negativo justifique!

3. O processo de construção de uma estrutura com 10 colunas obedece à seguinte regra:
– A altura a_k de uma coluna k deve ser igual à seguinte média ponderada de alturas de colunas mais próximas:

$$a_k = \frac{1}{2}(a_{k-1} - a_{k+1}) + a_{k-2}.$$

Tendo já sido construídas as duas primeiras colunas, com alturas $a_1 = 4,\ a_2 = 8$, bem como a última, com $a_{10} = 180$, determine qual a altura da coluna a_5.

4. Considere X, Y dois espaços de Banach, e sejam $A : X \to Y,\ B : Y \to X$ dois operadores que verificam

$$||Ax_1 - Ax_2||_Y \leq K_A||x_1 - x_2||_X \ , \forall x_1, x_2 \in X \ ; \ \ ||By_1 - By_2||_X \leq K_B||y_1 - y_2||_Y \ , \forall y_1, y_2 \in Y$$

Dê condições em K_A e K_B para que o método do ponto fixo $x_{n+1} = BAx_n$ convirja, qualquer que seja $x_0 \in X$.

Encontre um exemplo para $X, Y = \mathbb{R}^d$, em que $K_A > 1$, e em que $x_{n+1} = BAx_n$ converge, $\forall x_0 \in X$.

5. Considere a matriz ($a \in [0, \frac{\pi}{2}]$)

$$A = \begin{bmatrix} 9 + \sin(a) & 3 & 1 \\ \cos(a) & 1 + \cos(a) & 0 \\ 0 & 1 & -2 \end{bmatrix}$$

a) Determine uma localização dos valores próprios usando o T. Gerschgorin. A matriz é sempre invertível? Justifique.

b) Seja $a = 0$. Aproxime o valor próprio dominante usando duas iterações do método das potências.

c) Considere o seguinte método, baseado na aplicação do teorema de Cayley-Hamilton, para encontrar o polinómio característico de uma matriz A de dimensão d :

- Calcular A^k, para $k = 2, ..., d$
- Determinar os coeficientes α_i tais que $\alpha_0 I + \alpha_1 A + ... + \alpha_{d-1}A^{d-1} + A^d = 0$.

Indique uma estimativa do número de operações (*, /) necessário a esse cálculo.
Use este método para determinar a equação característica de A com $a = 0$.

6. Considere o método SOR (Gauss-Seidel generalizado) em que

$$C_\omega = (D + \omega L)^{-1}[(1 - \omega)D - \omega U].$$

Mostre que é necessário que $\omega \in]0,2[$ para que haja convergência do método, qualquer que seja a iterada inicial. *Sugestão:* Mostrar que $\rho(C_\omega) \geq |\omega - 1|$.

7. Pretende-se encontrar qual de três figuras melhor aproxima cada uma de outras duas, no sentido dos mínimos quadrados. Consideramos ϕ_1, ϕ_2, ϕ_3 funções base e $f_1,\ f_2$ as funções a aproximar no sentido dos mínimos quadrados. Temos a tabela com os respectivos valores:

$i=$	1	2	3	4	5	6
ϕ_1	0	1	1	1	1	0
ϕ_2	0	0	1	1	1	1
ϕ_3	0	1	0	1	1	0
f_1	6	20	13	17	13	17
f_2	6	20	13	6	13	17

Determine qual das funções base tem maior componente na aproximação de f_i e associe essa componente à figura pretendida.

Use a decomposição de Cholesky para resolver o sistema normal.

8.1.5 2a. Avaliação (98)

1. Mostre que existe um único ponto fixo da gaussiana $g(x) = e^{-(x-25)^2/2}$ e determine uma sua aproximação com um erro absoluto inferior a 10^{-50}. Indique também o erro relativo. *Sugestão:* Para a estimativa de erro use o teorema do valor médio.

2. Considere a equação em $\mathbb{R}$

$$\det \begin{bmatrix} x^3-10 & 2 & 1 \\ 1 & x^3+6 & 0 \\ 0 & 1 & x^3 \end{bmatrix} = 0 \tag{8.4}$$

a)$_{[1.5]}$ Aplicando o Teorema de Gerschgorin, mostre que existem e são apenas três, as raízes reais de (8.4) e que se tem

$$z_1 \in [\sqrt[3]{9}, \sqrt[3]{11}],\ z_2 \in [-\sqrt[3]{7}, -\sqrt[3]{5}],\ z_3 \in [-1,1].$$

b)$_{[1.5]}$ Determine uma aproximação da maior raiz positiva de (8.4) pelo método da secante mostrando a convergência para as iteradas iniciais escolhidas, de forma a garantir um erro relativo inferior a 10^{-2}.

c)$_{[1.5]}$ Decomponha a matriz

$$B = \begin{bmatrix} 16 & -2 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 6 \end{bmatrix}$$

na forma $B = LU$ usando o método de Doolittle, e resolva os sistemas $Bx = (0,1,0)$ e $Bx = \frac{-1}{13}(13, 96, 16)$.

d)$_{[2.0]}$ Calcule duas iteradas pelo método das iterações inversas para obter uma aproximação da maior raiz negativa de (8.4), indicando uma estimativa do erro.

3. Considere a matriz companheira do polinómio com coeficientes reais $p(x) = a_0 + a_1x + ... + a_{n-1}x^{n-1} + x^n$

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}$$

que tem p como polinómio característico.

a)$_{[1.5]}$ Mostre que se $|a_{n-1}| > 1 + M$, com $M = \max\{|a_0|, |a_1| + 1, ..., |a_{n-2}| + 1\}$ então existe uma e uma só raiz real dominante em $[-a_{n-1} - 1, -a_{n-1} + 1]$, e que as restantes se encontram na bola $\{|z| \leq M\}$.

b)$_{[1.5]}$ Considere $p(x) = 2 - 6x^2 + 4x^3 - 16x^4 + 2x^5$.

Localize as raízes dominante num intervalo de comprimento 2 e as restantes numa bola de raio 1.

Determine aproximadamente a raiz dominante usando duas iterações do método das potências.

4. Considere os operadores $A_m : C[0, 1] \to C[0, 1]$, com $m \in \mathbb{N}$

$$A_m f(x) = \int_0^x (f(t))^m dt$$

a)$_{[1.0]}$ Mostre que a derivada de Fréchet de A_m é $A'_{m,f}(h) = m \int_0^x f(t)^{m-1}h(t)dt$.

b)$_{[2.0]}$ Mostre que no espaço $X = \{f \in C[0, 1] : ||f||_\infty \leq \frac{3}{4}\}$ existe uma e uma só solução f de

$$\int_0^x (f(t))^5 dt + \int_0^x (f(t))^2 dt + 4f(x) = \phi(x), \tag{8.5}$$

onde $||\phi||_\infty < 1$. Indique um método iterativo que convirja, qualquer que seja a iterada inicial em X.

c)$_{[1.0]}$ Determine um número de iteradas suficiente pelo método do ponto fixo começando com $f_0 = 0$, para que f_n, uma aproximação da solução da equação (8.5) verifique $||e_n||_\infty < 10^{-2}$.

5.$_{[2.5]}$ Considere a função $f(x, y) = (\, x\cos(xy) - 3y,\ y\sin(x + y) - 4x)$ e a equação em $\mathbb{R}^2$:

$$f(x, y) = (-1, 1) \tag{8.6}$$

Encontre uma função iteradora g que garanta a convergência do método do ponto fixo para qualquer $(x_0, y_0) \in S$, em que $S = [-1, 1] \times [-1, 1]$. Mostre que existe uma única solução de (8.6) em $\mathbb{R}^2$.

6. Pretende-se aproximar a função $f(x) = \sin(\pi x)$ no intervalo $[0, 1]$ por funções $g(x) = a + b(x - \frac{1}{2})^2$.

a)$_{[1.0]}$ Sabendo que $\int_0^1 \sin(\pi x)(x^2 - x)dx = -\frac{4}{\pi^3}$, mostre que os parâmetros da função g que melhor aproxima f no sentido dos mínimos quadrados são soluções do sistema:

$$\begin{bmatrix} 1 & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{80} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{2}{\pi} \\ \frac{\pi^2 - 8}{2\pi^3} \end{bmatrix}$$

b)$_{[1.5]}$ Tomando $\pi \sim 3.14$ no segundo membro, o sistema terá uma solução aproximada $(\tilde{a}, \tilde{b})$, majore o erro relativo $||\delta_{(a,b)}||_\infty$.

8.1.6 Teste (99)

1. Considere $f(x) = \cos(\frac{x}{2}) + 2x + \sin(x)$.

a) Mostre que existe uma única raiz real z da equação

$$e^{200.48\, f(x)} - 1 = 0$$

e que z pertence ao intervalo $[-1, 1]$. Escolha um método iterativo para aproximar z e determine ao fim de quantas iterações garante um erro relativo menor que 10%.

b) Considere a área negativa A^- definida pelo gráfico da função f no intervalo $[-1, 1]$ e $\tilde{A}^-$ uma sua aproximação ao considerar o cálculo do integral com $\tilde{z}$ ao invés de z. Discuta o condicionamento do cálculo de A^- face ao erro relativo de z.

2. a) Mostre que qualquer equação em $\mathbb{R}$, da forma

$$ax^6 + bx^4 + cx^2 = 1 - |x|d$$

com $a > 0$, $b, c, d \geq 0$ tem somente duas raízes reais, uma positiva e outra negativa. Sabendo que $b \geq c + d$, indique intervalos que as contenham (dependendo de a e b).

b) Determine uma aproximação da raiz dominante de $x^3 + 10x^2 + 2x - 1 = 0$ calculando 3 iterações do método de Bernoulli. Apresente um majorante do erro absoluto.

c) Usando a aproximação de b) (se não resolveu, considere -9.8) e sabendo que a raiz dominante de $x^3 - 2x^2 - 10x - 1 = 0$ é aproximadamente 4.351 apresente aproximações para as raízes reais e complexas de

$$x^6 + 10x^4 + 2x^2 = 1.$$

3. Seja $a > 0$, e considere o operador $P : C[0, a] \to C[0, a]$

$$(Pf)(x) = \int_0^x g(f(t))\, dt,$$

em que $g \in C^2(\mathbb{R})$.

a) Mostre que $||P'_f||_{\mathcal{L}(C[0,a],C[0,a])} \leq a||g' \circ f||_\infty$.

b) Querendo resolver a equação diferencial $y'(2y-1) = \cos(y)$, obtivemos a equação integral em $C[0,a]$

$$f(x)^2 - f(x) = \int_0^x \cos(f(t))\, dt$$

Determine valores de a para os quais garanta a existência e unicidade de solução da equação integral no conjunto $X = \{f \in C[0,a] : ||f||_\infty \leq \frac{1}{4}\}$ e determine duas iteradas usando o método do ponto fixo com $f_0(x) \equiv 0$. Determine um majorante do erro (com $a = 0.1$).

8.1.7 1a. Avaliação (99)

1. Considere o corte de um recipiente com largura unitária e secção constante. Uma das paredes do recipiente é vertical e a outra é descrita por uma função f estritamente crescente. Pretende-se determinar o ponto mais elevado $(z, f(z))$ situado na parede do recipiente que será submerso por um volume de água V.

a) Reduza o problema ao cálculo da raiz da equação (mostre a unicidade)

$$zf(z) - F(z) + F(0) = V,$$

em que $F' = f$.

b) Considere $f(x) = \frac{1}{6}(x^4 + 3x^2) + 1$ e $V = 1$. Determine aproximadamente z de forma a garantir um erro absoluto inferior a 0.01.

2. Após a aplicação de uma certa transformação T, *algumas* partículas que se encontravam em $[-1,1] \times [-1,1]$ ficaram posicionadas numa parte desse conjunto.

a) Sabendo que

$$T(x,y) = \frac{1}{4}\left(\sin(\frac{x-y}{2}) + \cos(\frac{xy}{2}), \cos(x+y) - \sin(xy)\right)$$

mostre que apenas um ponto ficou invariante com a transformação. O que acontece aos restantes pontos se repetirmos a transformação?

b) Estabeleça um algoritmo que dê uma aproximação do ponto atractor referido em a) com um erro absoluto inferior a 10^{-8}.

3. Considere a matriz (com a real)

$$A_a = \begin{bmatrix} -\sin(a^2) & \cos(a) & 0 \\ 0 & 0 & \cos(a^2) \\ -1-\cos(a^2) & -1-\cos(a) & 4 \end{bmatrix}$$

a) Mostre que existe um valor próprio real dominante. No caso $a = 0$ determine uma sua aproximação usando três iterações do método das potências.

b) Use três iterações do método de Bernoulli para aproximar a raiz dominante da equação $x^3 - 4x^2 + 2x + 2 = 0$. Compare o resultado obtido com o da alínea anterior e comente.

c) Seja $a = 0$. Aplique o método das iterações inversas para aproximar um valor próprio próximo de 1, começando com $(0, 1, 0)$ como vector inicial e calculando duas iterações através da factorização LU.

4. Verificou-se que uma população tinha a seguinte fórmula de crescimento: o número de organismos novos era 3 vezes superior ao número de organismos actuais, mas verificava-se também que cada organismo da geração anterior aniquilava dois dos novos organismos, e os da geração ainda anterior aniquilavam um organismo antes de perecerem. Obteve-se assim a fórmula

$$u_{n+1} = 4u_n - 2u_{n-1} - 2u_{n-2}.$$

a) Sabendo que $a = -0.481194, b = 1.31111, c = 3.17009$, temos

$$\begin{bmatrix} 1 & 1 & 1 \\ 1/a & 1/b & 1/c \\ 1/a^2 & 1/b^2 & 1/c^2 \end{bmatrix}^{-1} \sim \begin{bmatrix} 0.0354 & -0.159 & 0.147 \\ -0.516 & 1.387 & 0.787 \\ 1.48 & -1.23 & -0.934 \end{bmatrix}.$$

Determine ao fim de quantas gerações pode obter uma população superior a 1 milhão, se começar com $u_0 = 2$ organismos apenas.

b) Comente a diferença entre começar com $u_0 = 12,\ u_{-1} = 10,\ u_{-2} = 6$, ou com $u_{-2} = 5$ relativamente ao condicionamento.

5. Seja $\alpha > 2$. Para resolver $(\alpha - y')y' = 1 - y^2$, $y(0) = 0$, considerou-se a sucessão de funções $f_0 \equiv 0$,

$$f_{n+1}(x) = \frac{1}{\alpha} \int_0^x 1 - f_n(t)^2 + f_n'(t)^2 \, dt.$$

a) Mostre que a derivada de Fréchet do operador

$$\begin{array}{rccc} A: & C^1[0,a] & \longrightarrow & C^1[0,a] \\ & f & \longmapsto & \frac{1}{\alpha}\int_0^x 1 - f(t)^2 + f'(t)^2 \, dt \end{array}$$

na norma $||f||_{C^1} = ||f||_\infty + ||f'||_\infty$ é igual a

$$A_f' h = \frac{2}{\alpha} \int_0^x f'(t)h'(t) - f(t)h(t)dt.$$

b) Estabeleça que

$$||A_f'||_{L(C^1[0,a],C^1[0,a])} \leq \frac{2(a+1)}{\alpha} ||f||_{C^1}$$

c) Considere $a = 1, \alpha = 8$. Mostre que se $||f||_{C^1} \leq 1$, existe um único ponto fixo do operador A e determine um majorante do erro de f_2.

8.2 Resoluções

8.2.1 1a. Avaliação (97)

1. a) Como $g(I), h(I) \subset I$, temos $h(g(I)), g(h(I)) \subset I$.

Por outro lado se $x \in I$, como $g(x), h(x) \in I$ usando as hipóteses, temos:

$$|(h \circ g)'(x)| = |h'(g(x))|\ |g'(x)| \leq L_1 L_2 < 1$$

$$|(g \circ h)'(x)| = |g'(h(x))|\ |h'(x)| \leq L_2 L_1 < 1$$

Usando o corolário do teorema do ponto fixo, isto garante a existência e unicidade de solução das equações $x = h(g(x))$ e $x = g(h(x))$ no intervalo I.

Apenas falta ver que z é solução dessas equações. Ora, $f(z) = 0$ se e só se $z = g(z), z = h(z)$, portanto também $z = h(g(z)),\ z = g(h(z))$. Como garantimos a existência e unicidade dos pontos fixos de $h \circ g, g \circ h$ no intervalo I, fica provado. Por outro lado, como as funções iteradoras $h \circ g, g \circ h$ estavam nas condições do corolário garantimos ainda a convergência para z.

Uma outra função iteradora poderá ser g ou h já que $0 \leq L_1 L_2 < 1$ implica $0 \leq L_1 < 1$ ou $0 \leq L_2 < 1$.

1. b) $a < 0, b \geq 0$ implica $0 \in I$ e podemos escolher $x_0 = 0$, e escolhendo a função iteradora $h \circ g$, obtemos $x_1 = h(g(0))$ e temos:

$$|z - x_0| \leq \frac{1}{1 - L_1 L_2}|x_1 - x_0| \ \Leftrightarrow\ |z| \leq \frac{1}{1 - L_1 L_2}|h(g(0))|.$$

Podemos fazer o mesmo para $g \circ h$, e tiramos $|z| \leq \frac{1}{1-L_1L_2}|g(h(0))|$. Agora basta escolher o menor majorante, considerando o $\min\{|g(h(0))|, |h(g(0))|\}$.

1. c) Sim. Basta reparar que em a), para mostrar que existia ponto fixo de $g \circ h$ não utilizamos a hipótese de g e h terem o mesmo ponto fixo.

1. d) Considerando $g(x) = \frac{1}{2}(\cos(x/2) + \sin(x))$, $h(x) = \frac{1}{x^2+1}$ vemos que podemos escrever a primeira equação na forma $x = g(h(x))$ e a segunda na forma $x = h(g(x))$. De acordo com o que vimos basta ver que estes g e h verificam as hipóteses, para considerarmos $g \circ h$ função iteradora que converge para a solução da primeira equação e $h \circ g$ para a solução da segunda.

Ora, seja $I = [0, 1]$. Se $x \in I$, temos $\cos(x/2), \sin(x) \geq 0$ e é claro que $g(x) \leq 1$, assim $g(I) \subseteq I$. Por outro lado

$$|g'(x)| = |\frac{-1}{4}\sin(x/2) + \frac{1}{2}\cos(x)| \leq \frac{3}{4} = L_1$$

(podiamos mesmo majorar por $\frac{1}{2}$)

Por outro lado, como $x^2 + 1 \geq 1$, é claro que $h(I) \subseteq I$, e se $x \in I$

$$|h'(x)| = |\frac{-2x}{(x^2+1)^2}| = \frac{2x}{(x^2+1)^2} \leq \frac{1}{2} = L_2$$

porque atinge o máximo quando $x = 1$, visto ser crescente:

$$\left(\frac{2x}{(x^2+1)^2}\right)' = \frac{2(x-1)^2}{(x^2+1)^2} \geq 0$$

Assim $L_1 L_2 = \frac{3}{8} < 1$, e podemos concluir pelo que vimos anteriormente em a) e c).

(Se não usassemos a sugestão, seriamos envolvidos em cálculos muito mais complicados...)

2. *Começamos por observar que a função g é solução de uma equação diferencial que verifica condições que asseguram existência e unicidade para aquela condição inicial, e até podemos ver que $g \in C'^{\infty}$, mas isto sai fora do nosso âmbito.*

Vemos que $|g'(x)| = \frac{1}{2}|x|\ |\cos(g(x))|^2 \leq \frac{1}{2}$, por outro lado $g'(x) \geq 0$, e g é crescente. Como $g(0) = \frac{1}{2} \in [0,1]$, se $g(1) \in [0,1]$, ficam verificadas as condições do corolário do teorema do ponto fixo, e podemos assegurar existência e unicidade do ponto fixo. Ora, como pelo teorema do valor médio:

$$g(1) - g(0) = g'(\xi) = \frac{1}{2}\xi\cos^2(g(\xi)) \in [0, \frac{1}{2}[$$

pois $\xi \in]0,1[$, temos $g(1) \in [\frac{1}{2}, 1[\subset [0,1]$.

3. a) Vimos, como consequência do teorema de Gershgorin, que sendo a matriz simétrica, basta ver que a matriz tem a diagonal positiva e estritamente dominante para que seja definida positiva. Por outro lado, para mostrar que o método de Gauss-Seidel converge, também basta ver que tem a diagonal estritamente dominante.

Ora, todos os elementos da diagonal são positivos e $|\sin(a^2)| + |\cos(b)| \leq 2 < 3$, $|\sin(a^2)| + |\sin(a+b)| \leq 2 < 2 + e^{b/\pi}$, $|\cos(b)| + |\sin(a+b)| \leq 2 \leq \frac{5}{2}$.

3. b) No caso $b = 0$ temos

$$A = \begin{bmatrix} 3 & \sin(a^2) & 1 \\ \sin(a^2) & 3 & \sin(a) \\ 1 & \sin(a) & \frac{5}{2} \end{bmatrix}$$

e como a matriz é definida positiva, os valores próprios λ são reais e positivos. Pelo teorema de Gerschgorin, $|\lambda - 3| \leq 1 + |\sin(a^2)| \leq 2$, $|\lambda - 3| \leq |\sin(a)| + |\sin(a^2)| \leq 2$, $|\lambda - \frac{5}{2}| \leq 1 + |\sin(a)| \leq 2$.

Concluimos assim que $\lambda \in [\frac{1}{2}, 5]$.

3. c) Neste caso, $e^{b/\pi} > 5$ e pelo T. de Gershgorin, a segunda linha garante que $|\lambda - e^{b/\pi} - 2| \leq 2 \Leftrightarrow (5 <)\ e^{b/\pi} \leq \lambda \leq e^{b/\pi} + 4$. Portanto existe um valor próprio tal que $\lambda > 5$, por outro lado, a primeira e última linha garantem $|\lambda - 3| \leq 2$ e $|\lambda - \frac{5}{2}| \leq 2$, ou seja vão existir dois valores próprios ≤ 5.

Concluimos assim que há um valor próprio dominante > 5.

Quando $a = 0,\ b = 7\pi$

$$A = \begin{bmatrix} 3 & 0 & -1 \\ 0 & e^7 + 2 & 0 \\ -1 & 0 & \frac{5}{2} \end{bmatrix},$$

e reparamos que $\lambda = e^7 + 2$ é valor próprio associado ao vector próprio $(0, 1, 0)$. (Esta conclusão sai imediatamente aplicando o T. Gerschgorin).

4. a) $\mathcal{L}(E, E)$ é espaço de Banach com a norma $||.||_{\mathcal{L}(E,E)}$ que designamos $||.||$ para abreviar. Ora, $(\lambda I - A)^{-1}$ é a solução de $(\lambda I - A)X = I \Leftrightarrow X = \lambda^{-1}I + \lambda^{-1}AX = G(X)$, e vemos que G é contractiva em $\mathcal{L}(E, E)$. Pois, para quaisquer $X, Y \in \mathcal{L}(E, E)$,

$$||G(X) - G(Y)|| = ||\lambda^{-1}AX - \lambda^{-1}AY|| = \lambda^{-1}||A(X - Y)|| \leq \lambda^{-1}||A||\ ||X - Y||$$

e $L = \lambda^{-1}||A|| < 1$, por hipótese. Aplicamos o teorema do ponto fixo a $\mathcal{L}(E, E)$ (que é fechado, pois é o próprio espaço) e concluímos a existência e unicidade de solução $Z = (\lambda I - A)^{-1}$. Usando $X_0 = 0$ obtemos $X_1 = G(X_0) = \lambda^{-1}I$ e

$$||(\lambda I - A)^{-1}|| = ||Z - X_0|| \leq \frac{1}{1 - L}||X_1 - X_0|| = \frac{1}{1 - \lambda^{-1}||A||}||\lambda^{-1}I||$$

como $||I|| = 1$, obtemos a desigualdade.

Finalmente, considerando $X_1 = \lambda^{-1}I$ obtemos

$$X_{n+1} = \sum_{k=0}^{n} \frac{A^k}{\lambda^{k+1}}.$$

Provamos por indução. Para $n = 0$ é imediato, já que se convenciona $A^0 = I$. Supondo que a fórmula é válida para X_n

$$X_{n+1} = \lambda^{-1}I + \lambda^{-1}AX_n = \lambda^{-1}I + \lambda^{-1}A\left(\sum_{k=0}^{n-1} \frac{A^k}{\lambda^{k+1}}\right) =$$

como A é linear,

$$= \lambda^{-1}I + \left(\sum_{k=0}^{n-1} \frac{A^{k+1}}{\lambda^{k+2}}\right) = \lambda^{-1}I + \left(\sum_{k=1}^{n} \frac{A^k}{\lambda^{k+1}}\right)$$

e obtemos a expressão de X_{n+1}.

4. b) Podemos aplicar a alínea a) já que a matriz pertence a $\mathcal{L}(\mathbb{R}^3, \mathbb{R}^3)$.

No exercício 3b) escrevemos a matriz e é fácil ver que $||A||_1 \leq 5$. Portanto, pela alínea anterior, considerando $\lambda = 6 > 5 \geq ||A||_1$, sabemos que $6I - A$é invertível e que

$$||(6I - A)^{-1}|| \leq \frac{1}{6 - ||A||_1} \leq 1$$

5. a) As funções base são $\phi_0(x) = 1,\ \phi_1(x) = x,\ \phi_2(x) = x^3$, que são polinómios linearmente independentes para $x_0 = -2,\ x_1 = -1,\ x_3 = 0,\ x_4 = 1,\ x_5 = 2$.

$$(\phi_0, \phi_0) = \sum_{k=0}^{4} \phi_0(x_k)^2 = \sum_{k=0}^{4} 1 = 5$$

$$(\phi_0,\phi_1)=\sum_{k=0}^{4}\phi_0(x_k)\phi_1(x_k)=\sum_{k=0}^{4}1.x_k=-2-1+0+1+2=0$$

$$(\phi_0,\phi_2)=0,(\phi_1,\phi_1)=4+1+0+1+4=10,$$

$$(\phi_1,\phi_2)=16+1+0+1+16=34,\ (\phi_2,\phi_2)=64+1+0+1+64=130$$

e

$$(\phi_0,f)=\sum_{k=0}^{4}\phi_0(x_k)f(x_k)=\sum_{k=0}^{4}1.f(x_k)=-10+-4-2+1+5=-10$$

$$(\phi_1,f)=-2\times-10+(-1\times-4)+0+1\times1+2\times5=35,$$

$$(\phi_2,f)=-8\times-10+(-1\times-4)+0+1\times1+8\times5=125$$

5. b) $g(x)=-2+(25x+5x^3)/12=0$, e verifica-se que $g\in C^2([0,1])$, pois é um polinómio. Como $g(0)g(1)=-2\times(30/12-2)=-1<0$, $g'(x)=(25+15x^2)/12>0$, $g''(x)=30x/12\geq 0$ e $g(x_0)>0$ para $x_0=1$, verificam-se as condições suficientes que asseguram a convergência do método de Newton:

$$x_1=x_0-\frac{-24+25x_0+5x_0^3}{25+15x_0^2}=1-\frac{6}{40}=\frac{17}{20}=0.85$$

$$x_2=x_1-\frac{-24+25x_1+5x_1^3}{25+15x_1^2}=0.841053....$$

5. c) A fórmula do erro do M. Newton dá :

$$|w-x_2|=\frac{|g''(\xi)|}{2|g'(x_1)|}e_1^2\leq\frac{2.5e_1^2}{5.9729}<0.41856e_1^2$$

e da mesma forma $|e_1|\leq\frac{2.5}{6.666...}|e_0|^2=0.375\times1$, pois $|e_0|<|1-0|$. Portanto $|w-x_2|<0.05886$, logo

$$|z-x_2|\leq|z-w|+|w-x_2|\leq0.01+0.05886=0.06886.$$

8.2.2 2a. Avaliação (97)

1. a) Como $N\tilde{p}/D=1.62$ temos que resolver:

$$1.62=x\frac{e^x}{e^x-1}\Leftrightarrow x=1.62\frac{e^x-1}{e^x}=g(x).$$

Basta verificar que existe um e um só ponto fixo de g no intervalo $[1,2]$.

Ora $g'(x)=e^{-x}>0$, portanto g é crescente e temos $g(1)=1.024035$, $g(2)=1.40076\in[1,2]$, logo $g([1,2])\subseteq[1,2]$.

Por outro lado, em $[1,2]$, $|g'(x)|=1.62e^{-x}\leq1.62e^{-1}=0.59596...<1$, logo podemos aplicar o teorema do ponto fixo e concluir acerca da existência e unicidade.

1. b) Se repararmos que $Np/D = 1.62$, como a equação (2) aproxima (1), podemos experimentar os valores obtidos:

$x_a = g(1) \Rightarrow t_a = x_a/N \sim 0.005689$ e $x_b = g(2) \Rightarrow t_b = x_b/N \sim 0.007782$

Sendo

$$f(t) = p - tD\left(1 + \frac{1}{(1+t)^N - 1}\right)$$

verificamos que $f(t_a) = 1.082$ e $f(t_a) = -13.45$, pelo que existe raiz neste intervalo, que podemos determinar usando, por exemplo, o método da secante em que $t_{-1} = t_a = 0.005689$, $t_0 = t_b = 0.007782$ e obtemos sucessivamente:

$$t_1 = 0.0058449359;\ t_2 = 0.0058505816$$

e temos $|t_2 - t_1| < 0.001$.

Portanto concluimos que $t \sim 0.00585 \Rightarrow T \sim 7.02\%$.

2. a) Considerando $h(x) = x - g(x)$ temos $h \in C^1(\mathbb{R})$ e $h(z_1) = h(z_2) = 0$. Como não existe mais nenhuma raiz entre z_1 e z_2, concluimos que ou $h(x) > 0$, $\forall x \in]z_1, z_2[$ ou $h(x) < 0$, $\forall x \in]z_1, z_2[$.

No nosso caso, $h'(z_1) = 1 - g'(z_1) > 0$, e como $h'(z_1) = \lim_{x \to z_1^+} \frac{h(x)}{x - z_1}$ (porque h tem derivada contínua e $h(z_1) = 0$), concluimos que $h(x) > 0$ num intervalo $]z_1, z_1 + \varepsilon[$. Portanto $h(x) > 0$ em $]z_1, z_2[$, logo

$$h'(z_2) = \lim_{x \to z_2^-} \frac{h(x)}{x - z_2} \leq 0$$

pois $\frac{h(x)}{x - z_2} < 0$ quando $x < z_2$. Assim, $g'(z_2) = 1 - h'(z_2) \geq 1$.

(*Podemos ver que quando uma função iteradora C^1 converge para uma raiz, ela não vai convergir para outra, consecutiva*)

2. b) Como $|g'(x)| > 1$ em I, nesse intervalo a função é injectiva, pois ou $g'(x) > 1 > 0$ ou $g'(x) < -1 < 0$, e g será estritamente monótona.

Faz pois sentido falar na inversa, g^{-1}, e é claro que $g(z_2) = z_2 \Leftrightarrow z_2 = g^{-1}(z_2)$.

Vejamos pois que se verificam as condições de convergência do teorema do ponto fixo para g^{-1} em I.

Se $x \in I$, como $I \subseteq g(I)$, existe $y \in I : x = g(y)$, logo $y = g^{-1}(x)$ e portanto $g^{-1}(x) \in I$. Isto implica $g^{-1}(I) \subset I$.

Por outro lado, para qualquer $x \in I$:

$$|(g^{-1})'(x)| = \frac{1}{|g'(g^{-1}(x))|} < 1$$

porque $y = g^{-1}(x) \in I$ e assim, pela hipótese, $|g'(y)| > 1$.

2. c) No caso de $p = 1$, temos raizes simples, e sabemos que o método de Newton converge quadraticamente, pelo que podemos usar $x - f(x)/f'(x)$ como função iteradora. Para a convergência linear, podemos usar g^{-1}, atendendo álínea b).

No caso de existirem raizes múltiplas, sabemos que o método de Newton converge linearmente, e para obtermos convergência quadrática usamos a modificação $x - p\, f(x)/f'(x)$.

(*Repare-se que o facto de o método de Newton convergir para qualquer raiz, não contradiz a alínea a) porque, no caso do método de Newton, não há diferenciabilidade da função iteradora, já que pelo teorema de Rolle entre dois zeros de f a derivada anula-se.*

Conclui-se que para termos uma função iteradora que convirja para qualquer raiz é necessário haver descontinuidades.)

3. a) Reparamos que, por linhas, $|\lambda - 10| \leq |3 - 2\cos(b)| + |\cos(b)| \leq 6$, $|\lambda - 25| \leq |5\sin(a)| + 1 \leq 6$, $|\lambda - 50| \leq 1 + |5\sin(a) + \sin(b)| \leq 7$.

Por colunas, conseguimos ainda $|\lambda - 10| \leq 2$, $|\lambda - 50| \leq 6$, o que significa $\lambda_1 \in [8, 12]$, $\lambda_2 \in [19, 31]$, $\lambda_3 \in [44, 56]$, porque os três circulos são disjuntos, e logo existe apenas um valor próprio em cada um deles que terá que ser real, pois os coeficientes da matriz são reais, e o polinómio caracteristico teria raizes complexas conjugadas (o que contradiria a existência de um único valor próprio).

3. b) Da alínea a) retiramos que a matriz é definida positiva, logo basta que a matriz seja simétrica para podermos obter a decomposição de Cholesky $A = LL^T$. Para isso, basta que $3 - 2\cos(b) = 1$; $\cos(b) = 1$; $5\sin(a) + \sin(b) = 5\sin(b)$, isto significa que $\cos(b) = 1$ e $\sin(b) = 0$, ou seja, $b = 2k\pi$.

3. c) Como a matriz A tem a diagonal estritamente dominante, podemos aplicar o método de Jacobi para qualquer $h \in \mathbb{R}^3$.

No caso do método de Jacobi, temos

$$||C||_\infty = || \begin{bmatrix} 0 & .3 - .2\cos(b) & 0.1\cos(b) \\ 0.04 & 0 & 0.2\sin(a) \\ 0.02 & 0.1\sin(a) + 0.02\sin(b) & 0 \end{bmatrix} ||_\infty \leq 0.6$$

Como

$$||e^{(n)}||_\infty \leq \frac{L^n}{1-L}||x^{(1)} - x^{(0)}||_\infty$$

com $L = ||C||_\infty$. Ora $x^{(0)} = 0 \Rightarrow x^{(1)} = (h_1/10, h_2/25, h_3/50)$, logo $||x^{(1)}||_\infty \leq ||h||_\infty/10$ e portanto:

$$||e^{(n)}||_\infty \leq \frac{0.6^n}{4}||h||_\infty$$

3. d) Utilizando o Teorema de Gerschgorin, já dissemos, em a), que se a matriz for real e os círculos forem disjuntos então os valores próprios são reais. Vejamos que as hipóteses implicam que os circulos sejam disjuntos:

Se $|\lambda - a_{ii}| \leq r_i$

$$|\lambda - a_{jj}| \geq |a_{ii} - a_{jj}| - |\lambda - a_{ii}| > r_i + r_j - r_i = r_j$$

(para $j \neq i$), o que significa que λ não está em nenhum outro circulo.

4. a) Vamos aplicar o corolário do Teorema do ponto fixo. Como o conjunto X é fechado e convexo, basta mostrar que $G(X) \subset X$ e que $||G'_f||_{\mathcal{L}(E,E)} \leq L < 1$, $\forall f \in X$, onde

$$G(f)(x) = \int_0^x (f(t))^2 dt - \phi(x).$$

Ora se $f \in X$, temos

$$||G(f)||_\infty = \max_{x\in[0,a]} |\int_0^x (f(t))^2 dt - \phi(x)|$$

$$\leq \max_{x\in[0,a]} \max_{t\in[0,x]} |f(t)|^2 \int_0^x 1\, dt + \max_{x\in[0,a]} |\phi(x)| \leq a||f||_\infty^2 + ||\phi||_\infty.$$

Como $||f||_\infty \leq 1$, $||\phi||_\infty \leq a$, concluimos que $||G(f)||_\infty \leq 2a < 1$, logo $G(f) \in X$.

Por outro lado, $G'_f = A'_f$ porque é imediato que a derivada de Fréchet de ϕ em ordem a f é zero, e temos

$$||A'_f||_{\mathcal{L}(E,E)} = \sup_{||h||_\infty \leq 1} ||(A'_f)h||_\infty \leq \sup_{||h||_\infty\leq 1} 2a||h||_\infty||f||_\infty \leq 2a||f||_\infty.$$

Portanto em X, temos $||G'_f||_{\mathcal{L}(E,E)} \leq 2a||f||_\infty \leq 2a < 1$.

4. b) Basta reparar que definindo $f_{n+1} = G(f_n)$ com $f_0 = 0 \in X$, obtemos $f_1 = -\phi$ e portanto, como $L = 2a$,

$$||f - f_0||_\infty \leq \frac{1}{1-L}||f_1 - f_0||_\infty = \frac{1}{1-2a}||\phi||_\infty \leq \frac{a}{1-2a}$$

5. Basta reparar que as funções base são 1 e f', logo:

$$(1,1) = \int_{-1}^1 1dx = 2;\ \ (1,f') = \int_{-1}^1 f'(x)dx = f(1) - f(-1);\ \ (f',f') = \int_{-1}^1 |f'(x)|^2 dx$$

e no segundo membro:

$$(1,f) = \int_{-1}^1 f(x)dx;\ \ (f',f) = \int_{-1}^1 f(x)f'(x)dx = \frac{1}{2}(f(1)^2 - f(-1)^2)$$

Chamando aos quatro valores referidos, respectivamente, a, b, c, d, obtemos:

$$\begin{bmatrix} 2 & c-d \\ c-d & a \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} b \\ (c^2-d^2)/2 \end{bmatrix}$$

Portanto, α e β são determinados pelos referidos valores.

No caso da função ser par temos $c = d$ o que implica $\beta = 0, \alpha = b/2$, ou seja, a melhor função será:

$$g(x) = \frac{1}{2}\int_{-1}^1 f(x)dx = \int_0^1 f(x)dx$$

No caso de ser impar, obtemos $b = \int_{-1}^1 f(x)dx = 0$, e como $d = f(-1) = -f(1) = -c$, temos o segundo membro do sistema nulo, portanto $\alpha = \beta = 0$ e a melhor função será $g = 0$.

8.2.3 Teste (98)

1. Considerando o algoritmo (que corresponde a implementar o método do ponto fixo):

$$x_0 = a;\ x_1 = g(x_0);\ \ldots x_{20} = g(x_{19})$$

e é claro que $g^{20}(a) = x_{20}$. Por outro lado, cada operação $x_{k+1} = g(x_k)$ implica

$$\delta_{x_{k+1}} \sim \frac{x_k g'(x_k)}{g(x_k)} \delta_{x_k} \tag{8.7}$$

e isto significa que podemos apresentar a estimativa aproximada

$$|\delta_{x_{k+1}}| \leq \frac{\max_{x\in[1,2]} |x| \max_{x\in[1,2]} |g'(x)|}{\min_{x\in[1,2]} |g(x)|} |\delta_{x_k}| \leq \frac{2 \cdot \frac{1}{4}}{1} |\delta_{x_k}| = \frac{1}{2} |\delta_{x_k}|$$

e sai imediatamente $|\delta_{x_{20}}| \leq (\frac{1}{2})^{20} |\delta_{x_0}| \sim 0.954 \times 10^{-9}$, ou melhor... $|\delta_{x_{20}}| \leq (\frac{1}{2})^{20} \frac{1.9}{1.1} |\delta_{x_0}| \sim 0.824 \times 10^{-9}$. Isto indica que este problema é (muito) bem condicionado, pois obtemos um erro relativo que é um milionésimo do inicial! Aproveitamos para reparar que no método do ponto fixo, obtemos sempre (8.7), e quando nos aproximamos do ponto fixo $x_k \sim z = g(z) \sim g(x_k)$, logo $\delta_{x_{k+1}} \sim g'(z)\delta_{x_k}$, e portanto de um modo geral basta haver convergência ($\Rightarrow |g'(z)| < 1$) para que o problema seja bem condicionado.

2.a)

A desigualdade é equivalente a $g(x) = (2f(x) - \cos(f(\frac{x}{2})) + 1)\frac{1}{3} \leq x$. Podemos concluir que a igualdade se verifica apenas nos pontos fixos de g. Em $\mathbf{R}$ há apenas um ponto fixo porque $|g'(x)| = \frac{1}{3}|2f'(x) + \frac{1}{2}f'(\frac{x}{2})\sin(f(\frac{x}{2}))| \leq \frac{1}{3}\frac{5}{2} = \frac{5}{6} < 1, \forall x \in \mathbb{R}$. Esse ponto fixo é zero pois $g(0) = 0 - 1 + 1 = 0$. Nos outros casos, ou $g(x) > x$ ou $g(x) < x$. Como $g'(0) < 1$ implica $g(x) < x$ para x próximo de zero, positivo, então estamos no segundo caso.

b)

Usando a) para $f(x) = \sin(x)$, que verifica as condições, obtemos imediatamente para $x \geq 0$:

$$-\Big(2\sin(x) - \cos(\sin(\frac{x}{2})) - 3x + 1\Big) + 2\sin(x) = 4x - 2$$

ou seja, equivalentemente, $x = g(x)$ com

$$g(x) = 1 + \cos(\sin(\frac{x}{2})).$$

Imediatamente, $g'(x) = -\frac{1}{2}\cos(\frac{x}{2})\sin(\sin(\frac{x}{2}))$ e $|g'(x)| \leq \frac{1}{2} < 1, \forall x \geq 0$, o que implica a contractividade em $[0, +\infty[$ conjunto fechado. Por outro lado $g(x) \in [0, 2]$, implica $g([0, +\infty[) \subseteq [0, +\infty[$, estando nas condições do teorema do ponto fixo, sabemos que há apenas um ponto fixo, e a convergência é assegurada, sendo alternada porque $g'(x) < 0$ para $x \in [0, 2]$, já que o coseno e o seno são positivos em [0,1] (e $\sin(x) \leq x$).

Assim sendo, podemos prever *a priori* que começando com $x_0 = 1$, ($|e_0| < 1$), sejam necessárias n iteradas: $|e_n| \leq (\frac{1}{2})^n < 10^{-2} \Rightarrow n > 2\log 10/\log 2 \sim 6.64.. \Rightarrow n \geq 7$. Mas *a posteriori* vemos que são precisas menos iteradas, porque

$$x_0 = 1, x_1 = 1.88726..., x_2 = 1.68972..., x_3 = 1.73313...$$

e como a convergência é alternada sabemos que $z \in [x_2, x_3]$ e temos mesmo $|e_3| \leq \frac{1}{4}|x_3 - x_2| = 0.0109...$ pelo que ainda é necessária mais uma iterada para podermos assegurar um erro absoluto inferior a 10^{-2}, e obtemos $x_4 = 1.7233804...$ Como informação, o valor exacto seria 1.725163...

3.

A equação característica associada a $x_{n+2} - \frac{9}{2}x_{n+1} + 2x_n = 0$ é $r^2 - \frac{9}{2}r + 2 = 0$, que tem como raízes $r_1 = \frac{1}{2}, r_2 = 4$, o que dá como solução geral da equação (homogénea):

$$x_n = A(\frac{1}{2})^n + B(4)^n$$

para obtermos $x_0 = 2, x_1 = 1$, resolvemos o sistema

$$\begin{cases} A + B = x_0 = 2 \\ A\frac{1}{2} + B \cdot 4 = x_1 = 1 \end{cases}$$

o que dá $A = 2, B = 0$, e a solução é $x_n = 2(\frac{1}{2})^n$ que converge para zero.

Deixamos como observação importante que se o dado inicial estivesse afectado de um pequeno erro, por exemplo, $x_0 = 2 + 10^{-4}$, já obteriamos uma sucessão divergente como solução (trata-se portanto de um problema mal condicionado)!

4a).

Começamos por ver que A é injectivo em X, porque $Ax = Ay$ implica $||x - y|| \leq \frac{1}{L}||Ax - Ay|| = 0$, logo $x = y$. Assim, existe aplicação inversa definida em $A(X)$, e

$$A^{-1} : A(X) \to X$$

é uma bijecção. Vamos ver que estamos nas condições do teorema de Banach para A^{-1} em $A(X)$:

i) $A(X)$ é não vazio (porque se $x \in X \neq \emptyset, Ax \in A(X)$) e fechado. Foi admitido, mas podemos provar:

Porque se $y_n \in A(X)$ e $y_n \to y$, então sendo $y_n = Ax_n$, com $x_n \in X$, temos $||x_n - x_m|| \leq \frac{1}{L}||Ax_n - Ax_m|| = \frac{1}{L}||y_n - y_m||$. Mas como (y_n) é sucessão de Cauchy (porque converge), então (x_n) também é, e converge para $x \in X$. Ora sendo A contínuo, $y_n = Ax_n \to Ax$, a unicidade do limite implica $y = Ax$, logo $y \in A(X)$.

ii) $A^{-1}(A(X)) \subseteq A(X)$, porque $A^{-1}(A(X)) = X$.

iii) A^{-1} é uma contracção em $A(X)$. Dados quaisquer $y_1, y_2 \in A(X)$, escrevemos $y_1 = Ax_1, y_2 = Ax_2$, e temos

$$||A^{-1}y_1 - A^{-1}y_2|| = ||x_1 - x_1|| \leq \frac{1}{L}||Ax_1 - Ax_2|| = \frac{1}{L}||y_1 - y_2||$$

em que $\frac{1}{L} < 1$.

Pelo Teorema de Banach $\exists^1 z \in A(X) : A^{-1}z = z$, mas também $Az = z$, pelo que $z \in X$. Por outro lado $y_{n+1} = A^{-1}y_n$ converge se $y_0 \in A(X)$. Equivalentemente, escrevendo $y_n = Ax_n$, temos $Ax_{n+1} = A^{-1}(Ax_n) = x_n$, com $x_0 \in X$.

4b)

Sendo X aberto contendo o ponto fixo z, existe sempre uma bola aberta $B(z,r) \subset X$, com $r > 0$, e uma bola mais pequena, fechada, $\bar{B}(z,r') \subset X$, com $r' < r$. Vemos que aí se cumprem as condições do T. de Banach:

i) $z \in \bar{B}(z,r') \neq \emptyset$, e $\bar{B}(z,r')$ é fechado

ii) A é contractivo em $\bar{B}(z,r')$, porque é contractivo em $X \supset \bar{B}(z,r')$.

iii) Para além disso $A(\bar{B}(z,r')) \subseteq \bar{B}(z,r')$ porque

$$x \in \bar{B}(z,r') \Rightarrow ||Ax - z|| = ||Ax - Az|| \leq L||x - z|| \leq Lr' < r' \Rightarrow Ax \in \bar{B}(z,r')$$

8.2.4 1a. Avaliação (98)

1.a) Começamos por reparar que $\sin(x) + x \geq 0$ se $x \geq 0$, pois $\sin(x)$ é positivo para $0 < x < 1$.

Portanto, se $x \geq 0$, temos $\sin(x) + 3x \geq 0$, logo $|\sin(x) + 3x| + \sin(x) = 2\sin(x) + 3x \geq 2(\sin(x) + x) \geq 0$.

Se $x \leq 0$, como o seno é ímpar, o resultado anterior dá-nos $\sin(x) + x \leq 0$, logo $|\sin(x) + 3x| + \sin(x) = -\sin(x) - 3x + \sin(x) = -3x \geq 0$.

Como a radiciação é bijectiva de $\mathbb{R}^+ \to \mathbb{R}^+$,

$$\sqrt{|\sin(x) + 3x| + \sin(x)} = 1 \Longleftrightarrow |\sin(x) + 3x| + \sin(x) = 1.$$

E, como vimos, podemos distinguir dois casos:

Se $x \leq 0$ então $|\sin(x) + 3x| + \sin(x) = 1 \Leftrightarrow -3x = 1 \Leftrightarrow x = -1/3$. Portanto só existe uma raiz negativa.

Se $x \geq 0$ então $|\sin(x) + 3x| + \sin(x) = 1 \Leftrightarrow 2\sin(x) + 3x - 1 = 0$. Temos duas maneiras de chegar ao resultado:

i) Designando $f(x) = 2\sin(x) + 3x - 1$, é uma função diferenciável em $\mathbb{R}$ e temos $f'(x) = 3 - 2\cos(x) > 0$. Portanto, existindo uma raiz, ela será única, pois a função f é estritamente crescente em $\mathbb{R}$. É fácil ver que a raiz existe, e é positiva, porque, por exemplo, podemos aplicar o T. valor intermédio em $[0, \frac{1}{2}]$ já que $f(0) = -1$ e $f(\frac{1}{2}) \sim 1.459$.

ii) Designando $g(x) = \frac{1}{3}(1 - 2\sin(x))$. Pelo teorema do ponto fixo em $\mathbb{R}$, como $|g'(x)| = |\frac{2}{3}\cos(x)| \leq \frac{2}{3} < 1$, podemos garantir que existe uma única raiz em $\mathbb{R}$, que é positiva porque $g([0, \frac{1}{2}]) \subset [0, \frac{1}{2}]$, porque g é decrescente nesse intervalo e $g(0) = 1/3$, $g(\frac{1}{2}) = .013716 \in [0, \frac{1}{2}]$.

1.b) Consideremos $f(x) = 2\sin(x) + 3x - 1$, $f \in C^2[0, \frac{1}{2}]$. Vamos utilizar o intervalo $[0, \frac{1}{2}]$, em que já mostrámos que existia uma raiz. Condições do M. de Newton:

i) $f(0)f(\frac{1}{2}) = (-1) \times 1.459 \leq 0$; ii) $f'(x) = 2\cos(x) + 3 \neq 0$, pois $\cos(x) = -\frac{3}{2}$ é impossível. (Já tinham sido verificadas em 1.a))

iii) $f''(x) = -2\sin(x) \leq 0$, se $x \in [0, \frac{1}{2}]$. iv) $|\frac{f(0)}{f'(0)}| = |\frac{-1}{5}| \leq 0.5$; $|\frac{f(0.5)}{f'(0.5)}| = \frac{1.459}{4.755} = 0.307 \leq 0.5$.

Consideramos $x_0 = 0.25$, imediatamente temos $|e_0| \leq 0.25$. Obtemos $x_1 = 0.2004219$, logo $|e_1| \leq \frac{\max|f''(x)|}{2|f'(x_0)|}|e_0|^2 \leq \frac{2\sin(\frac{1}{2})}{4.937}\ 0.25^2 = 0.01216...$

Com $x_2 = .2005366$, e o erro $|e_2| \leq \frac{2\sin(\frac{1}{2})}{2|f'(x_1)|}|e_1|^2 \leq \frac{0.95885}{4.9598}\, 0.01216^2 = 2.8586 \times 10^{-5}$ e como $f(0.2) < 0$, $|\delta_2| = \frac{|e_2|}{|z|} \leq \frac{0.29 \times 10^{-4}}{0.2} < 10^{-3}$.

1.c) Consideremos a sucessão $x_{n+1} = \frac{1}{3} - \frac{2}{3}\sin(x_n)$ com $x_0 = 0$. A sucessão converge devido à alínea a), por aplicação do T. do Ponto Fixo, e temos também a seguinte estimativa de erro $|x-x_n| \leq L^n|x-x_0| \leq \frac{1}{2}(\frac{2}{3})^n$. Por outro lado a sucessão (y_n) também converge porque sendo $g(y) = \frac{1}{3} - \frac{2}{3}(y - \frac{1}{6}y^3)$, temos também $g([0, \frac{1}{2}]) \subseteq [0, \frac{1}{2}]$ e $|g'(y)| = |-\frac{2}{3}(1 - \frac{1}{2}y^2)| \leq \frac{2}{3}$.

(i) Podemos obter rapidamente um majorante do erro usando

$$|x - y_n| \leq |x - y| + |y - y_n|.$$

Porque como $|y - y_n| \leq (\frac{2}{3})^n|y - y_0|$ basta reparar que $x - y = \frac{2}{3}(\sin(x) - y + \frac{1}{6}y^3)$ portanto, usando o T. Lagrange:

$$\begin{aligned} |x - y| &= \frac{2}{3}|\sin(x) - \sin(y) + \sin(y) - y + \frac{1}{6}y^3| \leq \\ &\leq \frac{2}{3}|\sin(x) - \sin(y)| + \frac{2}{3}|\sin(y) - y + \frac{1}{6}y^3| \leq \frac{2}{3}|x - y| + \frac{2}{3}\frac{|y|^5}{120} \end{aligned}$$

e daqui concluimos que $\frac{1}{3}|x - y| \leq \frac{2}{3}\frac{|y|^5}{120}$ o que nos dá então (notar que $y \in [0, \frac{1}{2}]$)

$$|x - y_n| \leq \frac{|y|^5}{60} + (\frac{2}{3})^n|y - y_0| \leq \frac{1}{1920} + (\frac{2}{3})^n\frac{1}{2}.$$

(ii) Outro processo, semelhante... mas 'mais correcto', seria utilizar desigualdade triangular

$$|x - y_n| \leq |x - x_n| + |x_n - y_n|.$$

Falta-nos apenas a estimativa para $|x_n - y_n|$. Observando que

$$|x_{n+1} - y_{n+1}| = \left|\frac{1}{3} - \frac{2}{3}\sin(x_n) - \left(\frac{1}{3} - \frac{2}{3}(y_n - \frac{1}{6}y_n^3)\right)\right| = \frac{2}{3}|\sin(x_n) - (y_n - \frac{1}{6}y_n^3)| \leq (*)$$

como sabemos que $|\sin(y_n) - (y_n - \frac{1}{6}y_n^3)| \leq \frac{|y_n|^5}{120}$, somando e subtraindo $\sin(y_n)$ obtemos

$$(*) \leq \frac{2}{3}|\sin(x_n) - \sin(y_n)| + \frac{2}{3}\frac{|y_n|^5}{120} \leq \frac{2}{3}|\cos(\xi_n)|\,|x_n - y_n| + \frac{2}{3}\frac{|y_n|^5}{120},$$

usando o T. de Lagrange para o seno. Assim, como $|\cos(\xi_n)| \leq 1$ e como $|y_n| \leq 0.5$ temos

$$\epsilon_{n+1} \leq \frac{2}{3}\epsilon_n + K$$

designando $\epsilon_n = |x_n - y_n|$ e $K = \frac{2}{3}\frac{0.5^5}{120}$. Podemos finalmente obter a expressão de a_n usando a equação às diferenças $a_{n+1} = \frac{2}{3}a_n + K$ que tem como solução particular: a constante $3K$, e como solução global $a_n = 3K + A(\frac{2}{3})^n$. Daqui sai $0 = a_0 = 3K + A \Rightarrow A = -3K$, portanto $a_n = 3K(1 - (\frac{2}{3})^n)$ e

$$\epsilon_n \leq 3K(1 - (\frac{2}{3})^n) \leq 3K = \frac{0.5^4}{120} = \frac{1}{2^7 \cdot 15} = 0.52 \cdot 10^{-3}$$

ou seja também ficamos com $|x - y_n| \leq \frac{1}{2}(\frac{2}{3})^n + \frac{1}{1920}$.

1.d) A coroa circular é definida por $\frac{1}{1+5} < |z| < 1 + \frac{5}{3}/\frac{1}{9}$, ou seja $\frac{1}{6} < |z| < 16$. Pela regra de Descartes há duas (ou zero) raízes positivas e uma negativa. Como $p(1) = -\frac{4}{3} + \frac{1}{9} < 0$ e $p(0) = \frac{1}{3} > 0$, concluimos que são todas reais e irão ficar nos intervalos $]-16, 0[,\ ]0, 1[,\ ]1, 16[$.

Há uma raiz dominante porque $z_1 + z_2 + z_3 = 0$ (o coeficiente em x^2 é zero) logo $z_1 = -z_3$ sse $z_2 = 0$, o que não acontece. Assim, o método de Bernoulli converge. Fazendo $y_0 = 0,\ y_1 = 0,\ y_2 = 1$, com

$$y_{n+3} = 15y_{n+1} - 3y_n$$

obtemos $y_3 = 0, y_4 = 15, y_5 = -3, y_6 = 225, y_7 = -90$ e encontramos $y \sim \frac{-90}{225} = -\frac{2}{5}$ que ainda se encontra muito longe da raiz dominante (que está próximo de -5).

2. Se $g(x)$ é uma função racional podemos escrevê-la como fracção de dois polinómios de coeficientes inteiros

$$g(x) = \frac{p(x)}{q(x)}.$$

Caso existisse uma tal função iteradora, π seria ponto fixo de g, logo $\pi = \frac{p(\pi)}{q(\pi)} \Leftrightarrow q(\pi)\pi - p(\pi) = 0$, o que significava que π era raiz de uma equação polinomial com coeficientes inteiros, ou seja um número algébrico, o que é falso, pois sabemos que π é transcendente.

3. Reparamos que a regra enunciada corresponde a uma equação às diferenças homogénea

$$a_{k+1} + 2a_k - a_{k-1} - 2a_{k-2} = 0.$$

Associando a equação característica $r^3 + 2r^2 - r - 2 = 0$, é fácil ver que $1, -1$, e -2 são as soluções dessa equação, e assim temos a solução global

$$a_k = A + B(-1)^n + C(-2)^n.$$

As constantes podem ser determinadas a partir do sistema

$$\left\{ \begin{array}{c} A + B + C = a_0 = 4 \\ A - B - 2C = a_1 = 8 \\ A + (-1)^9 B + (-2)^9 C = a_9 = 180 \end{array} \right. \Leftrightarrow \left[\begin{array}{ccc} 1 & 1 & 1 \\ 1 & -1 & -2 \\ 1 & -1 & -512 \end{array} \right] \left[\begin{array}{c} A \\ B \\ C \end{array} \right] = \left[\begin{array}{c} 4 \\ 8 \\ 180 \end{array} \right]$$

obtendo-se $C = -172/510 = -0.33726,\ B = -1.49412,\ A = 5.83137$.

Assim, $a_4 = 5.83137 - 1.49412 \cdot (-1)^4 - 0.33726 \cdot (-2)^4 = -1.05883$. Neste caso a coluna seria uma espécie de estalactite....

4. Como $BA : X \to X$, em que X é espaço de Banach (logo fechado, não-vazio), basta ver qual a condição a impôr a K_A e K_B para que haja contractividade e concluir acerca da convergência. Ora,

$$||BAx_1 - BAx_2||_X \leq K_B||Ax_1 - Ax_2||_Y \leq K_B K_A||x_1 - x_2||_X\ ,\ \forall x_1, x_2 \in X,$$

portanto será suficiente que $K_A K_B < 1$.

Considerando $A = 2I$ em que I é a matriz identidade em $\mathbb{R}^d$, temos claramente $||Ax_1 - Ax_2|| = 2||x_1 - x_2||$, logo $K_A = 2$. Por outro lado com $B = \frac{1}{3}I$ obtemos $BA = \frac{2}{3}I$ e como é

claro temos contractividade. Neste caso simples, podemos mesmo ver que $x_n = (\frac{2}{3})^n x_0$ que converge para zero.

5. a) Como $\sin(a) \in [0,1]$, pelo T.Gerschgorin (aplicado a linhas e colunas) os valores próprios λ têm que pertencer à reunião de bolas definidas por

$$|\lambda - 9 - \sin(a)| \leq |\cos(a)| \, ; |z - 1 - \cos(a)| \leq |\cos(a)| \, ; \; |\lambda + 2| \leq 1$$

por outro lado podemos aplicar a desigualdade triangular, $|\lambda - 9| - |\sin(a)| \leq |\lambda - 9 - \sin(a)| \leq 1 \Rightarrow |\lambda - 9| \leq 2$, e da mesma forma $|\lambda - 1| - |\cos(a)| \leq |\lambda - 1 - \cos(a)| \leq \cos(a) \Rightarrow |\lambda - 1| \leq 2|\cos(a)|$. Como não há intersecção das bolas (a única possibilidade seria em $\lambda = -1$ caso tivessemos $\cos(a) = -1$... o que não acontece), os três valores próprios que têm que ser reais (a matriz é real) e pertencem a $[-3,-1]$, a $[-1,3]$, e a $[7,11]$.

Sabendo agora que são reais, vemos que $\lambda \in [7,11]$ pode ser melhorado para

$$\lambda \in [9 + \sin(a) - \cos(a), 9 + \sin(a) + \cos(a)] \subseteq [8 + \sin(a), 10 + \sin(a)] \subset [8,11].$$

Da mesma forma, podemos melhorar $\lambda \in [-1,3]$ para $\lambda \in [1 + \cos(a) - \cos(a), 1 + \cos(a) + \cos(a)] = [1, 1 + 2\cos(a)] \subset [1,3]$ intervalo ao qual zero não pertence. Como não pode haver nenhum valor próprio nulo, a matriz é invertível.

Outra possibilidade para concluir da invertibilidade é ver que a matriz tem a diagonal estritamente dominante por linhas, porque $|9 + \sin(a)| > 4$, $|1 + \cos(a)| > 1$ e $|-2| > 1$.

b) Existe valor próprio dominante que está no intervalo $[8,11]$, todos os valores próprios são distintos, podemos aplicar o método das potências. Começamos a iteração com o vector $x^{(0)} = (1,0,0)$. Portanto,

$$x^{(1)} = \sigma_0 \frac{Ax^{(0)}}{||Ax^{(0)}||_\infty} = + \frac{(9,1,0)}{9} = (1, \frac{1}{9}, 0),$$

e

$$x^{(2)} = + \frac{(9 + 1/3, 1 + 2/9, 1/9)}{9 + 1/3} = (1, 11/84, 1/84).$$

Calculando $Ax^{(2)} = (\frac{395}{42}, \frac{53}{42}, \frac{3}{28})$ obtemos $\lambda \sim 9.4048$.

c) O cálculo de A^k pode ser efectuado guardando A^{k-1}. Para calcular o produto de duas matrizes efectuamos d^3 operações $(*, /)$, logo necessitamos de $d-1$ vezes d^3 operações $\sim d^4$. Podemos melhorar este valor elevado, determinando apenas a linha (ou coluna) $A^k_{(i)}$ de A^k que nos interessa para resolver o sistema, já que $A^k_{(i)} = A^{k-1}_{(i)} A$ envolve apenas d^2 operações, reduzindo assim o número de operações para $d-1$ vezes d^2.

Agora, determinar os coeficientes $\alpha_0, ..., \alpha_{d-1}$ é resolver o sistema

$$\alpha_0 I + \alpha_1 A + ... + \alpha_{d-1} A^{d-1} + A^d = 0$$

que tem solução caso as matrizes I, A, ..., A^{d-1} sejam linearmente independentes. No caso de se obter uma linha (ou coluna) independente para I, A, ..., A^{d-1}, é possível determinar esses coeficientes em $\sim d^3/3$ operações (caso contrário poderíamos ter que verificar $d^2 - d$ sistemas...). No caso favorável, o número total de operações $(*, /)$ será $\sim d^3 - d + \frac{1}{3}d^3 \sim \frac{4}{3}d^3$.)

No caso concreto de A com $a = 0$, basta reparar que a primeira linha de $A^2 = (84, 34, 7)$ e a de A^3 é $(790, 327, 70)$. Obtemos

$$\begin{bmatrix} 1 & 9 & 84 \\ 0 & 3 & 34 \\ 0 & 1 & 7 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} -790 \\ -327 \\ -70 \end{bmatrix},$$

e como também sabemos que o traço de A é $-\alpha_2$ obtemos imediatamente $\alpha_2 = -9$, assim

$$\alpha_1 = -70 - 7\alpha_2 = -7,\ \alpha_0 = -790 - 9\alpha_1 - 84\alpha_2 = -790 + 819 = 29.$$

O polinómio característico é $\lambda^3 - 9\lambda^2 - 7\lambda + 29$.

6. Sejam $\lambda_1, ..., \lambda_d$ os valores próprios de C_ω, sabemos que $|\det(C_\omega)| = |\lambda_1 \cdots \lambda_d|$. Reparemos que

$$\begin{aligned} \det(C_\omega) &= \det(D + \omega L)^{-1} \det((1-\omega)D - \omega U) = \\ &= \det(D^{-1}) \det((1-\omega)D) = |1-\omega|^d \end{aligned}$$

porque as matrizes L e U têm diagonais nulas. Assim, $\rho(C)^d \geq |\lambda_1| \cdots |\lambda_d| = |1-\omega|^d$, o que prova a sugestão. Sendo necessário que $\rho(C_\omega) < 1$ pela teoria, logo $|1-\omega| \leq \rho(C_\omega) < 1$ implica (com ω real) $\omega \in]0, 2[$.

7. (Identificação de caracteres) Basta reparar que $(\phi_m, \phi_n) = \sum_{i=1}^{6} \phi_m(i)\phi_n(i)$. Assim, p.ex.

$$\begin{aligned} (\phi_1, \phi_1) &= 0 + 1 + 1 + 1 + 1 + 0 = 4, \\ (\phi_1, \phi_2) &= 0 \cdot 0 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 = 3,\ etc... \end{aligned}$$

obtendo-se os sistemas normais (respectivamente para f_1 e f_2) :

$$\begin{bmatrix} 4 & 3 & 3 \\ 3 & 4 & 2 \\ 3 & 2 & 3 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \\ c_1 \end{bmatrix} = \begin{bmatrix} 63 \\ 60 \\ 50 \end{bmatrix};\ \begin{bmatrix} 4 & 3 & 3 \\ 3 & 4 & 2 \\ 3 & 2 & 3 \end{bmatrix} \begin{bmatrix} a_2 \\ b_2 \\ c_2 \end{bmatrix} = \begin{bmatrix} 52 \\ 49 \\ 39 \end{bmatrix}$$

usando a decomposição da matriz $A = LL^T = \begin{bmatrix} 2 & 0 & 0 \\ \frac{3}{2} & \frac{\sqrt{7}}{2} & 0 \\ 1 & \frac{-1}{2\sqrt{7}} & \sqrt{\frac{5}{7}} \end{bmatrix} \begin{bmatrix} 2 & \frac{3}{2} & 1 \\ 0 & \frac{\sqrt{7}}{2} & \frac{-1}{2\sqrt{7}} \\ 0 & 0 & \sqrt{\frac{5}{7}} \end{bmatrix}$ (a matriz do sistema normal está sempre nas condições de aplicabilidade do método de Cholesky), obtemos de $Lg = (63, 60, 50)$, $g = (31.5, 9.63809, 5.40899)$, e de $L^t x = g$ assim $x = (4.8, 6.2, 6.4)$. A maior componente é a terceira, pelo que a figura pretendida que melhor aproxima f_1 é a associada a ϕ_3.

Finalmente, para f_2, de $Lg = (52, 49, 39)$, $g = (23, 7.59929, 1.69031)$, e de $L^t x = g$ obtemos $x = (7,\ 6, 2)$, e a figura pretendida é a associada a ϕ_1.

8.2.5 2a. Avaliação (98)

1.

Queremos resolver $x = g(x)$. Como $0 < |e^{-(x-25)^2/2}| \leq 1$, então o ponto fixo tem que estar no intervalo $[0, 1]$.

Sendo $f(x) = x - g(x)$, $f'(x) = 1 + (x-25)e^{-(x-25)^2/2}$, e $f''(x) = -(x-25)^2 e^{-(x-25)^2/2} \leq 0$.

Como $f'(0) = 1 - 25e^{-25^2/2} \sim 1$, $f'(1) = 1 - 24e^{-24^2/2} \sim 1$, então a função f é estritamente crescente e como temos

$$f(0)f(1) = -e^{-25^2/2}(1 - e^{-24^2/2}) < 0,$$

concluímos que existe uma e uma só raiz em $[0, 1]$.

Usando o teorema de Lagrange, temos $|e_0| = |z - 0| \leq \frac{|0-f(0)|}{\min |f'(x)|}$, e daqui

$$|e_0| \leq e^{-25^2/2} < 10^{-25^2/6}.$$

Quando a aproximação é 0 o erro relativo é sempre 100%, porque $|\delta| = \frac{|z-0|}{|z|} = 1$.

2.a)

Trata-se de determinar as raizes do polinomio característico de

$$A = \begin{bmatrix} 10 & -2 & -1 \\ -1 & -6 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \text{ ou seja, encontrar } \lambda = x^3 : \det(\lambda I - A) = 0$$

Basta agora localizar os λ's usando o T. Gerschgorin. Isso dá-nos, por colunas, $\lambda \in \bar{B}(10, 1) \cup \bar{B}(-6, 3) \cup \bar{B}(0, 1)$, e como as bolas são disjuntas podemos concluir que existe um valor próprio em cada uma, que será sempre real, pois a matriz também o é. Por linhas obtemos $\lambda \in \bar{B}(10, 3) \cup \bar{B}(-6, 1) \cup \bar{B}(0, 1)$.

Intersectando a informação, podemos ordenar $\lambda_1 \in [9, 11]$, $\lambda_2 \in [-7, -5]$, $\lambda_3 \in [-1, 1]$, o que significa que

$$z_1 \in [\sqrt[3]{9}, \sqrt[3]{11}],\ z_2 \in [-\sqrt[3]{7}, -\sqrt[3]{5}],\ z_3 \in [-1, 1].$$

2.b)

Calculando o determinante temos $p(\lambda) = (\lambda - 10)(\lambda + 6)\lambda - (2\lambda - 1) = 0 \Leftrightarrow p(\lambda) = \lambda^3 - 4\lambda^2 - 62\lambda + 1 = 0$.

i) No intervalo $I = [9, 11]$ temos $p(9) = -152$, $p(11) = 166$. ii) Por outro lado $p'(\lambda) = 3\lambda^2 - 8\lambda - 62$ é sempre positivo em I porque $p''(\lambda) = 6\lambda - 8 > 0$ em I, logo p' é crescente e como $p'(9) = 109 > 0$, fica provado. iii) Já vimos que $p''(\lambda) > 0$, basta escolher duas iteradas iniciais positivas. Escolhendo $x_{-1} = 11, x_0 = 10.5$, como $p(11) = 166$, $p(10.5) = 66.625$, fica provado iv)a).

Fazemos agora as iterações não esquecendo que $\delta_{\sqrt[3]{x_n}} \sim \frac{1}{3}\delta_{x_n}$ e portanto bastará encontrar $|\delta_{x_n}| < 3 \cdot 10^{-2}$, e como $x_n > 9$, basta $|e_n| < 0.27$

$$x_1 = x_0 - p(x_0)\frac{x_0 - x_{-1}}{p(x_0) - p(x_{-1})} = 10.5 - 66.625\frac{0.5}{99.375} = 10.16477$$

Podemos mesmo ver que $p(10) = -19$, logo $|e_{-1}| \leq 1$, $|e_1| \leq 0.5$. Como sabemos que o erro verifica

$$|e_1| \leq \frac{\max |p''(x)|}{2 \min |p'(x)|} |e_0| \, |e_{-1}| \leq \frac{58}{218} 0.5 < 0.111.$$

2.c)

$$\begin{bmatrix} 16 & -2 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 6 \end{bmatrix} \longrightarrow \begin{bmatrix} 16 & -2 & -1 \\ \frac{-1}{16} & (0 & 0) \\ 0 & (-1 & 6) \end{bmatrix} \longrightarrow \begin{bmatrix} 16 & -2 & -1 \\ \frac{-1}{16} & -\frac{1}{8} & -\frac{1}{16} \\ 0 & 8 & (6) \end{bmatrix} \longrightarrow \begin{bmatrix} 16 & -2 & -1 \\ \frac{-1}{16} & -\frac{1}{8} & -\frac{1}{16} \\ 0 & 8 & 6 + \frac{1}{2} \end{bmatrix}$$

Assim,

$$\begin{bmatrix} 16 & -2 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{-1}{16} & 1 & 0 \\ 0 & 8 & 1 \end{bmatrix} \begin{bmatrix} 16 & -2 & -1 \\ 0 & -\frac{1}{8} & -\frac{1}{16} \\ 0 & 0 & \frac{13}{2} \end{bmatrix}$$

Resolvemos $Ly = (0, 1, 0)$ e obtemos $y_1 = 0, y_2 = 1, y_3 = -8$, depois $Ux = (0, 1, -8)$ dá $x_3 = \frac{-16}{13}, x_2 = -8(1 - \frac{1}{13}) = \frac{-96}{13}, x_1 = -\frac{208}{16 \cdot 13} = -1$.

De forma análoga teríamos para o segundo sistema $x = (1248, 9310, 1517)/169$

2.d)

Escolhemos $\lambda = 6$ como aproximação e temos

$$A + 6I = \begin{bmatrix} 16 & -2 & -1 \\ -1 & 0 & 0 \\ 0 & -1 & 6 \end{bmatrix}$$

que é a matriz do sistema anterior! Escolhendo $x^{(0)} = (0, 1, 0)$, estamos nas condições pretendidas em c).

Rapidamente obtemos

$$x^{(1)} = \sigma_1 \frac{(A + 6I)^{-1} x^{(0)}}{||(A + 6I)^{-1} x^{(0)}||_\infty} = - \frac{\frac{-1}{13}(13, 96, 16)}{96/13} = (\frac{13}{96}, 1, \frac{16}{96})$$

e daqui

$$x^{(2)} = \sigma_1 \frac{(A + 6I)^{-1} x^{(1)}}{||(A + 6I)^{-1} x^{(1)}||_\infty} = \frac{(1248, 9310, 1517)/169}{9310/169} = (0.134049, 1, 0.162943)$$

portanto, como $Ax^{(2)} = (-0.822448, -6.13404, -1)$, obtemos $\lambda^{(2)} = -6.13404$, e daqui $z_2 \sim -1.83055$.

Vejamos que $p(\lambda^{(2)}) = 0.002563865$, e como $p'(-7) = 141, p'(-5) = 53$, com p' crescente, temos

$$|z_2 - \lambda^{(2)}| \leq \frac{|p(z_2) - p(\lambda^{(2)})|}{\min_{x \in [-7, -5]} \, |p'(x)|} \leq \frac{0.002563865}{53}$$

3.a)

Basta ver que pelo T.Gerschgorin por colunas temos $\lambda \in \bar{B}(0, |a_0|) \cup_k \bar{B}(0, |a_k| + 1) \cup \bar{B}(|a_{n-1}|, 1)$.

Esta última bola é disjunta das restantes porque

$$|\lambda_k| \leq \max_k\{|a_0|, |a_k| + 1\} = M,$$

por outro lado $|\lambda_{n-1} + a_{n-1}| \leq 1$, logo $|\lambda_{n-1}| \geq |a_{n-1}| - 1$.

Assim como $|\lambda_k| \leq M < |a_{n-1}| - 1 \leq |\lambda_{n-1}|$, não há intersecção das bolas!

3.b)

Neste caso

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ -1 & 0 & 3 & -2 & 8 \end{bmatrix}$$

e portanto pelo teorema anterior temos $|\lambda_5 - 8| \leq 1$. Por outro lado como por linhas $|\lambda_i| \leq 1$, e $|\lambda_5 - 8| \leq 6$ não intersecta a outra, concluímos que estão todas as outras na bola $\bar{B}(0, 1)$.

Como a raiz dominante é tal que $|\lambda_5 - 8| \leq 1$,escolhemos $x^{(0)} = (0, 0, 0, 0, 1)$.

$$\begin{aligned} x^{(1)} &= \sigma_0 \frac{Ax^{(0)}}{||Ax^{(0)}||_\infty} = \frac{(0, 0, 0, 1, 8)}{8} = (0, 0, 0, \frac{1}{8}, 1); \\ x^{(2)} &= \sigma_1 \frac{Ax^{(1)}}{||Ax^{(1)}||_\infty} = (0, 0, \frac{2}{31}, \frac{4}{31}, 1) \end{aligned}$$

Assim $\lambda^{(2)} = \frac{(Ax^{(2)})_5}{x_5^{(2)}} = \frac{6-8}{31} + 8$.

4.a)

Para determinar a derivada precisamos de calcular

$$\begin{aligned} &A_m(f + h) - A_m(f) = \int_0^x (f(y) + h(y))^m dy - \int_0^x (f(y))^m dy = \\ &= \int_0^x mf(y)^{m-1}h(y)dy + \int_0^x P_{f,h}(y)h(y)^2 dy = \\ &= \int_0^x mf(y)^{m-1}h(y)dy + o(||h||_\infty) \end{aligned}$$

portanto $A'_{m,f}(h) = m \int_0^x f(y)^{m-1}h(y)dy$

4.b)

X é fechado e convexo, vamos provar as outras condições do Corolário do T. Ponto Fixo de Banach. Começamos por ver que

$$||A'_{m,f}||_{\mathcal{L}(.,.)} = \sup_{h \neq 0} \frac{||A'_{m,f}(h)||_\infty}{||h||_\infty} \leq m \int_0^1 |f(x)|^{m-1} dx < mK^{m-1}$$

e portanto como temos

$$f(x) = \frac{1}{4}\phi(x) - \frac{1}{4}\int_0^x (f(y))^5 dy + \frac{1}{4}\int_0^x (f(y))^2 dy = Bf(x)$$

então $B'_f = \frac{1}{4}A'_{5,f} + \frac{1}{4}A'_{2,f}$ e temos para $f \in X$

$$||B'_f||_{\mathcal{L}(.,.)} = ||\frac{1}{4}A'_{5,f} + \frac{1}{4}A'_{2,f}||_{\mathcal{L}(.,.)} \leq \frac{1}{4}(5(\frac{3}{4})^4 + 2(\frac{3}{4})^2) < 1$$

Falta ver que $B(X) \subseteq X$. Ora,

$$||Bf(x)|| \leq \frac{1}{4}||\phi(x)|| + \frac{1}{4}||\int_0^x (f(y))^5 dy|| + \frac{1}{4}||\int_0^x (f(y))^2 dy|| \leq \frac{3}{4}.$$

4.c)

Começando com $f_0 = 0$, temos

$$f_1 = \frac{1}{4}\phi$$

portanto

$$||e_0|| \leq \frac{1}{1-L}||f_1 - f_0|| = 4||\frac{1}{4}\phi|| \leq 1$$

como $||e_n|| \leq (\frac{3}{4})^n||e_0||$, basta que $(\frac{3}{4})^n < \frac{1}{100}$

5.a)

Vamos aplicar o Teorema do Ponto Fixo ao conjunto fechado e convexo $S = [-1,1] \times [-1,1]$ com a função g obtida a partir da equivalência:

$$\begin{array}{l} f(x,y) = (-1,1) \Leftrightarrow \left\{ \begin{array}{c} x\cos(xy) - 3y + 1 = 0 \\ y\sin(x+y) - 4x - 1) = 0 \end{array} \right. \Leftrightarrow \\ \Leftrightarrow (x,y) = (\frac{y}{4}\sin(x+y) - \frac{1}{4},\ \frac{x}{3}\cos(xy) + \frac{1}{3}) = g(x,y) \end{array}$$

Vemos que $g(S) \subseteq S$, porque se $x, y \in S$ então

$$|\frac{y}{4}\sin(x+y) - \frac{1}{4}| \leq \frac{1}{2},\ |\frac{x}{3}\cos(xy)) + \frac{1}{3}| \leq \frac{2}{3},$$

o que implica $g(S) \subseteq [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{2}{3}, \frac{2}{3}]$. Como

$$J_g(x,y) = \left[\begin{array}{cc} \frac{y}{4}\cos(x+y) & \frac{1}{4}\sin(x+y) + \frac{y}{4}\cos(x+y) \\ \frac{-xy}{3}\sin(xy) + \frac{1}{3}\cos(xy) & \frac{-x^2}{3}\sin(xy) \end{array} \right]$$

então para qualquer $x, y \in X$

$$\begin{array}{l} ||J_g(x,y)||_1 = \max\left\{|\frac{y}{4}\cos(x+y)| + |\frac{-xy}{3}\sin(xy) + \frac{1}{3}\cos(xy)|,\right. \\ \left. |\frac{1}{4}\sin(x+y) + \frac{y}{4}\cos(x+y)| + |\frac{-x^2}{3}\sin(xy)|\right\} \leq \\ \leq \max\left\{\frac{1}{4} + \frac{1}{3} + \frac{1}{3}, \frac{1}{4} + \frac{1}{4} + \frac{1}{3}\right\} = \frac{11}{12} < 1 \end{array}$$

Daqui concluímos pelo T. Ponto Fixo que existe um e um só ponto fixo de g em X, o que é equivalente a ser solução da equação $f(x,y) = 0$.

Não pode existir nenhuma raiz se $|x| > 1$ ou $|y| > 1$, porque: **(i)** se $|x| > 1$ então

$$\begin{array}{rcl} |x| & = & |\frac{y}{4}\sin(x+y) - \frac{1}{4}| \leq \frac{|y|+1}{4} = \frac{1}{4}|\frac{x}{3}\cos(xy) + \frac{1}{3}| + \frac{1}{4} \leq \\ & \leq & \frac{1}{4}|\frac{x}{3}| + \frac{1}{3} + \frac{1}{4} = |\frac{x}{12}| + \frac{7}{12} \end{array}$$

o que implica $\frac{11}{12}|x| \leq \frac{7}{12} \Rightarrow |x| \leq \frac{7}{11} < 1$, o que dá uma contradição; **(ii)** da mesma forma se $|y| > 1$ então

$$\begin{aligned} |y| &= |\frac{x}{3}\cos(xy) + \frac{1}{3}| \leq \frac{|x|+1}{3} = \frac{1}{3}|\frac{y}{4}\sin(x+y) - \frac{1}{4}| + \frac{1}{3} \leq \\ &\leq \frac{1}{3}|\frac{y}{4}| + \frac{1}{4} + \frac{1}{3} = |\frac{y}{12}| + \frac{7}{12} \end{aligned}$$

e isto dá a mesma contradição porque implica que $|y| < 1$.

6. As funções base podem ser $\phi_0(x) = 1, \phi_1(x) = (x - \frac{1}{2})^2$ o que dá o sistema normal

$$\begin{bmatrix} 1 & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{80} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \frac{2}{\pi} \\ \frac{\pi^2-8}{2\pi^3} \end{bmatrix}$$

porque

$$\begin{aligned} (\phi_0, \phi_0) &= \int_0^1 1^2 dx = 1, \\ (\phi_0, \phi_1) &= \int_0^1 (x - \frac{1}{2})^2 dx = \frac{1}{3}(x - \frac{1}{2})^3|_{x=0}^{x=1} = \frac{1}{12}, \\ (\phi_1, \phi_1) &= \frac{1}{5}(x - \frac{1}{2})^5|_{x=0}^{x=1} = \frac{1}{80}, \end{aligned}$$

e

$$\begin{aligned} (\phi_0, f) &= \int_0^1 \sin(\pi x) dx = -\frac{1}{\pi}\cos(\pi x)|_{x=0}^{x=1} = \frac{2}{\pi}, \\ (\phi_1, f) &= \int_0^1 (x^2 - x + \frac{1}{4})\sin(\pi x) dx = -\frac{4}{\pi^3} + \frac{1}{4}\frac{2}{\pi}. \end{aligned}$$

Resolvendo o sistema obtemos $(a, b) = (0.980162, -4.12251)$

8.2.6 Teste (99)

1. a)

Temos $e^{200.48\, f(x)} = 1 \Leftrightarrow f(x) = 0 \Leftrightarrow x = -\frac{1}{2}(\cos(\frac{x}{2}) + \sin(x)) = g(x)$. Por outro lado, $|g'(x)| \leq \frac{3}{4} < 1$, pois

$$|g'(x)| = |\frac{1}{2}\frac{1}{2}\sin(\frac{x}{2}) - \frac{1}{2}\cos(x)| \leq \frac{1}{4} + \frac{1}{2} = \frac{3}{4} = L < 1.$$

Aplicando T.P.Fixo, em que $E = \mathbb{R}$ fechado, convexo, não vazio, com $g(\mathbb{R}) \subseteq \mathbb{R}$,temos existência e unicidade em $\mathbb{R}$.

Por outro lado,

$$-1 = \frac{-1}{2}(1+1) \leq -\frac{1}{2}(\cos(\frac{x}{2}) + \sin(x)) \leq \frac{-1}{2}(-1-1) = 1$$

ou seja $g(\mathbb{R}) \subseteq [-1,1]$, logo $z = g(z) \in [-1,1]$.

Sabemos que $|e_n| \leq L^n|z-x_0|$, fazendo $x_0 = 0$, temos $\frac{|e_n|}{|z|} \leq L^n = (\frac{3}{4})^n$ (note que $z \neq 0$), portanto se $(\frac{3}{4})^n < 0.1$ teremos $|\delta_n| = \frac{|e_n|}{|z|} < 0.1$.

Isso acontece se $n\log(0.75) < \log(0.1) \Leftrightarrow n > -1/\log_{10}(0.75) = 8.0039$, ou seja, basta efectuar 9 iterações.

((Outro processo, seria considerar um $\tilde{z} : |\tilde{z}| \leq |z|$, e assim $|\delta_n| = \frac{|e_n|}{|z|} \leq \frac{|e_n|}{|\tilde{z}|}$ e bastaria exigir $|e_n| \leq 0.1|\tilde{z}|$ com as estimativas habituais.))

b) Como $f(-1) < 0$ a função é negativa em $[-1, z[$ e a área negativa é dada por

$$A(x) = \int_{-1}^{x} f(t)dt, \text{ tendo-se } |\tilde{\delta}_A| = \frac{xA'(x)}{A(x)}\delta_x = \frac{xf(x)}{A(x)}\delta_x$$

e quando $x = z$, como $f(z) = 0$ e $A(z) \neq 0$ (porque a área é mesmo negativa!) então $|\tilde{\delta}_A| = 0$ e o problema é *muito* bem condicionado próximo de z.

2. a) Vamos usar a regra de Descartes. Basta reparar que se $x > 0$ então temos

$$ax^6 + bx^4 + cx^2 + dx - 1 = 0,$$

o que corresponde a $++++-$ e há uma variação mínima de sinal, logo existe uma e uma só raiz positiva (reparando também que b, c, d nulos não alteram as variações mínimas).

Se $x < 0$ então $p(x) = ax^6 + bx^4 + cx^2 - dx - 1$, e assim $p(-x)$ dá também $+++++-$, i.e: uma variação de sinal, e existe uma só raiz negativa.

Para determinar um intervalo, usamos a regra do máximo, reparando que $|\pm d| = d$, e que $M = \max\{b,c,d,1\} = \max\{1,b\}$, bem como $\tilde{M} = \max\{a,b,c,d\} = \max\{a,b\}$

$$\text{a função é par, logo } z^+ = -z^-, \ \frac{1}{1+\max\{a,b\}} < z^+ < 1 + \frac{\max\{b,1\}}{a},$$

b) Começando com $x_0 = 0, x_1 = 0, x_2 = 1$, temos $x_3 = -10x_2 - 2x_1 + x_0 = -10$.

Sucessivamente $x_4 = -10x_3 - 2x_2 + x_1 = 100 - 2 = 98$, $x_5 = -10 \cdot 98 + 20 + 1 = -959$.

Ou seja, $y_5 = \frac{x_5}{x_4} = -959/98 = -9.78571...$

Sendo $q(x) = x^3 + 10x^2 + 2x - 1$, para avaliarmos o erro, vemos que $q(y_5) = -0.0509838$, que $q(y_4) = q(-9.59) = 17.527 > 0$,

que $q'(x) = 3x^2 + 20x + 2$ varia entre 93.567 e 82.104 no intervalo $I = [y_5, y_4]$ onde estará a raiz z.

(Note que $q'' = 6x + 20 < 0$ se $x < \frac{-10}{3}$). Usamos a estimativa a posteriori

$$|e_5| = |z - y_5| \leq \frac{|f(y_5)|}{\min_{x\in I}|3x^2 + 20x + 2|} = \frac{0.0509838}{82.104} = 0.000621.$$

c) Isto significa que uma raiz de $x^3 + 10x^2 + 2x - 1$ será $\sim \frac{1}{4.351} = 0.229832$, e portanto temos as raízes reais de $(*)$ iguais a $\pm\sqrt{0.229832} = \pm 0.479408$.

Da alínea b) surgem as raízes complexas $\pm 3.12812\mathbf{i}$.

Para determinar as restantes poderiamos dividir x^3+10x^2+2x-1 por $(x+9.78571)$ e por $(x-0.229832)$, usando a regra de Ruffini (o que daria $\sim x+0.444122$).

Mas, muito mais fácil é reparar que $-z_1z_2z_3=-1$, logo

$$z_2 \sim \frac{1}{0.229832\cdot(-9.78571)} = -0.444628$$

Portanto as restantes raízes são $\pm 0.666805\mathbf{i}$.

(Os valores com 6 decimais, seriam $\pm 0.479403, \pm 0.666830\mathbf{i}, \pm 3.12812\mathbf{i}$, ou seja, foram obtidos excelentes resultados!)

3. a) Para calcular a derivada de Fréchet,

$$P(f+h)-Pf = \int_0^x g(f(t)+h(t))\,dt - g(f(t))\,dt$$

como $g \in C^2(\mathbf{R})$, usamos a fórmula de Taylor $g(y+\epsilon)-g(y) = g'(y)\epsilon + \frac{1}{2}g''(\xi)\epsilon^2$, com $\xi \in]y; y+\epsilon[$, aplicada aos valores $y=f(t), \epsilon = h(t)$.

Obtemos

$$P(f+h)-Pf = \int_0^x g'(f(t))h(t) + \frac{1}{2}g''(\xi_t)\,h(t)^2 dt,$$

em que $\xi_t \in]f(t), f(t)+h(t)[$. Assim, $P'_f h = \int_0^x g'(f(t))h(t)dt$ que é um operador linear contínuo, pois

$$\begin{aligned} ||P(f+h)-Pf-\int_0^x g'(f(t))h(t)||_\infty &= ||\int_0^x \frac{1}{2}g''(\xi_t)\,h(t)^2 dt||_\infty \leq \\ &\leq \frac{a}{2}\max_{\xi\in S}|g''(\xi)|\;||h||_\infty^2 = o(||h||_\infty), \end{aligned}$$

em que designámos

$$S = [\min_{t\in[0,a]}\{f(t), f(t)+h(t)\}, \max_{t\in[0,a]}\{f(t), f(t)+h(t)\}]$$

com $||h||_\infty \leq 1$.

Querendo calcular $||P'_f||_{\mathcal{L}(.,.)}$, reparamos que

$$\begin{array}{l} ||P'_f||_{\mathcal{L}(.,.)} = \sup_{h\neq 0}\frac{||P'_f h||}{||h||}, \text{ e que} \\ \frac{||P'_f h||_\infty}{||h||_\infty} = \frac{1}{||h||_\infty}\max_{x\in[0,a]}|\int_0^x g'(f(t))h(t)dt| \leq \\ \leq \int_0^x \max_{t\in[0,a]}\frac{|g'(f(t))|\,|h(t)|}{||h||_\infty}dt \leq a||g'\circ f||_\infty \end{array}$$

b) Designando $(Bf)(x) = f(x)^2$ e $(Pf)(x) = \int_0^x \cos(f(t))\,dt$, temos

$$f = Bf - Pf \Leftrightarrow f = Af, \text{ designando } A = B - P.$$

Podemos tentar aplicar o corolário do teorema do ponto fixo, reparando que $(B'_f h)(x) = 2f(x)h(x)$, pois

$$B(f+h)-Bf = (f+h)^2 - f^2 = 2fh + h^2$$

Assim $||B'_f||_{\mathcal{L}(.,.)} = \sup_{h\neq 0} \frac{||2fh||}{||h||} = 2||f||_\infty$, e pela alínea anterior $||P'_f||_{\mathcal{L}(.,.)} \leq a||-\sin(f)||_\infty \leq a$, portanto

$$||A'_f||_{\mathcal{L}(.,.)} \leq ||B'_f||_{\mathcal{L}(.,.)} + ||P'_f||_{\mathcal{L}(.,.)} \leq 2||f||_\infty + a.$$

Sabemos assim que para $f \in X$ temos $||(B-P)'_f||_{\mathcal{L}(.,.)} \leq 2\frac{1}{4} + a < 1$ ao impôr $a < \frac{1}{2}$.

X é uma bola fechada, logo um conjunto convexo, fechado e não-vazio.

Falta ver que se $f \in X$ então $Af \in X$.

$$||Af||_\infty = ||f^2 - \int_0^x \cos(f(t))\,dt||_\infty \leq (\frac{1}{4})^2 + a||\cos(f)||_\infty = \frac{1}{16} + a,$$

portanto basta que $\frac{1}{16} + a \leq \frac{1}{4}$, ou seja, basta que $a \leq \frac{3}{16} < \frac{1}{2}$ para que todas as condições estejam verificadas.

$f_0 = 0,\ f_1 = Af_0 = 0 - \int_0^x \cos(0)\,dt = -x,\ f_2 = Af_1 = (-x)^2 - \int_0^x \cos(-t)\,dt = x^2 - \sin(x)$.

Para o majorante com $a = 0.1$, temos $L = \frac{1}{2} + 0.1$, logo $||e_2||_\infty \leq L^2 \frac{1}{1-L}|| - x - 0||_\infty$, e fica $||e_2||_\infty \leq \frac{0.6^2}{0.4}\,0.1 = 0.09$.

8.2.7 1a. Avaliação (99)

1. a)

Basta reparar que a área=volume=V_z é dado pela área do rectângulo= $zf(z)$ subtraída da área do gráfico de f que é= $\int_0^z f(x)dx = F(z) - F(0)$.

Quanto à unicidade, basta reparar que sendo $h(x) = xf(x) - F(x) + F(0) - V$, temos $h'(x) = f(x) + xf'(x) - f(x) = xf'(x) > 0$ se $x > 0$.

Portanto, como consequência do Teorema de Rolle haverá no máximo uma raiz.

b) $F(x) = \frac{1}{6}(\frac{1}{5}x^5 + x^3) + x$. Assim obtemos a equação

$$\frac{1}{6}(z^5 + 3z^3) + z - \frac{1}{6}(\frac{1}{5}z^5 + z^3) - z = 1 \Leftrightarrow \frac{4}{5}z^5 + 2z^3 = 6 \Leftrightarrow 2z^5 + 5z^3 - 15 = 0$$

que podemos aproximar usando o método de Newton.

Notamos que sendo $h(x) = 2x^5 + 5x^3 - 15$, temos $h(1) = -8,\ h(2) = 89$, e o único zero existe em $[1, 2]$. Como $h''(x) = 40x^3 + 30x \geq 0$ em $[1, 2]$ temos as condições suficientes de convergência satisfeitas se começarmos com $x_0 = 2$, $x_1 = 1.59545$, $x_1 = 1.34315$, $x_2 = 1.24487$, $x_3 = 1.23164$.

Podemos usar a estimativa a posteriori

$$|e_3| \leq \frac{|h(x_3)|}{\min_{[1,2]}|h'(x)|} = \frac{0.00993061}{25} \leq 0.0004.$$

2. a) Basta ver que estamos nas condições do T. ponto fixo. Calculando a norma da jacobiana no convexo $D = [-1, 1] \times [-1, 1]$

$$J_T(x,y) = \frac{1}{4}\left\|\begin{bmatrix} \cos(\frac{x-y}{2}) - \frac{y}{2}\sin(xy/2) & \cos(\frac{x-y}{2}) - \frac{x}{2}\sin(xy/2) \\ -\sin(x+y) - y\cos(xy) & -\sin(x+y) - x\cos(xy) \end{bmatrix}\right\|_\infty \leq$$
$$\leq \frac{1}{4}\max\{\tfrac{3}{2} + \tfrac{3}{2}, 1 + 1\} = \tfrac{3}{4}$$

Por outro lado é claro pela figura que se verifica $T(D) \subset D$, mas devemos provar pois só algumas partículas foram representadas. Assim,

$$||T(x,y)||_\infty \leq \frac{1}{4}\max\{2,2\} = \frac{1}{2}, \text{ para } (x,y) \in D.$$

As condições do T. P. Fixo estão verificadas, e existe um e um só $(x,y) : T(x,y) = (x,y)$. Se repetirmos a transformação obtemos uma sucessão pelo método do ponto fixo que converge para o ponto fixo.

b) Sabendo que $||e_n||_\infty \leq (\frac{3}{4})^n||e_0||_\infty \leq (\frac{3}{4})^n$ temos $(\frac{3}{4})^n < 10^{-8}$ se $n > \frac{8}{-\log_{10} 0.75} = 64.0314$.

Logo, um algoritmo pode ser

x=0.0; y=0.0; Do[w=N[sin($\frac{x-y}{2}$)+cos($\frac{xy}{2}$)]/4;y=N[cos($x+y$)−sin(xy)]/4;x=w,{k,1,65}]; Print[″(″,N[x,16],″,″,N[y,16],″)″]

que dará $(0.2556140945030951, 0.2099772200205733)$, números que têm todos os 16 dígitos correctos!

3. a) Analisando por colunas, valores próprios estão localizados em $\bar{B}(-\sin(a^2), |1+\cos(a^2)|), \bar{B}(0, |1+\cos(a)| + |\cos(a)|), \bar{B}(4, |\cos(a^2)|)$.

Para vermos que a última bola não intersecta as restantes, temos $|\lambda - 4| \leq |\cos(a^2)| \Rightarrow |\lambda| \geq 4 - |\cos(a^2)| \geq 3$.

Por outro lado $|\lambda + \sin(a^2)| \leq |1+\cos(a^2)| \Rightarrow |\lambda| \leq 1 + |\cos(a^2)| + |\sin(a^2)| < 3$, e também $|\lambda| \leq |1+\cos(a)| + |\cos(a)| \leq 3$.

A única hipótese de intersecção é $|\lambda| = 3$, mais concretamente $\lambda = 3$, quando $\cos(a) = 1$.

Nesse caso basta ver que $\lambda = 3$ não é valor próprio. A matriz fica

$$A_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -2 & 4 \end{bmatrix}, \det(A_0 - 3I) = \begin{vmatrix} -3 & 1 & 0 \\ 0 & -3 & 1 \\ -2 & -2 & 1 \end{vmatrix} = -3(-3+2)-(0+2) = 3-2 = 1 \neq 0.$$

Calculando agora duas iterações

$$A_0[0\ 0\ 1]^T = [0\ 1\ 4]^T, \text{logo } u^{(1)} = [0\ \tfrac{1}{4}\ 1]^T$$
$$A_0[0\ \tfrac{1}{4}\ 1]^T = [\tfrac{1}{4}\ 1\ \tfrac{7}{2}]^T, \text{ logo } u^{(2)} = [\tfrac{1}{14}\ \tfrac{2}{7}\ 1]^T$$

e basta ver que $\lambda^{(2)} = [A_0 u^{(2)}]_3 = -\frac{1}{7} - \frac{4}{7} + 4 = \frac{23}{7}$.

b) Sendo $u_{n+3} = 4u_{n+2} - 2u_{n+1} - 2u_n$, temos para $u_0 = 0, u_1 = 0, u_2 = 1$ que $u_3 = 4, u_4 = 16 - 2 = 14, u_5 = 56 - 8 - 2 = 46$.

Obtemos $z \sim u_5/u_4 = 23/7$ o que é igual à aproximação anterior. Isto explica-se porque a matriz de a) é a matriz companheira do polinómio anterior, e vimos que a aplicação do método das potências à matriz companheira, corresponde a aplicar o método de Bernoulli.

c) Tendo

$$A_0 - I = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ -2 & -2 & 3 \end{bmatrix} \mapsto \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & -2 & 3 \end{bmatrix} \mapsto \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & \frac{-2-2}{-1} = 4 & 3 \end{bmatrix} \mapsto \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 4 & -1 \end{bmatrix}$$

ficamos com

$$\begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ -2 & -2 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 4 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}$$

e resolvendo $Ly = [010]$, dá $y = [01-4]$, e $Uw = [01-4]$ dá $w = [334]$, ou seja $[\frac{3}{4}\frac{3}{4}1]$.

4. a) O polinómio característico associado à equação às diferenças é $x^3 - 4x^2 + 2x + 2$, e verifica-se que as suas raízes são a, b, c.

Logo

$$u_n = Aa^n + Bb^n + Cc^n$$

e de $u_0 = 2, u_{-1} = 0, u_{-2} = 0$ retiramos o sistema

$$\begin{bmatrix} 1 & 1 & 1 \\ 1/a & 1/b & 1/c \\ 1/a^2 & 1/b^2 & 1/c^2 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 0.0354 & -0.159 & 0.147 \\ -0.516 & 1.387 & 0.787 \\ 1.48 & -1.23 & -0.934 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.0708 \\ -1.032 \\ 2.96 \end{bmatrix}$$

e concluímos que

$$u_n = 0.0708(-0.481194)^n - 1.032(1.31111)^n + 2.96(3.17009)^n$$

o termo que domina na sucessão quando $n \to \infty$ é a potência de 3.17... pelo que podemos prever que $u_n \sim 2.96(3.17009)^n > 10^6$ quando

$$n > \frac{\log(\frac{10^6}{2.96})}{\log 3.17009} = 11.033... \text{ ou seja quando } n \geq 12.$$

Na realidade, obtemos $u_{11} = 962112$ e $u_{12} = 3050016$.

b) Se começarmos com $u_0 = 12,\ u_{-1} = 10,\ u_{-2} = 6$, obtemos

$$\begin{bmatrix} 0.0354 & -0.159 & 0.147 \\ -0.516 & 1.387 & 0.787 \\ 1.48 & -1.23 & -0.934 \end{bmatrix} \begin{bmatrix} 12 \\ 10 \\ 6 \end{bmatrix} = \begin{bmatrix} -0.28 \\ 12.4 \\ -0.125 \end{bmatrix}$$

e portanto $u_n = -0.27(-0.481194)^n + 12.4(1.31111)^n - 0.125(3.17009)^n$ converge para $-\infty$.

Se começarmos com $u_0 = 12,\ u_{-1} = 10,\ u_{-2} = 5$, obtemos $u_n = -0.43(-0.481194)^n + 11.6(1.31111)^n + 0.81(3.17009)^n$ e a sucessão irá convergir para $+\infty$.

Portanto, com uma mudança mínima nas condições iniciais (não poderia ser menor se consideramos números naturais) há uma mudança radical no comportamento da sucessão. Essa mudança é aliás óbvia quando efectuamos $z = ax^n$, já que $\delta_z = \delta_a + n\delta_x$ e mesmo que admitissemos $\delta_x = 0$ iríamos ter $\delta_z = \delta_a$ o que neste caso é bastante grave, já que se $z \to \infty$ basta que $\delta_a \neq 0$ para que $e_z = z\delta_z \to \infty$.

5. a)

$$T(f+h)-Tf=\int_0^x f'(t)^2+2f'(t)h'(t)+h'(t)^2\,dt-\int_0^x f'(t)^2=\int_0^x 2f'(t)h'(t)+h'(t)^2\,dt$$
$$(T(f+h)-Tf)'=2f'(x)h'(x)+h'(x)^2$$

portanto $T'_f h=2\int_0^x f'h'$ já que

$$||T(f+h)-Tf-T'_f h||_\infty+||(T(f+h)-Tf-T'_f h)'||_\infty=||\int_0^x h'(t)^2dt||+||h'(x)^2||\leq(a+1)||h'||^2$$

Por outro lado,

$$U(f+h)-Uf=\int_0^x 2f(t)h(t)+h(t)^2\,dt$$
$$(U(f+h)-Uf)'=2f(x)h(x)+h(x)^2$$

e assim $U'_f h=\int_0^x 2fh$ pois

$$||U(f+h)-Uf-U'_f h||_\infty+||(U(f+h)-Uf-U'_f h)'||_\infty=||\int_0^x h(t)^2dt||+||h(x)^2||\leq(a+1)||h||^2$$

Portanto a F-derivada de $\frac{1}{\alpha}(x-Uf+Tf)$ é $\frac{1}{\alpha}(-U'_f+T'_f)$.

b) Temos

$$||T'_f||_{L(C^1)}=\sup\frac{2||\int_0^x f'h'||+2||f'h'||}{||h||+||h'||}\leq 2\frac{(a+1)||f'||\,||h'||}{||h||+||h'||}\leq 2(a+1)||f'||_\infty$$

e também

$$||U'_f||_{L(C^1)}=\sup\frac{2||\int_0^x fh||+2||fh||}{||h||+||h'||}\leq 2\frac{(a+1)||f||\,||h||}{||h||+||h'||}\leq 2(a+1)||f||_\infty$$

Assim

$$||\frac{1}{\alpha}(T'_f-U'_f)||_{L(C^1)}=\frac{2(a+1)}{\alpha}||f||_{C^1}$$

e convém que

$$\frac{2(a+1)}{\alpha}||f||_{C^1}<1$$

c) Considere $a=1,\alpha=8$. Mostre que se $||f||_{C^1}\leq 1$, existe um único ponto fixo do operador A e determine um majorante do erro de f_2.

Da alínea anterior temos

$$||A'_f||\leq\frac{1}{2}$$

Como

$$||Af||_{C^1}\leq\frac{a+1}{\alpha}(||f||^2_{C^1}+1)$$

porque

$$||Af||_{C^1}=\tfrac{1}{\alpha}||\int_0^x 1-f(t)^2+f'(t)^2\,dt||_\infty+\tfrac{1}{\alpha}||1-f(t)^2+f'(t)^2||_\infty\leq$$
$$\leq\tfrac{1}{\alpha}(a+a||f||^2_\infty+a||f'||^2_\infty+1+||f||^2_\infty+||f'||^2_\infty)\leq\tfrac{a}{\alpha}(||f||+||f'||)^2+\tfrac{1}{\alpha}(||f||+||f'||)^2$$

temos

$$||Af||_{C^1}\leq\frac{a+1}{\alpha}(||f||^2_{C^1}+1)\leq\frac{2}{8}2=\frac{1}{2}.$$

8.3 Trabalhos computacionais

8.3.1 Trabalhos computacionais (97)

Problema 1 : Determine todas as raízes da equação

$$x\cos(x^2) + 1 = 0$$

no intervalo $[-3, 3]$ com um erro relativo inferior a 10^{-10}.

Indique o método que utilizou em cada caso, mostrando a convergência e as estimativas de erro. Indique os valores de todas as iteradas e compare, pelo menos num caso, dois métodos distintos, relativamente à rapidez de convergência. Aplique um método do ponto fixo (com convergência linear) para determinar uma das raízes.

Problema 2 : Considere a seguinte equação algébrica:

$$p(x) = \frac{a}{2}x^5 - ax^4 + 4x - a^3 = 0$$

onde a é um parâmetro real.

a) Baseando-se nos resultados vistos nas aulas, relativamente à localização de raízes, trace os gráficos dos limites, inferior e superior do anel onde se prevê encontrar as raízes quando a varia em $[-5, 5]$.

b) Usando um método iterativo, coloque no plano complexo as 5 curvas que correspondem à localização das 5 raízes quando a varia entre $[-5, 5]$. (Não precisa de justificar a convergência, no caso de raízes complexas.)

c) Considere a sucessão $u_{n+5} = 2u_{n+4} - 2u_{n+1} + 32u_n$. Verifique experimentalmente que $u_{n+1}/u_n \to 2.593068...$, considerando $u_i = 0$ se $i = 0, 1, 2, 3$ e $u_4 = 1$. Justifique a convergência para aquele valor.

Problema 3 : Considere a equação:

$$e^{ax+y} + \frac{y^3 + bx}{c} = d$$

em que $a \geq -1, b, c > 0, d \in \mathbb{R}$, são valores fixos. Esta equação define implicitamente y como função de x, ou seja, dado um certo x, obtemos o valor $y(x)$ resolvendo a equação em y. Para determinar aproximadamente os valores de $y(x)$ utilize o método de Newton.

a) Mostre que qualquer que seja $x \in \mathbb{R}$ as condições suficientes para a convergência do método de Newton estão asseguradas.

b) Considere três diferentes valores para a, b, c, d. Trace o gráfico de y como função de x, no intervalo $[-5, 5]$ (calcule no mínimo 20 pontos). Use como critério de paragem $|y_{n+1} - y_n| \leq 0.001$.

c) Para os valores a, b, c, d considerados em b) mostre que existe um único ponto fixo de $y(x)$ em $\mathbb{R}$ e determine-o com um erro absoluto inferior a 10^{-8}.

Problema 4 : Considere as curvas em $\mathbb{R}^2$ parametrizadas por

$$\begin{aligned} C : [0, 2\pi] &\to \mathbb{R}^2 \\ t &\to (r(t), s(t)) \end{aligned}$$

onde $r, s \in C^1[0, 2\pi]$. Considere as seguintes curvas fechadas:

$$\begin{aligned} \mathbf{C}_1 &: \quad r(t) = 1 + t(t - 2\pi), \;\; s(t) = 1 + 2\sin(t)e^{2-t/2} \\ \mathbf{C}_2 &: \quad r(t) = \sin(t), \;\; s(t) = \cos(t - 4) + 4 \end{aligned}$$

a) Usando um método iterativo, determine os pontos do plano onde as curvas C_1 e C_2 se intersectam.

b) Considere agora curvas da forma

$$\begin{aligned} C(r) : [0, 2\pi] &\rightarrow \mathbb{R}^2 \\ t &\rightarrow r(t)(\cos(t), \sin(t)) \end{aligned}$$

onde $r(t) = \cos(t/2)(2\pi - t)t/a + b$.

Definindo $r^{n+1}(t) = r(r^n(t)), r^0(t) = t$, considere as curvas $C(r^n)$.

Mostre que, no limite, quando $n \rightarrow \infty$, as curvas $C(r^n)$ convergem para um circulo se $a = 8, b = 2$. Determine o valor da área desse circulo.

c) Estabeleça condições gerais sobre uma função $r(t)$ de forma a que, no limite, as curvas $C(r^n)$, consideradas em b), sejam circulos.

d) Se considerar $a = 4$ e $b = 2$, na alínea b), para n suficientemente grande (p.ex: $n > 20$), trace as curvas $C(r^n)$ e $C(r^{n+1})$.

Determine aproximadamente os valores de t para os quais as curvas se intersectam, isto é: $C^n(t) = C^{n+1}(t)$, quando n tende para infinito.

Determine aproximadamente os pontos de interseção da curva $C(r^{32})$ com a recta $y = 6x - 9$ se $a = 4, \; b = 0.5$.

Problema 5 : Considere duas funções $f, g \in C^1(\mathbb{R})$. Pretendem-se determinar pontos $(x, f(x))$ que estão à menor distância (euclidiana) de pontos $(y, g(y))$.

a) Considere primeiro o caso trivial, em que f intersecta g, e dê um exemplo de duas funções diferenciáveis em $\mathbb{R}$ que se intersectam em todos os pontos do intervalo $[-2, 2]$.

Indique também um exemplo em que a distância entre f e g é um ínfimo em $\mathbb{R}$ e consequentemente não é atingida em nenhum ponto.

b) No caso em que f não intersecta g demonstre que em pontos em que a distância é mínima, ou seja:

$$||(x, f(x)) - (y, g(y))||_2 = \min_{a,b \in \mathbb{R}} ||(a, f(a)) - (b, g(b))||_2$$

temos $f'(x) = g'(y)$ e as normais coincidem.

Sugestão: Verifique primeiro que se um ponto da função $(z, f(z))$ está uma distância mínima de um outro ponto qualquer (a, b), então a normal ao ponto $(z, f(z))$ intersecta (a, b).

c) Aplique o resultado anterior para determinar esses pontos (através de um sistema) no caso de ter funções f, g não triviais... um exemplo: $f(x) = e^x, \; g(x) = x + cos(x) - 4$.

d) Usando as funções $f(x) = e^x, \; g(x) = x + cos(\mu x) - 4$, trace um gráfico da distância em função de μ. Determine, justificando, para que o valor tende a distância quando $\mu \rightarrow \infty$.

Observação: Este problema corresponde, por exemplo, a encontrar o local onde efectuar uma ponte com o mínimo comprimento...

Problema 6 : Considere um túnel cujas paredes são definidas por duas funções f_1 e f_2, positivas, num intervalo $[0, R]$ e com uma distância mínima $h > 0$ entre elas (ver figura). No ponto A supõe-se existir uma fonte emissora de um projectil segundo um ângulo $\alpha \in [\frac{-\pi}{4}, \frac{\pi}{4}]$, e supõe-se que o impacto desse projectil nas paredes do túnel causa uma reflexão perfeita (ou seja o ângulo de reflexão é simétrico ao ângulo de incidência, relativamente à normal).

a) Construa, a partir de 3 pares distintos de funções que definam essas paredes, gráficos que relacionem o ângulo de emissão do projectil com o ponto de impacto quando $x = R$ (ver figura 2).

(Um desses pares deverá ser: $f_1 = \sin(x) + 1$, $f_2 = \cos(x) + 3.5$ no intervalo $[0, 10]$)

Para efectuar este trabalho terá que determinar cada ponto de impacto, resolvendo uma equação não linear do tipo $f_i(x) = ax + b$, usando um método iterativo.

b) Efectue também gráficos que indiquem o comprimento da trajectória e o ângulo de impacto, em função do ângulo de emissão, para os exemplos estudados em a).

c) Considerando $f_1 = \sin(ax + b) + 1$ e $f_2(x) = \cos(cx + d) + 3.5$, construa um procedimento que lhe permita tentar identificar os parâmetros a, b, c, d, a partir do conhecimento dos gráficos obtidos em a).

Ou seja, a partir dos resultados das experiências com a emissão de projécteis, pretende-se tentar determinar a forma do túnel. Na prática, deve considerar valores de a, b, c, d aleatórios (que corresponderão ao túnel a determinar) e construir um processo que lhe permita reencontrar esses valores. Isso deve ser feito, considerando certos valores iniciais a_0, b_0, c_0, d_0 e a partir da comparação dos dois resultados obter novos valores a_1, b_1, c_1, d_1 que melhor se ajustem... até que este processo iterativo permita obter os verdadeiros valores.

8.3.2 Trabalhos computacionais (98)

Problema 1 : Determine todas as raízes reais da equação

$$x^5 = 1 - x\cos(x^p)$$

assegurando um erro relativo inferior a 10^{-10}, para $p = 0, 2, 4, 6, 8$.

a) Indique o método que utilizou em cada caso, mostrando a convergência e as estimativas de erro. Indique os valores de todas as iteradas e compare, pelo menos num caso, dois métodos distintos, relativamente à rapidez de convergência.

b) Determine uma aproximação para as raízes complexas no caso $p = 0$, utilizando 3 iterações do método de Newton, começando com $x_0 = -0.8 + 0.7i$ e com $x_0 = 0.4 + 0.9i$.

c) Determine a raiz real z de $f(x) = x^{51} + 50xe^x = 50$, usando o método de Newton, com um erro relativo inferior a 10^{-10}. Se calcularmos $f(\tilde{z})$, com $\tilde{z} = z + 0.1$, o valor será próximo de 0? Justifique!

Problema 2 : Considere as curvas em $\mathbb{R}^2$ parametrizadas por

$$\begin{array}{rcl} C : [-\pi^{1/3}, \pi^{1/3}] & \rightarrow & \mathbb{R}^2 \\ t & \rightarrow & (r(t), s(t)) \end{array}$$

onde $r, s \in C^1[-\pi^{1/3}, \pi^{1/3}]$.

a) Considere as seguintes curvas fechadas:

$$\mathbf{C}_1: \quad r(t) = (\cos(t))^2 + t^2 - 1, \quad s(t) = \sin(t^3 - 1) + (\sin(t^3))^2$$
$$\mathbf{C}_2: \quad r(t) = (\sin(t))^2 - t^2 + 1, \quad s(t) = 1 - \sin(t^3 - 1) - (\sin(t^3))^2$$

Usando um método iterativo, determine os pontos do plano onde as curvas C_1 e C_2 se intersectam.

b) Sendo $r(t) = \frac{1}{a}\sin(t/2)(2\pi - t)t + b$, definimos agora curvas da forma

$$\begin{aligned} C(r) : [0, 2\pi] &\rightarrow \mathbb{R}^2 \\ t &\rightarrow r(t)(\cos(t), \sin(t)) \end{aligned}$$

Definindo $r^{n+1}(t) = r(r^n(t)), r^0(t) = t$, considere as curvas $C(r^n)$.

Mostre que, no limite, quando $n \rightarrow \infty$, as curvas $C(r^n)$ convergem para um círculo se $a = 10, b = 2$. Determine o valor da área desse círculo.

c) Estabeleça condições gerais sobre uma função $r(t)$ de forma a que, no limite, as curvas $C(r^n)$, consideradas em b), sejam círculos.

d) Se considerar $a = 3$ e $b = 1$, na alínea b), para $n = 1000$, trace as curvas $C(r^n)$ e $C(r^{n+1})$. Determine aproximadamente os valores de t para os quais as curvas se intersectam, isto é: $C(r^n(t)) = C(r^{n+1}(t))$.

Trace $C(r^n), C(r^{n+1}), C(r^{n+2}), C(r^{n+3})$, para $n > 100$, e justifique os gráficos obtidos.

Problema 3 :

a) Considere duas funções $f, g \in C^1(\mathbb{R})$. Pretendem-se determinar pontos $(x, f(x))$ que estão à menor distância (euclidiana) de pontos $(y, g(y))$. No caso em que f não intersecta g mostre que, em pontos em que a distância é mínima, ou seja:

$$||(x, f(x)) - (y, g(y))||_2 = \min_{a,b \in \mathbb{R}} ||(a, f(a)) - (b, g(b))||_2$$

temos $f'(x) = g'(y)$ e as normais coincidem.

Sugestão: Verifique primeiro que se um ponto da função $(z, f(z))$ está uma distância mínima de um outro ponto qualquer (a, b), então a normal ao ponto $(z, f(z))$ intersecta (a, b).

Aplique este resultado para determinar esses pontos (através de um sistema) no caso de ter funções f, g não triviais, por exemplo: $f(x) = e^x$, $g(x) = x^3 + cos(x) - 16$.

b) Considere uma função $f \in C^3(\mathbb{R})$. A cada ponto do seu gráfico $(z, f(z))$ podemos associar uma recta tangente e uma recta normal. Se $f'(z) \neq 0$, a equação da recta normal nesse ponto é dada por $y = \frac{-1}{f'(z)}(x - z) + f(z)$. Dado um outro ponto w, suficientemente próximo de z, determine o ponto de intersecção das rectas normais.

i) Considerando o limite quando $w \rightarrow z$, verifique que a fórmula para esse ponto de intersecção limite (x_z, y_z) é dada por

$$(x_z, y_z) = (z, f(z)) + \frac{L_f(z)^2}{f''(z)}(-f'(z), 1)$$

onde $L_f(z) = \sqrt{1+f'(z)^2}$ representa a unidade elementar de comprimento de arco. Ao valor $\rho(z) = \frac{L_f(z)^3}{f''(z)}$ chamamos raio de curvatura, e uma circunferência com centro em (x_z, y_z) e raio $|\rho(z)|$ será a circunferência tangente que melhor aproxima o gráfico da função f numa vizinhança de z.

ii) Para as funções $f(x) = x^2, f(x) = x^3 - 3x + 1, f(x) = \sin(x), f(x) = \sqrt{a^2 - x^2/b^2}$, etc... trace os gráficos dos centros de curvatura (x_z, y_z).

iii) Determine os pontos $z : x'_z = 0, y'_z = 0$, e interprete geometricamente.

Problema 4 : Considere a equação

$$\sin(x) = 2t^2x + 2x - \frac{1}{3}(\cos(t+x))^3$$

a) Mostre que ela define implicitamente x como função de t, para qualquer $t \in \mathbb{R}$, e trace o gráfico da função em $[-5, 5]$.

b) Seja M uma matriz de dimensão $d = 1000$, que varia com o valor de t :

$$M = \begin{bmatrix} x(t)^2 & x(t)+1 & 0 & 0 & ... & 0 \\ x(t)-1 & x(t)^2 & x(t)+1 & 0 & ... & 0 \\ 0 & x(t)-1 & x(t)^2 & x(t)+1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & x(t)-1 & x(t)^2 \end{bmatrix}$$

Determine o valor da componente x_{500} da solução do sistema $Mx = (1, 0, ..., 0, 1)$, e trace o gráfico dessa componente em função de $t \in [-5, 5]$. Indique um majorante para o erro cometido.

Problema 5 : Considere a equação integral

$$f(x) - \lambda \int_a^x K(x,y) f(y) dy = \phi(x), \quad \text{para } x \in [a, b],$$

em que $\lambda \in \mathbb{R}$, ϕ é uma função contínua em $[a, b]$ e K é uma função contínua em $[a, b] \times [a, b]$.

a) Mostre que a equação tem uma e uma só solução contínua no intervalo $[a, b]$, qualquer que seja λ.

b) Construa numericamente uma aproximação, usando um método de ponto fixo, para a solução da equação integral em $[0, 1]$:

$$f(x) + \frac{1}{2} \int_0^x \frac{f(y)}{x^2 + y^2 + 1} \, dy = \cos(x)$$

Para esse efeito considere a seguinte fórmula dos trapézios para aproximar o valor do integral (com n suficientemente grande):

$$\int_a^b f(x)\, dx \sim \frac{b-a}{n} \left(\frac{f(a) + f(b)}{2} + \sum_{k=1}^{n-1} f(a + k\frac{b-a}{n}) \right)$$

Apresente numa figura os gráficos das últimas aproximações de $f(x)$ obtidas, apresentando as majorações de erro. Discuta essas majorações de erro considerando que o erro da fórmula de integração é $|E_f| \leq \frac{(b-a)^3}{12n^2}||f''||_\infty$.

c) Considere agora uma equação em que a incógnita é a sucessão (x_n), verificando-se

$$x_n = \lambda \sum_{k=1}^{30} \frac{\sin(x_k)}{n+k} + \cos(n)$$

Mostre que existe uma única sucessão em l^∞ que verifica esta equação, e aproxime os seus primeiros 10 termos, indicando uma majoração para o erro cometido.

Problema 6 : Considere um circuito fechado, cujos limites estão definidos pelas curvas interior C_1, e exterior C_2, dadas com a parametrização:

$$\begin{array}{rccc} C_i: & [0,2\pi] & \longrightarrow & \mathbb{R}^2 \\ & t & \longmapsto & r_i(t)(\cos(t), \sin(t)) \end{array}$$

em que $0 < r_1 < r_2$ são funções em $C^1([0,2\pi])$ com uma distância mínima $h > 0$ entre elas.

Sendo $A_c = ((1-c)r_1(0) + cr_2(0), 0)$, em que $0 < c < 1$, o ponto de emissão de um projéctil segundo a direcção $(0,1)$ e supondo que o impacto desse projéctil nas paredes do túnel causa uma reflexão perfeita (ou seja o ângulo de reflexão é simétrico ao ângulo de incidência, relativamente à normal):

a) Construa, a partir de 3 pares distintos de funções que definam essas paredes, gráficos que relacionem o ponto de emissão do projéctil com o primeiro ponto de retorno ao segmento $[r_1(0), r_2(0)]$. Trata-se portanto de um gráfico de uma função $T : [r_1(0), r_2(0)] \to [r_1(0), r_2(0)]$.

(Estude o caso em que $r_i(t)$ são constantes. Um outro par poderá ser: $r_1(t) = \frac{1}{2}\cos(4t) + \frac{3}{2}$, $r_2(t) = \cos(4t) + \frac{7}{2}$)

Para efectuar este trabalho deverá determinar cada ponto de impacto, resolvendo uma equação não linear do tipo $f_i(x) = ax+b$, usando um método iterativo. Caso não o projéctil fique preso em reflexões sucessivas, não regressando ao segmento de partida, pode terminar o ciclo ao fim de um número suficientemente grande de iterações (p.ex: 200).

b) Para alguns valores de $x_0 \in [r_1(0), r_2(0)]$, construa a sucessão $x_{n+1} = Tx_n$, determinando um número razoável de iterações (p.ex: 10). Apresente conclusões acerca do comportamento destas sucessões.

c) Considerando

$$\begin{aligned} r_1(t) &= \sin(2t) + \sin(a)\sin(t) + \sin(b)\sin(2t) + \tfrac{5}{2} \\ r_2(t) &= \cos(2t) + \cos(c)\sin(2t) + \cos(d)\sin(3t) + \tfrac{9}{2} \end{aligned}$$

construa um programa que lhe permita tentar identificar os parâmetros a, b, c, d, a partir do conhecimento de gráficos semelhantes aos obtidos em a).

Ou seja, a partir dos resultados das experiências com a emissão de projécteis, pretende-se tentar determinar a forma do circuito!!

Na prática, deve considerar valores de a, b, c, d aleatórios (que corresponderão ao túnel a determinar) e construir um processo que lhe permita reencontrar esses valores. Isso deve ser feito, considerando certos valores iniciais a_0, b_0, c_0, d_0 e a partir da comparação dos dois resultados obter novos valores a_1, b_1, c_1, d_1 que melhor se ajustem... de forma a que este processo iterativo permita aproximar os verdadeiros valores.

8.3.3 Trabalhos computacionais (99)

Problema 1 : Considere a equação

$$x^3 = 3x + 3 + a\cos(x).$$

a) Discuta a existência e unicidade de solução real em função do parâmetro $a \in \mathbb{R}$. Trace o gráfico da raiz em função de a, nos casos em que seja raiz única.

b) Suponha que o valor de a é substituído sucessivamente pelo valor da raiz, de forma a que se obtenha uma sucessão de raízes z_n.

Considerando $z_0 = a = 1$, determine o valor da raiz z_1 usando um método do ponto fixo com convergência linear, assegurando um erro absoluto inferior a 10^{-6}.

Use o valor z_1 como novo a, ou seja $a = z_1$ para determinar z_2 da mesma forma. Atendendo a que já existia um erro no parâmetro a determine um majorante para o erro absoluto de z_2.

c) A sucessão z_n converge para que valor? Determine-o pelo método da secante assegurando um erro inferior a 10^{-8}.

d) Verifique experimentalmente que a sucessão $x_{n+1} = h(x_n)$ em que $h(x) = -2x - \frac{1}{2} + x^3$ não converge (excepto para três iteradas iniciais exactas, quais?) traçando os gráficos das funções $h(x), h(h(x))$, etc... Mostre que, no entanto, $h(I) \subseteq I$ para um certo intervalo, determinando-o exactamente.

Problema 2 : Considere os seguintes métodos numéricos

$$(R) \begin{cases} x_0 \\ x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)f'(x_n) - f(x_n)f''(x_n)} \end{cases}$$

e

$$(S) \begin{cases} x_0 \\ x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{f(x_n)f(x_n)f''(x_n)}{2f'(x_n)f'(x_n)f'(x_n)} \end{cases}$$

para a resolução de equações $f(x) = 0$.

a) Obtenha a expressão do método (R) calculando a nova iterada através do zero da função $r(x) = \frac{ax+b}{cx+1}$ em que os coeficientes a, b, c são obtidos resolvendo o sistema $r(x_n) = f(x_n),\ r'(x_n) = f'(x_n),\ r''(x_n) = f''(x_n)$. Da mesma forma obtenha a expressão do método (S), agora considerando $s(y) = ay^2 + by + c$, e resolvendo o sistema $s(f_n) = x_n,\ s'(f_n) = (f^{-1})'(x_n),\ s''(f_n) = (f^{-1})''(x_n)$, em que $f_n = f(x_n)$, notando que $y = 0$ corresponde $x_{n+1} = s(0) = c$.

Interprete geometricamente os métodos e trace os gráficos das funções iteradoras para alguns exemplos de f.

b) Mostre a convergência local de ambos os métodos. Mostre que ambos os métodos apresentam pelo menos convergência local cúbica sob certas condições (que deve explicitar) e calcule os coeficientes assimptóticos de convergência. Comente os resultados obtidos.

c) Utilize os métodos anteriores para obter aproximações de $\sqrt[p]{a}$, para vários valores reais de $p, a > 1$. Compare os valores obtidos pelos métodos (R), (S) e também pelo método de Newton, relativamente à rapidez de convergência face à iterada inicial.

(Para esse efeito, para além do estudo da ordem de convergência, esboce os gráficos das iteradas $x_1, x_2, x_3, ...$ em função da iterada inicial x_0.

Considere $a = 209, p = 3$, e outros dois pares de valores, tomando valores $x_0 \in [-15, 15]$)

d) Efectue o mesmo estudo, considerando agora outras funções, como seja $f(x) = x^3 - x\cos(x) - 1$. Interprete os gráficos obtidos.

Problema 3 : Uma corda presa pelas extremidades $a = -1, b = 1$ vibra verificando a equação (harmónica no tempo):

$$u''(x) + k^2 u(x) = 0.$$

Através de uma divisão do intervalo $[-1, 1]$ em $m + 1$ subintervalos de comprimento $h = 2/(m + 1)$ é possível aproximar o valor de u nos pontos $u_n = u(-1 + nh)$ através da aproximação de

$$u''(x_n) \sim \frac{u_{n+1} - 2u_n + u_{n-1}}{h^2}.$$

a) Obtenha uma fórmula explícita de u_n em função de $u_a = u(-1)$ e $u_b = u(1)$ usando uma equação às diferenças.

b) Estude o problema de estabilidade face ao erro dos valores das raízes da equação característica.

c) Considerando valores concretos de u_a, u_b e k trace o gráfico que une os pontos u_n para vários valores de m. Comente os resultados.

d) Altere os valores de u_a e u_b usando como novos valores os valores obtidos para u_1 e u_m, respectivamente. Repita este processo pelo menos m vezes e trace um gráfico tridimensional que mostre a variação da solução ao longo deste processo.

Problema 4 : Considere o seguinte método para determinar as raízes de polinómios de grau m.

Usando a decomposição

$$p(x) = a_0 + a_1 x + ... + a_m x^m = a_m(x - z_1)...(x - z_m)$$

é possível escrever o sistema não linear

$$(S) \begin{cases} (-1)^m z_1 ... z_m = a_0/a_m \\ \vdots \\ -z_1 - ... - z_m = a_{m-1}/a_m. \end{cases}$$

Aplicando à resolução do sistema (S) um método iterativo obtêm-se sucessivas aproximações às raízes $z_1, ..., z_m$.

a) Mostre que o sistema (S) tem uma e uma só solução em $\mathbb{C}^m$.

b) Considere $m = 3, m = 5$ e escreva explicitamente o sistema não linear (S). Explicite também os cálculos necessários à implementação do método de Newton para a resolução numérica de (S). Discuta a possibilidade de aproximação das raízes relativamente às iteradas iniciais consideradas (por exemplo, sendo reais ou complexas).

c) Aplique 10 iterações do método de Newton a (S) para aproximar as raízes de algumas equações algébricas, entre as quais

$$p_1(x) = x^5 + x^4 - 2x^3 + x^2 + 2x + 1 = 0, \quad p_2(x) = x^5 + x^4 - 6x^3 - 4x^2 + 2x + 1 = 0,$$

escolhendo convenientemente as iteradas iniciais (e justificando essa escolha, baseando-se em resultados teóricos).

Apresente majorações do erro absoluto para as aproximações das raízes de $p_2(x)$.

Problema 5 : Uma matriz M de dimensão $n \times n$ diz-se *matriz de Toeplitz* se tiver todas as subdiagonais constantes (em particular, uma matriz tridiagonal com as três diagonais constantes é uma matriz de Toeplitz).

a) Estabelecer um processo explícito de calcular a inversa de matrizes de Toeplitz usando equações às diferenças.

b) Indicar a fórmula geral para a matriz inversa de uma matriz de Toeplitz tridiagonal cujas diagonais têm valores a, b, c reais (a é o valor constante da subdiagonal inferior e c o da superior).

c) Calcular os valores próprios de qualquer matriz de Toeplitz tridiagonal (como em (b)).

Ou seja, usando ainda equações às diferenças, mostre que se $a\,c > 0$, então

$$\lambda_k = b + 2\sqrt{ac}\,\cos(\theta_k)$$

em que $\theta_k = \frac{k\pi}{n+1}$, para $k = 1, ..., n$.

Deduza uma fórmula semelhante para o caso $a\,c < 0$.

Problema 6

a) Implementação computacional de um algoritmo que permita obter a factorização QR de uma qualquer matriz.

b) Implementação do método QR para a determinação de valores próprios da matriz companheira de um polinómio.

c) Aplicando o método de (b), apresente um gráfico com as curvas (no plano complexo) que correspondem à localização das raízes dos polinómios

$$p_a(x) = x^7 + ax^6 + a^2x^5 - ax^4 + x^3 + 1,$$

ao variar $a \in [-2, 2]$. Apresente e discuta as aproximações que considerar e os resultados obtidos.

Efectue o mesmo para $p_a(x) = x^7 + 2a^2x^6 + 3a^3x^5 - 2ax^4 + 3ax^3 + 1$, com $a \in [-2, 2]$.

8.4 Glossário

.

A
algarismos significativos, 6(rodapé)
algébrico (número), 75
algoritmo, 12
arredondamento simétrico/ por corte, 7
Aitken (aceleração de), 56

B
Banach
... espaço de, 101
... teorema do ponto fixo de, 106
Bauer-Fike (teorema de), 187
Bernoulli (método de), 79
bissecção (método da), 37
Bolzano (teorema do valor intermédio de), 35
Brouwer (teorema do ponto fixo de), 215
Budan-Fourier (regra de), 77

C
cancelamento subtractivo, 12
Cauchy, sucessão de, 4, 101
Cauchy-Riemann, condições de, 217
Cholesky (método de), 155
completo (espaço), 101
condicionamento, 15, 146
conjunto
... convexo, 111
... estrelado, 216
contractiva (função), 48
contractivo (operador), 105
convergência monótona/alternada, 47
coordenadas (método das), 210
correcção residual (método da), 158
Crout (método de), 155

D
declive máximo (método do), 202
derivada de Fréchet, 109
diagonal estritamente dominante, 137
dígitos correctos, 6
dígitos de guarda, 9
Doolittle (método de), 154

E
equações às diferenças, 219
erro absoluto/ relativo, 7
estável (numericamente), 16
estimativa a priori/ posteriori, 26

F
falsa posição (método da), 40
falsa posição modificado (método da), 46
fracção contínua, 51
Fréchet (derivada de), 109
função
... convexa, 200
... de descida, 200
... holomorfa, 217
... iteradora, 47

G
Gauss (método de eliminação de), 149
Gauss-Seidel (método de), 133
Gerschgorin (teorema de), 170
gradiente (método do), 201
gradiente conjugado (método do), 207

H
Hamilton-Cayley (teorema de), 165
Hörner (forma de), 78

I
iteração de Picard, 47, 105, 119
iterações inversas de Wielandt (método das), 181

J
Jacobi (método de), 132
Jordan (forma de), 166

K
Kahan (teorema de), 142
Kantorovich (teorema de), 130
Krylov (método de), 190

L
Lagrange (teorema do valor médio de), 35
Levenberg-Marquardt (método de), 209
Liouville (teorema de), 209
LR de Rutishauser (método), 183

M
matriz
... companheira, 188
... hermitiana, 166
... hessiana, 215
... de Householder, 184
... irredutível, 138
... jacobiana, 215
... semelhante, 165
métodos de descida, 200
mínimos quadrados (método dos), 193
Müller (método de), 73

N
Neumann (série de), 117
Newton (método de)
... em $\mathbb{R}$, 58
... em $\mathbb{C}$, 85
... em $\mathbb{R}^N$, 129
... em *Banach's,* 116
... em minimização, 209
Newton-Kantorovich (método de), 116
norma, espaço normado, 95
normas equivalentes, 98
norma matricial induzida, 123
notação científica, 4
número de *condição*
... para funções, 11
... para matrizes, 146
número de operações elementares:
... no método de eliminação de Gauss, 151
... para a inversão de matrizes, 156

O
operador
... compacto, 230
... contínuo, 103
... integral, 230
... linear, 104
ordem de convergência
... linear, 29, 108
... supralinear, 29, 31, 115
... quadrática/ cúbica, 31, 115
Ostrowski-Reich (teorema de), 143
overflow, 7

P
Picard (método de), 105, 119
polinómio característico, 168
ponto fixo, 47
ponto fixo (método do)
... em $\mathbb{R}$, 47
... em $\mathbb{C}$, 83
... em $\mathbb{R}^N$, 127
... em *Banach's*, 105
ponto flutuante (sistema de), 5
ponto de mínimo, 198
potências de Von Mises (método das), 174
pré-condicionamento, 149
precisão simples/ dupla, 5

Q
QR de Francis (método), 184
QR com shift (método), 186

R
raio espectral, 125
raiz (de uma equação), 23
Rayleigh (quóciente de), 170
Rolle (teorema de), 35
Rouché (teorema de), 218

S
Schur (forma de), 166
secante (método da), 69
secção dourada (método da), 201
sistema normal, 195
SOR (método), 141
Steffenson (método de), 57
supressão de zeros, 63

T
Taylor (fórmula de), 215
tempo de cálculo, 33
teorema da convergência global (para métodos de descida), 200
teorema fundamental da álgebra, 75
transcendente (número), 75

U
underflow, 7
unidade de arredondamento, 8

V
valor, vector e subespaço próprio, 164
valores singulares de uma matriz, 167
variações de sinal (número mínimo/máximo), 76

W
Weierstrass (teorema de), 216
Wronskiano, 223

Z
zero de uma função, 23

Bibliografia

[1] Atkinson K. E.; An Introduction to Numerical Analysis. *Wiley & Sons*, New York, 1989.

[2] Bahder T. B.; *Mathematica* for Scientists and Engineers. *Addison-Wesley*, Reading, 1995.

[3] Burden R., Faires J.; Numerical Analysis, 5th. ed. *PWS Publishers*, Boston, 1993.

[4] Carmo J., Sernadas A, Sernadas C., Dionisio F. M., Caleiro C.; Introdução à programação em *Mathematica*, IST Press, 1999.

[5] Carpentier M.; Análise Numérica. *Secção de Folhas*, A.E.I.S.T., 1996.

[6] Cartan H.; Cours de Calcul Différentiel. *Hermann*, Paris, 1990.

[7] Ciarlet P. G.; Introduction à l'analyse numérique matricielle et à l'optimisation. *Masson*, Paris, 1988.

[8] Clarke F. H.; Optimization and Nonsmooth Analysis. *SIAM*, Philadelphia, 1990.

[9] Conte, S. D., de Boor C.; Elementary numerical analysis. *McGraw-Hill*, Singapore, 1981.

[10] Démidovitch B., Maron I.; Eléments de calcul numérique. *MIR*, Moscovo, 1973.

[11] Ferreira, J. Campos.; Introdução à Análise Matemática. *Fundação Calouste Gulbenkian*, Lisboa, 1987.

[12] Flanigan F. J.;. Complex Variables. Harmonic and Analytic Functions. *Dover*, New York, 1983.

[13] Golub G. H., Van Loan C.F.; Matrix Computations. *John Hopkins*, Baltimore, 1985.

[14] Guerreiro, J. S. C.; Curso de Matemáticas Gerais, vol. I II e III. *Escolar Editora*, Lisboa, 1981.

[15] Henrici P.; Elements of numerical analysis. *Wiley & Sons*, New York, 1964.

[16] Householder A. S.; The theory of matrices in numerical analysis. *Dover*, New York, 1975.

[17] Isaacson E., Keller H. B.; Analysis of numerical methods. *Wiley & Sons*, New York, 1966.

[18] Kress R.; Numerical Analysis. *Springer-Verlag*, New York, 1998.

[19] Kurosh A.; Cours d'Algèbre Supérieure. *MIR*, Moscovo, 1973.

[20] Lima P.; Métodos Numéricos da Álgebra. *Secção de Folhas*, A.E.I.S.T., 1997.

[21] Loura L. C.; Tópicos de Análise Numérica. *Monografia*, I.S.T., 1990.

[22] Luenberger, D. G.; Linear and Nonlinear Programming. *Addison-Wesley*, Reading-Massachussets, 1984.

[23] Magalhães, L.;. Álgebra Linear. *Monografia*, A.E.I.S.T., 1988.

[24] Nougier J. P.; Méthodes de calcul numérique, 2nd. ed. *Masson*, Paris, 1985.

[25] Ortega J. M.; Numerical Analysis, a second course. *SIAM*, Philadelphia, 1990.

[26] Ortega J.M., Rheinboldt, W.; Iterative solution of nonlinear equations in several variables. *Academic Press*, New York, 1970.

[27] Pina H.; Métodos numéricos. *McGraw-Hill*, Lisboa, 1995.

[28] Scheid F.; Análise numérica. *McGraw-Hill*, Lisboa, 1991.

[29] Stoer J., Bulirsch R.; Introduction to Numerical Analysis, 2nd ed. *Springer Texts in Appl. Math.*, 1993.

[30] Tavares L.V., Correia F.N.; Optimização Linear e não Linear. *Fundação Calouste Gulbenkian*, Lisboa, 1987.

[31] Valença M. R.; Métodos numéricos. *Livraria Minho*, Braga, 1993.

[32] Wilkinson J. H.; The Algebraic Eigenvalue Problem. *Clarendon Press*, Oxford, 1988.

[33] Wilkinson J. H.; Rounding errors in algebraic processes. *Dover*, New York, 1994.

[34] Zeidler E.; Nonlinear Functional Analysis and its Applications. *Springer-Verlag*, New York, 1993.