Análise Numérica Funcional e Optimização. Teoria

Carlos J. S. Alves

Instituto Superior Técnico

2012

Sumário

Prefácio

Estas folhas seguem essencialmente os cursos de Análise Numérica Funcional e Optimização leccionados entre 2010 e 2012.

Capítulo 1

Espaços funcionais e Método do Ponto Fixo

O *Método do Ponto Fixo* é talvez o método numérico mais simples e eficaz para a resolução de quaisquer equações, dado o seu âmbito generalizável.

Ao escrever uma equação na forma

$$x = g(x),$$

a ideia do método do ponto fixo consiste em considerar apenas a iteração

$$x_{n+1} = g(x_n)$$

partindo de um valor inicial x_0.

A validade do método resultará da continuidade da função g, e da unicidade do limite, porque existindo limite da sucessão, $x_n \to z$ então

$$x_{n+1} \to z, \quad g(x_n) \to g(z) \quad \Longrightarrow z = g(z).$$

Uma solução z é assim designada *ponto fixo* de g porque se mantém invariante perante a aplicação da função iteradora g. A convergência deste método depende muito da escolha da função g. Com efeito é importante notar que podemos escrever $x = g(x)$ com diferentes funções, bastando ver que

$$x = g(x) \Leftrightarrow x = \frac{x}{2} + \frac{1}{2}g(x) = G(x),$$

e a aplicação do método do ponto fixo à nova função iteradora $G = \frac{x}{2} + \frac{1}{2}g(x)$ não produzirá os mesmos resultados que g.

A teoria para resolução de equações algébricas em $\mathbb{R}$ faz parte de cursos introdutórios a métodos numéricos.

Exemplo 1. Podemos lembrar, como exemplo, a resolução de uma equação $x^2 = a$, que pode ser reescrita de forma equivalente como $x = \frac{a}{x}$ (se $x \neq 0$), ou ainda

$$x = g(x) = \frac{x}{2} + \frac{a}{2x}$$

e daqui definir o método do ponto fixo

$$x_{n+1} = \frac{x_n}{2} + \frac{a}{2x_n}$$

começando com $x_0 = 1$. Automaticamente teremos

$$x_1 = \frac{a+1}{2}, \quad x_2 = \frac{a+1}{4} + \frac{a}{a+1}, \ldots$$

e para certos valores de a esta sucessão converge para $+\sqrt{a}$. Por exemplo, se $a = 2$, temos

$$x_1 = \frac{3}{2} = 1.5,\ x_2 = \frac{17}{12} = 1.4166..,\ x_3 = \frac{17}{12} = 1.4142..$$

e ao fim de 3 iterações temos uma aproximação de $\sqrt{2}$ com 5 dígitos correctos, usando apenas somas e divisões.

No entanto se esta escolha de g permite estes resultados notáveis, se tivessemos deixado a equação apenas na forma equivalente $x = g(x) = \frac{a}{x}$, o método $x_{n+1} = \frac{a}{x_n}$ levaria de $x_0 = 1$ a $x_1 = a$, e de novo $x_2 = 1$, $x_3 = a$, não se saindo deste ciclo. Ou seja, há escolhas de g que funcionam e outras não.

O que determina o sucesso do Método do Ponto Fixo é o comportamento da função g escolhida. Para a resolução de qualquer equação algébrica $f(x) = 0$, uma das melhores escolhas, quando possível, é a utilização da função iteradora $g(x) = x - f(x)/f'(x)$, que leva ao chamado Método de Newton (que é um caso particular de método de ponto fixo):

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

e foi essa escolha que fizémos no exemplo anterior, em que $f(x) = x^2 - a$.

Idealmente, o método de Newton procura que $g'(z) = 0$, para obter convergência mais rápida (quadrática), mas sabemos que o Método do Ponto Fixo converge localmente quando a função g é contractiva no ponto fixo, ou seja, quando

$$|g'(z)| < 1,$$

se a derivada estiver definida.

Começamos por generalizar a aplicação do método do ponto fixo à resolução de equações num contexto geral em que as incógnitas não são apenas números reais.

1.1 Motivação

Tomemos como exemplo uma equação em que a incógnita é uma função f contínua tal que

$$f(x) = 1 - \int_0^x f(t)dt$$

Figura 1.1:

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.

1.2 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 1. Seja E um espaço vectorial, em que o corpo dos escalares é $\mathbb{R}$ ou $\mathbb{C}$.

Uma aplicação $||.||$ designa-se *norma* se verificar:

- $||.||\ :\ E \to [0, +\infty)$
- $||\alpha x|| = |\alpha|\ ||x||\ ,\ \forall x \in E,\ \forall \alpha \in \mathbb{R}$ ou $\mathbb{C}$,
- $||x + y|| \leq ||x|| + ||y||$, $\forall x, y \in E$ (desigualdade triangular),
- $||x|| = 0 \Leftrightarrow x = 0$.

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 $(E, ||.||)$.

Observação 1. **(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 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).

1.2.1 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 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 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*.

Observação 2. 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. Os conjuntos E e $\emptyset$ são simultaneamente 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. Uma sucessão (x_n) num espaço normado E converge para $x \in E$, e escrevemos $x_n \to x$, se

$$||x_n - x|| \overset{n\to\infty}{\longrightarrow} 0$$

Observação 3. 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 $x = y$.

1.2.2 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 4. Duas normas $||.||$ e $|||.|||$, num mesmo espaço vectorial E, dizem-se equivalentes se existirem $C_1, C_2 > 0$ tais que:

$$C_1||x|| \leq |||x||| \leq C_2||x|| \ , \ \forall x \in E \tag{1.1}$$

Observação 4. 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 1. *Seja E um espaço normado de dimensão finita. Então, qualquer que seja a norma $|||.|||$ em E, existe $R > 0$:*

$$|||x||| \leq R, \ \ \forall x \in \{||x||_\infty \leq 1\}.$$

[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 extrair nenhuma subsucessão convergente da sucessão $u^{(k)}$ porque os elementos constituem uma base do espaço l^∞.

Figura 1.2:

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 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 2. 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 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. 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 $x = 1$).

[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.

Figura 1.3:

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...

1.3 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 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 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 5. 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 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 4. 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 2. *Se o conjunto A é fechado em E então*

$$\forall (x_n) \subseteq A : \ x_n \to x \implies x \in A$$

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. 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 4. Verifique que qualquer um dos espaços normados apresentados no exercício1 é um espaço de Banach.

Observação 6. (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 descontí nua! Concluí mos 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).

1.3.1 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 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

$$\forall (x_n) \subseteq X,\ x_n \to x \ \ \Rightarrow \ \ Ax_n \to Ax.$$

Exemplo 5. 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 $|||x_n|| - ||x||| \ = ||x - x_n + x_n|| - ||x_n|| \leq ||x - x_n|| + ||x_n|| - ||x_n||$.

[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 contí nuos em espaços de dimensão finita)... o contrário nem sempre será válido – essa é a principal precaução que se deve ter sempre!

Exercício 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)

1.3.2 Operadores Lineares

De entre os operadores contínuos, são especialmente importantes aqueles que sejam *lineares*, ou seja, que verifiquem as propriedades

$$A(x + y) = Ax + Ay,\ \forall x, y \in E$$
$$A(\alpha x) = \alpha Ax,\ \forall x \in E, \forall \alpha \in \mathbb{R} \text{ (ou } \mathbb{C})$$

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

$$||A||_{\mathcal{L}(E,F)} = \sup_{||x||_E \leq 1} ||Ax||_F = \sup_{x \neq 0} \frac{||Ax||_F}{||x||_E} \tag{1.2}$$

é um espaço de Banach. É claro que, para qualquer $x \in E$,

$$||Ax||_F \leq ||A||_{\mathcal{L}(E,F)} ||x||_E \ .$$

Exercício 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.*

[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.

1.4 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 domí nio, 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 8. Um operador $A : X \subseteq E \to E$, num espaço de Banach E diz-se *contractivo* em X, se existir $0 \leq L < 1$ (denominada *constante de contractividade*), tal que

$$||Ax - Ay|| \leq L||x - y|| \ , \ \forall x, y \in X.$$

Proposição 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 $x_n \to x$

$$||Ax_n - Ax|| \leq L||x_n - x|| \to 0 \ \Rightarrow \ Ax_n \to Ax.$$

□

Estamos agora em condições de demonstrar o teorema do ponto fixo de Banach.

Teorema 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 $x_0 \in X$.

[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.

iii) Verificam-se as desigualdades:

$$\begin{aligned} ||z - x_n|| &\leq L||z - x_{n-1}|| \leq L^n||z - x_0||, \\ ||z - x_n|| &\leq \frac{1}{1-L}||x_{n+1} - x_n||, \\ ||z - x_n|| &\leq \frac{L^n}{1-L}||x_1 - x_0||, \end{aligned}$$

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:

$$||z - x_n|| \leq ||Az - Ax_{n-1}|| \leq L||z - x_{n-1}|| \leq ... \leq L^n||z - x_0||$$

$$||z - x_n|| \leq ||z - x_{n+1}|| + ||x_{n+1} - x_n|| \leq L||z - x_n|| + ||x_{n+1} - x_n||$$

e daqui saiem facilmente as restantes. □

Observação 7. (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].

Observação 8. 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 6. 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}.

Exemplo 7. Consideremos o sistema em $\mathbb{R}^3$

$$\begin{cases} 3x_1 + x_2 = 1 \\ 2x_1 + 4x_2 + x_3 = 0 \\ x_2 + 2x_3 = 2 \end{cases} \Leftrightarrow \begin{cases} x_1 = 1/3 - x_2/3 \\ x_2 = -x_1/2 - x_3/4 \\ x_3 = 1 - x_2/2 \end{cases}$$

[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'.

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 $||.||_\infty$:

$$||Ax - Ay||_\infty = || \begin{bmatrix} \frac{1}{3}(x_2 - y_2) \\ \frac{1}{2}(x_1 - y_1) + \frac{1}{4}(x_3 - y_3) \\ \frac{1}{2}(x_2 - y_2) \end{bmatrix} ||_\infty$$

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 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.

1.5 Derivação de Fréchet

A derivação em espaços abstractos tem aspectos não triviais, que omitiremos deliberadamente (para uma compreensão aprofundada ver, por exemplo [**?**]). 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 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:

$$||A(x+h) - Ax - Th||_F = o(||h||_E) \quad \text{quando } ||h||_E \to 0 \tag{1.3}$$

Caso o operador T exista, é chamado *derivada de Fréchet em* x, e escrevemos A'_x, tendo-se

$$\begin{array}{rccl} A' : & X & \longrightarrow & \mathcal{L}(E,F) \\ & x & \longmapsto & A'_x : E \to F \quad \text{(operador linear)} \end{array}$$

Se A for F-diferenciável em todos os pontos $x \in X$ diremos que A é F-diferenciável em X.

Definição 11. 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 4. *Se A'_x existir é único.*

Demonstração. 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: $T = U$. □

[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 $\tilde{X}$.

Exercício 7. Verifique que se A é Fréchet-diferenciável, então A é contínuo.

Exemplo 8. 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],

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \dots & \frac{\partial f_1}{\partial x_N}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1}(x) & \dots & \frac{\partial f_N}{\partial x_N}(x) \end{bmatrix},$$

que é uma aplicação linear $\mathbb{R}^N \to \mathbb{R}^N$. Isto é uma consequência da fórmula de Taylor em $\mathbb{R}^N$,

$$f(y) = f(x) + \nabla f(x)(y - x) + o(||y - x||)$$

– Por exemplo, se $A(x_1, x_2) = (x_1^2 + x_2, x_1 e^{x_2})$ temos

$$A'_{(x_1,x_2)} = \begin{bmatrix} 2x_1 & 1 \\ e^{x_2} & x_1 e^{x_2} \end{bmatrix}.$$

Observação 9. A derivada de Fréchet de qualquer operador constante é o operador nulo.

- Sendo A um operador linear, a sua derivada de Fréchet em qualquer ponto x é sempre o próprio operador linear (sendo assim constante em x). Convém 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.

- 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:

Exercício 8. 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

$$(B \circ A)'_x = B'_{Ax} \circ A'_x \tag{1.4}$$

onde $(B \circ A)'_x \in \mathcal{L}(E, G)$. Para além disso, é claro que

$$||(BA)'_x||_{\mathcal{L}(E,G)} \leq ||B'_{Ax}||_{\mathcal{L}(F,G)} ||A'_x||_{\mathcal{L}(E,F)}.$$

[9]Por vezes também é designada matriz jacobiana a transposta desta. Essa escolha implicaria escrever $[\nabla f]^\top v$, quando quisessemos efectuar o produto por v, o que tornaria as notações mais pesadas.

1.5.1 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 12. Um conjunto não vazio $X \subseteq E$ diz-se *convexo* se verificar

$$x, y \in X \Rightarrow \forall t \in [0,1],\ x + t(y - x) \in X. \tag{1.5}$$

Observação 10. 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. *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

$$||Ax - Ay||_F \leq L||x - y||_E\ ,\ \forall x, y \in X.$$

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 (1.4), obtemos:

$$B'_t = A'_{x+t(y-x)}(y - x)$$

e vamos usar uma generalização da fórmula dos acréscimos finitos 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. □

Lema 2. *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 $||f(b) - f(a)||_F \leq K(b - a)$.*

[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.

Figura 1.4:

Demonstração. (e.g. [?]). □

Corolário 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*

$$||A'_x||_{\mathcal{L}(E,F)} \leq L < 1, \ \ \forall x \in X$$

as condições do Teorema do Ponto Fixo de Banach estão verificadas.

Observação 11. (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 consiste em considerar $g(x) = a\cos(x)$ com $|a| < 1$.

1.5.2 Comportamento assimptótico da convergência.

Estamos agora em condições de estudar o comportamento do erro obtido pela iteração do ponto fixo.

Proposição 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

$$\frac{||e_{n+1} - A'_z e_n||}{||e_n||} \longrightarrow 0.$$

□

Observação 12. Este resultado mostra 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 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

$$\frac{||e_{n+1}||}{||e_n||} \leq \frac{||e_{n+1} - A'_z e_n|| + ||A'_z e_n||}{||e_n||} = \varepsilon_n + ||A'_z(\frac{e_n}{||e_n||})|| \leq \varepsilon_n + ||A'_z||.$$

□

Observação 13. 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 9. 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.

Figura 1.5:

Figura 1.6:

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

$$||\nabla g(z)||_\infty = \max\{0.9\,|\sin 0.6065|\,, 0.9\,|\cos 0.7395|\} = \max\{0.513, 0.665\} = 0.665...$$

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.

1.5.3 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 13. 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 $\tilde{K}_\infty \leq \frac{1}{2}||A''_z||$.

1.6 Método de Newton

Neste contexto geral dos espaços de Banach, apenas fazemos uma breve referência ao método de Newton, já que iremos ver 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 invertí vel numa vizinhança da solução. Nessas condições, podemos estabelecer a equivalência:

$$Ax = 0 \Leftrightarrow (A'_x)^{-1}(Ax) = 0,$$

porque o inverso do operador linear contí nuo A'_x será um operador linear contí nuo, 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

$$x_{n+1} = x_n - (A'_{x_n})^{-1}(Ax_n), \tag{1.6}$$

que nesta generalização também é designado como *método de Newton-Kantorovich*.

[11] A norma $||A''_z||$ é a norma das aplicações bilineares contínuas, definida por

$$||B|| = \sup_{v,w \neq 0} \frac{||B(v,w)||}{||v||\,||w||}.$$

Observação 14. 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.

1.7 Ponto Fixo - Complementos

O método do ponto fixo de Banach, quando enunciado para espaços de Fréchet (espaços métricos completos), assume a seguinte forma:

Teorema 4. *Seja $K \subseteq E$ um subconjunto não vazio e fechado de um espaço de Fréchet E. Sendo $G : K \to K$ uma aplicação contractiva, isto é, existe uma constante $L \in [0, 1[$:*

$$d(G(u), G(v)) \leq Ld(u, v) \quad \forall u, v \in K,$$

então existe um e um só ponto fixo $z \in K : z = G(z)$, e são válidas as estimativas de erro do método do ponto fixo:

$$d(z, x_n) \leq L^n d(z, x_0); \quad d(z, x_n) \leq \frac{1}{1-L} d(x_{n+1}, x_n).$$

Demonstração. Exercício (é semelhante à demonstração do método do ponto fixo em espaços de Banach), usar a desigualdade triangular para métricas: $d(a, b) \leq d(a, c) + d(c, b)$. □

Observação 15. Para verificar que um conjunto K é fechado, é útil usar a propriedade das funções contínuas que estabelece que se a imagem é um fechado a pré-imagem é também um fechado.
Por exemplo, sendo a norma uma aplicação contínua $||.|| : K \subseteq E \to I \subseteq \mathbb{R}_0^+$ então se a imagem I é um fechado, a pré-imagem $K = f^{-1}(I)$ também é um fechado.

Teorema 5. *(de Brouwer). Seja $K \subset \mathbb{R}^N$ um conjunto homeomorfo à bola unitária $\bar{B}(0, 1)$, e seja $\mathcal{G} : K \to K$ uma função contínua, então existe pelo menos um ponto fixo de $\mathcal{G}$.*

Demonstração. O teorema original de Brouwer foi estabelecido em 1912 para $\bar{B}(0,1)$ e a sua demonstração não é construtiva. Apresentamos apenas a justificação de que é extensível a conjuntos homeomorfos a essa bola, que é simples.
Nesse caso, existe um homeomorfismo $\mathcal{H} : K \to \bar{B}(0,1)$, tal que $\bar{B}(0,1) = \mathcal{H}(K)$. Consideramos por isso $G = \mathcal{H} \circ \mathcal{G} \circ \mathcal{H}^{-1}$

$$G: \quad \bar{B}(0,1) \underset{\mathcal{H}^{-1}}{\to} K \underset{\mathcal{G}}{\to} K \underset{\mathcal{H}}{\to} \bar{B}(0,1)$$

A aplicação G é composição de funções contínuas, logo é contínua, e aplica-se o teorema original de Brouwer na bola unitária, existindo um ponto fixo $z = G(z) = \mathcal{H} \circ \mathcal{G} \circ \mathcal{H}^{-1}(z) \in \bar{B}(0,1)$. Resta agora notar que $x = \mathcal{H}^{-1}(z) \in K$, é ponto fixo de $\mathcal{G}$ em K, pois

$$\mathcal{G}(x) = \mathcal{G} \circ \mathcal{H}^{-1}(z) = \mathcal{H}^{-1} \circ \mathcal{H} \circ \mathcal{G} \circ \mathcal{H}^{-1}(z) = \mathcal{H}^{-1}(G(z)) = \mathcal{H}^{-1}(z) = x.$$

□

Observação 16. Muitas vezes o teorema de Brouwer é generalizado apenas para convexos de $\mathbb{R}^N$, o que é um caso particular, já que todos os convexos de $\mathbb{R}^N$ são homeomorfos à bola unitária. Esta outra formulação inclui outro tipo de conjuntos, por exemplo, todos os estrelados.

Observação 17. No caso em que o conjunto K não é homeomorfo à bola unitária, por exemplo no caso da circunferência, ou de uma coroa circular (em $\mathbb{R}^2$), podemos definir uma rotação dos seus pontos de forma a excluir o seu centro, que seria o ponto fixo, não sendo válido o Teorema de Brouwer.

O resultado de Brouwer é apenas aplicável a dimensão finita. Para estendermos o resultado a outros espaços, de Banach, ou mesmo apenas normados, necessitamos de uma hipótese de compacidade, que permite essa redução a dimensão finita. Começamos por recordar a definição geral de conjunto compacto, aplicada a cobertura por bolas.

Definição 14. Seja $K \subset E$ um subconjunto de um espaço normado E. Dizemos que K é compacto, quando dada uma qualquer cobertura infinita por bolas abertas $B(x,\epsilon)$, ou seja $K \subset \cup_{x \in K} B(x,\epsilon)$ é possível extrair uma cobertura finita

$$K \subset \cup_{j=1}^{N} B(x_j, \epsilon).$$

Num espaço normado, isto é equivalente a exigir que as sucessões em K têm subsucessões convergentes cujo limite está em K. Diz-se que o conjunto é relativamente compacto se o fecho for compacto.

Observação 18. Note-se que as bolas só são compactas quando o espaço é de dimensão finita. Basta reparar que a sucessão definida pelos vectores da base canónica em $\mathbb{R}^N$ está na bola unitária, no entanto quando $N \to \infty$ isso continuaria a sucessão, não podendo ser extraída uma subsucessão convergente.

Teorema 6. *(de Schauder). Seja $K \subset E$ um conjunto compacto e convexo de um espaço normado E. Seja $\mathcal{G} : K \to K$ uma função contínua, então existe pelo menos um ponto fixo de $\mathcal{G}$.*

Demonstração. A demonstração usa o teorema de Brouwer, reduzindo à dimensão finita pelo número finito de bolas que cobrem o conjunto compacto K. Apenas notamos que definindo

$$g(x) = \frac{\sum_{j=1}^{N} x_j g_j(x)}{\sum_{j=1}^{N} g_j(x)} \quad \text{com } g_j(x) = (\epsilon - ||x - x_j||)\chi_{\bar{B}(x_j,\epsilon)}(x)$$

em que χé a função característica, trata-se de uma função contínua de K para o envelope convexo definido pelo número finito de centros na cobertura das bolas, sendo possivel reduzir a um problema de dimensão finita no envelope convexo definido pelos centros das bolas. Aí é possível aplicar o teorema de Brouwer.

□

1.8 Métodos Iterativos para Sistemas de Equações Não Lineares

Iremos agora apresentar métodos iterativos que permitem aproximar a solução de sistemas de equações. Começamos por apresentar o caso geral, em que se supõe que o sistema pode ou 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.

Através das 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 $\mathbb{R}^N$

$$\begin{cases} f_1(x_1, \ldots, x_N) = 0, \\ \vdots \\ f_N(x_1, \ldots, x_N) = 0, \end{cases}$$

que podemos escrever abreviadamente $F(x) = 0$, estabelecemos uma equivalência com o sistema na forma $x = G(x)$, ou seja,

$$\begin{cases} x_1 = g_1(x_1, \ldots, x_N), \\ \vdots \\ x_N = g_N(x_1, \ldots, x_N), \end{cases}$$

Sendo $\nabla G(x)$ 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 3. *(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) $||\nabla G(x)|| \leq L < 1, \ \forall x \in D$
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

$$\begin{aligned} ||z - x^{(n)}|| &\leq L||z - x^{(n-1)}|| \leq L^n||z - x^{(0)}|| \\ ||z - x^{(n)}|| &\leq \frac{1}{1-L}||x^{(n+1)} - x^{(n)}|| \\ ||z - x^{(n)}|| &\leq \frac{L^n}{1-L}||x^{(1)} - x^{(0)}|| \end{aligned}$$

Exemplo 10. Consideremos o sistema linear

$$\left\{ \begin{array}{c} 3x_1 + x_2 = 1 \\ 2x_1 + 4x_2 + x_3 = 0 \\ x_2 + 2x_3 = 2 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{c} x_1 = 1/3 - x_2/3 \\ x_2 = -x_1/2 - x_3/4 \\ x_3 = 1 - x_2/2 \end{array} \right.$$

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 $||\nabla G(\mathbf{x})||_\infty = ||\mathbf{A}||_\infty = 5/6 < 1$, 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 11. Consideremos agora o sistema não-linear:

$$\left\{ \begin{array}{l} 2x - \cos(x+y) = 2 \\ 3y - \sin(x+y) = 6 \end{array} \right.$$

Vamos ver que existe uma e uma só solução em $\mathbb{R}^2$ e que ela está em $X = [\frac{1}{2}, \frac{3}{2}] \times [\frac{5}{3}, \frac{7}{3}]$. Com efeito, se considerarmos

$$G(x, y) = (\cos(x+y)/2 + 1, \sin(x+y)/3 + 2),$$

a matriz jacobiana de G vem

$$\nabla G(x, y) = \begin{bmatrix} -\sin(x+y)/2 & -\sin(x+y)/2 \\ \cos(x+y)/3 & \cos(x+y)/3 \end{bmatrix}$$

Aplicando o corolário do T. Ponto Fixo, vemos que $||\nabla G(x,y)||_1 \leq 5/6 < 1$ e concluímos que existe uma e uma só solução em $\mathbb{R}^2$ (repare-se que se escolhessemos a norma $||.||_\infty$ teríamos apenas $||\nabla G(x,y)||_\infty \leq 1$, 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 $(0.549322733, 2.144360661)$.

1.8.1 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

$$F(x) = 0 \Leftrightarrow [\nabla F(x)]^{-1} F(x) = 0 \Leftrightarrow x = x - [\nabla F(x)]^{-1} F(x)$$

e a função iteradora será, portanto, $G(x) = x - [\nabla F(x)]^{-1} F(x)$.

Dado $x^{(0)} \in \mathbb{R}^N$, o método consistiria na iteração

$$x^{(n+1)} = x^{(n)} - [\nabla F(x^{(n)})]^{-1} F(x^{(n)}). \tag{1.7}$$

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:

$$[\nabla F(x^n)]v = -F(x^{(n)}) \tag{1.8}$$

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 12. Consideremos o sistema do exemplo anterior. A matriz jacobiana de F vem

$$\nabla F(x,y) = \begin{bmatrix} 2 + \sin(x+y)/2 & \sin(x+y)/2 \\ -\cos(x+y)/3 & 3 - \cos(x+y)/3 \end{bmatrix}$$

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 6. *(convergência local). Seja $F \in C^1(V_z)$, em que V_z é uma vizinhança de uma solução z, onde $\det(\nabla F(x)) \neq 0$, $\forall x \in V_z$. Então o método de Newton converge para z, desde que a vizinhança seja suficientemente pequena e $x_0 \in V_z$.*

Demonstração. Exercício. □

Teorema 7. *Seja $F \in C^2(V_z)$, em que a solução z não é um ponto crítico*[12]*. 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 $f : \mathbb{R}^N \to \mathbb{R}$:

$$f(x+h) = f(x) + \nabla f(x) \cdot h + \frac{1}{2} h \cdot \nabla^2 f(x + \xi h)\, h, \text{ para um certo } \xi \in]0,1[$$

onde $\nabla^2 f(y) = [\frac{\partial^2 f}{\partial x_i \partial x_j}]$ 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

$$F(x+h) = F(x) + \nabla F(x) \cdot h + \frac{1}{2} h \cdot \nabla^2 f_i(x + \xi_i h)\, h, \text{ para certos } \xi_i \in]0,1[,$$

onde o termo $\frac{1}{2} h \cdot \nabla^2 f_i(x + \xi_i h) \cdot h$ é um vector, que está apresentado na componente i.

Aplicando este resultado ao método de Newton, obtemos

$$0 = F(z) = F(x^{(n)}) + \nabla F(x^{(n)}) \cdot e^{(n)} + \frac{1}{2} e^{(n)} \cdot \nabla^2 f_i(x^{(n)} + \xi_i e^{(n)})\, e^{(n)}$$

em que $e^{(n)} = z - x^{(n)}$ é o erro na iterada n. Reparando que, no método de Newton, $\nabla F(x^{(n)}) \cdot (x^{(n+1)} - x^{(n)}) = -F(x^{(n)})$, ao somar e subtrair z, ficamos com

$$-F(x^{(n)}) = \nabla F(x^{(n)}) \cdot (x^{(n+1)} - z + z - x^{(n)}) = -\nabla F(x^{(n)}) \cdot e^{(n+1)} + \nabla F(x^{(n)}) \cdot e^{(n)},$$

obtendo-se

$$\nabla F(x^{(n)}).e^{(n+1)} = -\frac{1}{2} e^{(n)} \cdot \nabla^2 f_i(x^{(n)} + \xi_i e^{(n)})\, e^{(n)}.$$

Como $f \in C^2(V_z)$, supomos agora que $||\nabla^2 f_i(x)|| \leq M_2$, e que $|| \, [\nabla F(x_n)]^{-1}|| \leq \frac{1}{M_1}$, numa vizinhança da solução[13]. Obtemos a estimativa pretendida,

$$||e^{(n+1)}|| \leq \frac{M_2}{2M_1} ||e^{(n)}||^2.$$

□

[12] Ou seja, $\det(\nabla F(z)) \neq 0$.

[13] 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.

Observação 19. (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 $[\nabla F(x_n)]^{-1}$, a sua norma pode ser facilmente calculada, e nesse caso podemos escrever a estimativa

$$||e^{(n+1)}|| \leq \frac{1}{2} \max_{x \in V} ||\nabla^2 F(x)|| \; || \; [\nabla F(x_n)]^{-1}|| \; ||e^{(n)}||^2, \tag{1.9}$$

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 $||\nabla^2 F(x)||$ deve ser entendido como o máximo das normas matriciais $\max_i ||\nabla^2 f_i(x)||$.

1.8.2 Complementos

Há ainda a possibilidade de apresentar uma condição suficiente para a convergência, semelhante à obtida no caso escalar, e que também poderá servir de critério em $\mathbb{R}$. Enunciamos apenas o resultado, cuja demonstração pode ser encontrada em [**?**]:

Teorema 8. *(Kantorovich). Seja $D \subset \mathbb{R}^N$ um conjunto aberto e convexo e $F \in C^1(D)$.*
Se
(i) $\exists M_1 > 0 : || \, [\nabla F(x)]^{-1} \, || \leq \frac{1}{M_1}, \; \forall x \in D$,
(ii) $\exists M_2 > 0 : ||\nabla F(x) - \nabla F(y)|| \leq M_2 ||x - y||, \; \forall x, y \in D$,
(iii) existe $x_0 \in D$, *tal que* $\varepsilon_0 = 2 \, || \, [\nabla F(x_0)]^{-1} F(x_0) \, ||$ *verifica* $\frac{M_2}{M_1} \varepsilon_0 < 1$,
(iv) $\bar{B}(x_0, \varepsilon_0) \subset D$,
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).*

Observação 20. 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 21. (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. [**?**]).

Observação 22. (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 ∇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 $\nabla F(x^{(n)})$ 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 23. 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.

1.8.3 Quasi-Newton usando diferenças divididas

A computação do Método de Newton envolve cálculos morosos quando o valor da função tem um tempo de cálculo longo, e para além disso exige-se a computação da derivada, o que pode ser inexequível, bem como a resolução de um sistema linear (ou a inversão da matriz jacobiana) que também pode ser moroso para dimensão grande. Assim foram surgindo métodos que procuram aliviar esse cálculo, substituindo principalmente o cálculo das derivadas na computação de

$$J_{f(x_n)}\Delta x_n = -f(x_n).$$

Um processo simples de evitar o cálculo da matriz jacobiana em $\mathbb{R}^N$ é considerar a aproximação ($e_k$são os vectores da base canónica)

$$J_{f(x)} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_N} \end{bmatrix} \approx \begin{bmatrix} f_1[x-\frac{h}{2}e_1, x+\frac{h}{2}e_1] & \cdots & f_1[x-\frac{h}{2}e_N, x+\frac{h}{2}e_N] \\ \vdots & \ddots & \vdots \\ f_N[x-\frac{h}{2}e_1, x+\frac{h}{2}e_1] & \cdots & f_N[x-\frac{h}{2}e_N, x+\frac{h}{2}e_N] \end{bmatrix}$$

usando diferenças centradas, que converge em $O(h^2)$ quando $h \to 0$, podendo escolher-se h próximo da precisão da máquina, desde que não afecte demasiado os cálculos por arredondamento. No entanto reparamos que isto exige um número de novas avaliações da função exagerado ($2N^2$), o que torna a sua aplicabilidade reduzida em grandes matrizes. É

preferível recorrer às diferenças progressivas $\frac{\partial f_k}{\partial x_j} \approx f_k[x, x + he_j]$ já que apesar da aproximação ser apenas $O(h)$, requer metade dos cálculos, ou seja N^2, que seria também o número de cálculos a efectuar para as derivadas na matriz jacobiana.

Ainda que a resolução do sistema seja preferível à computação da inversa, podemos usar uma aproximação da inversa iterativamente (usando o Método de Newton). Começamos com $X_0 = [J_{f(x_{n-1})}]^{-1}$ e iteramos

$$X_{k+1} = 2X_k - X_k J_{f(x_n)} X_k$$

o que nos garante convergência quadrática $X_k \to [J_{f(x_n)}]^{-1}$ quando $||I - X_0 J_{f(x_n)}|| < \frac{1}{2}$, o que acontece se $J_{f(x_n)}[J_{f(x_{n-1})}]^{-1} \approx I$, ou seja quando $x_n \approx x_{n-1}$.

Vemos de seguida um outro método, de Broyden, que generaliza o método da secante, a fim de evitar estes problemas.

1.8.4 Método de Broyden

O método de Broyden é uma implementação alternativa do método de Newton, apenas válido para sistemas, em dimensão finita. O método pode evitar o sucessivo cálculo da matriz jacobiana, bem como o da sua inversa aplicando a identidade de Sherman-Morrison.

Lema 3. *(Fórmula de Sherman-Morrison). Seja A uma matriz invertível, e u, v vectores quaisquer (todos com a mesma dimensão), temos:*

$$(A + uv^*)^{-1} = A^{-1} - \frac{A^{-1}uv^*A^{-1}}{1 + v^*A^{-1}u} \tag{1.10}$$

Demonstração. Basta confirmar que é a inversa:

$$(A + uv^*)\left(A^{-1} - \frac{A^{-1}uv^*A^{-1}}{1 + v^*A^{-1}u}\right) = I + uv^*A^{-1} - (A + uv^*)\frac{A^{-1}uv^*A^{-1}}{1 + v^*A^{-1}u}$$

$$= I + uv^*A^{-1} - \frac{AA^{-1}uv^*A^{-1} + uv^*A^{-1}uv^*A^{-1}}{1 + v^*A^{-1}u}$$

$$= I + uv^*A^{-1} - \frac{u(1 + v^*A^{-1}u)v^*A^{-1}}{1 + v^*A^{-1}u} = I$$

□

Consideramos a aplicação desta fórmula à expressão do método de Newton para sistemas:

$$x_{n+1} = x_n - J^{-1}_{f(x_n)} f(x_n)$$

Usamos a notação abreviada J_k para designar a aproximação que fazemos de $J^{-1}_{f(x_k)}$.

Assim, com u_k e v_k são vectores convenientemente escolhidos, de forma a que

$$J_{k+1} = J_k + u_k v_k^*$$

e podemos escrever J_{k+1}^{-1} a partir da inversa de J_k usando a fórmula de Sherman-Morrison. Como explicaremos na Observação 25, Broyden propôs usar $u_k = \frac{\Delta f_k - J_k \Delta x_k}{||\Delta x_k||^2}$ e $v_k = \Delta x_k$, (abreviamos $f_k = f(x_k)$) o que corresponde a escrever

$$J_{k+1} = J_k + \frac{\Delta f_k - J_k \Delta x_k}{||\Delta x_k||^2} (\Delta x_k)^\top \tag{1.11}$$

e usando a fórmula de Sherman-Morrison, esta aproximação permite calcular $J_n^{-1} \approx J_{f(x_n)}^{-1}$ a partir de J_{n-1}, que depois substituímos na expressão do método de Newton:

$$x_{n+1} \approx x_n - J_n^{-1} f(x_n)$$

Através desta aproximação, é possível reduzir significativamente o número de operações.

Observação 24. Podemos explicitar melhor a inversa J_n^{-1} obtida pela fórmula de Sherman-Morrison, usando os u_k e v_k apresentados por Broyden:

$$J_{k+1}^{-1} = (J_k + u_k v_k^\top) = J_k^{-1} - \frac{J_k^{-1} \frac{\Delta f_k - J_k \Delta x_k}{||\Delta x_k||^2} (\Delta x_k)^\top J_k^{-1}}{1 + (\Delta x_k)^\top J_k^{-1} \frac{\Delta f_k - J_k \Delta x_k}{||\Delta x_k||^2}}$$

$$= J_k^{-1} - \frac{J_k^{-1} (\Delta f_k - J_k \Delta x_k)(\Delta x_k)^\top J_k^{-1}}{||\Delta x_k||^2 + (\Delta x_k)^\top J_k^{-1} \Delta f_k - (\Delta x_k)^\top J_k^{-1} J_k \Delta x_k}$$

simplificando,

$$J_{k+1}^{-1} = J_k^{-1} - \frac{J_k^{-1} \Delta f_k - \Delta x_k}{(\Delta x_k)^\top J_k^{-1} \Delta f_k} (\Delta x_k)^\top J_k^{-1} \tag{1.12}$$

Observação 25. A expressão (1.11) pode ser justificada pela semelhança com o Método da Secante.
No caso do método da secante escolhíamos x_{n+1} de forma a que $f_{n+1} = 0$, logo $\Delta f_n = -f_n$, escolhendo-se J tal que $\Delta x_n = J^{-1} \Delta f_n$ de onde $J = \frac{\Delta f_n}{\Delta x_n}$ (a razão incremental).
De forma semelhante, para substituir a iteração exacta de Newton $J_{f(x_k)} \Delta x_k = -f_k$, queremos encontrar agora uma matriz J_{k+1} tal que

$$\Delta x_k = J_{k+1}^{-1} \Delta f_k$$

o que é verificado pela expressão (1.11), pois

$$J_{k+1} \Delta x_k = \left(J_k + \frac{\Delta f_k - J_k \Delta x_k}{||\Delta x_k||^2} (\Delta x_k)^\top \right) \Delta x_k = J_k \Delta x_k + \Delta f_k - J_k \Delta x_k = \Delta f_k.$$

Notamos ainda que há outras variantes do método de Broyden.

Exemplo 13. Aplicamos o Método de Broyden ao sistema

$$\begin{cases} x_1^2 + 2x_1 + x_3 = 1 \\ x_1 x_2 + x_3 = 1 - x_3^2 \\ x_1 x_2 x_3 = x_2^2 - x_2 \end{cases}$$

com a função $\mathbf{f}(\mathbf{x}) = (x_1^2 + 2x_1 + x_3 - 1, x_1x_2 + x_3 + x_3^2 - 1, x_1x_2x_3 + x_2 - x_2^2) = 0$.

Começando com $\mathbf{x}^{(0)} = (0, 0, 0)$, calculamos a matriz jacobiana só na 1ª iterada (que é igual à do método de Newton)

$$J_{\mathbf{f}(x^{(0)})} = \begin{bmatrix} 2x_1 + 2 & 0 & 1 \\ x_2 & x_1 & 2x_3 + 1 \\ x_2x_3 & x_1x_3 - 2x_2 + 1 & x_1x_2 \end{bmatrix}_{\mathbf{x}=\mathbf{x^{(0)}}} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = J_0$$

$$\implies J_0^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \implies \mathbf{x}^{(1)} = \mathbf{x}^{(0)} - J_0^{-1} f(\mathbf{x}^{(0)}) = \mathbf{0} - \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

A 2ª iterada e restantes serão diferentes, como $\mathbf{f}_1 = \mathbf{f}(\mathbf{x}^{(1)}) = (0, 1, 0)$ e $\Delta\mathbf{x}^{(0)} = (0, 0, 1)$

$$\mathbf{u}_0 = \frac{\Delta\mathbf{f}_0 - J_0\Delta\mathbf{x}^{(0)}}{||\Delta\mathbf{x}^{(0)}||^2} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_0^\top = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

portanto

$$J_1 = J_0 + \mathbf{u}^{(0)}(\mathbf{v}^{(0)})^\top = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 2 \\ 0 & 1 & 0 \end{bmatrix}$$

e pela fórmula de Sherman-Morrison

$$J_1^{-1} = J_k^{-1} - \frac{J_k^{-1}\mathbf{u}^{(0)}(\mathbf{v}^{(0)})^\top J_k^{-1}}{1+(\mathbf{v}^{(0)})^\top J_k^{-1}\mathbf{u}^{(0)}} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} - \frac{\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}}{1+\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

assim, obtemos

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - J_1^{-1} f(\mathbf{x}^{(1)}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} \\ 0 \\ \frac{1}{2} \end{bmatrix}$$

Na terceira iterada obteríamos $\mathbf{x}^{(3)} = (\frac{3}{17}, 0, \frac{61}{102}) \approx (0.17647, 0, 0.59804)$ o que já é um valor próximo da solução $\mathbf{z}$ =(0.17557.., 0, 0.618034..), valor que seria obtido (nesta precisão) na 4ª iterada do Método de Newton, mas apenas na 6ª iterada do Método de Broyden.

Capítulo 2

Espaços Funcionais

O contexto de espaços de Banach é adequado a espaços de funções contínuas ou diferenciáveis no sentido clássico, mas os espaços$C^m[a,b]$ sendo completos para a norma do máximo

$$||u||_{C^m[a,b]} = ||u||_\infty + ||u'||_\infty + ... + ||u^{(m)}||_\infty,$$

não são completos quando se considera o produto interno clássico, definido em $L^2(a,b)$ pela integração de Lebesgue. Aí é possível obter completude para essas funções, mas perdemos a possibilidade de considerar derivadas clássicas. É nesse sentido que vamos rever alguns resultados em espaços de Hilbert, e introduzir espaços de Sobolev, $H^m(a,b)$ que estão definidos através de uma noção de derivação generalizada.

2.1 Resultados em Espaços de Hilbert

Um espaço vectorial com produto interno $\langle .,. \rangle$ é denominado pré-hilbertiano (ou euclidiano), notando que no caso em que o corpo de escalares é complexo temos

$$\langle \alpha u, v \rangle = \bar{\alpha} \langle u, v \rangle \text{ e também } \langle v, u \rangle = \overline{\langle u, v \rangle},$$

por isso convencionamos que a conjugação se efectua no primeiro termo. Associa-se a norma definida por $||u||^2 = \langle u, u \rangle$, e temos (caso complexo)

$$||u+v||^2 = ||u||^2 + 2 \text{ Re} \langle u, v \rangle + ||v||^2 \tag{2.1}$$

e se $\langle u, v \rangle = 0$ obtemos o teorema de Pitágoras $||u+v||^2 = ||u||^2 + ||v||^2$, e daqui é consequência imediata a igualdade do paralelogramo:

$$||\frac{u+v}{2}||^2 + ||\frac{u-v}{2}||^2 = \frac{1}{2}(||u||^2 + ||v||^2).$$

A existência de produto interno num espaço normado pode ser avaliada pela igualdade do paralelogramo[1]. Outro resultado conhecido do produto interno é a desigualdade de

[1]A igualdade do paralelogramo não é verificada para a norma do máximo, basta ver este contra-exemplo em $\mathbb{R}^2$, como $u = (1,0), v = (0,1)$

$$||\frac{u+v}{2}||^2_\infty + ||\frac{u-v}{2}||^2_\infty = \frac{1}{2} \neq 1 = \frac{1}{2}(||u||^2_\infty + ||v||^2_\infty).$$

Cauchy-Schwarz

$$|\langle u, v\rangle| \leq ||u||\,||v||.$$

Na teoria de funções considera-se muito habitualmente o produto interno e a norma associada em $L^2(a,b)$,

$$\langle f, g\rangle_{L^2(a,b)} = \int_a^b f(\bar{t})g(t)dt, \quad ||f||_{L^2(a,b)} = \left(\int_a^b f(t)^2 dt\right)^{1/2}.$$

Quando o espaço pré-hilbertiano é completo (as sucessões de Cauchy convergem), designa-se espaço de Hilbert H. Relembramos que este é o caso de todos os espaços de dimensão finita (isomorfos a $\mathbb{R}^n$), ou das funções em $L^2(a,b)$. No entanto, se considerarmos$C[a,b]$, e apesar do produto interno L^2 estar bem definido, as sucessões de Cauchy nessa norma $L^2(a,b)$ podem convergir para uma função $L^2(a,b)$ que não é contínua...

2.1.1 Sistema Normal

Consideramos a aproximação num subespaço vectorial de H gerado por funções base, $S = \langle \varphi_1, \cdots, \varphi_n\rangle$, que tem dimensão finita n, onde está definido um produto interno. Relativamente à distância definida nesse espaço de Hilbert, pelo produto interno $dist(u,v) = ||u-v|| = \langle u-v, u-v\rangle^{1/2}$, a melhor aproximação que é neste caso única, é dada pela resolução do sistema normal. Relembramos que dado $f \in H$ isso corresponde a encontrar g tal que

$$||f-g|| = \inf_{\varphi \in S} ||f-\varphi||$$

como S tem dimensão finita existe um mínimo, podemos encontrar uma condição para mínimo através da derivada de Fréchet de $d(g) = ||f-g||^2$, fixo f. Usando (2.1),

$$d(g+h) - d(g) \;=\; ||g+h-f||^2 - ||g-f||^2 = 2\,\mathrm{Re}\,\langle g-f, h\rangle + \underbrace{||h||^2}_{o(||h||)},$$

no caso real conclui-se que $d'_g(h) = 2\langle g-f, h\rangle$. Procurando g tal que $d'_g \equiv 0$, e restringindo o problema a S (subespaço fechado), trata-se de encontrar $g \in S$ tal que

$$\langle f-g, h\rangle = 0, \forall h \in S$$

De facto, escrevendo $g = a_1\varphi_1 + ... + a_n\varphi_n$ basta verificar a condição para as funções base e assim $\langle \varphi_k, f-g\rangle = 0$ leva ao sistema normal

$$\langle \varphi_k, g\rangle = \langle \varphi_k, f\rangle \Leftrightarrow \begin{bmatrix} \langle \varphi_1, \varphi_1\rangle & \cdots & \langle \varphi_1, \varphi_n\rangle \\ \vdots & \ddots & \vdots \\ \langle \varphi_n, \varphi_1\rangle & \cdots & \langle \varphi_n, \varphi_n\rangle \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \langle \varphi_1, f\rangle \\ \vdots \\ \langle \varphi_n, f\rangle \end{bmatrix}$$

notando que a matriz $\Phi = [\langle \varphi_i, \varphi_j\rangle]_{ij}$ é simétrica (hermitiana) e definida positiva (pois $\mathbf{a}^*\Phi\mathbf{a} = \sum_k \bar{a_k}\langle \varphi_k, g\rangle = \langle g, g\rangle = ||g|| > 0$, para qualquer $g \neq 0$).

A função $g \in S$ obtida pela solução do sistema normal é denominada projecção de f sobre S, escrevendo-se

$$g = \text{Proj}_S(f) = \sum_{k=1}^{n} a_k \varphi_k.$$

O erro da aproximação $||f-g||$, é determinado imediatamente pela norma, com a solução g obtida.

O espaço de Hilbert é separável se admite uma base ortonormada numerável $\hat{\phi}_1, \cdots, \hat{\phi}_n, \cdots$, e assim podemos escrever qualquer $f \in H$ através da expansão de Fourier

$$f = \sum_{k=1}^{\infty} \left\langle \hat{\phi}_k, f \right\rangle \hat{\phi}_k$$

uma generalização da expansão em série de Fourier. Este é o limite da sucessão de somas finitas

$$f_n = \sum_{k=1}^{n} \left\langle \hat{\phi}_k, f \right\rangle \hat{\phi}_k$$

que correspondem à solução do sistema normal limitando a base, já que no caso de base ortonormada $\left\langle \hat{\phi}_i, \hat{\phi}_j \right\rangle = \delta_{ij}$ e a matriz do sistema normal seria a identidade.

Teorema 9. *Num espaço de Hilbert definido pela base ortonormada* $(\hat{\phi}_n)$*, temos a desigualdade de Bessel*

$$||f||^2 \geq \sum_{k=1}^{n} \left| \left\langle \hat{\phi}_k, f \right\rangle \right|^2 = ||f_n||^2$$

verificando-se $||f - f_n||^2 = ||f||^2 - ||f_n||^2$*, o que no caso limite dá a igualdade de Parseval*

$$||f||^2 = \sum_{k=1}^{\infty} \left| \left\langle \hat{\phi}_k, f \right\rangle \right|^2.$$

2.1.2 Derivada generalizada

Quando f não é diferenciável, podemos definir uma generalização da noção, usando um formulação fraca em L^2. Para esse efeito consideramos *funções teste* que são diferenciáveis, e têm suporte compacto, por exemplo $v \in C_c^p(a,b)$

$$v \in C_c^p[a,b] = \{v \in C^p[a,b] : \text{supp}(v) \subset (a,b)\}$$

onde $\text{supp}(v) = \overline{\{x \in (a,b) : v(x) \neq 0\}}$.

Observação 26. Definimos também $C_c^p(\mathbb{R}) = \{v \in C^p(\mathbb{R}) : \exists (a_v, b_v), \text{supp}(v) \subset (a_v, b_v)\}$, generalizando o caso anterior para um intervalo infinito. Desta forma, como como o suporte está no interior de um intervalo $[a,b]$, será nula nos extremos $v^{(k)}(a) = v^{(k)}(b) = 0$ (com $k \leq p$).

Assim, para $v \in C_c^1(\mathbb{R})$, obtemos pela regra de integração por partes

$$\begin{aligned}\langle f', v\rangle &= \int_{\mathbb{R}} f'(t)v(t)dt = \int_a^b f'(t)v(t)dt = \\ &= [f(t)v(t)]_a^b - \int_a^b f(t)v'(t)dt = -\int_a^b f(t)v'(t)dt \\ &= -\langle f, v'\rangle .\end{aligned}$$

Reparamos que a expressão $-\langle f, v'\rangle$ tem sentido clássico, qualquer que seja f localmente integrável. Isto permite generalizar a noção de derivada, mesmo para funções não diferenciáveis. De forma recursiva, definimos as restantes derivadas. Para simplificar usamos o contexto generalizado apenas em $L^2(a, b)$ porque é esse que nos interessa no contexto dos espaços de Sobolev que são espaços de Hilbert.

Definição 15. Dizemos que $\phi \in L^2(a, b)$ é derivada generalizada de ordem p de $f \in L^2(a, b)$ se verificar

$$\langle \phi, v\rangle = (-1)^p \left\langle f, v^{(p)}\right\rangle , \quad \forall v \in C_c^\infty(a, b).$$

Esta derivada está bem definida, a menos de conjunto de medida nula.

Exemplo 14. Calcular a derivada de $f(x) = |x|$ em $\mathbb{R}$. Esta função não é diferenciável em 0, por isso vemos como pode ser definida a derivada generalizada. Dado qualquer $v \in C_c^\infty(\mathbb{R})$, com $\text{supp}(v) \subset (-r, r)$, para r suficientemente grande, temos

$$\langle f, v'\rangle = \int_{-r}^r |x|v'(x)dx = \int_{-r}^0 (-x)v'(x)dx + \int_0^r xv'(x)dx$$

(e prosseguimos no sentido clássico da integração por partes)

$$= [xv]_{-r}^0 - \int_{-r}^0 (-1)v(x)dx - [xv]_0^r - \int_0^r (+1)v(x)dx,$$

como $[xv]_{-r}^0 = 0$, $[xv]_0^r = 0$ (porque $v(\pm r) = 0$, já que o suporte está dentro do intervalo $(-r, r)$), obtemos

$$-\langle f, v'\rangle = \int_{-r}^r \text{sgn}(x)v(x)dx = \langle \text{sgn}, v\rangle$$

em que *sign* é a função sinal ($\text{sign}(x) = \pm 1$ se $\pm x > 0$), observando que o valor no ponto zero é irrelevante (conjunto de medida nula). Concluímos assim que no, sentido generalizado, a função sinal é a derivada do módulo, o que aliás coincide com a derivada em cada um dos troços diferenciáveis.

Podemos prosseguir para uma segunda derivada? De forma semelhante obtemos

$$\begin{aligned}-\langle \text{sgn}, v'\rangle &= -\int_{-r}^0 (-1)v'(x)dx - \int_0^r (+1)v'(x)dx \\ &= [v(0) - v(-r)] - [v(r) - v(0)] = 2v(0),\end{aligned}$$

o que não pode ser escrito na forma de um integral no sentido clássico[2]. O resultado pode ser expresso através de um funcional que é o delta de Dirac δ centrado em zero.

Definição 16. Definimos o funcional linear delta de Dirac, $\delta(v) = v(0)$, para $v \in C_c^\infty(\mathbb{R})$, ou simplesmente para $v \in C(\mathbb{R})$. Este pode ser representado na forma$\langle \delta, v \rangle$, entendendo o integral neste sentido generalizado $\int_{\mathbb{R}} \delta(t)v(t)dt = v(0)$. Por translação definimos ainda $\delta_x(v) = v(x)$, notando que deve considerar-se $\int_a^b \delta_x(t)v(t)dt = v(x)$ apenas quando $x \in (a, b)$, se $x \notin [a, b]$ o valor do integral deve ser considerado zero.

Pela definição, acabamos por determinar que $|x|'' = \text{sgn}' = 2\delta$, generalizando a noção da derivada ainda para além das funções L^2, introduzindo funcionais lineares em $\mathcal{D} = C_c^\infty(\mathbb{R})$. Esses funcionais lineares quando contínuos (na topologia adequada a $C_c^\infty(\mathbb{R})$) são denominados *distribuições* (que estão assim no dual$\mathcal{D}'$).

Exercício 9. Verificar que a derivada da função de Heaviside $H_y(x) = \begin{cases} 1 & (x \geq y) \\ 0 & (x < y) \end{cases}$ é o delta de Dirac δ_y.

Resolução: Para qualquer $v \in C_c^\infty(\mathbb{R})$

$$\langle H_y', v \rangle = -\langle H_y, v' \rangle = -\int_y^r v'(t)dt = -v(r) + v(y) = v(y) = \delta_y(v).$$

Observação 27. Quando trabalhamos com funções descontínuas, neste contexto, o valor pontual tem um significado desprezável. Assim, por exemplo, a aplicação da fórmula de Barrow

$$\int_a^b f'(t)dt = f(b) - f(a)$$

deve ser entendida usando a noção de traço, já que o seu valor nos extremos é irrelevante na medida de Lebesgue, por se tratar de um conjunto de medida nula. A igualdade pode ser entendida no sentido limite, pela densidade das funções C^∞ em $L^1(a, b)$ podemos considerar os diversos valores da função ou das derivadas na fronteira enquanto limite dessas aproximações.

2.1.3 Espaços de Sobolev (em $\mathbb{R}$)

Definição 17. Definimos

$$H^m(a, b) = \{v \in L^2(a, b) : v^{(p)}(a, b) \in L^2(a, b)\}$$

[2]Não existe nenhuma função δ integrável que permita a igualdade para qualquer $v \in C_c^\infty(-r, r)$, podemos ver pela desigualdade de Cauchy-Schwarz que não há em L^2, pois isso implicaria $v(0) = 0$,

$$|v(0)| = |\int_{\mathbb{R}} \delta(x)v(x)dx| \leq |\int_{-\varepsilon}^{\varepsilon} \delta(x)v(x)dx| \leq ||\delta||_{L^2(-\varepsilon,\varepsilon)}||v||_{L^2(-\varepsilon,\varepsilon)} \to 0,$$

notando que $||v||^2_{L^2(-\varepsilon,\varepsilon)} = \int_{-\varepsilon}^{\varepsilon} v(x)^2dx = 2\varepsilon v(\xi) \to 0$, quando $\varepsilon \to 0$. Há vários exemplos de funções $v \in C_c^\infty(-r, r)$ que não são nulas em zero, por exemplo $v(x) = e^{(\rho^2 - x^2)^{-1}}$ para $|x| < \rho < r$.

trata-se de um espaço de Hilbert onde podemos definir o produto interno

$$\langle u, v\rangle_{H^m(a,b)} = \int_a^b u(t)v(t)dt + ... + \int_a^b u^{(m)}(t)v^{(m)}(t)dt,$$

a que se associa a norma $||u||_{H^m(a,b)} = \langle u, u\rangle^{1/2}_{H^m(a,b)}$.

Teorema 10. *As funções $H^1(a,b)$ são contínuas. A aplicação identidade de $C[a,b] \to H^1(a,b)$, é denominada injecção, e é compacta.*

Demonstração. Considere-se $f \in H^1(a,b)$, então $||f'||_{L^2(a,b)} = C < \infty$, usando a fórmula de Barrow(*)

$$|f(y) - f(x)| = |\int_x^y f'(t)dt| = |\langle f', 1\rangle_{L^2(x,y)}| \leq ||f'||_{L^2(x,y)}||1||_{L^2(x,y)} \leq C|y-x|^{1/2}$$

o que significa que f é uniformemente contínua em $[a,b]$.

(*) A utilização da fórmula de Barrow pode ser feita no sentido que explicamos de seguida. Consideremos $x, y \in (a,b)$, podemos escrever

$$\begin{aligned} f(y) - f(x) &= \delta_y(f) - \delta_x(f) = \langle f, H'_y - H'_x\rangle \\ &= \langle f', H_x - H_y\rangle_{L^2(a,b)} = \langle f', 1\rangle_{L^2(x,y)}. \end{aligned}$$

Para evitar deltas de Dirac podemos ainda alternativamente considerar funções teste $w_\varepsilon \in C^\infty(a,b)$ que aproximem $H_x - H_y$ quando $\varepsilon \to 0$, a relação é então obtida no limite, pois

$$|\langle f, w'_\varepsilon\rangle| = |\langle f', w_\varepsilon\rangle| \leq C||w_\varepsilon||_{L^2(a,b)}$$

por um lado $\langle f, w'_\varepsilon\rangle \to f(y) - f(x)$, e por outro, $||w_\varepsilon||_{L^2(a,b)} \to |y-x|^{1/2}$.

□

Observação 28. As funções H^1 são contínuas em dimensão 1, mas não é assim em dimensão superior. Sendo contínuas o seu valor nas extremidades do intervalo está bem definido enquanto limite, e por isso definimos adicionalmente um seu subespaço fechado:

$$H^1_0(a,b) = \{v \in H^1(a,b) : v(a) = v(b) = 0\},$$

esta noção pode ser estendida para dimensões superiores, usando a noção do traço.

O dual do espaço $H^1_0(a,b)$ designa-se $H^{-1}(a,b)$, sendo o espaço das formas lineares contínuas, munido da norma associada

$$||F||_{H^{-1}(a,b)} = ||F||_{\mathcal{L}(H^1_0(a,b),\mathbb{R})} = \sup_{v\neq 0} \frac{|F(v)|}{||v||_{H^1(a,b)}}.$$

Exercício 10. Escreva o problema de Sturm Liouville, um problema de valores na fronteira em equações diferenciais ordinárias de 2ª ordem:

$$\begin{cases} -(pu')' + qu = f, & \text{em } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

(em que $p, q \in C[0,1]$, são funções positivas), no sentido generalizado em que $u \in H^1_0(a,b)$.

Resolução: Consideramos $v \in C_c^\infty(\mathbb{R})$, e aplicamos as regras de derivação generalizada:

$$\begin{aligned} \langle -(pu')' + qu, v\rangle = \langle f, v\rangle \quad &\Leftrightarrow \quad \langle -(pu')', v\rangle + \langle qu, v\rangle = \langle f, v\rangle \\ &\Leftrightarrow \quad \langle pu', v'\rangle + \langle qu, v\rangle = \langle f, v\rangle \end{aligned}$$

desta forma a igualdade fica estabelecida mesmo para funções $u \in H^1(a,b)$, pois como p, q são contínuas, temos $pu', qu \in L^2(a,b)$. Quanto a f, notamos que basta estar definido $\langle f, v\rangle$ o que se verifica quando $f \in L^2(a,b)$.

Iremos ver, pelo teorema de representação de Riesz, que qualquer forma linear se pode identificar com uma função pelo produto interno, e por isso podemos considerar mesmo $f \in H^{-1}(a,b)$.

2.2 Teorema de Representação de Riesz

Teorema 11. *Seja $M \subseteq H$ subconjunto (não vazio) convexo e fechado num espaço de Hilbert H. Dado $y \notin M$, existe um e um só $x \in M$ que minimiza a distância*

$$||y - x|| = \inf_{z \in M} ||y - z|| = d,$$

e habitualmente escrevemos $x = Proj_M(y)$.

Demonstração. Consideramos uma sucessão minimizante (z_n) de elementos de M tal que

$$d_n = ||y - z_n|| \to d$$

aplicando a igualdade do paralelogramo, obtemos com $a = y - z_n$, $b = y - z_m$,

$$||\underbrace{\frac{(y - z_n) - (y - z_m)}{2}}_{y - \frac{1}{2}(z_m + z_n)}||^2 + ||\underbrace{\frac{(y - z_n) - (y - z_m)}{2}}_{\frac{1}{2}(z_m - z_n)}||^2) = \frac{1}{2}\left(||y - z_n||^2 + ||y - z_m||^2\right) = \tfrac{d_n^2 + d_m^2}{2}$$

e como $||y - \frac{z_m + z_n}{2}|| \geq d$ (porque $\frac{z_m + z_n}{2} \in M$), obtemos

$$||\tfrac{z_m - z_n}{2}||^2 = \tfrac{d_n^2 + d_m^2}{2} - ||y - \frac{z_m + z_n}{2}||^2 \leq \tfrac{d_n^2 + d_m^2}{2} - d^2 \longrightarrow 0$$

ou seja, a sucessão (z_n) é de Cauchy, logo converge no espaço de Hilbert H, e como M é fechado, o limite da sucessão ainda pertence a M. Portanto $z_n \to x \in M$.

Este valor x é denominado a projecção de y sobre M, escrevendo-se $x = \mathrm{Proj}_M(y)$.

Até aqui apenas demonstra a existência. Para demonstrar a unicidade recorremos a um lema:

Lema: A tese do Teorema é equivalente a $\exists^1_{\cdot} x \in M : \langle y - x, v - x\rangle \leq 0\, (\forall v \in M)$.

demonstração: ($\Rightarrow$) Seja $v \in M$, como M é convexo, a linha de pontos entre v e x também pertence a M, ou seja, para $t \in [0,1]$

$$w_t = x + t(v - x) \in M$$

cuja distância a y será maior que x (ponto de mínimo):

$$||y - x||^2 \leq ||y - w_t||^2 = ||y - x - t(v - x)||^2 = ||y - x||^2 - 2t\,\langle y - x, v - x\rangle + t^2||v - x||^2$$

portanto $2t\,\langle y - x, v - x\rangle \leq t^2||v - x||^2$ o que implica

$$\langle y - x, v - x\rangle \leq \frac{t}{2}||v - x||^2 \underset{t \to 0}{\longrightarrow} 0.$$

porque a desigualdade foi demonstrada para um v qualquer, e $t \in]0,1]$ (o caso $t = 0$ seria trivial).

($\Leftarrow$) Inversamente, para qualquer $v \in M$: $||y-x||^2 - ||y-v||^2 = \langle (y-v)+(v-x), (y-v)+(v-x)\rangle - \langle y-v, y-v\rangle = 2\langle y-x, v-x\rangle - ||x-v||^2 \leq 0$ (verificar a igualdade), e portanto $||y-x|| \leq ||y-v||$. □

Resta demonstrar a unicidade, o que fazemos com base no lema. Supondo haver dois pontos de mínimo x_1, $x_2 \in M$ teríamos:

$\langle y - x_1, v - x_1\rangle \leq 0 \wedge v = x_2 \in M \implies \langle y - x_1, x_2 - x_1\rangle \leq 0$

$\langle y - x_2, v - x_2\rangle \leq 0 \wedge v = x_1 \in M \implies \langle y - x_2, x_1 - x_2\rangle \leq 0$, e somando obtemos $\langle x_1 - x_2, x_1 - x_2\rangle \leq 0$, o que implica $x_1 = x_2$. □

Corolário do Lema

Corolário 4. *Seja $S \subseteq H$ subespaço fechado num espaço de Hilbert H. Dado $y \notin S$, o único $x \in S$ que minimiza a distância de y a S é dado pela condição*

$$\langle y - x, v\rangle = 0,\ \forall v \in S,$$

ou ainda $\langle y - Proj_S(y), v\rangle = 0$, $\forall v \in S$.

Demonstração. O subespaço vectorial é convexo, logo pelo teorema anterior, existe x o ponto de mínimo, e temos pelo Lema: $\langle y - x, w - x\rangle \leq 0$, pelo que consideramos $w = v + x \in M$, obtendo $\langle y - x, v\rangle \leq 0$ e da mesma maneira, tomando $w = x - v \in M$, retiramos $\langle y - x, -v\rangle \leq 0$, ou seja, juntando ambas, $0 \leq \langle y - x, v\rangle \leq 0$. □

Proposição 7. *Seja M subespaço vectorial de um espaço de Hilbert H, então $H = M \oplus M^\perp$.*

Demonstração. (Exercício). Dado $y \in H$, consideramos $x = \mathrm{Proj}_M(y) \in M$ (que é único), logo pelo corolário $\langle y - x, v\rangle = 0$, $\forall v \in M$, e assim $y - x \in M^\perp$. Ou seja, estabelecemos a soma directa

$$y = \underbrace{\mathrm{Proj}_M(y)}_{\in M} + \underbrace{y - \mathrm{Proj}_M(y)}_{\in M^\perp}.$$

□

Proposição 8. *Seja M subespaço vectorial de um espaço de Hilbert H, então $(M^\perp)^\perp = \bar{M}$. Em particular, se $M^\perp = \{0\}$, como $\{0\}^\perp = H$, temos $\bar{M} = H$. Este resultado permite estabelecer que um subespaço vectorial é denso no espaço todo, verificando a ortogonalidade de cada elemento.*

Relembramos que um funcional é um operador que transforma elementos de um espaço de Hilbert em valores no seu corpo de escalares (consideramos normalmente ser $\mathbb{R}$), quando se trata de um funcional linear contínuo $F \in \mathcal{L}(H, \mathbb{R}) = H'$ (H' designa-se dual topológico de H), tem associado a norma habitual

$$||F||_{H'} = \sup_{v \neq 0} \frac{|F(v)|}{||v||_H}$$

relembrando que como $|F(v) - F(w)| \leq ||F||_{H'}||v - w||_H$ se for limitado, essa limitação da norma implica a continuidade.

Teorema 12. *(de Representação de Riesz). Seja H um espaço de Hilbert. Dado um funcional $F \in H'$, existe um e um só $f \in H$:*

$$F(v) = \langle f, v \rangle \quad \forall v \in H,$$

e esta correspondência é uma isometria, ou seja $||f||_H = ||F||_{H'}$.

Demonstração. Seja $M = \mathrm{Ker}(F) = F^{-1}(\{0\}) \subseteq H$, como F é contínuo M será fechado (pré-imagem do fechado $\{0\}$), e como F é linear M será um subespaço vectorial. No caso em que $M = \mathrm{Ker}(F) = H$ isto significaria $F \equiv 0$, e o resultado é trivial com $f = 0$. Caso contrário, dado $g \notin M$ (logo não é nulo), definimos

$$g^\perp = g - \mathrm{Proj}_M(g), \quad \hat{g}^\perp = \frac{g^\perp}{||g^\perp||}.$$

É claro que para qualquer $w \in M$, $\left\langle g^\perp, w \right\rangle = \langle g - \mathrm{Proj}_M(g), w \rangle = 0$, pelo corolário anterior, e portanto $g^\perp \in M^\perp$.

Dado qualquer $v \in H$, definimos $\lambda = F(v)/F(\hat{g}^\perp)$ e $w = v - \lambda \hat{g}^\perp \in M$, podemos escrever trivialmente

$$v = \lambda \hat{g}^\perp + w,$$

e como, $0 = \left\langle \hat{g}^\perp, w \right\rangle = \left\langle \hat{g}^\perp, v - \lambda \hat{g}^\perp \right\rangle = \left\langle \hat{g}^\perp, v \right\rangle - \lambda$ (porque $||\hat{g}^\perp|| = 1$), ou seja

$$\frac{F(v)}{F(\hat{g}^\perp)} = \lambda = \left\langle \hat{g}^\perp, v \right\rangle \Leftrightarrow F(v) = F(\hat{g}^\perp) \left\langle \hat{g}^\perp, v \right\rangle$$

e portanto definindo $f = F(\hat{g}^\perp)\hat{g}^\perp$ temos $F(v) = \langle f, v \rangle$ (notando que isto foi demonstrado para qualquer $v \in H$).

Finalmente, aplicando a desigualdade de Cauchy-Schwarz $| \langle f, v \rangle | \leq ||f||_H ||v||_H$

$$||F||_{H'} = \sup_{v \neq 0} \frac{|F(v)|}{||v||_H} = \sup_{v \neq 0} \frac{| \langle f, v \rangle |}{||v||_H} \leq ||f||_H$$

e por outro lado (tomando $v = f \in H$) temos

$$||f||_H = \frac{| \langle f, f \rangle |}{||f||_H} \leq \sup_{v \neq 0} \frac{| \langle f, v \rangle |}{||v||_H} = ||F||_{H'},$$

juntando as desigualdades $||f||_H \leq ||F||_{H'} \leq ||f||_H$ conclui-se a isometria. □

Exercício 11. Mostre que o problema de Sturm-Liouville

$$\begin{cases} -(pu')' + qu = f, & \text{em } (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

(em que $p, q \in C[0,1]$, são funções positivas), tem solução única em $u \in H_0^1(0,1)$, qualquer que seja $f \in H^{-1}(0,1)$.

2.2.1 Transformada de Fourier e soluções fundamentais

A transformada de Fourier contínua é definida em todo $\mathbb{R}$ como um integral

$$\mathcal{F}(f)(\xi) = \int_{\mathbb{R}} f(x)e^{-ix\xi}dx$$

e a sua inversa é dada de forma semelhante,

$$\mathcal{F}^{-1}(F)(x) = \frac{1}{2\pi}\int_{\mathbb{R}} F(\xi)e^{ix\xi}d\xi.$$

Verificando-se propriedades semelhantes para o produto de convolução contínuo,

$$\mathcal{F}(f * g) = \mathcal{F}(f)\mathcal{F}(g).$$

Proposição 9. $\mathcal{F}(\delta) = 1$, *e o delta de Dirac é o elemento neutro do produto de convolução.*

No caso de derivadas temos ainda a importante relação

$$\mathcal{F}(f^{(m)}) = (i\xi)^m\mathcal{F}(f),$$

o que permite retirar a seguinte propriedade da Transformada de Fourier para operadores diferenciais da forma

$$Du = a_0u + a_1u' + ... + a_mu^{(m)},$$

na forma da multiplicação por um polinómio semelhante:

$$\begin{aligned}\mathcal{F}(Du) &= p_D(i\xi)\mathcal{F}(u)\\ &= (a_0 + a_1(i\xi) + ... + a_m(i\xi)^m)\,\mathcal{F}(u).\end{aligned}$$

2.2.2 Solução Fundamental

Associada a um operador diferencial D, definido atrás, designamos que ϕ é uma *solução fundamental* de D se verificar

$$D\phi = \delta,$$

onde δ é o delta de Dirac.

Em particular já vimos que $\phi(x) = |x|/2$ é a solução fundamental de $Du = u''$.

Notamos que em termos da Transformada de Fourier isto significa $\mathcal{F}(D\phi) = 1 \Leftrightarrow p_D(i\xi)\mathcal{F}(\phi) = 1$. Portanto uma solução fundamental pode ser obtida pela inversão

$$\phi = \mathcal{F}(p_D(i\xi)^{-1}),$$

quando estiver definida.

Quando pretendemos apresentar uma solução de uma equação diferencial, com o termo não homogéneo:

$$Du = f$$

basta fazer a convolução com a solução fundamental, ou seja

$$u = \phi * f,$$

pois $Du = f$ implica

$$\begin{aligned} \mathcal{F}(Du) &= \mathcal{F}(f) \implies p_D\mathcal{F}(u) = \mathcal{F}(f) \implies \underbrace{p_D\mathcal{F}(\phi)}_{=1}\mathcal{F}(u) = \mathcal{F}(\phi)\mathcal{F}(f) \\ &\implies \mathcal{F}(u) = \mathcal{F}(\phi * f). \end{aligned}$$

Capítulo 3

Optimização não linear sem restrições

3.1 Noções básicas e resultados

3.1.1 Aspectos gerais

Estamos interessados na minimização de um funcional

$$f : H \to \mathbb{R}$$

onde H é espaço de Hilbert (ou de Banach).

Focaremos o problema de minimização, porque o problema de maximização é o mesmo, substituindo f por $-f$.

No caso de minimização sem restrições, o mínimo é procurado em todo o espaço funcional H, mas começamos por definir as noções fundamentais para um conjunto $\Omega \subseteq H$. Estas noções são em tudo semelhantes às existentes quando $H = \mathbb{R}^N$.

Definição 18. Dizemos que $z \in H$ é um *ponto de mínimo absoluto* de f em $\Omega \subseteq H$ se

$$f(z) \leq f(y), \quad \forall y \in \Omega$$

(dizemos ainda ser *estrito* se $f(z) < f(y)$, $\forall y \in \Omega$, $y \neq z$). Neste caso escreve-se

$$z = \arg\min_{y \in \Omega} f(y), \quad \left(\text{quando } f(z) = \min_{y \in \Omega} f(y) \right).$$

Por outro lado, dizemos que $z \in H$ é um *ponto de mínimo relativo* de f se existir uma vizinhança $V_z \ni z$, tal que

$$f(z) \leq f(y), \quad \forall y \in V_z \subset H$$

(da mesma forma, dizemos ser *estrito* se $f(z) < f(y)$, $\forall y \in V_z$, $y \neq z$).

Observação 29. Recordamos a expansão generalizada de Taylor em temos das derivadas de Fréchet em x

$$f(x+h) = f(x) + f'_x(h) + \tfrac{1}{2} f''_x(h,h) + o(||h||^2), \quad \text{quando } ||h|| \to 0, \tag{3.1}$$

onde $f'_x \in \mathcal{L}(H, \mathbb{R})$ é a derivada de Fréchet, uma forma linear, e onde $f''_x \in \mathcal{B}(H \times H, \mathbb{R})$ é a segunda derivada de Fréchet, uma forma bilinear.

Quando $H = \mathbb{R}^N$ temos $f'_x(h) = h^\top \nabla f(x)$ (a derivada é o gradiente) e $f''_x(h, h) = h^\top \nabla^2 f(x) h$ (a segunda derivada é a matriz Hessiana $\nabla^2 f(x)$), ficando

$$f(x + h) = f(x) + h^\top \nabla f(x) + \tfrac{1}{2} h^\top \nabla^2 f(x) h + o(||h||^2)$$

Definição 19. Dizemos que z é um *ponto crítico* de f se $f'_z \equiv 0$.

Definição 20. Dizemos que a forma bilinear b é *semidefinida positiva* se

$$b(h, h) \geq 0, \quad \forall h \in H.$$

e dizemos que é *definida positiva* se $b(h, h) > 0$ para $h \neq 0$.

Se existir $\alpha > 0$ tal que $b(h, h) \geq \alpha ||h||^2$, diz-se que b é coerciva (todas as formas bilineares coercivas são definidas positivas).

Teorema 13. (condição *suficiente* para ponto de máximo relativo estrito)

Seja f duas vezes Fréchet-diferenciável em V_z vizinhança num ponto crítico z, tal que f''_z é uma forma bilinear definida positiva. Então x é um ponto de mínimo relativo estrito de f.

Demonstração. Como z é ponto crítico $f'_z \equiv 0$, e temos da expansão de Taylor (3.1), para $h \in V_z$,

$$f(z + h) - f(z) = \tfrac{1}{2} f''_z(h, h) + o(||h||^2).$$

Considerando $h = ||h|| \hat{h}$, temos para $h \neq 0$, quando $||h|| \to 0$,

$$\frac{f(z + h) - f(z)}{||h||^2} = \tfrac{1}{2} f''_z(\hat{h}, \hat{h}) + o(1) > 0,$$

porque f''_z é definida positiva, logo $f''_z(\hat{h}, \hat{h}) > 0$ e não depende de $||h||$ (pois $||\hat{h}|| = 1$). Note-se que $||h|| \to 0$ permite tornar $o(1)$ tão pequeno, não influenciando na soma o sinal positivo de $f''_z(\hat{h}, \hat{h})$.

Assim, $f(z + h) - f(z) > 0$, para todo $h \neq 0$, $z + h \in V_z$, sendo z um mínimo relativo estrito de f. □

Teorema 14. (condição *necessária* para ponto de mínimo relativo)

Se f é Fréchet diferenciável numa vizinhança V_z onde z é um ponto de mínimo relativo, então $f'_z \equiv 0$.

Se f for duas vezes Fréchet diferenciável em V_z então f''_z é ainda uma forma bilinear semidefinida positiva.

Demonstração. Como z é um ponto de mínimo relativo $f(z + h) - f(z) \geq 0$ para todo $z + h \in V_z$, e usando a expansão (3.1)

$$0 \leq \frac{f(z + h) - f(z)}{||h||} = f'_z(\hat{h}) + o(1) \longrightarrow f'_z(\hat{h}),$$

com $h = ||h||\hat{h}$. Isto significa $f'_z(\hat{h}) \geq 0$, mas também usando $-\hat{h}$, temos $f'_z(\hat{h}) \leq 0$, logo

$$f'_z(\hat{h}) = 0,\ (\forall \hat{h} : ||\hat{h}|| = 1) \implies f'_z(h) = 0\ (\forall h)$$

O outro resultado é análogo, porque como $f'_z(\hat{h}) = 0$, temos

$$0 \leq \frac{f(z+h) - f(z)}{||h||^2} = f''_z(\hat{h}, \hat{h}) + o(1) \longrightarrow f''_z(\hat{h}, \hat{h}),$$

o que implica $f''_z(\hat{h}, \hat{h}) \geq 0$ para qualquer h. □

3.1.2 Convexidade

Definição 21. Dizemos que f é *convexa* em $\Omega \subseteq H$ (Ω convexo) se

$$f((1-\theta)x + \theta y) \leq (1-\theta)f(x) + \theta f(y), \quad \forall x, y \in \Omega, \theta \in [0,1]$$

(e dizemos ser *estritamente convexa* se a desigualdade for estrita para $x \neq y$ com $\theta \in (0,1)$).

Exemplo 15. O exemplo mais simples, em $H = \mathbb{R}$, corresponde a funções $f''(x) \geq 0$.

Para mostrarmos isso, usamos a fórmula para $f \in C^2$ sendo $z = (1-\theta)x + \theta y$,

$$\begin{aligned} f(z) &= (1-\theta)f(x) + \theta f(y) - \frac{1}{2}f''(\xi)\underbrace{(1-\theta)\theta(y-x)^2}_{\geq 0} \quad (\xi \in [x;y]) \\ &\leq (1-\theta)f(x) + \theta f(y) \quad (\text{se e só se } f'' \geq 0) \end{aligned}$$

NOTA: A fórmula acima, resulta de considerar a aproximação do funcional $A(f) = f(z)$ com

$$Q(f) = (1-\theta)f(x) + \theta f(y)$$

o que é uma fórmula de grau 1 com erro $A(f) - Q(f) = A(f - p_1)$, ou seja,

$$A(f - p_1) = f(z) - p_1(z) = f[x, y, z](z - x)(z - y)$$

notando que $z - x = \theta(y - x),\ z - y = (1-\theta)(x - y)$, e portanto

$$A(f) - Q(f) = -\frac{1}{2}f''(\xi)(1-\theta)\theta(y-x)^2.$$

Observação 30. Num contexto mais geral, podemos ver que a Hessiana é semidefinida positiva se e só se f for convexa.

Proposição 10. *Se f for convexa em Ω, um ponto de mínimo relativo é um ponto de mínimo absoluto em Ω. Para além disso, se for estritamente convexa, haverá um único ponto de mínimo absoluto, estrito.*

Demonstração. Tomando qualquer $y \in \Omega$, se $z \in \Omega$ é um ponto de mínimo relativo, existe um $\theta > 0$ suficientemente pequeno tal que

$$\tilde{y} = (1-\theta)z + \theta y \in V_z$$

tendo-se $f(z) \leq f(\tilde{y})$. Por convexidade,

$$f(z) \leq f(\tilde{y}) \leq f((1-\theta)z + \theta y) \leq (1-\theta)f(z) + \theta f(y),$$

subtraindo, isto implica $\theta f(z) \leq \theta f(y)$, logo $f(z) \leq f(y)$. Por isso trata-se de um mínimo absoluto, global.

Por outro lado, se x, z fossem pontos de mínimo absoluto distintos, com $f(x) = f(z)$, então

$$f(\theta x + (1-\theta)z) < \theta f(z) + (1-\theta)f(z) = f(z),$$

mas $f(z)$ deveria ser mínimo, o que é contradição. □

Exemplo 16. A norma é uma função convexa, porque sendo $f(x) = ||x||$,

$$f((1-\theta)x + \theta y) = ||(1-\theta)x + \theta y|| \leq (1-\theta)||x|| + \theta||y|| = (1-\theta)f(x) + \theta f(y).$$

Exemplo 17. Seja

$$f(x) = c + a(x) + b(x, x),$$

onde a é uma forma linear, b é forma bilinear simétrica e semidefinida positiva.

Então f é convexa, porque

$$\begin{aligned} f((1-\theta)x + \theta y) &= c + a((1-\theta)x + \theta y) + b((1-\theta)x + \theta y, (1-\theta)x + \theta y) \\ &= (1-\theta)(c + a(x) + b(x, x)) + \theta(c + a(y) + b(y, y)) - \theta(1-\theta)b(x-y, x-y) \\ &= (1-\theta)f(x) + \theta f(y) - \theta(1-\theta)b(x-y, x-y) \\ &\leq (1-\theta)f(x) + \theta f(y) \end{aligned}$$

uma vez que $\theta(1-\theta)b(x-y, x-y) \geq 0$.

3.2 Equações com pontos críticos

Uma consequência dos resultados anteriores é que para um funcional estritamente convexo, o seu ponto de mínimo absoluto pode ser encontrado resolvendo a equação com a derivada

$$f'_x = 0.$$

Numa situação geral para funcionais regulares (F-diferenciáveis), isto é também uma condição necessária. Por isso, uma estratégia para resolver o problema de minimização será procurar os valores x que tenham derivada de Fréchet nula.

No caso geral isso corresponde a resolver

$$z \in \Omega : \quad f'_z(h) = 0, \quad \forall h \in H.$$

3.2.1 Exemplo - Mínimos Quadrados

O problema de mínimos quadrados consiste na determinação da melhor aproximação $x \in S \subset H$ (onde S é um subespaço de dimensão finita do espaço de Hilbert H), que minimiza a distância a uma certa norma $\phi \in H$. Isto corresponde a minimizar $||x - \phi||$, ou o seu quadrado, o funcional $f : S \subset H \to \mathbb{R}$

$$f(x) = ||x - \phi||^2 = \langle x - \phi, x - \phi \rangle .$$

Neste caso f é convexa, pois é uma forma quadrática (considere $f(x) = \langle x - \phi, x - \phi \rangle$ no Exemplo17), por resultar da norma.

A derivada de Fréchet é $f'_x(h) = 2 \langle x - \phi, h \rangle$, (exercício), e assim

$$f'_x(h) = 0 \Leftrightarrow \langle x - \phi, h \rangle = 0$$

para todo o $h \in S$, e sendo $S = \langle g_1, \cdots, g_N \rangle$, ao escrever $x = \sum_{j=1}^{N} x_j g_j$ isto leva ao sistema normal

$$\langle x - \phi, g_k \rangle = 0 \Leftrightarrow \sum_{j=1}^{N} x_j \langle g_j, g_k \rangle = \langle \phi, g_k \rangle .$$

Este é o problema de mínimos quadrados, no caso linear. Contudo a dependência nos coeficientes desconhecidos x_j pode ser não linear e a derivada de Fréchet pode não ser tão simples.

3.2.2 Exemplo - dimensão finita

Quando $H = \mathbb{R}^N$ isto resume-se a verificar a igualdade para a base canónica $h = \mathbf{e}_1, \ldots, \mathbf{e}_N$, porque a dimensão é finita.

Em $\mathbb{R}^N$ temos $f'_x(h) = h^\top \nabla f(x)$, e isso corresponde a resolver $\mathbf{e}_k^\top \nabla f(x) = 0$, para cada $k = 1, \ldots, N$.

Abreviadamente, isto corresponde a resolver o sistema

$$\nabla f(x) = 0.$$

Este sistema pode ser resolvido usando os métodos habituais: Newton, Broyden, ou iteração do ponto fixo.

Por exemplo, para o Método de Newton, começando com $x^{(0)}$, encontramos $x^{(1)} = x^{(0)} + h$, resolvendo o sistema linear

$$[\nabla^2 f(x^{(k)})]h = -\nabla f(x^{(k)})$$

Esta abordagem terminaria por aqui o assunto, já que o enquadraríamos no contexto da resolução de equações.

No entanto, esta não é a melhor abordagem. O método de Newton poderá ser usado para encontrar a direcção de descida, mas ao longo dessa direcção não é claro que o valor dado pelo método habitual seja a melhor escolha.

Os métodos de optimização são assim diferentes dos métodos de resolver equações, ainda que possam ser relacionáveis, e notamos ainda que, no caso em que $f(x) = 0$ tenha solução,

- minimizar $||f(x)||$ ou $||f(x)||^2$ equivale a encontrar a raiz, pois o mínimo será zero, e o ponto de mínimo verifica

$$||f(x)|| = 0 \Leftrightarrow f(x) = 0.$$

3.3 Limitação computacional na optimização global

No caso de funções estritamente convexas vimos que o mínimo absoluto é único, mas quando a função f é suficientemente geral, pode ser computacionalmente impossível encontrar o ponto de mínimo.

Para simplificarmos a abordagem, tomemos o caso unidimensional, em que queremos apenas encontrar

$$f(z) = \min_{y \in [a,b]} f(y)$$

Para esse efeito, consideramos um algoritmo geral $\mathcal{A}$ que devolve a iterada x_{k+1} baseado na função e num conjunto de pontos anteriores.

$$x_{k+1} = \mathcal{A}(f, x_k, \cdots, x_0),$$

e este algoritmo pode incluir informação das derivadas, $f^{(m)}(x_j)$, também.

Apenas excluímos aqui algoritmos triviais que produzem uma sucessão densa de pontos, para varrer o intervalo, e que são ineficazes. São ineficazes, porque por exemplo por bissecção sucessiva, isso implicaria considerar $1 + 2^M$ cálculos para avaliar o intervalo $[0, 1]$ com espaçamento 2^{-M}. Assim, para uma inspecção com erro inferior a 0.001 seriam precisos 1000 cálculos de f, o que é extremamente ineficaz se o cálculo de f for moroso.

Observação 31. De qualquer forma, os algoritmos triviais podem ser usados para uma função genérica f que tenha um cálculo rápido. Por exemplo, se for possível computar um milhão de vezes f em menos de um segundo, então a simples avaliação de $f(x_n)$ no intervalo $[0, 1]$ com

$$x_n = n\, 10^{-6} \quad (n = 0, \ldots, N = 10^6)$$

poderá permitir encontrar o ponto de mínimo com erro inferior a 10^{-6}. Apesar de serem altamente ineficazes, este algoritmos não devem deixar de ser considerados, quando pouco se sabe de f.

No entanto, mesmo este procedimento não garante que

$$f(x_m) = \min_{k=0}^{N} f(x_k)$$

seja uma boa aproximação de $\min_{a \le x \le b} f(x)$ porque esse mínimo pode estar longe do x_m obtido, especialmente quando f tem um grande comportamento oscilatório e a lista de pontos é pequena.

De qualquer forma, o algoritmo de construir uma lista densa de pontos converge, desde que a função seja contínua. Simplesmente é demasiado ineficaz, e só serve normalmente para detectar um intervalo onde se aplica outro método mais eficaz.

Teorema 15. *Não há nenhum algoritmo eficaz $\mathcal{A}$ que permita obter o mínimo de f para qualquer $f \in C^\infty[a,b]$.*

Demonstração. Considere a sucessão de pontos (x_n) dada pelo algoritmo $\mathcal{A}$ e assuma que $x_n \to x$, onde $x \in [a,b]$ é um mínimo absoluto de uma função $f \in C^\infty[a,b]$. Então podemos considerar uma função $\tilde{f}$ que coincide com f em todos os pontos da sucessão (x_n) até às derivadas de ordem m, mas que tem um ponto de mínimo absoluto em $\tilde{x} \neq x$.

Esta função $\tilde{f}$ pode ser facilmente considerada diferente em $[\tilde{x}-\epsilon, \tilde{x}+\epsilon] \subset [a,b]$ tal que $x_n \notin [\tilde{x}-\epsilon, \tilde{x}+\epsilon]$ (porque o algoritmo é eficaz e não gera um conjunto denso em $[a,b]$). Por isso como $\tilde{f} = f$ excepto no intervalo $(\tilde{x}-\epsilon, \tilde{x}+\epsilon)$, tomando $\tilde{f}(\tilde{x}) < f(x)$, o algortimo produzirá os mesmos resultados, ignorado o intervalo $(\tilde{x}-\epsilon, \tilde{x}+\epsilon)$, e cairá no ponto de mínimo de f que não o é de $\tilde{f}$. Ou seja, como $x_n \notin [\tilde{x}-\epsilon, \tilde{x}+\epsilon]$, a sucessão

$$x_{k+1} = \mathcal{A}(\tilde{f}, x_k, \cdots, x_0) = \mathcal{A}(f, x_k, \cdots, x_0)$$

converge para x, que não é $\tilde{x}$, ponto de mínimo para a função $\tilde{f}$. □

Observação 32. Exclui-se da situação anterior o caso em que a função é convexa, porque não haveria dois mínimos relativos, e exclui-se também o caso em que a função é analítica. No caso em que a função é analítica, a sucessão de pontos (x_n) definiria a função de forma única, e não seria possível construir o contraexemplo.

3.4 Problemas de optimização unidimensional

Para além do interesse próprio, é importante o problema de minimização em $\mathbb{R}$ ou num intervalo, pois a maioria dos métodos irá usar a pesquisa linear (*line search*), o que significa encontrar um mínimo na linha dada por uma certa direcção, o que se trata de um problema unidimensional.

É claro que poderíamos considerar vários métodos no caso unidimensional, por exemplo procurando os pontos críticos,

$$x : f'(x) = 0,$$

por um método habitual (bissecção, secante, ponto fixo ou Newton), mas iremos considerar aqui a Pesquisa pela Secção de Ouro, e a Aproximação Quadrática, que evitam esse cálculo.

3.4.1 Pesquisa seccional

A ideia destes métodos é semelhante à do método da bissecção. Assumimos que a função contínua f tem um único mínimo relativo estrito $z \in (a,b)$, o que pode ser garantido reduzindo o intervalo suficientemente, até que a função seja aí estritamente convexa.

Construímos uma sucessão que converge para esse ponto de mínimo absoluto z.

Algoritmos Seccionais

Consideramos $a_0 = a$, $b_0 = b$, e um qualquer $c_0 = c \in (a,b)$ que terá menor valor (porque os extremos não são o mínimo).

Definimos o tripleto (a_k, c_k, b_k), onde o ponto de mínimo deste conjunto está em c_k.

A iteração consiste em tomar um novo $d_k \in (a_k, b_k) \backslash \{c_k\}$ e testá-lo:

- Caso $c_k < d_k$.
  - Se $f(d_k) < f(c_k)$ então $(a_{k+1}, c_{k+1}, b_{k+1}) = (c_k, d_k, b_k)$, senão $(a_{k+1}, c_{k+1}, b_{k+1}) = (a_k, c_k, d_k)$,
- Caso $d_k < c_k$.
  - Se $f(d_k) < f(c_k)$ então $(a_{k+1}, c_{k+1}, b_{k+1}) = (a_k, d_k, c_k)$, senão $(a_{k+1}, c_{k+1}, b_{k+1}) = (d_k, c_k, b_k)$.

Isto significa que ou d_k é um novo ponto de mínimo e fica no meio do tripleto, ou então passa a ser fronteira do novo intervalo. Em qualquer caso, reduzimos a dimensão do intervalo, o que permite provar a convergência.

Proposição 11. *A sucessão (c_k) converge para o único ponto de mínimo estrito $z \in (a, b)$.*

Demonstração. Basta ver que $z \in (a_k, b_k)$, pois por construção $c_k \in (a_k, b_k)$ e $|a_k - b_k| \to 0$, portanto $z - c_k \to 0$.

Com efeito, se $z \notin (a_k, b_k)$ há também um ponto de mínimo relativo em (a_k, b_k), pois $f(c_k)$ é menor, e isso contradiz a hipótese de haver apenas um ponto de mínimo relativo. □

Observação 33. A escolha de d_k poderia ser qualquer, uma hipótese poderia ser usar a bissecção. Por exemplo, no intervalo $[0, 1]$ começávamos com o tripleto $(0, \frac{1}{2}, 1)$, e definindo $d_0 = \frac{1}{4}$ poderíamos ter a sorte de $f(\frac{1}{4}) < f(\frac{1}{2})$ com um novo tripleto $(0, \frac{1}{4}, \frac{1}{2})$ num intervalo de comprimento 0.5, mas caso contrário o tripleto seria $(\frac{1}{4}, \frac{1}{2}, 1)$ com um comprimento 0.75. A questão de considerar uma escolha que mantenha o comprimento dos intervalos em ambas as circunstâncias é respondida pelo número de ouro!

Algoritmo da Secção de Ouro *(Golden section search)*

Para a melhor opção na secção convém que o comprimento do novo intervalo não dependa do resultado do teste.

Para simplificar suponhamos que $[a, b] = [0, 1]$, e que $c > d$.

Então o novo intervalo tem comprimento c ou $1-d$, e queremos que sejam iguais $c = 1-d$, e se mantenha a proporção $\frac{c}{1} = \frac{d}{c}$, ou seja:

$$c = \frac{d}{c} = \frac{1-c}{c} \implies \frac{1-c}{c} = c \Leftrightarrow c^2 + c - 1 = 0$$

pelo que se obtém o número de ouro

$$c = \frac{-1+\sqrt{5}}{2} = \rho = 0.618034...$$

Por isso, iremos considerar se $d_k > c_k$

$$c_k = a_k + (1-\rho)(b_k - a_k), \quad d_k = a_k + \rho(b_k - a_k),$$

Com esta escolha, como $b_k - c_k = \rho(b_k - a_k) = d_k - a_k$, temos

$$|a_{k+1} - b_{k+1}| = \rho|a_k - b_k| \quad \implies \quad |a_n - b_n| = \rho^n |a - b|$$

ou seja, uma convergência linear com coeficiente assimptótico ρ.

3.4.2 Aproximação Quadrática

Uma outra estratégia consiste em considerar que f se comporta como uma função quadrática próximo do mínimo, e aproximar por interpolação em três pontos calculados antes

$$(x_{n-2}, f(x_{n-2})),\ (x_{n-1}, f(x_{n-1})),\ (x_n, f(x_n))$$

o polinómio interpolador fica

$$p_2(x) = \frac{x-x_{n-1}}{x_{n-2}-x_{n-1}}\frac{x-x_n}{x_{n-2}-x_n}f(x_{n-2}) + \frac{x-x_{n-2}}{x_{n-1}-x_{n-2}}\frac{x-x_n}{x_{n-1}-x_n}f(x_{n-1}) + \frac{x-x_{n-2}}{x_{n-1}-x_{n-2}}\frac{x-x_{n-1}}{x_n-x_{n-1}}f(x_n).$$

Resolvendo $p_2'(x_{n+1}) = 0$, obtemos

$$x_{n+1} = \frac{1}{2}\frac{f(x_n)(x_{n-1}^2 - x_{n-2}^2) + f(x_{n-1})(x_{n-2}^2 - x_n^2) + f(x_{n-2})(x_n^2 - x_{n-1}^2)}{f(x_n)(x_{n-1} - x_{n-2}) + f(x_{n-1})(x_{n-2} - x_n) + f(x_{n-2})(x_n - x_{n-1})}$$

que é o ponto de mínimo de p_2 e que servirá para aproximar o ponto de mínimo de f.

De certa forma é semelhante ao método da secante para f', mas evita o cálculo das derivadas, e tal como o método da secante apresenta uma convergência supralinear. Também há formas de circunscrever as iterações a um intervalo, como é o caso do método de Brent, por adaptação do Método *Regula Falsi*.

3.5 Métodos de Descida

Vamos agora considerar uma grande classe de métodos denominados *métodos de descida.* De um modo geral, podemos considerar um método dado por um algoritmo com uma função $A : \Omega \to \Omega$, e um procedimento iterativo,

$$x_{n+1} = A(x_n)$$

começando com um x_0 inicial. Este é um procedimento semelhante à iteracão do ponto fixo, mas agora associando uma função de descida $Z : \Omega \to \mathbb{R}$, tal que

$$Z(A(x)) < Z(x), \quad x \neq z,$$

onde z são pontos de mínimo estritos em Ω. A função de descida pode coincidir com f, e no caso dos pontos de mínimo de Z são os mesmos de f. Uma outra possibilidade será considerar $Z = |\nabla f|^2$, mas nesse caso os pontos de mínimo de Z serão os pontos críticos de f (notando que o ponto de mínimo coincide com o ponto crítico quando f é convexa).

Teorema 16. *(*convergência global). *Se A é contínua e Ω compacto, então a sucessão (x_n) dada pelo método de descida, tem subsucessões que convergem para pontos de mínimo estritos.*

Demonstração. Como (x_n) é uma sucessão num compacto Ω então podemos extrair subsucessões convergentes.

Sendo (x_{n_k}) uma dessas subsucessões temos $x_{n_k} \to x$, e como A é contínua,

$$x_{n_k} = A(x_{n_k-1}) \to A(x),$$

e por unicidade do limite, $A(x) = x$. Portanto $Z(A(x)) = Z(x)$, e x será um ponto de mínimo estrito, pois $Z(A(x)) < Z(x)$ nos restantes casos. □

Exemplo 18. The golden ratio search is a descent method. Consider $Z = f$, and we may consider A to be the algorithm that constructs the sequence (c_n), since $f(c_{n+1}) = f(A(c_n)) < f(c_n)$ if $c_n \neq z$, where z is the unique minimum point in $X = [a, b]$.

Definição 22. *Algoritmos de Pesquisa Linear.* Considere-se a iteração

$$x_{n+1} = x_n + \omega_n d_n,$$

onde d_n é denominada *direcção de descida.* Escolhendo $\omega_n > 0$ para aproximar o ponto de mínimo de

$$g(\omega) = f(x_n + \omega d_n)$$

temos um método de descida, baseado numa *pesquisa linear* ao longo da semi-recta

$$S(x_n, d_n) = \{x_n + \omega d_n : \omega \geq 0\}.$$

A pesquisa é *exacta* se procurarmos que ω_n seja o ponto de mínimo em $S(x_n, d_n)$, caso contrário, diz-se *inexacta.*

Observação 34. Estes algoritmos de pesquisa linear são métodos de descida (com $Z = f$), porque

$$f(x_{n+1}) = f(x_n + \omega_n d_n) = g(\omega_n) < g(0) = f(x_n).$$

A única condição que precisamos para garantir uma descida é que o ω_n escolhido verifique $g(\omega_n) < g(0)$. Notamos que a determinação exacta do mínimo pode ser ineficaz computacionalmente.

Proposição 12. *Para uma função Fréchet-diferenciável, se $f'_{x_n}(d_n) < 0$ então d_n é uma direcção de descida.*

Demonstração. Basta notar que pela expansão de Taylor

$$f(x_{n+1}) = f(x_n + \omega d_n) = f(x_n) + \omega f'_{x_n}(d_n) + o(||\omega d_n||)$$

e assim, como $\omega > 0$, quando $\omega \to 0$

$$\frac{f(x_{n+1}) - f(x_n)}{\omega} = f'_{x_n}(d_n) + o(||d_n||) < 0$$

o que significa que $f(x_{n+1}) < f(x_n)$. □

Observação 35. No que se segue iremos considerar apenas o caso $H = \mathbb{R}^N$, e portanto, exigimos apenas que

$$\nabla f(x_n)^\top d_n < 0.$$

3.5.1 Método do Gradiente

O exemplo mais conhecido de método de descida é o Método do Gradiente ou do Máximo Declive (*steepest descent*), que corresponde a escolher a direcção do gradiente como descida

$$d_n = -\nabla f(x_n),$$

fazendo a pesquisa linear nessa direcção.

Proposição 13. *O Método do Gradiente é um método de descida.*

Demonstração. Basta notar que $\nabla f(x_n)^\top d_n = -\nabla f(x_n)^\top \nabla f(x_n) = -||\nabla f(x_n)||^2 < 0$, até que $\nabla f(x_n) = 0$, sendo aí ponto crítico e x_n seria o mínimo. □

Exemplo 19. Considere a função $f(x) = e^{2x_1} + e^{-x_1-x_2} + 4x_2$, onde temos

$$\nabla f(x) = (2e^{2x_1} - e^{-x_1-x_2}, -e^{-x_1-x_2} + 4).$$

Neste caso é fácil determinar os pontos críticos porque de $\nabla f(x) = (0,0)$ temos

$$\begin{cases} 2e^{2x_1} - e^{-x_1-x_2} &= 0 \\ 4 - e^{-x_1-x_2} &= 0 \end{cases} \begin{cases} 2e^{2x_1} &= 4 \\ e^{-x_1-x_2} &= 4 \end{cases} \begin{cases} x_1 &= \frac{1}{2}\log 2 \\ -\frac{1}{2}\log 2 - x_2 &= 2\log 2 \end{cases} \begin{cases} x_1 &= \frac{1}{2}\log 2 \\ x_2 &= -\frac{3}{2}\log 2 \end{cases}$$

Este $x = \frac{1}{2}(1,-3)\log 2$ é um ponto de mínimo estrito, conforme podemos ver pela matriz Hessiana

$$\nabla f(x) = \begin{bmatrix} 4e^{2x_1} + e^{-x_1-x_2} & e^{-x_1-x_2} \\ e^{-x_1-x_2} & e^{-x_1-x_2} \end{bmatrix} \underbrace{=}_{x=\frac{1}{2}(1,-3)\log 2} \begin{bmatrix} 12 & 4 \\ 4 & 4 \end{bmatrix}$$

que é definida positiva.

Consideremos agora o Método do Gradiente com pesquisa exacta começando com $x^{(0)} = (0,0)$. Então

$$x^{(1)} = x^{(0)} - \omega_0 \nabla f(x^{(0)}) = -\omega_0(1,3)$$

e ω_0 minimiza $g(\omega) = f(-\omega, -3\omega) = e^{-2\omega} + e^{4\omega} - 12\omega$. Como

$$g'(\omega) = -2e^{-2\omega} + 4e^{4\omega} - 12 = 0$$

é equivalente a $4\zeta^2 - 2/\zeta - 12 = 0$, com $\zeta = e^{-2\omega}$.

3.5.2 Método de Newton

Consideramos agora um novo desenvolvimento de Taylor para o método de Newton

$$f(x^{(n+1)}) \;=\; f(x^{(n)}) + \omega d^{(n)} \cdot \nabla f(x^{(n)}) + \frac{\omega^2}{2}(d^{(n)})^\top \nabla^2 f(x^{(n)}) d^{(n)} + o(\omega^2)$$

e tomamos a direcção do Método de Newton

$$d^{(n)} = -[\nabla^2 f(x^{(n)})]^{-1} \nabla f(x^{(n)}),$$

o que é equivalente a resolver o sistema linear

$$[\nabla^2 f(x^{(n)})]d^{(n)} = -\nabla f(x^{(n)}).$$

Isto leva à seguinte igualdade ($\omega > 0$),

$$\begin{aligned}
\frac{f(x^{(n+1)}) - f(x^{(n)})}{\omega} &= d^{(n)} \cdot \nabla f(x^{(n)}) + \frac{\omega}{2}(d^{(n)})^\top \nabla^2 f(x^{(n)}) d^{(n)} + o(\omega) \\
&= -\nabla f(x^{(n)})^\top ([\nabla^2 f(x^{(n)})]^{-1})^\top \nabla f(x^{(n)}) \\
&\quad -\frac{\omega}{2}(d^{(n)})^\top \nabla^2 f(x^{(n)})[\nabla^2 f(x^{(n)})]^{-1} \nabla f(x^{(n)}) + o(\omega) \\
&= -\nabla f(x^{(n)})^\top ([\nabla^2 f(x^{(n)})]^{-1})^\top \nabla f(x^{(n)}) \\
&\quad +\frac{\omega}{2}\nabla f(x^{(n)})^\top ([\nabla^2 f(x^{(n)})]^{-1})^\top \nabla f(x^{(n)}) + o(\omega) \\
&= (-1 + \frac{\omega}{2})\nabla f(x^{(n)})^\top ([\nabla^2 f(x^{(n)})]^{-1})^\top \nabla f(x^{(n)}) + o(\omega)
\end{aligned}$$

Quando a Hessiana é uma matriz definida positiva, também é a sua inversa e temos

$$\nabla f(x^{(n)})^\top ([\nabla^2 f(x^{(n)})]^{-1})^\top \nabla f(x^{(n)}) > 0,$$

quando $x^{(n)}$ não é um ponto crítico. Assim, quando $0 < \omega < 2$, suficientemente pequeno, obtemos um método de descida.

De novo podemos usar um algoritmo de pesquisa linear para obter $\omega_n > 0$, que minimize

$$g(\omega) = f(x^{(n)} + \omega d^{(n)}),$$

ou pelo menos, tal que $f(x^{(n+1)} = g(\omega_n) < g(0) = f(x^{(n)})$, para garantir a descida no caso da pesquisa linear inexacta.

Proposição 14. *Para $f \in C^2$, o método de Newton com pesquisa linear tem pelo menos convergência quadrática.*

Demonstração. Trata-se de uma simples consequência da convergência quadrática do Método de Newton clássico, quando aplicado a $\nabla f(x) = 0$. A iteração de Newton clássica é

$$x^{(n+1)} = x^{(n)} - [\nabla^2 f(x^{(n)})]^{-1} \nabla f(x^{(n)}),$$

o que corresponde a tomar $\omega_n = 1$. Por isso se considerarmos ω_n tal que $g(\omega_n) < \min\{g(0), g(1)\}$, temos um método de descida melhor que o clássico que tinha já convergência quadrática. □

No Método de Newton para garantir a descida precisamos que a matriz Hessiana seja definida positiva nos pontos de cálculo. Se isto é esperado perto do ponto de mínimo estrito, pode não ser verdade em pontos mais distantes, por isso vamos considerar uma variante, que corresponde ao Método de Levenberg-Marquardt.

3.5.3 Método de Levenberg-Marquardt

No Método de Levenberg-Marquardt consideramos um parâmetro $\varepsilon_n \geq 0$ e a direcção

$$d^{(n)} = -[\varepsilon_n I + \nabla^2 f(x^{(n)})]^{-1} \nabla f(x^{(n)}).$$

Notamos que se $\varepsilon_n = 0$ é o Método de Newton, e quando $\varepsilon_n \to \infty$ comporta-se como o Método do Gradiente, porque

$$\varepsilon_n d^{(n)} = -[I + \frac{1}{\varepsilon_n} \nabla^2 f(x^{(n)})]^{-1} \nabla f(x^{(n)}) \approx -\nabla f(x^{(n)}).$$

Por isso, o Método de Levenberg-Marquardt pode ser visto como uma combinação convexa dos métodos do gradiente e de Newton. Será melhor considerar $\varepsilon_n = 0$ quando estamos próximos do ponto de mínimo, para beneficiar da convergência quadrática do Método de Newton, caso contrário é melhor tomar um $\varepsilon_n > 0$ tal que a matriz

$$M = \varepsilon_n I + \nabla^2 f(x^{(n)})$$

seja definida positiva.

Para esse efeito podemos efectuar um decomposição de Cholesky da matriz $M = LL^\top$, e usá-la para resolver o sistema

$$[\varepsilon_n I + \nabla^2 f(x^{(n)})] d^{(n)} = -\nabla f(x^{(n)}) \iff LL^\top d^{(n)} = -\nabla f(x^{(n)}).$$

Observação 36. Uma outra possibilidade é considerar a diagonal da matriz hessiana ao invés de I.

Exercício 12. Mostre que o Método de Levenberg-Marquardt é também um método de descida.

3.6 Pesquisa Linear Inexacta

É computacionalmente ineficaz efectuar uma pesquisa linear exacta em cada passo da iteração, já que só acidentalmente o mínimo iria estar ao longo dessa direcção. Em vez disso, podemos prosseguir usando uma pesquisa não exacta, procurando apenas aproveitar a descida nessa direcção.

3.6.1 Regra de Armijo

Retomando a função $g(\omega) = f(x^{(k)} + \omega d^{(k)})$, um valor $\omega > 0$ é considerado "não demasiado grande" se verificar a Regra de Armijo:

$$g(\omega) \leq g(0) + \varepsilon g'(0)\omega,$$

dado um $\varepsilon \in (0, 1)$ fixo. Podemos começar com $\omega_0 > 0$, e definir um factor multiplicativo $\eta > 1$, tal que ω_n não seja demasiado pequeno.

O algoritmo reduz-se a:

• Se

$$g(\omega_k) \leq g(0) + \varepsilon g'(0)\omega_k,$$

tomamos $\omega_{k+1} = \eta\omega_k$, até que a regra não se verifique (aí usamos $\omega_n = \omega_k$).

Por outro lado se começarmos com ω_0 que não verifica a regra de Armijo, tomamos $\eta < 1$ e procedemos de forma semelhante

$$\omega_{k+1} = \eta\omega_k,$$

até que a regra seja verificada.

Observação 37. **Teste de Goldstein.** É semelhante à regra de Armijo agora com $\varepsilon \in (0, \frac{1}{2})$, e com

$$g(0) + (1-\varepsilon)g'(0)\omega < g(\omega) \leq g(0) + \varepsilon g'(0)\omega.$$

A primeira desigualdade é adicionada para que ω não seja demasiado pequeno.

Exemplo 20. Vejamos um exemplo para aplicação da Regra de Armijo na pesquisa linear inexacta associada ao método do gradiente.

Seja

$$f(x) = (x_1 + 2)^2 + (x_2 - 1)^2 + 3x_1 + 2,$$

e começamos com $x^{(0)} = (0, 0)$. Como

$$\nabla f(x) = \begin{bmatrix} 7 + 2x_1 \\ -2 + 2x_2 \end{bmatrix},$$

a direcção do gradiente é $d^{(0)} = -\nabla f(x^{(0)}) = \begin{bmatrix} -7 \\ 2 \end{bmatrix}$. Temos assim

$$g(\omega) = f(x^{(0)} + \omega d^{(0)}) = f(-7\omega, 2\omega) = (-7\omega + 2)^2 + (2\omega - 1)^2 - 21\omega + 2.$$

Aqui é fácil obter $g'(\omega) = 53(2\omega - 1)$, dando $\omega = \frac{1}{2}$ como solução e levando a

$$x^{(1)} = x^{(0)} + \omega d^{(0)} = (-\frac{7}{2}, 1)$$

que é imediatamente o ponto de mínimo da forma quadrática f.

Porém aqui o objectivo é ilustrar a Regra de Armijo.

Neste caso $g'(0) = -53$, e tomando $\varepsilon = 0.5$, começando com $\omega_0 = 1$ obtemos

$$g(\omega_0) - (g(0) + \varepsilon g'(0)\omega_0) = 26.5 > 0,$$

o teste falha e consideramos $\omega_1 = \eta\omega_0$ com $\eta = 0.5 < 1$, levando a $\omega_1 = 0.5$, e esta é já a solução.

3.6.2 Teste de Wolfe

O tese de Wolfe usa alternativamente

$$g'(\omega) \geq (1-\varepsilon)g'(0)$$

para assegurar que ω não é demasiado pequeno.

Note-se que quando consideramos $g(0) \approx g(\omega) + g'(\omega)(0-\omega)$ então

$$g(\omega) - g(0) \approx g'(\omega)\omega$$

e a primeira desigualdade fica

$$g(0) + (1-\varepsilon)g'(0)\omega \quad < \quad g(\omega) \Leftrightarrow (1-\varepsilon)g'(0)\omega < g(\omega) - g(0)$$

o que leva (aproximadamente) ao teste de Wolfe

$$(1-\varepsilon)g'(0)\omega < g(\omega) - g(0) \approx g'(\omega)\omega.$$

3.7 Sistemas lineares e Direcções Conjugadas

Podemos relacionar a minimização de uma função quadrática à solução de sistemas lineares, pois o mínimo de

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x + c$$

é a solução do sistema linear $Ax = b$, onde A é um matriz simétrica e definida positiva.

3.7.1 Método do gradiente para sistemas lineares

Neste caso temos

$$d^{(n)} = -\nabla f(x^{(n)}) = b - Ax^{(n)}$$

e a escolha optimal para ω_n é dada pela solução de

$$\begin{aligned} 0 = \frac{d}{d\omega} f(x^{(n)} + \omega d^{(n)}) &= d^{(n)} \cdot \nabla f(x^{(n)} + \omega d^{(n)}) \\ &= d^{(n)} \cdot (b - A(x^{(n)} + \omega d^{(n)})) \\ &= d^{(n)} \cdot (\underbrace{b - Ax^{(n)}}_{d^{(n)}} - \omega d^{(n)} \cdot Ad^{(n)}) \end{aligned}$$

de onde temos $\omega d^{(n)} \cdot Ad^{(n)} = d^{(n)} \cdot d^{(n)}$ e por isso

$$\omega_n = \frac{||d^{(n)}||^2}{||d^{(n)}||_A^2}.$$

Assim, o método do gradiente aplicado à minimização de formas quadráticas leva-nos à expressão

$$x^{(n+1)} = x^{(n)} + \frac{||d^{(n)}||^2}{||d^{(n)}||_A^2} d^{(n)}$$

onde $d^{(n)} = b - Ax^{(n)}$ no caso de sistemas lineares. Este $d^{(n)} = b - Ax^{(n)}$ é habitualmente designado "resíduo do sistema".

Observação 38. O valor dado para ω_n no caso de sistemas lineares pode ser adoptado como iterada inicial no método do gradiente (versão exacta ou inexacta), usando

$$d^{(n)} = -\nabla f(x^{(n)}),\ A = \nabla^2 f(x^{(n)}),$$

pois na vizinhança do ponto de mínimo a função comporta-se como uma forma quadrática onde a matriz hessiana é A.

Assim, por vezes o método do gradiente é mais simplesmente tomado como

$$x^{(n+1)} = x^{(n)} - \tilde{\omega}\nabla f(x^{(n)})$$

usando

$$\tilde{\omega} = \frac{||\nabla f(x^{(n)})||^2}{||\nabla f(x^{(n)})||^2_{[\nabla^2 f(x^{(n)})]}}$$

como aproximação do valor exacto ω_n, algo que poderá ser melhorado usando os algoritmos de pequisa linear inexacta (Armijo, Goldstein, ou Wolfe).

Proposição 15. *O método do gradiente para minimização quadrática verifica a estimativa de erro*

$$||e^{(n+1)}||_A \leq \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} ||e^{(n)}||_A$$

exibindo convergência linear.

Demonstração. [ver Luenberger] □

3.7.2 Método das direcções conjugadas

Ao invés de usarmos as direcções do gradiente, podemos usar direcções ortogonais relativamente ao produto interno definido por

$$\langle u, v \rangle_A = u^\top A v,$$

já que A é uma matriz definida positiva e simétrica.

Definição 23. Quando $\langle u, v \rangle_A = 0$ as direcções u e v são designadas A-conjugadas (ou A-ortogonais).

Tal como no processo de ortonormalização de Gram-Schmidt, podemos definir direcções A-conjugadas

$$d^{(0)}, \ldots, d^{(N-1)}$$

basedas nas direcções do gradiente, que são resíduos ($n = 0, \ldots, N-1$),

$$r^{(n)} = -\nabla f(x^{(n)}) = b - Ax^{(n)}.$$

De forma semelhante, dada a direcção conjugada $d^{(n)}$, o optimal ω_n para a minimização quadrática é dado por

$$0 = \frac{d}{d\omega} f(x^{(n)} + \omega d^{(n)}) = d^{(n)} \cdot (\underbrace{b - Ax^{(n)}}_{r^{(n)}} - \omega d^{(n)} \cdot A d^{(n)})$$

ou seja $\omega_n = \frac{d^{(n)}.r^{(n)}}{||d^{(n)}||_A^2}$, e o Método das direcções conjugadas fica

$$x^{(n+1)} = x^{(n)} + \frac{d^{(n)}.r^{(n)}}{||d^{(n)}||_A^2} d^{(n)}.$$

Teorema 17. *O método das direcções conjugadas atinge a solução depois de N iterações (onde N é a dimensão da matriz A).*

Demonstração. Considere o erro na iterada $e^{(n)} = x - x^{(n)}$, o que dá imediatamente

$$e^{(n+1)} = e^{(n)} - \omega_n d^{(n)} = e^{(n)} - \frac{d^{(n)}.r^{(n)}}{||d^{(n)}||_A^2} d^{(n)},$$

ou seja

$$e^{(k)} = e^{(k-1)} - \omega_{k-1} d^{(k-1)} = \cdots = e^{(0)} - \omega_0 d^{(0)} - \cdots - \omega_{k-1} d^{(k-1)}.$$

Por outro lado podemos escrever $e^{(0)}$ na base A-conjugada

$$e^{(0)} = p_0 d^{(0)} + \cdots + p_{N-1} d^{(N-1)},$$

onde p_k é a A-projecção

$$p_k = \frac{\left\langle e^{(0)}, d^{(k)} \right\rangle_A}{||d^{(k)}||_A^2}.$$

Para concluir que $e^{(N)} = 0$, fica suficiente mostrar que $p_k = \omega_k$,

$$\begin{aligned} p_k &= \frac{\left\langle e^{(0)}, d^{(k)} \right\rangle_A}{||d^{(k)}||_A^2} = \frac{\left\langle e^{(k)} + \omega_0 d^{(0)} + \cdots + \omega_{k-1} d^{(k-1)}, d^{(k)} \right\rangle_A}{||d^{(k)}||_A^2} \\ &= \frac{\left\langle e^{(k)}, d^{(k)} \right\rangle_A + 0}{||d^{(k)}||_A^2} = \frac{d^{(k)} \cdot A e^{(k)}}{||d^{(k)}||_A^2} = \frac{d^{(k)} \cdot (b - A x^{(k)})}{||d^{(k)}||_A^2} = \frac{d^{(k)} \cdot r^{(k)}}{||d^{(k)}||_A^2} = \omega_k, \end{aligned}$$

para $k = 0, \ldots, N-1$. □

Observação 39. Para construir as direcções conjugadas a partir das direcções do gradiente $r^{(k)}$, usamos o processo de Gram-Schmidt, começando com

$$d^{(0)} = r^{(0)},$$

$$d^{(k)} = r^{(k)} - \sum_{j=0}^{k-1} \frac{\left\langle r^{(j)}, d^{(j)} \right\rangle_A}{||d^{(j)}||_A^2} d^{(j)}$$

o que leva a

$$d^{(k)} = r^{(k)} + \frac{||r^{(k)}||^2}{||r^{(k-1)}||^2} d^{(k-1)}.$$

3.8 Métodos dos Gradientes Conjugados

As direcções conjugadas podem ser aplicadas ao método do gradiente, usando

$$r^{(k)} = -\nabla f(x^{(k)}),$$

mas agora notamos que algumas expressões que seriam equivalentes no caso quadrático podem deixar de o ser neste caso mais geral. Dão assim origem a métodos diferentes, por exemplo:

- as variantes Fletcher-Reeves e Polak-Ribière.

3.8.1 Métodos de Fletcher-Reeves e Polak-Ribière

O Método de Fletcher-Reeves usa as direcções definidas recursivamente por

$$d^{(k)} = r^{(k)} + \frac{||r^{(k)}||^2}{||r^{(k-1)}||^2} d^{(k-1)},$$

enquanto o Método de Polak-Ribière usa a expressão

$$d^{(k)} = r^{(k)} + \frac{\left(r^{(k)} - r^{(k-1)}\right) \cdot r^{(k)}}{||r^{(k-1)}||^2} d^{(k-1)},$$

que era equivalente à primeira no caso quadrático.

Observação 40. Uma outra forma equivalente leva ao método de Hestenes-Stiefel

$$d^{(k)} = r^{(k)} + \frac{\left(r^{(k)} - r^{(k-1)}\right) \cdot r^{(k)}}{-\left(r^{(k)} - r^{(k-1)}\right) \cdot d^{(k-1)}} d^{(k-1)},$$

3.8.2 Implementação como métodos de descida

Os métodos anteriores apenas nos dão as direcções de descida, seguindo as expressões que se obtinham para o caso quadrático.

Essas direcções devem ser consideradas na forma

$$x^{(k+1)} = x^{(k)} + \omega_k d^{(k)}$$

onde o ω_k deve ser encontrado pela pesquisa linear com a função habitual

$$g(\omega) = f(x^{(k)} + \omega d^{(k)}),$$

usando um pesquisa exacta ou inexacta, conforme descrito anteriormente.

3.9 Métodos Quasi-Newton

Há diversas formas de evitar o cálculo da matriz hessiana para a implementação do método de Newton. Assim, não teremos o método de Newton, mas suas aproximações, que são chamadas Métodos Quasi-Newton.

Tal como no caso do Método de Broyden para resolver sistemas não lineares, podemos considerar uma aproximação do "tipo secante", usando uma matriz B em vez da hessiana $H = \nabla^2 f(x^{(k)})$, que verifica

$$\nabla f(x^{(k+1)}) = \nabla f(x^{(k)}) + H(x^{(k+1)} - x^{(k)}) + o(||h_x^{(k)}||)$$

com $h_x^{(k)} = x^{(k+1)} - x^{(k)}$.

O método de Newton clássico toma $x^{(k+1)}$ tal que $\nabla f(x^{(k+1)}) = 0$.

Assim, em vez de fixarmos a hessiana H, procuramos uma matriz B_k tal que

$$\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = B_k(x^{(k+1)} - x^{(k)}).$$

Duas escolhas habituais para B_k surgem dos algoritmos DFP e BFGS, que são casos particulares destes métodos de Broyden.

3.9.1 Métodos BFGS e DFP

Considere-se $y^{(k)} = \nabla f(x^{(k)})$, e $h_y^{(k)} = y^{(k+1)} - y^{(k)}$.

A variante DFP (Davidon-Fletcher-Powell) considera

$$B_{k+1} = \left(I - \frac{h_y^{(k)}[h_x^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}}\right) B_k \left(I - \frac{h_x^{(k)}[h_y^{(k)}]^\top}{[h_x^{(k)}]^\top h_y^{(k)}}\right) + \frac{h_y^{(k)}[h_y^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}}$$

onde a sua inversa é dada pela fórmula de Sherman-Morrison

$$B_{k+1}^{-1} = B_k^{-1} + \frac{h_x^{(k)}[h_x^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}} - \frac{B_k^{-1}h_y^{(k)}[B_k^{-1}h_y^{(k)}]^\top}{[h_y^{(k)}]^\top B_k^{-1} h_y^{(k)}}.$$

Isto pode ser usado numa pesquisa linear

$$x^{(k+1)} = x^{(k)} + \omega_k B_k^{-1}(-\nabla f(x^{(k)})),$$

onde a direcção é

$$d^{(k)} = -B_k^{-1}\nabla f(x^{(k)}),$$

e onde ω_k é procurado minimizar $g(\omega) = f(x^{(k)} + \omega d^{(k)})$, por pesquisa exacta ou inexacta.

Alternativamente podemos considerar a variante BFGS (Broyden-Fletcher-Goldfarb-Shanno)

$$B_{k+1} = B_k + \frac{h_y^{(k)}[h_y^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}} - \frac{B_k h_x^{(k)}[B_k h_x^{(k)}]^\top}{[h_x^{(k)}]^\top B_k h_x^{(k)}},$$

com inversa dada pela fórmula de Sherman-Morrison

$$B_{k+1}^{-1} = \left(I - \frac{h_x^{(k)}[h_y^{(k)}]^\top}{[h_x^{(k)}]^\top h_y^{(k)}} \right) B_k^{-1} \left(I - \frac{h_y^{(k)}[h_x^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}} \right) + \frac{h_x^{(k)}[h_x^{(k)}]^\top}{[h_y^{(k)}]^\top h_x^{(k)}}$$

e um processo semelhante é considerado.

Observação 41. Estes dois métodos podem ser agrupados numa combinação convexa (família de Broyden)

$$B_{k+1} = (1 - \theta_k) B_{k+1}^{BFGS} + \theta_k B_{k+1}^{DFP}$$

com um parâmetro fixo ou variável $\theta_k \in [0, 1]$. A combinação convexa tanto pode ser usada directamente em B_k ou na sua inversa B_k^{-1}, dependendo na forma de avaliação de $d^{(k)}$ (ou seja, resolver um sistema, ou calcular a inversa).

3.9.2 Método de Gauss-Newton (mínimos quadrados não lineares)

Um caso conhecido é o problema de mínimos quadrados com uma função não linear

$$F : \mathbb{R}^N \to \mathbb{R}^M$$

e onde o objectivo é minimizar

$$f(x) = \frac{1}{2}||F(x)||^2 = \frac{1}{2} \langle F(x), F(x) \rangle$$

Exemplo 21. Aproximar $g(t)$ com funções da forma $x_1 e^{x_2 t}$, ou seja temos que minimizar

$$F(x_1, x_2)(t_k) = x_1 e^{x_2 t_k} - g(t_k)$$

para pontos $t_1, \ldots, t_M$, no caso da norma ℓ^2 discreta

$$f(x_1, x_2) = \sum_{k=1}^{M} (x_1 e^{x_2 t_k} - g(t_k))^2.$$

Como a dependência nos coeficientes x_1, x_2 deixou de ser linear, o procedimento dos mínimos quadrados habituais deixa de ser possível.

- O processo geral é ainda usar a derivada de Fréchet e obtemos

  $$f'_x(h) = \langle F'_x(h), F(x) \rangle$$

  porque

  $$\begin{aligned} \langle F(x+h), F(x+h) \rangle - \langle F(x), F(x) \rangle &= \langle F(x+h) - F(x), F(x+h) + F(x) \rangle \\ &= \langle F'_x(h) + o(||h||), F'_x(h) + o(||h||) + 2F(x) \rangle \\ &= 2 \langle F'_x(h), F(x) \rangle + o(||h||) \end{aligned}$$

  notando que $\langle F'_x(h), F'_x(h) \rangle = ||F'_x(h)||^2 \leq (||F'_x||_{\mathcal{L}(H)} ||h||)^2 = O(||h||^2) = o(||h||)$.

- No caso em que $H = \mathbb{R}^N$ e em que $F(x) \in \mathbb{R}^M$, temos

$$0 = f'_x(h) = \langle F'_x(h), F(x) \rangle = (\nabla F(x)h) \cdot F(x),$$

e usando a base canónica com $h = \mathbf{e}_k$, obtemos

$$\nabla F(x)^\top F(x) = 0.$$

Ou seja

$$\begin{bmatrix} \frac{\partial F}{\partial x_1}(t_1) & \cdots & \frac{\partial F}{\partial x_1}(t_M) \\ \vdots & \ddots & \vdots \\ \frac{\partial F}{\partial x_N}(t_1) & \cdots & \frac{\partial F}{\partial x_N}(t_M) \end{bmatrix} \begin{bmatrix} F(t_1) \\ \vdots \\ F(t_M) \end{bmatrix} = 0$$

Exemplo 22. No exemplo considerado antes, com $F(x_1, x_2)(t) = x_1 e^{x_2 t} - g(t)$, obtemos

$$\nabla F(x) = \begin{bmatrix} e^{x_2 t_1} & x_1 t_1 e^{x_2 t_1} \\ \vdots & \vdots \\ e^{x_2 t_M} & x_1 t_M e^{x_2 t_M} \end{bmatrix}, \quad -F(x) = \begin{bmatrix} g(t_1) - x_1 e^{x_2 t_1} \\ \vdots \\ g(t_M) - x_1 e^{x_2 t_M} \end{bmatrix}$$

$$0 = f'_x = \nabla F(x)^\top F(x)$$

ficando

$$\nabla F(x)^\top F(x) = \begin{bmatrix} \sum_{k=1}^M (g(t_k) - x_1 e^{x_2 t_k}) e^{x_2 t_k} \\ \\ \sum_{k=1}^M (g(t_k) - x_1 e^{x_2 t_k}) x_1 t_k e^{x_2 t_k} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

- Aplicando o Método de Newton à função $\Phi(x) = \nabla F(x)^\top F(x)$, temos

$$\nabla \Phi(x) = \nabla(\nabla F(x)^\top F(x)) = \nabla^2 F(x) : F(x) + \nabla F(x)^\top \nabla F(x).$$

O ***Método de Gauss-Newton*** consiste em ignorar a parte das 2ª derivadas,$\nabla^2 F(x) : F(x)$, e trabalhar apenas com

$$\nabla \Phi(x) \approx \nabla F(x)^\top \nabla F(x)$$

na resolução do sistema linear de Newton $\nabla \Phi(x^{(k)})\mathbf{h} = -\Phi(x^{(k)})$, ficando assim

$$\text{[Método de Gauss-Newton]} \quad J^\top J \ \mathbf{h} = -J^\top \mathbf{f} \tag{3.2}$$

onde $J = \nabla F(x^{(k)})$, $\mathbf{f} = F(x^{(k)})$.

Esta atribuição do valor de $\mathbf{h}$ pode ser tanto encarada como a escolha directa para $x^{(k+1)}$

$$x^{(k+1)} = x^{(k)} + \mathbf{h},$$

como também uma direcção de descida, e nesse caso $x^{(k+1)} = x^{(k)} + \omega \mathbf{h}$, onde ω é determinado pelos métodos de pesquisa linear.

Observação 42. Pelo método de Gauss-Newton

$$x^{(k+1)} - x^{(k)} = \mathbf{h} = -(J^\top J)^{-1} J^\top \mathbf{f}$$

e portanto pode ser visto como uma aplicação do método do ponto fixo usando

$$\begin{aligned} x &= G(x) = x - (J^\top J)^{-1} J^\top \mathbf{f} \\ &= x - (\nabla F(x)^\top \nabla F(x))^{-1} \nabla F(x)^\top F(x) \end{aligned}$$

Observação 43. Uma forma mais conhecida de aplicação do Método de Levenberg-Marquardt é aplicada ao Método de Gauss-Newton

$$\left(J^\top J + \varepsilon I\right) \mathbf{h} = -J^\top \mathbf{f}$$

sendo também habitual substituir I pela diagonal da matriz $J^\top J$.

Capítulo 4

Optimização com restrições

Dada uma função $f : \mathbb{R}^N \to \mathbb{R}$ considere-se o problema de minimização (P):

$$\min_{x \in \bar{\Omega} \subseteq \mathbb{R}^N} f(x)$$

onde $\bar{\Omega}$ é um conjunto admíssivel, definido agora por restrições c_i :

- restrições de igualdade, $c_i(x) = 0$, onde $i \in E$ (E é o conjunto de índices de igualdade)
- restrições de igualdade, $c_i(x) \geq 0$, onde $i \in D$ (D é o conjunto de índices de deigualdade)

Portanto o conjunto admissível é dado por

$$\bar{\Omega} = \{x \in \mathbb{R}^N : c_i(x) = 0 \ (i \in E), c_i(x) \geq 0 \ (i \in D)\}.$$

Definição 24. Seja f contínua, dizemos que $z \in \bar{\Omega}$ é um ponto de mínimo local se existir uma vizinhança V_z :

$$f(x) \geq f(z), \forall x \in \Omega \cap V_z;$$

e dizemos ser estrito se $f(x) > f(z)$ quando $x \neq z, \ x \in \Omega \cap V_z$.

No caso de minimização com restrições o ponto de mínimo pode estar no interior, $z \in \Omega$, ou num ponto da fronteira, $z \in \partial\Omega$.

Definição 25. A restrição c_i diz-se *Activa* em $x \in \bar{\Omega}$, se $c_i(x) = 0$. As restrições de desigualdade dizem-se inactivas se $c_i(x) > 0$. Para cada $x \in \bar{\Omega}$, o conjunto de índices das restrições activas é:

$$A(x) = E \cup \{i \in D : c_i(x) = 0\}.$$

Para evitar condições redundantes no mesmo ponto, quando c_i são diferenciáveis, também assumimos independência linear dos gradientes.

Definição 26. (LICQ - *Linear independence constraint qualification*). A condição LICQ é válida em z, se o conjunto de funções

$$\{\nabla c_i(z) : i \in A(z)\}$$

for linearmente independente.

4.1 Condições KKT

Ao problema de minimização com restrições associamos a *função lagrangiana*

$$\mathcal{L}(x,\lambda) = f(x) - \sum_{i\in E\cup D} \lambda_i c_i(x),$$

onde $\lambda_i \in \mathbb{R}$ são os multiplicadores de Lagrange, definidos para cada $i \in E \cup D$.

Teorema 18. *(Condições necessárias de 1ª ordem - KKT - Karush-Kuhn-Tucker).*
Se z é um ponto de mínimo local verificando LICQ, então existe um multiplicador de Lagrange (vectorial) λ que verifica as condições (KKT):

$$\begin{cases} \nabla f(z) - \sum_{i\in A(z)} \lambda_i \nabla c_i(z) = 0 & \\ c_i(z) = 0 & (i \in E) \\ c_i(z) \geq 0 & (i \in D) \\ \lambda_i \geq 0 & (i \in D) \\ \lambda_i c_i(z) = 0 & (i \in E \cup D) \end{cases}$$

As condições KKT levam a um sistema onde o número de equações coincide com o número de incógnitas. Esse sistema pode ser resolvido no contexto habitual de métodos para sistemas de equações, no entanto há outras abordagens - nomeadamente as que levam a uma adaptação a métodos para problemas sem restrições.

Observação 44. Ao número de incógnitas N correspondente às coordenadas do ponto z acresce o número de incógnitas $M_E + M_D$ (o cardinal de E somado ao de D) resultante das restrições. Essas incógnitas passam a ser os multiplicadores de Lagrange, um por cada restrição. A primeira condição tem N equações, a que juntam as equações M_E. As inequações M_D passam a equações activas pela última condição quando os λ_i não são nulos. Devem discutir-se os casos conforme os λ_i sejam ou não nulos (ao definir um $\lambda_i = 0$ fica determinada uma das incógnitas). Se assumirmos todos os λ_i nulos vamos encontrar mínimos de f sem ter em conta as restrições (isso só acontece quando o mínimo é interno ao conjunto admissível).

Exemplo 23. (Restrições de igualdade)
Consideremos $f(x) = 4 - x_1^2 - 4x_1x_2 - x_2^2$ sujeito à restrição $x_1^2 + x_2^2 = 1$.
Podemos escrever $c_1(x) = 1 - x_1^2 - x_2^2 = 0$, e temos

$$\mathcal{L}(x,\lambda) = 4 - x_1^2 - 4x_1x_2 - x_2^2 - \lambda(1 - x_1^2 - x_2^2)$$

o que nos dá

$$\nabla\mathcal{L}(x,\lambda) = \begin{bmatrix} -2x_1 - 4x_2 + 2\lambda x_1 \\ -4x_1 - 2x_2 + 2\lambda x_2 \end{bmatrix} = 0$$

e podemos resolver este sistema pelo método de Newton.

O método de Newton envolve aqui a matriz Hessiana de $\mathcal{L}$ que pode ser escrita, no caso de restrições de igualdade, na forma seguinte

$$H_{\mathcal{L}}(x,\lambda) = \begin{bmatrix} \nabla_x^2\mathcal{L} & -\nabla c \\ (\nabla c)^\top & 0 \end{bmatrix}$$

onde ∇c é a matriz jacobiana relativa às condições c, e onde $\nabla^2_x \mathcal{L}$ representa a matriz hessiana apenas em termos das condições sobre x.

Neste caso, como

$$\nabla^2_x \mathcal{L} = \begin{bmatrix} -2+2\lambda & -4 \\ -4 & -2+2\lambda \end{bmatrix}, \ \nabla c = [2x_1 \ 2x_2]^\top$$

obtemos

$$H_{\mathcal{L}}(x,\lambda) = \begin{bmatrix} -2+2\lambda & -4 & 2x_1 \\ -4 & -2+2\lambda & 2x_2 \\ 2x_1 & 2x_2 & 0 \end{bmatrix}$$

o que também poderia ser obtido por cálculo directo.

Assim, o método de Newton consistiria na resolução sucessiva dos sistemas

$$H_{\mathcal{L}}(x^{(n)}, \lambda^{(n)})\mathbf{s}^{(n)} = -\nabla \mathcal{L}(x^{(n)}, \lambda^{(n)})$$

que definem a nova iterada $(x^{(n+1)}, \lambda^{(n+1)}) = (x^{(n)}, \lambda^{(n)}) + \mathbf{s}^{(n)}$.

No caso mais geral, devemos ainda considerar a possibilidade de restrições de desigualdade, e iremos ver como estas são usadas.

Definição 27. Definimos o conjunto

$$F_1(z) = \left\{ \alpha d : \ \alpha > 0, \ \begin{cases} d^* \nabla c_i(z) = 0, & (i \in E) \\ d^* \nabla c_i(z) \geq 0, & (i \in D \cap A(z)) \end{cases} \right\},$$

que corresponde ao cone tangente à região admissível (sendo LICQ válida);

e também o conjunto $F_2(\lambda, z) \subseteq F_1(z)$:

$$F_2(\lambda, z) = \{d \in F_1(z) : d^* \nabla c_i(z) = 0 \ (i \in D \cap A(z), \lambda_i > 0)\},$$

subespaço tangente das restrições activas.

Proposição 16.

(i) Se z verifica as condições KKT e $w \in F_1(z)$, então $w^* \nabla f(z) \geq 0$.

(ii) Se z verifica as condições KKT e $w \in F_2(\lambda, z)$, então $w^* \nabla f(z) = 0$.

Teorema 19. *(Condições necessárias de 2ª ordem). Se z é solução local do problema de minimização (P), e sendo $f, c \in C^2(V_z)$, verificando-se a condição LICQ em z, então são válidas as condições KKT e a matriz Hessiana do Lagrangiano é semidefinida positiva em $F_2(\lambda, z)$:*

$$w^* H_{\mathcal{L}}(z) w \geq 0, \ \forall w \in F_2(\lambda, z).$$

De forma análoga estabelecem-se condições de 2ª ordem, suficientes

Teorema 20. *(Condições suficientes de 2ª ordem). Sejam $f, c \in C^2(V_z)$, tal que existe um vector multiplicador de Lagrange λ associado ao ponto z onde se verificam as condições KKT e a matriz Hessiana do Lagrangiano é definida positiva em $F_2(\lambda, z)$:*

$$w^* H_{\mathcal{L}}(z) w > 0, \ \forall w \in F_2(\lambda, z), w \neq 0;$$

então z é solução local estrita do problema de minimização (P).

Exemplo 24. Consideramos o problema de minimização de

$$f(x) = (x_1 - 2)^2 + 2x_2^2$$

com as restrições de desigualdade

$$c_1(x) = 1 - x_1 - x_2,\ c_2(x) = x_2 - x_1 + 1,\ c_3(x) = x_1.$$

Nota: aqui sabemos imediatamente que o ponto de mínimo é $z = (1, 0)$.

O gradiente do Lagrangiano é dado por

$$\begin{aligned} \nabla\mathcal{L}(z,\lambda) &= \nabla f - \lambda_1\nabla c_1 - \lambda_2\nabla c_2 - \lambda_3\nabla c_3 \\ &= \begin{bmatrix} 2(x_1-2) \\ 4x_2 \end{bmatrix} - \lambda_1 \begin{bmatrix} -1 \\ -1 \end{bmatrix} - \lambda_2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \lambda_3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{aligned}$$

Pelo que as condições KKT ficam:

$$\begin{bmatrix} 2(x_1-2) + \lambda_1 + \lambda_2 - \lambda_3 \\ 4x_2 + \lambda_1 - \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

com as restrições

$$x_1 - 1 \leq x_2 \leq 1 - x_1;\ x_1 \geq 0$$

com $\lambda_1, \lambda_2, \lambda_3 \geq 0$, e com

$$\begin{cases} \lambda_1(1 - x_1 - x_2) = 0 \\ \lambda_2(x_2 - x_1 + 1) = 0 \\ \lambda_3 x_1 = 0 \end{cases}$$

Podemos ver os vários casos.

1º) $\lambda_1, \lambda_2, \lambda_3 = 0$, corresponde a não impor restrições, e o mínimo seria interno. Obteríamos imediatamente $x = (2, 0)$, mas a condição c_2 ou seja $x_2 \leq 1 - x_1$ não seria verificada $0 \leq 1 - 2$ é falso. Concluímos que o mínimo sem restrições não é o mínimo do problema (P).

2º) $\lambda_1, \lambda_2 = 0, \lambda_3 \neq 0$, corresponde a considerar apenas uma restrição activa em c_3, obtemos logo $x_1 = 0$, mas isso implica $-4 - \lambda_3 = 0$, ou seja $\lambda_3 < 0$, o que não pode acontecer (por KKT).

Se considerarmos $\lambda_1 \neq 0, \lambda_2 = \lambda_3 = 0$, é semelhante, mas apenas com a restrição activa em c_1, obtemos (resolvendo o sistema) $x = (4/3, -1/3)$ que não pertence à região admissível, pois $c_2(x) = -\frac{1}{3} - \frac{4}{3} + 1 < 0$.

Finalmente, $\lambda_2 \neq 0, \lambda_1 = \lambda_3 = 0$, resulta numa falha da condição c_1.

3º) $\lambda_1, \lambda_2 \neq 0, \lambda_3 = 0$, corresponde a considerar as restrições c_1 e c_2 activas. A resolução do sistema leva imediatamente à solução pretendida $x = (1, 0)$, com $\lambda = (1, 1, 0)$.

Assim verificam-se as condições KKT, e a matriz Hessiana do Lagrangiano é definida positiva

$$H_{\mathcal{L}}(x) = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

Notamos ainda que as condições activas, sendo c_1 e c_2 levam aos vectores gradientes

$$\nabla c_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \nabla c_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

que são linearmente independentes, sendo válida a condição LICQ.

Observação 45. Fazemos notar que no caso em que há apenas restrições lineares, ou seja em que a função c é linear (ou afim) as segundas derivadas são nulas, tendo-se uma coincidência das matrizes hessianas: $H_{\mathcal{L}} = H_f$.

Podemos ainda considerar a matriz jacobiana das restrições activas

$$\mathcal{R} = [\nabla c_i(z)]_{(i \in A(z), \lambda_i > 0)}$$

que será uma matriz $M \times N$ em que $M \leq N$ é o número das restrições activas consideradas. Se a condição LICQ for verificada a característica da matriz $\mathcal{R}$ será M e o núcleo terá dimensão $N - M$.

Verifica-se ainda facilmente que $Ker(\mathcal{R}) = F_2(\lambda, z)$.

Definindo uma matriz $\mathcal{S}$ em que as suas colunas formam uma base de $Ker(\mathcal{R})$, ou seja $\mathcal{RS} = 0$, e escrevendo então os vectores de $F_2(\lambda, z)$ na forma $w = \mathcal{S}v$, podemos reescrever as condições de 2ª ordem, sobre a matriz Hessiana do Lagrangiano, numa forma diferente:

$$(\mathcal{S}v)^* H_{\mathcal{L}}(z)(\mathcal{S}v) \geq 0, \forall v \in \mathbb{R}^{N-M}$$

que corresponde à condição de $H_{\mathcal{L}}$ ser definida positiva em $F_2(\lambda, z)$; e de forma semelhante a condição definida positiva, com a desigualdade estrita (... > 0, se $v \neq 0$).

No caso em que $M = N$, temos $Ker(\mathcal{R}) = F_2(\lambda, z) = \{0\}$ e a matriz hessiana é aí trivialmente definida positiva, o que podemos resumir na seguinte proposição:

Proposição 17. *Se a condição KKT é satisfeita com o par (z, λ) e o conjunto dos gradientes com restrições activas é linearmente independente*

$$N = dim\{\nabla c_i(z) : \ i \in A(z), \lambda_i > 0 \ (i \in D)\}$$

então z é solução local estrita de (P).

4.2 Casos Especiais

4.2.1 Restrições lineares de igualdade

No caso em que c_i são apenas lineares/afins, as condições de igualdade podem ser escritas

$$Ax = b$$

decompômos a matriz $A = [B \vdots C]$ tal que B seja uma matriz $M \times M$ invertível (pode ser necessário uma prévia troca de linhas), e separando ainda $x = (x_B, x_C)$ obtemos

$$Ax = Bx_B + Cx_C$$

pelo que podemos resolver $Ax = b$ na forma

$$x_B = B^{-1}b - B^{-1}Cx_C.$$

Assim, a restrição é incorporada num problema sem restrições definindo uma nova função

$$g(x_c) = f(x) = f(x_B, x_C) = f(B^{-1}b - B^{-1}Cx_C, x_C)$$

podendo aplicar-se um método sem restrições em $\mathbb{R}^{N-M}$ para encontrar o mínimo de g, que será o mínimo de f sujeito às restrições lineares.

Observação 46. O mesmo processo pode ser aplicado noutros casos em que seja possível uma simplificação, escrevendo algumas variáveis em função das restantes. Por exemplo, sendo

$$f(x) = (x_1 + x_2)^2 + x_3^4 + e^{6x_4}$$

com restrições $c_1(x) = x_1 + x_2 - e^{x_4} = 0, c_2(x) = x_3 - x_1 - x_2 = 0$, podemos definir imediatamente $x_3 = x_1 + x_2 = e^{x_4}$, obtendo um problema de minimização a uma só variável

$$g(x_3) = x_3^2 + x_3^4 + x_3^6.$$

4.2.2 Caso quadrático com restrições de igualdade lineares

No caso em que f é uma função quadrática da forma

$$f(x) = \frac{1}{2}x^*Gx + x^*d$$

sujeito a condições $Ax = b$ com A matriz de característica M, o problema a resolver é linear, pois sendo $c(x) = Ax - b$, temos

$$\nabla f = Gx + d,\ \nabla c = A,$$

obtendo-se o sistema resultante da condição KKT, sobre o lagrangiano

$$\begin{bmatrix} G & -A^* \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -d \\ b \end{bmatrix}.$$

4.3 Métodos para optimização com restrições

Mencionamos apenas brevemente alguns tipos de métodos - penalização e barreira - que permitem lidar com o problema de optimização com restrições numa forma em que podem ser usados métodos de optimização sem restrições. A ideia consiste simplesmente em incorporar na minimização sem restrições uma função de custo adaptada ao domínio.

Um estudo mais detalhado destes e outros algoritmos não será apresentado nesta introdução à matéria.

4.3.1 Métodos de Penalização

Consideramos uma função de penalização P regular, não negativa, e tal que

$$P(x) = 0 \Leftrightarrow x \in \bar{\Omega}.$$

Definimos uma nova função para minimizar

$$F_\mu(x) = f(x) + \mu P(x)$$

onde $\mu > 0$ é um parâmetro suficientemente grande, que pode ser redimensionado a cada passo.

Exemplo 25. Por exemplo no caso em que procuramos

$$\min_{\bar{\Omega}} f(x)$$

com $\bar{\Omega}$ definido por $c_i(x) \geq 0, (i \in D)$, podemos definir

$$F_\mu(x) = f(x) + \mu \sum_{i \in D} P(c_i(x))$$

onde

$$P(c_i(x)) = \min\{0, c_i(x)\}^2.$$

Começamos com $\mu > 0$ e resolvemos o problema sem restrições para F_μ. A cada aumento do valor de μ podemos usar o valor obtido anteriormente.

4.3.2 Métodos de Barreira

Os métodos de barreira são semelhantes aos anteriores, mas consideram B não negativa tal que

$$B(x) \to \infty, \text{ when } x \to \partial\Omega,$$

e a função a minimizar sem restrições passa a ser $(\mu \to \infty)$

$$F_\mu(x) = f(x) + \mu^{-1} B(x).$$

Exemplo 26. Uma possibilidade usada é a barreira logarítmica. Por exemplo no caso em que procuramos

$$\min_{\bar{\Omega}} f(x)$$

com $\bar{\Omega}$ definido por $c_i(x) \geq 0, (i \in D)$, podemos definir

$$F_\mu(x) = f(x) + \mu^{-1} \sum_{i \in D} B(c_i(x))$$

onde $B(y) = -\log(y)$.

4.3.3 Lagrangiano Aumentado

Uma possibilidade usando o método de penalização para condições $c_i(x) = 0$ seria considerar

$$F_\mu(x) = f(x) + \mu \sum_{i \in E} c_i(x)^2$$

e a variante do Método do Lagrangiano Aumentado consiste em incorporar os termos do lagrangiano, na nova função sem restrições:

$$F_\mu(x) = f(x) + \frac{\mu}{2} \sum_{i \in E} c_i(x)^2 - \sum_{i \in E} \lambda_i c_i(x),$$

e onde os valores λ_i são actualizados com

$$\lambda_i - \mu c_i(\tilde{z})$$

onde $\tilde{z}$ é o ponto de mínimo obtido com o anterior valor de μ.

Referências Bibliográficas

[1] K. Atkinson, W. Han. *Theoretical Numerical Analysis: A Functional Analysis Framework*. Springer-Verlag 2001

[2] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Masson, Paris, 1982

[3] P. E. Gill, W. Murray, M. H. Wright. *Practical optimization.* Academic Press, London-New York 1981

[4] E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley, New York 1989

[5] R. Kress. *Linear integral equations*. App. Math. Sci. 82, Springer-Verlag 1999

[6] D. Luenberger, Y. Ye. *Linear and Nonlinear Programming*. Springer-Verlag 2008

[7] J. Nocedal, S. J. Wright. *Numerical Optimization*. Springer-Verlag 1999

[8] J. M. Ortega, W. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York 1970

[9] P. Pedregal. *Introduction to optimization*. Texts in Applied Mathematics, 46. Springer-Verlag, New York 2004

[10] E. Zeidler. *Nonlinear Functional Analysis and Its Applications*. Springer-Verlag, New York 1989