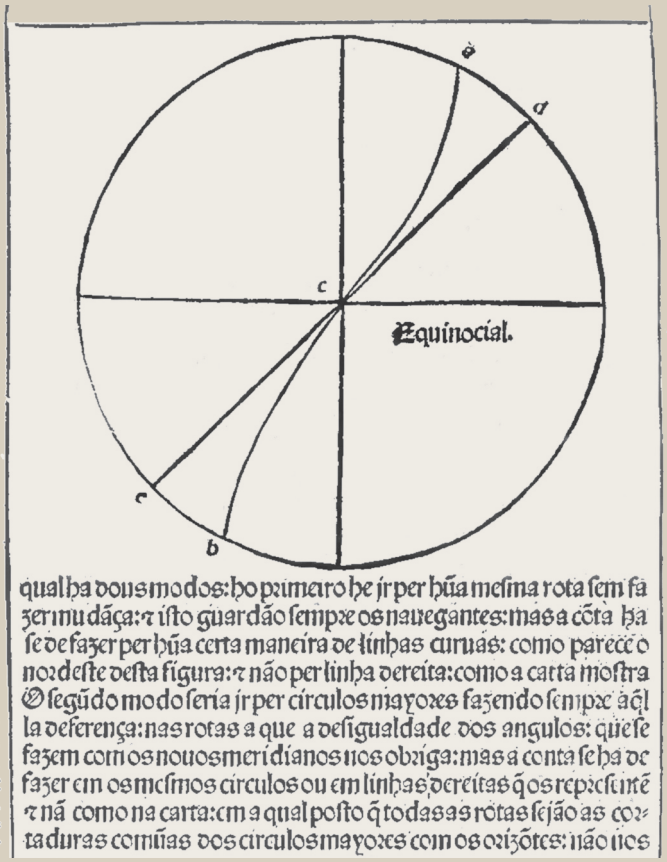
Boletim

da Sociedade Portuguesa de Matemática



Linha de rumo "acb" e Círculo máximo "dce"

Pedro Nunes, 1537

76

Número

DEZEMBRO 2018

4 euros

[semestral]

Boletim
da Sociedade Portuguesa de Matemática

Propriedade e Edição
Sociedade Portuguesa de Matemática, Av. da República, 45–3º E, 1050–187 Lisboa
☎ 217939785 FAX 217952349
Internet: `http://www.spm.pt`

Impressão e acabamento: Dossier | Comunicação e Imagem • Alameda dos Oceanos, lote 3.10.06 B • Parque das Nações • 1990–186 Lisboa

Tiragem: 1500 exemplares

Depósito Legal: nº 63134/93

ISSN: 0872–3672

ICS: 106437

aos leitores do Boletim
aos sócios da SPM

A partir do actual número 76, o Boletim iniciará uma nova fase, passando tendencialmente a ser publicado apenas em formato electrónico. Os próximos números ficarão disponíveis na página web do Boletim, em

`https://revistas.rcaap.pt/boletimspm`

Ana Jacinta Soares
Dezembro 2018

A Matemática e o Planeta Terra

José Francisco Rodrigues[*]
Departamento de Matemática e CMAFcIO
Faculdade de Ciências da Universidade de Lisboa
e-mail: jfrodrigues@ciencias.ulisboa.pt

Resumo: Neste artigo abordam-se alguns conceitos fundamentais da Matemática diretamente associados ao Planeta Terra numa perspetiva histórica. Da esfera à linha de rumo, das origens e certas aplicações do Cálculo Infinitesimal aos problemas com fronteira livre, das simetrias nos sólidos platónicos e nos cristais à calçada portuguesa, a Matemática é o elo de ligação entre temas aparentemente díspares associados ao Planeta Terra.

Abstract: In this article we discuss some fundamental concepts of Mathematics directly associated with Planet Earth in a historical perspective. From the sphere to the rhumb line from the origins and certain applications of the Infinitesimal Calculus to the problems with the free boundaries from the symmetries in the Platonic solids and in the crystals to the Portuguese sidewalk, Mathematics is the link between apparently disparate themes associated with Planet Earth.

palavras-chave: Matemática do Planeta Terra; História da Matemática; Divulgação da Matemática.

keywords: Mathematics of Planet Earth; History of Mathematics; Popularisation of Mathematics.

> *Sciencia nam eh outra cousa senão hum conhecimento habituado no entendimento: o qual se acquirio per demonstração: e demonstração he aquelle discurso que nos faz saber (...) Nem deue auer duuida no que nesta parte escreui: porque nenhua cousa he mais euidente: que ha demonstração mathematica: a que em nenhua maneyra se pode contrariar*
>
> **Pedro Nunes** (1537)

1 Esferas – da forma da Terra e do *Ceo*.

Já na Antiguidade Clássica era claro que a forma da Terra tinha que ser esférica. Em *Sobre os Céus*, no século IV Antes da Era Corrente (AC),

[*]O Autor agradece o apoio do CMAFcIO e da FCT através do Projeto UID/MAT/04561/2013. O financiamento da impressão a cores foi assegurado pelo Autor.

Aristóteles argumentou esse facto com a sombra da Terra sobre a Lua durante os eclipses e com a mudança da posição das estrelas e a alteração do círculo do horizonte com a variação da posição do observador para Norte ou para Sul.

Segundo a Definição 14 do Livro XI de *Os elementos* de Euclides[1], cerca 300 AC, *Esfera é a figura compreendida quando, o diâmetro do semicírculo permanecendo fixo, o semicírculo, tendo sido levado à volta, tendo retornado, de novo, ao mesmo lugar de onde começou a ser levado.* Tal é, também, a versão do "*Tratado da Sphera*"[2], traduzido do texto latino do século XIII de Sacrobosto pelo matemático e cosmógrafo português Pedro Nunes (1502-1578), publicado em 1537 com outros textos comentados e com os seus dois primeiros ensaios originais sobre a teoria matemática da navegação. Essa definição é aí completada com a versão de Teodósio, matemático helénico do século II AC, que a define como um "*corpo maciso recolhido debaixo de hua soo face*" equidistante de um ponto que se chama centro.

Figura 1: Partes do primeiro capítulo do "*Tratado da Sphera*"[2], da edição de 1537 de Pedro Nunes, com as definições da esfera e com uma gravura alusiva à "*redondeza da terra*", onde é patente a conceção geocêntrica da época.

A importância da esfera e das suas propriedades não se limita à Geometria euclidiana. Em grego antigo, *metron* e *geo* correspondem a medida e Terra e a referência à esfera foi central na Geografia, na Astronomia e na Cosmografia renascentistas. A esfera é o conceito instrumental de toda a Astronomia ptolemaica e não podia ser ignorado por um poeta da dimensão de Camões na sua obra épica *Os Lusíadas*[4]. Como foi demonstrado por Luciano Pereira da Silva[3], o "*Tratado da Sphera*", na tradução e com os comentários de Pedro Nunes, foi a principal fonte cosmográfica de Camões.

No canto X[4], na lição de Tétis aos navegadores portugueses sobre a Cosmografia alexandrina assente numa sobreposição de movimentos celestes circulares, uniformes e periódicos, os últimos versos das estâncias 77 e 78:

Aqui um globo vêm no ar, que o lume
Claríssimo por ele penetrava,
De modo que o seu centro está evidente,
Como a sua superfícia, claramente. (X.77)

não só descrevem as propriedades geométricas da esfera, como também através de

Volvendo, ora se abaxe, agora se erga,
Nunca s'ergue ou se abaxa, e um mesmo rosto
Por toda a parte tem; e em toda a parte
Começa e acaba, enfim, por divina arte, (X.78)

o trocadilho poético reflete as duas definições do "*Tratado da Sphera*", "volvendo" enquanto um sólido de revolução e transparente na equidistância do seu centro à sua superfície. Nesta síntese lírica, Camões evoca a constância da curvatura do globo descrevendo-a com "um mesmo rosto por toda a parte tem; e em toda a parte começa e acaba", a qual é completada no verso inicial da estância seguinte:

Uniforme, perfeito, em si sustido,
Qual, enfim, o Arquetipo que o criou. (X.79)

Como observara Luciano Pereira da Silva[3], o termo grego arquétipo, enquanto modelo primitivo, aparece também no parágrafo "*Da redondeza do ceo*", logo no início do "*Tratado da Sphera*": "*Que ho ceo seja redõdo ha três rezões. Semelhãça. proueito. e necessidade. Pella semelhança se proua ho ceo ser redondo porque este mundo sensiuel: he feito a semelhança do mundo archetypo: em ho qual nam ha principio nem fim.*" E, fiel à cosmografia noniana que é ptolemaica, prossegue o poeta:

«Vês aqui a grande máquina do Mundo,
Etérea e elemental, que fabricada
Assi foi do Saber, alto e profundo,
Que é sem princípio e meta limitada.
Quem cerca em derredor este rotundo
Globo e sua superfícia tão limada,
É Deus: mas o que é Deus, ninguém o entende,
Que a tanto o engenho humano não se estende. (X.80)

Sobre a esfera pouco existe em *Os elementos* de Euclides[1], para além da Proposição 18 do Livro XII, estabelecendo apenas que "*As esferas estão entre si em uma razão tripla da dos próprios diâmetros*", e do objetivo central do Livro XIII que consiste na construção, e sua inclusão numa esfera, dos cinco sólidos platónicos, a pirâmide regular ou tetraedro (Prop. 13), o octaedro (Prop. 14), o cubo (Prop. 15), o icosaedro (Prop. 16), e o dodecaedro (Prop. 17). Essa inclusão do poliedro na esfera significa a construção da esfera circunscrita e, portanto, a determinação da relação do seu diâmetro com as arestas dos respetivos poliedros, ou seja, uma vez e meia para o tetraedro, o dobro para o octaedro, o triplo para cubo, e proporções irracionais para o icosaedro, a chamada *menor*, e para o dodecaedro, o chamado *apótomo*[1].

Contudo, o cálculo do volume e da superfície da esfera apenas foram obtidos por Arquimedes (287-212 AC)[5] que, nos dois livros "*Sobre a esfera e o cilindro*", demonstrou que (i) a superfície de uma esfera é quádrupla da do seu círculo máximo, (ii) o volume do cilindro circunscrito na esfera com altura igual ao diâmetro desta é 3/2 do volume da esfera e (iii) a superfície desse cilindro circunscrito é também 3/2 da superfície da esfera. No entanto, como Arquimedes comunicou a Eratóstenes em "*O Método*", a descoberta destes resultados foi obtida previamente pelo método mecânico, usando considerandos físico-matemáticos da lei da alavanca e do "equilíbrio de planos". Em particular, na Proposição 2 de "*O Método*", onde obteve o volume da esfera, Arquimedes deduziu uma relação para o equilíbrio dos pesos de um cilindro, de um cone e de uma esfera inscrita no cilindro que, apesar de diferente, é equivalente à sugestão da variação moderna da Figura 2.

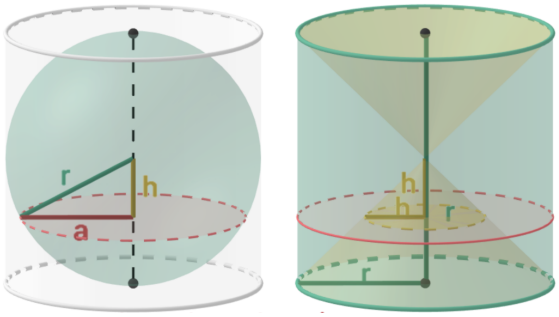


Figura 2: Como a área da coroa circular ($\pi\, r^2 - \pi\, h^2$) é, pelo teorema de Pitágoras, igual à área do círculo de raio a^2, que varre ao longo do diâmetro da esfera todo o seu volume, o volume do cilindro menos o dos dois cones será igual ao volume da esfera de raio r inscrita no cilindro, o que dá o volume da esfera $2\pi\, r^3 - 2/3\pi\, r^3 = 4/3\pi\, r^3$. Uma animação em *Geogebra* pode encontrar-se em http://www.gi2.pt/galerias/volume-de-uma-esfera/

Não sabemos se Eratóstenes (276-194 AC), matemático, geógrafo e bibliotecário em Alexandria, calculou o volume da Terra, mas, com o valor aproximado de π, $3\frac{10}{71} < \pi < 3\frac{1}{7}$, obtido por Arquimedes, poderia tê-lo feito a partir do cálculo que efetuou da sua circunferência, estimando-a em 250 000 estádios (Figura 3). Admitindo que um estádio tinha 157,5 metros, essa estimativa dá um valor de 39 690 km para o perímetro da Terra, ou seja, um valor aproximado com um erro de apenas 2%. No entanto, apesar

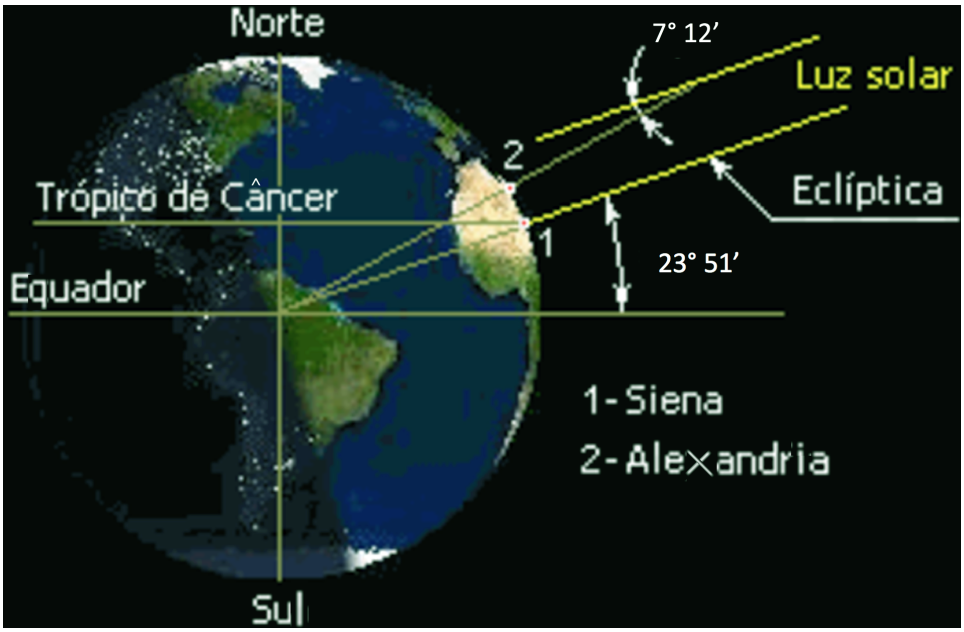


Figura 3: Eratóstenes mediu em Alexandria o ângulo de elevação do sol ao meio-dia, ou seja, a distância angular do zénite, no solstício de verão, e encontrou 1/50 de um círculo (cerca de 7° 12'), fazendo então a proporção, pois sabia que Siena (Assuão) estava situada no Trópico de Câncer e distava cerca de 5000 estádios de Alexandria. A obliquidade da eclíptica é a atribuída a Eratóstenes, e tem um excesso de 25' face ao valor médio atual. A sombra do *gnómon* também foi utilizada para medir o tempo, como se ilustra no interessante filme *Relógios de Sol, Matemática e Astronomia*[†].

de Ptolomeu, no século II da nossa era, ter atribuído a Eratóstenes a medição da inclinação do eixo da Terra em cerca 23° 51', conforme comenta Pedro Nunes no fim do primeiro capítulo do "*Tratado da Sphera*", "*há muita diversidade entre estes autores*" reconhecendo alguma incerteza nos valores "*da quantidade da terra*". Mas independentemente da precisão, o mais im-

[†] http://formas-formulas.fc.ul.pt/multimedia/filmes/

portante foi a invenção do método, que mostrou o extraordinário poder da Matemática na modelação do planeta Terra[6].

Também este famoso episódio histórico-geográfico não passou despercebido a Camões que refere Siena nuns versos de *Os Lusíadas* no canto III:

> ***Posto que o frio Fásis ou Siene,***
> ***Que pera nenhum cabo a sombra inclina,***
> ***O Bootes gelado e a linha ardente***
> ***Temessem o teu nome geralmente.*** (III.71)

O poeta faz aqui o contraste entre o *frio Fásis*, nome grego antigo do rio Rioni na atual Geórgia, com a cidade egípcia de Siena sobre "a linha ardente", ou seja, sobre o trópico de Câncer onde não há sombra ao meio dia do solstício de verão.

A esfera armilar, constituída por anéis ou armilas representando círculos máximos e paralelos, terá sido utilizada por Eratóstenes no século III AC para medir a obliquidade da eclíptica e foi escolhida pelo rei Manuel I de Portugal (1495-1521) como um símbolo associado às navegações portuguesas e integrada como um elemento artístico do estilo manuelino (Figura 4). Mais tarde, a esfera armilar foi integrada nas bandeiras do Reino Unido de Portugal, Brasil e Algarves (1815-1822), do Império do Brasil (1822-1889) e na de Portugal em 1911.

Figura 4: Esfera armilar num azulejo do Palácio Nacional de Sintra[‡] e num medalhão em pedra no claustro do Mosteiro dos Jerónimos em Lisboa[§].

[‡] http://cvc.instituto-camoes.pt/azulejos/tradisl.html
[§] http://www.mosteirojeronimos.pt/

2 Loxodrómias – da linha de rumo das navegações ao cálculo infinitesimal.

A conquista dos oceanos determinou a globalização da influência europeia a partir do século XVI e correspondeu ao início de uma grande transformação científica não só na Geografia, na História Natural, mas também na Astronomia e nas ciências físicas. A importância da náutica nesta transformação, levou o historiador inglês David Waters a defender a tese de que "*os inícios da revolução científica devem ser encontrados nos trabalhos dos portugueses e de outros eruditos que a Coroa de Portugal teve a perspicácia de empregar para resolver os novos problemas de navegação colocados pelas viagens oceânicas no século XV e inícios do XVI; e nos efeitos que as descobertas geográficas feitas pelos navegadores usando a nova ciência náutica dos portugueses tiveram sobre os intelectuais da Europa desse tempo, e sobre os homens nas práticas de governo, comércio e indústria*"[7].

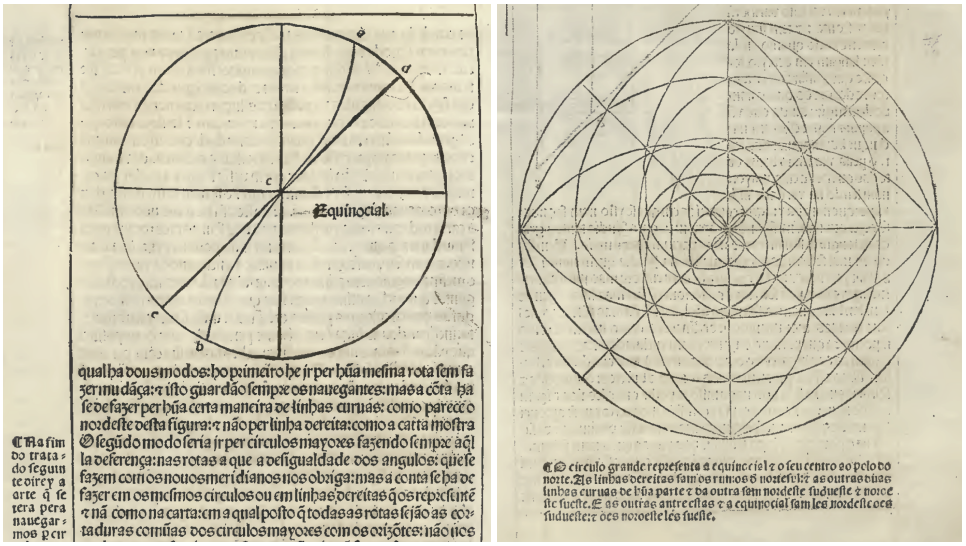


Figura 5: **(a)** A representação de 1537, por Pedro Nunes, da linha de rumo "acb" e do círculo máximo "dce". **(b)** A belíssima Roseta de Nunes do "*Tratado em defesam da carta de marear*" — na aplicação interativa "loxo" mostra-se que estes arcos não correspondem exatamente a nenhuma projeção das linhas de rumo de azimute ±45° e ±67° 30' aí representadas[¶].

Complementares e anexos ao "*Tratado da Sphera*" de 1537, publicou Pedro Nunes dois importantes ensaios originais que lançaram as bases da

[¶] http://formas-formulas.fc.ul.pt/interactive/loxo/pt/index_pt.html

teoria matemática da navegação: o "*Tratado (...) sobre certas duuidas da navegação*" e o "*Tratado (...) em defensam da carta de marear*", motivados por "certas dúvidas que trouxe Martim Afonso de Sousa quando veio do Brasil"[2], poucos anos antes. Com efeito, se um navegador seguir sempre com o leme fixo em direção ao seu objetivo descreve um arco de círculo máximo, que é afinal a "reta" sobre a superfície esférica. Esta é a linha mais curta ou geodésica sobre a esfera, que se veio a chamar ***ortodrómia***. No entanto, a orientação oceânica baseada na bússola ao manter constante o azimute, i.e. um ângulo constante e não nulo com o Norte, desvia-se da rota mais curta e constituiu motivo de confusões e de erros graves na navegação. Esta linha de rumo veio a ser chamada linha loxodrómica ou ***loxodrómia***, do grego *loxos* e *dromos*, que significam, respetivamente, torto e percurso. Estas curvas coincidem com os meridianos no caso limite do azimute a 0° e, nestes casos, também são ortodrómias. Nos rumos com azimute a 90° as loxodrómias correspondem aos paralelos e apenas no equador é uma ortodrómia.

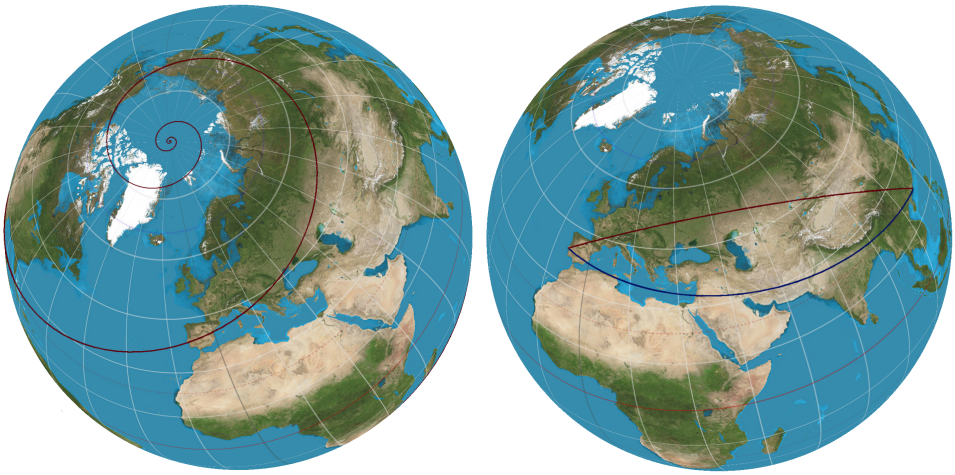


Figura 6: **(a)** *Loxodrómia com azimute de 76° a partir de Lisboa.* **(b)** A comparação do trajeto da Ortodrómia (vermelho, 11010 km) com o da Loxodrómia (azul, 11862 km) entre Lisboa e Macau, na aplicação interativa LOXO[‖], a qual dá uma diferença de cerca 850 km.

Se em 1537 Pedro Nunes pode ter admitido que as linhas de rumo atingiam os polos, num manuscrito que está em Florença, escrito poucos anos depois, clarifica "*que o rumo não chega ao polo (...), mas chega a qualquer ponto antes do polo, pois que de qualquer ponto se pode ir adiante pelo*

[‖] http://formas-formulas.fc.ul.pt/interactive/loxo/pt/index_pt.html

mesmo rumo" e, na sua *Opera*[8] em latim de 1566 publicada em Basileia, demonstra-o, não inserindo neste livro a sua roseta. No capítulo "*De Arte Atque Ratione Nauigandi*", Nunes apresenta um método para construir uma tábua de rumos, i.e., uma tabela de latitudes e longitudes das sete loxodrómias com azimutes múltiplos de 11° 15' "*para o traçado de quaisquer rumos num globo*", deixando-a para ser preenchida pelos "*moços aplicados*" que "*acharão os números que se devem escrever nela, estendendo-os quanto lhe aprouver.*"

A teoria da linha de rumo de Pedro Nunes teve um papel seminal na teoria matemática da loxodrómia e da cartografia e só recentemente começou a ser reconhecido[8] o verdadeiro alcance e o pioneirismo das suas ideias. Por exemplo, a linha quebrada que aproxima a loxodrómia por arcos de 1° de círculos máximos (Figura 7 **(b)**), a que chamaremos de ***noniodrómia***, é uma ideia natural, mas inovadora, que Pedro Nunes introduz para a construção da sua tábua de rumos aproximados, iterando soluções de sucessivos triângulos planos ao longo da linha correspondente sobre a esfera. Como cada um dos arcos de círculo máximo *bc*, *ce*, *eg* ou *gi*, é tangente à loxodrómia que passa no seu ponto inicial com o mesmo azimute V, e pode ser interpretado como um segmento de "reta" na superfície esférica, o método de Nunes é, essencialmente e traduzido para o cálculo infinitesimal, o mesmo que Euler introduziu dois séculos mais tarde, em 1768, para a resolução aproximada de uma equação diferencial de primeira ordem.

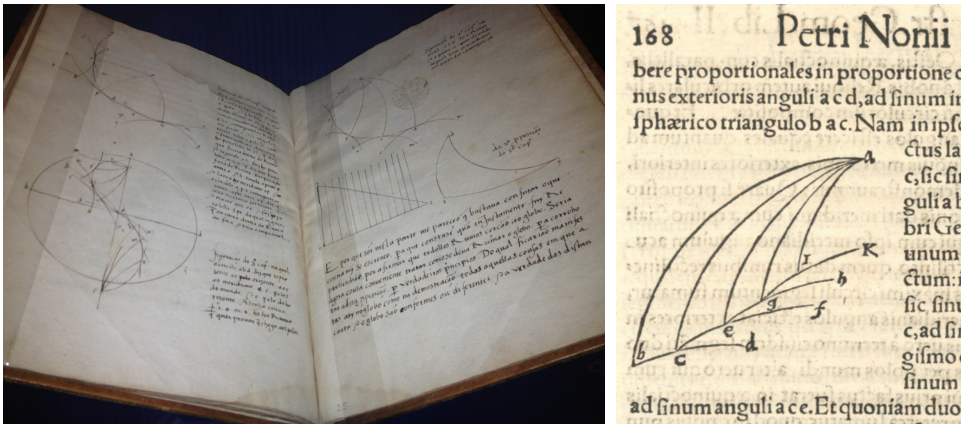


Figura 7: **(a)** Páginas do manuscrito de Pedro Nunes, cerca de 1540, com clarificações sobre a linha de rumo. **(b)** A ***noniodrómia*** é a linha *bcegi* composta por arcos de "retas" sobre a esfera fazendo o mesmo ângulo com os meridianos convergindo no polo *a* e aparece pela primeira vez em "*De Arte Atque Ratione Nauigandi*" de 1566.

Um dos problemas científicos mais importantes do século XVI era o problema matemático da navegação oceânica de como, pela rota mais curta, pela direção mais adequada e no tempo mais curto, se deve conduzir um navio entre dois lugares previamente assinalados e o problema correspondente da cartografia de representar os rumos como retas na carta de marear plana. Sem explicar detalhes, Pedro Nunes já havia imaginado e trabalhado um método para a retificação da loxodrómia no plano, antes de cartógrafo Gerardus Mercator (1512-1594) a ter realizado no seu *Mapa-múndi* em 1569, tendo-o anunciado três anos antes na introdução à sua *Opera* de 1566: "*Mas, porque era muito difícil e até inviável para os mareantes traçar nos globos linhas semelhantes a estas, os matemáticos imaginaram uma descrição plana do orbe, não só adaptada à arte de navegar que praticam, como também muitíssimo fácil. Nesta representação são desenhadas linhas retas em lugar dos rumos do mesmo nome; como são paralelas, fazem ângulos iguais com toda a linha meridiana ou rumo Norte-Sul*".

Apenas recentemente foi possível estabelecer que foi através de uma tabela de rumos que Mercator construiu a primeira projeção cilíndrica conforme num mapa[9], apesar de se saber que ele conhecia as obras de Nunes. Três décadas após a construção de Mercator, o matemático inglês Eduard Wright (1561 - 1615), que conheceu os problemas práticos da navegação oceânica numa viagem aos Açores, em 1589, e publicou em Londres, em 1599, o livro *Certain errors in navigation*, apresentou a primeira descrição detalhada e rigorosa da regra matemática subjacente àquela projeção. Neste livro, para ilustrar esta projeção, Wright utilizou a imagem duma superfície esférica a inchar, tal como acontece ao soprar uma bexiga mantendo o equador constante e em contacto com a superfície côncava de um cilindro fixo, de modo que os meridianos se tornam retas verticais e ortogonais aos paralelos, cujas distâncias entre si vão aumentando progressivamente na direção da geratriz do cilindro.

Mas existe uma infinidade de projeções cilíndricas, como a que se atribui a Arquimedes e projeta horizontalmente os pontos da esfera terrestre sobre o cilindro tangente a esta e que tem o efeito oposto de encurtar as distâncias entre paralelos no plano que se obtém desse cilindro depois de desdobrado e cortado ao longo de um meridiano. Outro exemplo, é o da projeção centográfica cilíndrica em que um ponto P, de latitude φ e longitude λ, da superfície do globo se projeta radialmente no cilindro, tangente ao seu equador, num ponto P' de coordenadas x e y. Neste caso tem-se $x/\lambda = y/\tan\varphi$, o que determina uma maior extensão das ordenadas com o aumento das latitudes. Mas esta "carta de latitudes crescidas" amplia demasiado as lati-

tudes e não é conforme, ou seja, não preserva os ângulos. Não é, portanto, a carta de marear proposta por Pedro Nunes, pois não transforma as loxodrómias em retas no cilindro, apesar de também transformar os meridianos em perpendiculares dos paralelos estendidos na superfície cilíndrica.

Seguindo as ideias de Nunes, em 1594, Wright observou que o mapa de Mercator aumenta a distância entre paralelos de tal modo que as pequenas "*partes do meridiano, em cada latitude, têm de crescer com a mesma proporção das respetivas secantes ou hipotenusas do arco, determinado pelos pontos de crescimento da latitude e da equinocial*". Assim, para construir a tabela publicada em 1599 para "*a verdadeira divisão do meridiano na carta marítima*", que é uma tabela de latitudes para $\Delta\varphi = 1'$, i.e., com variações de um minuto, Wright utilizou somas de secantes para obter as ordenadas de cada paralelo na carta de marear.

Em termos do Cálculo Infinitesimal, que só apareceu quase um século mais tarde, nesta projeção cilíndrica, em que o equador é o eixo dos x e o paralelo de latitude φ é a reta horizontal $y = l(\varphi)$, a proporção da distorção segundo a direção dos meridianos deve ser igual à proporção com o fator de escala igual a $\sec\varphi = 1/\cos\varphi$ segundo a direção dos paralelos. Então podemos concluir

$$\lim_{\Delta\varphi\to 0} \frac{\Delta \textit{ distância no mapa}}{\Delta \textit{ distância na esfera}} = \lim_{\Delta\varphi\to 0} \frac{l(\varphi+\Delta\varphi) - l(\varphi)}{\Delta\varphi} = l'(\varphi) = \sec\varphi.$$

Como observou o historiador da Matemática Florian Cajori[10], o processo utilizado por Wright corresponde a um cálculo numérico do integral da secante, ou seja, a utilização duma relação do tipo $\Sigma \sec\varphi\,\Delta\varphi$ para a aproximação do integral

$$l(\varphi) = \int_0^{\varphi} \sec\phi\, d\phi = \log\tan(\frac{\varphi}{2} + \frac{\pi}{4}).$$

É interessante observar que esta expressão também fornece a equação em coordenadas polares da loxodrómia de azimute V, i.e., tem-se

$$\lambda(\varphi) = \tan V\; l(\varphi) = \tan V\;\; \log\tan(\frac{\varphi}{2} + \frac{\pi}{4}).$$

Com efeito, se numa esfera de raio unitário, considerarmos um triângulo esférico com base infinitesimal num paralelo de latitude φ, correspondendo a uma variação infinitesimal de longitude $d\lambda$, essa base terá por comprimento $\cos\varphi\, d\lambda$, uma vez que a essa latitude o paralelo é uma circunferência de raio

$\cos\varphi$. Então para uma variação $d\varphi$ num ponto P, de coordenadas (λ, φ), sobre a loxodrómia, o quociente $\cos\varphi\, d\lambda/d\varphi$ dá o declive da tangente a essa linha nesse ponto e é igual à constante $\tan V$, como se pode facilmente concluir do respetivo triângulo incremental com vértice em P [11]. Então, obtemos a derivada da longitude sobre a loxodrómia na forma $d\lambda/d\varphi = \tan V \ \sec\varphi$ e, por integração, obtemos a expressão para a longitude $\lambda(\varphi)$ proporcional a uma tangente logarítmica da latitude.

Desse triângulo infinitesimal, podemos também concluir que o elemento da variação do comprimento de arco sobre a loxodrómia ds vem dado por $\cos V\, ds = d\varphi$ [12], e, integrando, obtemos imediatamente que a distância s_{12} entre dois pontos de latitude $\varphi_1 < \varphi_2$ vem dada por

$$s_{12} = \sec V\, (\varphi_2 - \varphi_1).$$

Em particular, podemos concluir que, apesar da loxodrómia de azimute V, $0 < V < \frac{\pi}{2}$, espiralar indefinidamente entre o polo Sul e o polo Norte sem nunca os atingir, tem comprimento total finito e igual a $\rho \sec V\, \pi$, representando por ρ o raio da Terra.

A cronologia da história da compreensão matemática da loxodrómia no século XVII[11] tem aspetos muito interessantes e paralelos à história do aparecimento do Cálculo Infinitesimal, incluindo a descoberta por Henri Bond, referida num compêndio inglês de navegação de 1645, que a "*linha meridiana era análoga a uma escala de tangentes logarítmicos de meio complemento de latitudes*" começando a 45°, por fortuita comparação da tabela de rumos de Wright com a tabela dos logaritmos naturais, introduzidos por J. Napier em 1614. Esta conjetura despertou o interesse dos matemáticos britânicos da época, tendo sido demonstrada por James Gregory, na sua *Exercitationes geometricae* de 1668, mostrando uma relação entre uma certa área sobre um cilindro com a área de um sector hiperbólico, portanto um cálculo logarítmico, e por Isaac Barrow que, nas suas *Lectiones geometricae* de 1770, encontrou a seguinte expressão[13], equivalente à anterior por conhecidas relações trigonométricas,

$$\int_0^{\varphi} \sec\phi\, d\phi = \frac{1}{2}[\log(1 + \sin\,\varphi) - \log(1 - \sin\,\varphi)].$$

Apesar de John Wallis, em 1686, utilizar a sua aritmética dos infinitos para substituir a "coleção de secantes", correspondendo a este integral, por um "agregado" de potências ímpares do seno da latitude divididos pela respetiva potência, a menos de um fator constante γ, tendo obtido uma série

logarítmica da forma $\gamma \sum_{k\geq 0} S^{2k+1}/(2k+1)$, para $S = \sin\varphi$, tal como Barrow, Wallis também não reconheceu na sua fórmula a relação com a tangente logarítmica.

Esta questão é retomada num importante trabalho de 1696 do matemático e astrónomo Edmond Halley (1656-1742)[14]. Reconhecendo a prioridade da demonstração de Gregory, ao pretender dar uma demonstração fácil da "analogia das tangentes logarítmicas com a linha meridiana ou soma de secantes", utiliza de um modo direto a nova "regra de Newton" do recém inventado Cálculo Infinitesimal. Obteve o incremento rigoroso das latitudes da carta de marear e provou que a projeção estereográfica da loxodrómia é a espiral logarítmica, concluindo que "*a soma total de todas as secantes infinitas sobre o arco* φ" é a "*tangente logarítmica, na forma neperiana, para o arco* $45° + \frac{1}{2}\varphi$", a qual obtém na forma de uma série[14],

$$\int_0^{\varphi} \sec\phi\, d\phi = \varphi + \frac{1}{6}\varphi^3 + \frac{1}{24}\varphi^5 + \frac{61}{5040}\varphi^7 + \frac{277}{72576}\varphi^9 + \&c.$$

Como Cajori observou[10], Halley determinou ainda a diferença das longitudes entre duas latitudes φ_1 e φ_2 sobre uma loxodrómia também sob a forma de uma série numérica, o que corresponde na notação moderna ao integral definido entre φ_1 e φ_2, podendo ser ambos os valores diferentes de zero, algo que é inédito até 1696. Halley neste pequeno tratado discute não só vários aspetos teóricos como os numéricos dos cálculos das loxodrómias e, com isso, considera mesmo completa a "doutrina destes rumos espiraliformes que são de grande importância na Arte da Navegação".

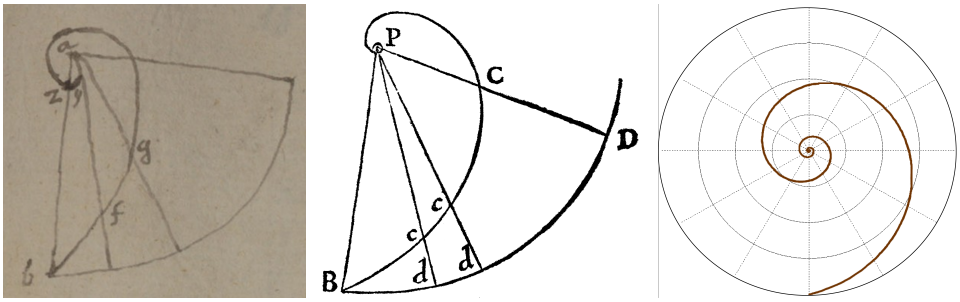


Figura 8: **(a)** Manuscrito de Harriot — De longitudine Helicis 45° (Leconfield MS 240, f 221)[16]. **(b)** A espiral logarítmica ("*proportional spiral*") de Halley de 1696[14]. **(c)** A projeção estereográfica da loxodrómia com azimute de 76°, obtida com o programa interativo LOXO**.

**http://formas-formulas.fc.ul.pt/interactive/loxo/pt/POINSOT2_pt.html

No entanto, os manuscritos de Thomas Harriot (1560-1621), matemático inglês contemporâneo de Wright e Neper, revelaram que os seus cálculos numéricos completos de 1614 das "partes meridionais", ou latitudes crescentes das linhas de rumo, se basearam na conformidade da projeção estereográfica da loxodrómia, na propriedade equiangular da espiral logarítmica e em fórmulas de interpolação, que conheceria, em parte, desde 1594 quando obteve uma resolução aproximada do problema de Nunes-Mercator para a carta de marear[15]. Apesar de não haver referências aos manuscritos de Harriot na segunda metade do século XVII, estes resultados foram conhecidos por alguns matemáticos ingleses, nomeadamente Collins e Newton, antes da publicação de Halley, cuja figura é particularmente semelhante à de um dos manuscritos (Figuras 8 **(a)** e 8 **(b)**).

As linhas de rumo, através da sua história e das suas profundas relações com a cartografia e navegação, constituem um importante exemplo de um problema prático, que foi instrumental na história da expansão europeia e foi relevante na evolução do cálculo infinitesimal e na fundamentação do estudo do planeta Terra pelas ciências matemáticas.

3 Fronteiras livres – dos ventos ao aquecimento da Terra.

A acumulação, durante o século XVII, de métodos particulares para o cálculo de tabelas, quadraturas, tangentes, centros de gravidade e extremos, nomeadamente por Kepler, Galileu, Cavalieri, Torricelli, Pascal, Fermat, Descartes, Wallis e Barrow, entre outros, preparou o salto qualitativo, dado por Isaac Newton (1642-1727) e Gottfried Leibniz (1646-1716) no último quartel desse século, com o estabelecimento da relação geral e da reversibilidade recíproca entre o cálculo diferencial e integral e as múltiplas aplicações aos problemas da Mecânica e da Geometria, que ocorreram nos séculos seguintes.

Se podemos datar num manuscrito de Newton de 1671, apenas publicado num tratado póstumo em 1736[17], o início de um tratamento de uma equação diferencial no seu "Problema II. *Dada a relação das fluxões, encontrar a relação das quantidades fluentes*", apesar da extensa tabela de integrais e de exemplos de curvas, expostos no Problema IX, e de referir aí que para tratar as curvas mecânicas se "devem primeiro transformar-se nas suas equivalentes geométricas", Newton não conhecia nem desenvolveu um método geral para o tratamento das equações diferenciais.

No final do primeiro artigo[18] de Cálculo Diferencial, publicado em 1684,

Leibniz resolve o célebre problema de De Beaune, colocado por este a Descartes em 1638, que consiste em determinar a curva cuja subtangente é dada por uma constante a. Como a subtangente é o segmento entre a intersecção da tangente, a um ponto de ordenada $w(x)$ na curva, com o eixo das abcissas e a projeção x desse ponto neste eixo, Leibniz observa a proporção "w está para a assim como o acréscimo dw está para dx", que se traduz na equação diferencial

$$\frac{dw}{dx} = \frac{1}{a}\, w \qquad \text{(equação de Leibniz).}$$

Na forma $dx = a\ dw/w$, com as variáveis separadas, a sua quadratura foi obtida por Leibniz apenas como a curva "*logarithmica*" na forma "se os x formam uma progressão aritmética, então os w formam uma progressão geométrica", ou seja, através da expressão $x = a\log(w)$. Mais tarde, com Euler (1707-1783), a equação de Leibniz veio a identificar-se com a equação da função exponencial e a aplicar-se a uma multiplicidade de modelos.

Este exemplo de um problema inverso das tangentes resolvido pelos novos métodos do cálculo infinitesimal, tal como muitos outros de caráter geométrico, como a cicloide ou as espirais, ou de caráter mecânico, como o da linha de rumo ou o da órbita dos planetas, está na génese de um enorme desenvolvimento científico que se irá aprofundar no século XVIII com as necessidades da revolução industrial e a formação de um mercado mundial, nomeadamente com os progressos da navegação, da construção naval, da técnica militar, das fontes de energia térmica e hidráulica, pelo lado prático, e com os problemas físico-matemáticos da Mecânica, da Astronomia e da Termodinâmica, pelo lado teórico.

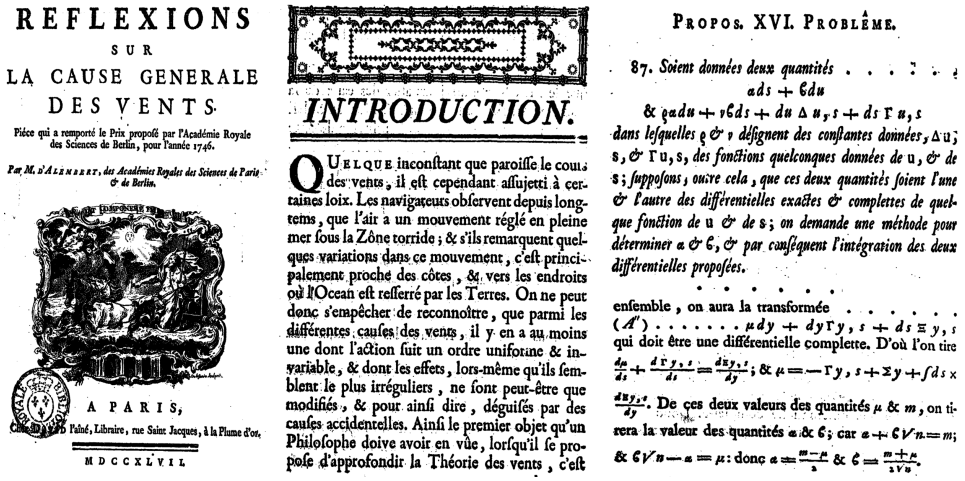
Mas a modelação matemática dos vários subsistemas do planeta Terra, a atmosfera, os oceanos, as superfícies geladas, a estrutura interna, a biosfera, etc., requer sistemas de equações mais complexos. Por exemplo, os sismos requerem a teoria das equações com derivadas parciais, em particular da equação das ondas, que teve origem nos anos 1740's com os trabalhos de Jean D'Alembert (1717-1783). Uma primeira equação das ondas apareceu no seu *Traité de dynamique*, em 1743, para tratar o problema da vibração de uma corda, e as primeiras soluções desse tipo de equações de segunda ordem, as "fórmulas de d'Alembert", foram publicadas em 1747, na monografia *Réflexions sur la cause générale des vents*, que venceu o prémio de 1746 da Academia de Ciências de Berlim[19].

Nesse tratado, d'Alembert visou estudar os ventos como vibrações da atmosfera sob o efeito da rotação da Terra e das atrações lunares e solares, algo que se verificou não ser fisicamente realista, mas desenvolveu a condição necessária das formas diferenciais exatas, utilizada por Euler e Clairaut,

na integração de equações diferenciais de primeira ordem, introduzindo as mudanças de variáveis que conduziram ao método das características. Sem nunca ter escrito nesse tratado a equação de segunda ordem com diferenciais parciais, formulou os problemas com duas expressões diferenciais para a determinação de duas funçõesa $= a(u, s)$ e $b = b(u, s)$, com a condição de cada uma delas ser um diferencial total, o que é equivalente a formular o problema, em notações modernas, na forma de um sistema de duas equações com derivadas parciais de primeira ordem

$$\frac{\partial a}{\partial u} = \frac{\partial b}{\partial s} \quad \text{e} \quad \nu \frac{\partial b}{\partial u} = \varrho \frac{\partial a}{\partial s} + h(u, s),$$

onde h é uma função dada de duas variáveis, ν e ϱ são constantes positivas. Destas equações retiramos a equação das ondas, que, para $\gamma^2 = \varrho/\nu$, se resolve explicitamente ao longo das características $s+\gamma u = C_1$ e $s-\gamma u = C_2$, com soluções da forma $a = f(s + \gamma u) + g(s - \gamma u)$, em função das duas condições iniciais.



REFLEXIONS
SUR
LA CAUSE GENERALE
DES VENTS.

Piéce qui a remporté le Prix proposé par l'Académie Royale des Sciences de Berlin, pour l'année 1746.

Par M. D'ALEMBERT, des Académies Royales des Sciences de Paris & de Berlin.

A PARIS,

MDCCXLVII

INTRODUCTION.

QUELQUE inconſtant que paroiſſe le cours des vents, il eſt cependant aſſujetti à certaines loix. Les navigateurs obſervent depuis longtems, que l'air a un mouvement réglé en pleine mer ſous la Zône torride; & s'ils remarquent quelques variations dans ce mouvement, c'eſt principalement proche des côtes, & vers les endroits où l'Ocean eſt reſſerré par les Terres. On ne peut donc s'empêcher de reconnoître, que parmi les différentes cauſes des vents, il y en a au moins une dont l'action ſuit un ordre uniforme & invariable, & dont les effets, lors-même qu'ils ſemblent le plus irréguliers, ne ſont peut-être que modifiés, & pour ainſi dire, déguiſés par des cauſes accidentelles. Ainſi le premier objet qu'un Philoſophe doive avoir en vûe, lorſqu'il ſe propoſe d'approfondir la Théorie des vents, c'eſt

PROPOS. XVI. PROBLÊME.

87. *Soient données deux quantités* . . . . .
$\alpha ds + \zeta du$
& $\varrho a du + \nu \zeta ds + du\,\Delta u,s + ds\,\Gamma u,s$
dans leſquelles ϱ *&* ν *déſignent des conſtantes données,* $\Delta u,s$, $\Gamma u,s$, *des fonctions quelconques données de* u, *&* s; *ſuppoſons, outre cela, que ces deux quantités ſoient l'une & l'autre des différentielles exactes & complettes de quelque fonction de* u *&* s; *on demande une méthode pour déterminer* α *&* ζ, *& par conſéquent l'intégration des deux différentielles propoſées.*

. . . . . .

enſemble, on aura la transformée . . . . . .
(A') . . . . . $\mu dy + dy\Gamma y,s + ds\,\Sigma y,s$
qui doit être une différentielle complette. D'où l'on tire $\frac{d\mu}{ds} + \frac{d\,\Gamma y,s}{ds} = \frac{d\Sigma y,s}{dy}$; & $\mu = -\Gamma y,s + \Sigma y + \int ds \times \frac{d\Sigma y,s}{dy}$. De ces deux valeurs des quantités μ & m, on tirera la valeur des quantités α & ζ; car $\alpha + \zeta\sqrt{n} = m$; & $\zeta\sqrt{n} - \alpha = \mu$: donc $\alpha = \frac{m-\mu}{2}$ & $\zeta = \frac{m+\mu}{2\sqrt{n}}$.

Figura 9: Frontespício da memória de 1747 de d'Alembert **(a)**, com a respetiva introdução à reflexão sobre a causa geral dos ventos **(b)** e a Proposição XVI, na página 164, onde desenvolve o método da resolução da equação das ondas **(c)**.

Este novo ramo da Análise Matemática foi rapidamente desenvolvido pelos matemáticos contemporâneos, nomeadamente Euler e Lagrange, e tornou-se instrumental nos desenvolvimentos da Física-Matemática dos séculos seguintes. No entanto, a famosa controvérsia sobre a natureza das condições iniciais admissíveis para a equação das ondas que opôs d'Alembert, que só aceitava cordas iniciais com formas regulares, a Euler, que a admitia

inicialmente com cantos, só obteve uma solução satisfatória no século XX com a introdução da noção de soluções generalizadas por Sergey Sobolev (1908-1989) em 1934, quando trabalhava no Instituto de Sismologia da Academia das Ciências da URSS em Leningrado, atual S. Petersburgo. Essa noção foi introduzida independentemente por Jean Leray, também em 1934, ao construir soluções irregulares do sistema de Navier-Stokes, que nomeou "soluções turbulentas", numa evocação ao escoamento das águas do rio Sena sob as pontes de Paris.

Uma outra questão célebre, conjeturada por Newton e Huygens, consistia na forma da Terra, enquanto planeta em rotação, ser um esferoide achatado nos polos pelo efeito da força centrífuga, o que só foi confirmado em meados de 1730, com uma expedição à Lapónia em que participou Clairaut. No seu livro *Théorie de la figure de la terre, tirée des príncipes de l'hydrostatique*, publicado em 1743, Clairaut introduziu integrais curvilíneos para estabelecer a equação da superfície livre de um fluido em rotação em torno do eixo dos x, com velocidade angular ω, e direção radial y, na forma

$$\frac{1}{2}\omega^2 y^2 + \int P\,dx + Q\,dy = \textit{constante}$$

sujeito a forças P e Q verificando a condição $\partial P/\partial y = \partial Q/\partial x$. Este problema, nomeadamente a estabilidade das figuras de equilíbrio de um líquido rodando uniformemente como um corpo rígido em torno de um eixo fixo tornou-se clássico e foi objeto de inúmeros trabalhos de física e de matemática.

Recentemente, estudando as equações de Navier-Stokes com tensão superficial e condições de fronteira livre cinemáticas, Vsevolod Solonnikov mostrou[20] que a positividade da segunda variação do funcional da energia num adequado espaço funcional é uma condição suficiente para a estabilidade de certas figuras de equilíbrio de fluidos viscosos em rotação, mesmo sem simetria, confirmando uma velha conjetura de Poincaré e Lyapunov do final do século XIX.

O interesse dos matemáticos pelos problemas planetários não se limitava aos problemas da Mecânica, como se verifica pelo testemunho de Fourier, autor do célebre tratado físico-matemático *Théorie Analytique de la Chaleur*, de 1822, que reconheceu ter tido a questão das temperaturas terrestres como "um assunto maior dos estudos cosmológicos", que teve em vista no estabelecimento da teoria matemática do calor desde os seus primeiros manuscritos de 1807[21]. Também nos anos 1820's, Fourier analisou o problema do arrefecimento do globo terrestre, concluindo que a Terra deveria ser mais fria se o seu aquecimento dependesse apenas do calor irradiado pelo Sol, considerando a possibilidade de outras fontes, nomeadamente aquela que

hoje identificamos como o efeito estufa, e foi precursor ao assinalar o possível efeito de fatores antropogénicos nas alterações climáticas ao referir que o "*estabelecimento e o progresso da sociedade humana, a ação das forças naturais podem mudar significativamente, e em grande parte, o estado da superfície do solo, a distribuição de águas e os grandes movimentos do ar*", poderiam ser causas possíveis da variação das temperaturas médias ao longo de vários séculos.

EXTRAIT D'UN MÉMOIRE

SUR LE

REFROIDISSEMENT SÉCULAIRE DU GLOBE TERRESTRE.

Bulletin des Sciences par la Société philomathique de Paris, p. 58 à 70; avril 1820.

avons déjà définies. Cela posé, les équations différentielles qui expriment le mouvement de la chaleur dans cette sphère sont

$$\frac{\partial v}{\partial t} = \frac{K}{CD}\left(\frac{\partial^2 v}{\partial x^2} + \frac{2}{x}\frac{\partial v}{\partial x}\right) \quad (7)$$

et

$$K\frac{\partial v}{\partial x} + hv = 0. \quad (8)$$

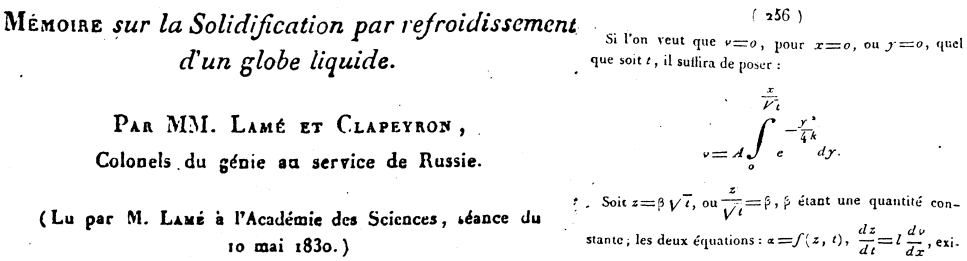
Ces deux équations et l'intégrale (9) que nous allons rapporter ont été données, pour la première fois, dans un Mémoire remis à l'Institut de France, le 21 décembre 1807 (p. 143, 144 et 150). Il est nécessaire

Figura 10: Extrato da memória de 1820 **(a)**, de Joseph Fourier (1768-1830), sobre o arrefecimento secular do globo terrestre com a equação do calor em coordenadas polares e a condição da troca de calor à superfície da Terra com a lei de Newton, estabelecendo que o fluxo de calor é proporcional à temperatura e que se traduz pela equação de Leibniz de 1684. Gravura do jovem Fourier **(b)** antes da sua participação na expedição de Napoleão ao Egito em 1798.

Problemas com fronteiras livres modelam interfaces, por exemplo curvas ou superfícies como a da gota, fixa ou em rotação, que são *a priori* desconhecidas e que separam diferentes regiões no espaço ou no tempo e aparecem normalmente em problemas com mudanças de fase ou com descontinuidades. No planeta Terra existe um grande número de problemas com fronteiras livres[22], como por exemplo a fronteira da sua superfície gelada que é essencial para os modelos climáticos a várias escalas. Estes conduziram a várias teorias, como a dos ciclos de Milankovich, de 1920, sobre o efeito da radiação solar em diferentes latitudes e sobre o albedo, motivaram visões sobre o

controlo do clima, como a de J. von Neumann em 1955, projetos ambiciosos dos anos 1990's de Jacques-Louis Lions com base em modelos matemáticos e simulações computacionais da dinâmica da atmosfera, dos oceanos e dos seus efeitos de acoplamento, e continuam a ser estudados no século XXI lançando novos desafios face à "crescente maré de dados científicos" e os avanços matemáticos e computacionais[23].

O primeiro problema deste tipo foi exatamente motivado por uma aplicação da teoria de Fourier ao arrefecimento do globo terrestre por Lamé e Clapeyron em 1830 e é, hoje em dia, conhecido como a resolução clássica do problema de Stefan a uma fase numa dimensão espacial: a profundidade da frente de solidificação da Terra, suposta homogénea e inicialmente líquida, é dada por $x = \beta\sqrt{t}$, i.e. a fronteira livre varia proporcionalmente a $\sqrt{t}$, e a temperatura $v(x,t)$ é solução da equação do calor na zona sólida variável e função apenas de $y = x/\sqrt{t}$.



Mémoire *sur la Solidification par refroidissement d'un globe liquide.*

Par MM. Lamé et Clapeyron,
Colonels du génie au service de Russie.

(Lu par M. Lamé à l'Académie des Sciences, séance du 10 mai 1830.)

(256)

Si l'on veut que $v=o$, pour $x=o$, ou $y=o$, quel que soit t, il suffira de poser :

$$v = A\int_0^{\frac{x}{\sqrt{t}}} e^{-\frac{y^2}{4k}}\,dy.$$

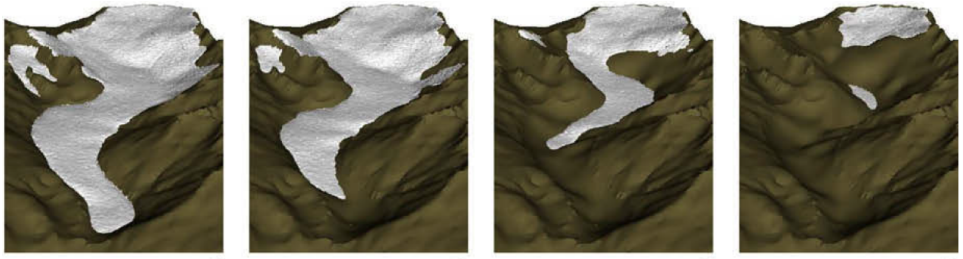
Soit $z=\beta\sqrt{t}$, ou $\frac{z}{\sqrt{t}}=\beta$, β étant une quantité constante ; les deux équations : $\alpha=f(z,t)$, $\frac{dz}{dt}=l\frac{dv}{dx}$, exi-

Figura 11: A Memória de Lamé e Clapeyron **(a)**, com a resolução explícita do modelo unidimensional para a solidificação de um globo líquido **(b)**, constituiu a primeira aplicação da teoria analítica do calor de Fourier a um problema com fronteira livre e foi apresentada na sessão da Academia das Ciências de Paris na semana anterior ao seu falecimento.

Esta classe de problemas com transição de fase, do tipo água-gelo com uma ou duas fases, ficou associado ao nome do físico-matemático esloveno Joseph Stefan, que publicou em 1889 uma série de quatro artigos, discutindo um modelo para a formação de gelo nos oceanos polares. Estes problemas tiveram desenvolvimentos matemáticos importantes no século XX, com a introdução de soluções generalizadas por S. Kamin e O. Oleinik, em 1958, e as soluções das inequações variacionais associadas a uma transformação (de Baiocchi), obtidas por G. Duvaut, em 1973 para o problema a uma fase, e por M. Frémond, em 1974 para o de duas fases. Uma impressionante bibliografia[24], contendo perto de seis mil referências só até ao ano 2000, reflete o grande número de variantes e aplicações dos problemas do tipo Stefan e a importância desses problemas matemáticos.

Utilizando os métodos das inequações variacionais e um modelo com uma aproximação do tipo gelo-raso (*shallow-ice*), foi possível mostrar que a diferença de escalas entre a direção do movimento de um glaciar conduz a problemas bem-postos ultraparabólicos do tipo Stefan[25], a uma fase para a descrição cinemática da superfície do glaciar e a duas fases para a descrição da temperatura. Mais recentemente, um modelo mais complexo num outro contexto da glaciologia, envolvendo as equações de Stokes para o escoamento secular de um manto de gelo em que a linha de contacto com a terra é a fronteira livre, permitiu mostrar que existem soluções matemáticas cujo movimento das linhas de assentamento na terra, correspondendo a um ângulo nulo de contato, podem ser obtidas de forma rigorosa e caraterizadas assintoticamente[26].

Uma modelação matemática e simulação numérica do glaciar do Reno[27], nos alpes suíços, permitiu reconstruir, com notável precisão, o recuo da massa gelada de gelo entre 1874 e 2007, através da resolução aproximada de um problema tridimensional, envolvendo uma variante não linear das equações de Stokes para um fluido viscoso com uma condição de deslizamento na base do glaciar, e da comparação com os registos históricos. Utilizando medições aéreas da geometria do glaciar em 2007 e utilizando três possíveis cenários na evolução das temperaturas, os cálculos computacionais deste modelo permitiram simular as posições do glaciar até 2010. Estas previsões, baseadas em hipóteses realistas para as mudanças climáticas futuras, preveem um significativo recuo do glaciar do Reno para as próximas décadas.



Vol.=1,92 km^3 Vol.=1,38 km^3 Vol.=0,62 km^3 Vol.=0,10 km^3

Figura 12: Visualizações gráficas das simulações do recuo do *Rhonegletscher* para os anos 2025, 2050, 2075 e 2100 num cenário intermédio da simulação de Jouvet et al.[27], mostrando a redução do volume do gelo com a evolução do clima. Um cenário mais severo do aquecimento do clima na Terra prevê o desaparecimento do glaciar no final do século XXI. O interessante filme de 5 minutos *The future of the galciers* mostra o diálogo entre uma glacióloga e um matemático na construção deste modelo e na análise das simulações, e pode ser visto em `https://vimeo.com/58023768`.

Institutos e Sociedades de Matemática em todo o mundo propuseram uma iniciativa global em 2013 sobre a *Matemática do Planeta Terra*. Na sequência de uma proposta de Christiane Rousseau[28], foram sugeridos vários temas para a divulgação matemática a vários níveis, desde os modelos matemáticos associados à descoberta geofísica do planeta, aos aspetos da vida na Terra, passando pela organização das sociedades humanas e os balanços planetários, como as epidemias ou as mudanças climáticas. Com o patrocínio da UNESCO, a 5 de março de 2013 foram apresentados em Paris cerca de uma dezena de módulos que iniciaram uma exposição virtual que continua em construção. Entre esses, os primeiros prémios foram atribuídos, respetivamente, a uma aplicação interativa sobre as deformações geométricas das projeções cartográficas e a dois módulos com problemas com fronteiras livres, uma simulação interativa da dispersão sobre a Europa de uma nuvem de cinzas provocada por um vulcão islandês e um filme sobre o futuro dos glaciares[28].

4 Simetrias – dos cristais à calçada portuguesa.

A primeira descrição dos cinco sólidos regulares aparece no clássico diálogo de *Timeu* de Platão, cerca 360 AC, associada à teoria da composição do mundo pelos quatro elementos, o fogo, a terra, o ar e a água, sendo o dodecaedro associado ao cosmos. Os poliedros platónicos (Figura 13) são os primeiros exemplos de simetria tridimensional e são os únicos regulares, ou seja, os únicos invariantes por uma transformação espacial do respetivo grupo de simetria. O tetraedro (4,6,4) tem 24 simetrias, o cubo (6,12,8) tem 48, tal como o seu dual, o octaedro (8,12,6), e o dodecaedro (12,30,20) e o icosaedro (20,30,12) ambos têm 120 simetrias. Em particular, todos verificam trivialmente a fórmula que Euler estabeleceu nos anos 1750's para todos os poliedros convexos:

$$F - A + V = 2 \qquad \text{(F=Faces, A=Arestas, V=vértices).}$$

O matemático Johannes Kepler (1571-1630) no seu *Mysterium Cosmographicum*, de 1596, a partir das observações astronómicas de Tycho Brahe, concluiu experimentalmente que "*a razão que existe entre os tempos* $[T]$ *de revolução de dois planetas quaisquer está em proporção precisamente sesquialtera com a razão das suas distâncias médias* $[d]$*, isto é, das suas orbes*", o que se traduz na relação $T \sim d^{3/2}$ e é a terceira lei das órbitas dos planetas. Também nessa obra, o astrónomo germânico especulando sobre "*a admirável proporção entre os corpos celestes*" e a simetria do cosmo, comparando as

razões das esferas inscritas e circunscritas nos cinco sólidos platónicos com as razões das órbitas dos seis planetas então conhecidos, e associando-as aos cinco intervalos das respetivas orbes, julgou ter encontrado a chave do universo, Figura 14 **(a)**.

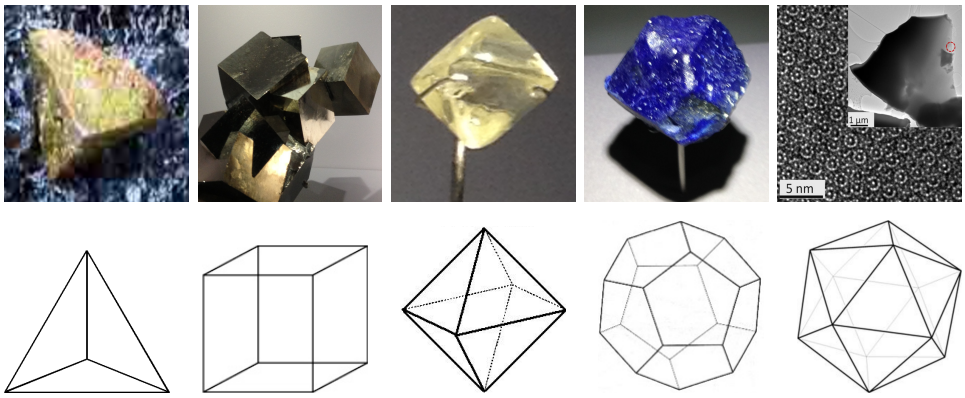


Figura 13: **(a)** A calcopirite ($CuFeS_2$) é o mineral de cobre mais frequente na natureza e cristaliza no sistema tetragonal originando cristais muito próximos do tetraedro e do octaedro. **(b)** A pirite (FeS_2) é um dissulfeto de ferro cujos cristais apresentam formas cúbicas, octaédricas e ainda (pseudo)dodecaédricas, que não tem a simetria pentagonal completa, pois as 30 arestas dividem-se em dois grupos de 24 e 6 com o mesmo comprimento — o piritoedro. **(c)** O diamante é o carbono cristalino (C), que pode aparecer em formas octaédricas como neste cristal originário do Cabo, África do Sul, e existente no MNHN de Paris. **(d)** O dodecaedro é uma forma ideal, podendo ser sugerida pela forma de vários cristais, como a lazurite, que aparece associada à calcite e pirite no lápis-lazuli. **(e)** O icosaedro é uma forma rara na natureza que está associada à simetria dos quasicristais, ou cristais quasiperiódicos, como a icosaedrite ($Al_{63}Cu_{24}Fe_{13}$) e o quasicristal natural $Al_{71}Ni_{24}Fe_5$ (na imagem), ambos existentes num meteorito descoberto em 2009 na Sibéria oriental[36].

A teoria clássica reduz os cristais a reticulados de pontos com simetrias de grupos, tendo por ordem de rotação apenas 2, 3, 4 ou 6 o que impede o icosaedro e o dodecaedro de representarem formas cristalinas com simetria periódica. No entanto, a difração de raios-X numa liga de alumínio-magnésio permitiu ao cientista Daniel Shechtman descobrir experimentalmente, em 1982, uma simetria aperiódica, do tipo da pavimentação de Penrose e inventada vinte anos antes por matemáticos, que deu lugar a uma nova classe de materiais, os quasicristais, e lhe valeu o prémio Nobel da Química de 2011 por ter "revelado um novo princípio para o empilhamento de átomos

e moléculas". Em 1992, a União Internacional de Cristalografia alterou a definição de cristal para incluir os quasicristais.

O problema do empilhamento otimal de esferas e o do seu número máximo num certo espaço, seja a solução piramidal das laranjas no mercado ou das bolas de canhão num compartimento, foi objeto de tratamento por Thomas Harriot nos finais do século XVI. Na sequência da correspondência entre Harriot e Kepler sobre ótica, que suscitaram questões e interpretações atomistas sobre a estrutura da matéria, o matemático germânico publicou em 1611 um pequeno livro *Strena Seu de Nive Sexangula*, no qual descreveu a simetria hexagonal dos flocos de neve e, seguindo a hipótese atomística, formulou o que ficou conhecido como a *conjetura de Kepler* sobre o empilhamento das esferas.

A conjetura de Kepler estabeleceu a não existência de um empilhamento ou empacotamento espacial de esferas com densidade superior ao empilhamento cúbico de faces centradas. Mas existem outros arranjos ótimos de esferas, i.e., com a maior fração possível de espaço ocupado pelas esferas, como o hexagonal, que também atinge a densidade máxima $\pi/3\sqrt{2} \simeq 74\%$, o que só foi provado por Gauss, em 1831. Este matemático também demonstrou que o acondicionamento hexagonal de discos, na forma de favos de mel, com densidade $\pi/2\sqrt{3} \simeq 90,7\%$, é o mais denso entre os reticulados em grelha, mas apenas em 1910 Axel Thue demonstrou que esse é, de facto, o mais denso entre todos os empilhamentos planos possíveis. Após

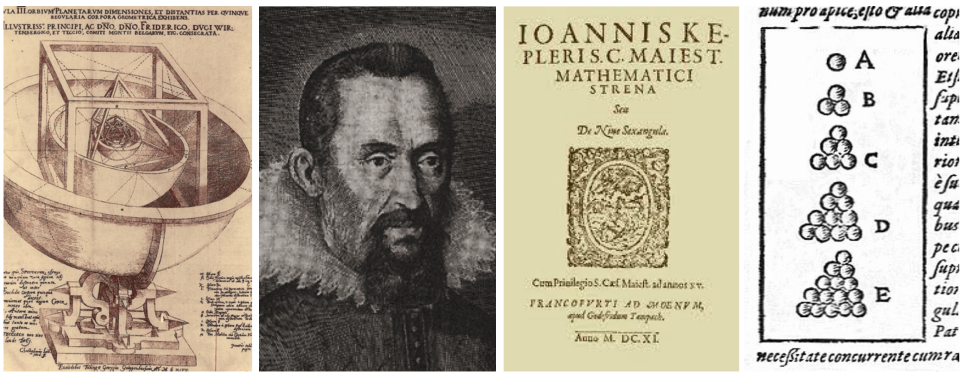


Figura 14: **(a)** Frontespício de ***Mysterium Cosmographicum*** (1596), de J. Kepler **(b)**, com os encaixes nas esferas do cubo (Saturno-Júpiter), tetraedro (Júpiter-Marte), dodecaedro (Marte-Terra), icosaedro (Terra-Vénus) e octaedro (Vénus-Mercúrio). **(c)** Frontespício de ***Strena seu de nive sexângula*** (1611), o livro sobre o "floco de neve de seis pontas", onde Kepler estabelece a sua conjetura de empilhamento otimal das esferas **(d)**, que só foi demonstrada em 1998[29].

quatro séculos de tentativas, uma demonstração da conjetura de Kepler, por exaustão de casos e envolvendo cálculos computacionais para resolução de um gigantesco número de problemas de programação linear, foi anunciada pelo matemático norte-americano Thomas Hales[29] em 1998 e só publicada em 2005, após um longo e difícil processo de revisão e arbitragem por a demonstração requerer a utilização do computador.

A maximização do empacotamento de esferas foi a terceira questão geométrica do 18.º problema, entre os 23 apresentados por David Hilbert (1862-1943) no Congresso Internacional de Matemáticos de 1900, em Paris, sobre o tema genérico da "construção de um espaço euclidiano com poliedros congruentes"[30]. A primeira parte do problema questionou se, no espaço euclidiano n-dimensional, apenas existiria um número finito de grupos de movimentos essencialmente diferentes com uma região fundamental compacta, questão que foi respondida afirmativamente por Bieberbach em 1910. Para a segunda questão, sobre a existência de poliedros que não são regiões fundamentais de grupos de movimentos, mas que por justaposição preenchem todo o espaço euclidiano, foram apresentados contraexemplos por Karl Reinhardt em 1928, para o espaço tridimensional, e por Heinrich Heesch em 1935 para o plano.

Estas questões aparentemente distintas, como a simetria da estrutura hexagonal dos flocos de neve, descoberta por Kepler, ou a sua conjetura sobre o problema do empilhamento de esferas, estimularam o desenvolvimento de modelos físico-matemáticos que permitiram compreender e simular a estrutura da matéria, em particular o alinhamento dos seus átomos e moléculas e os seus correspondentes estados físicos. Apesar dos progressos da Física contemporânea, a misteriosa formação dos flocos de neve e a sua multiplicidade de formas não está ainda completamente compreendida. No entanto, simulações computacionais recentes, baseadas em modelos matemáticos realistas, permitiram recriar os intrincados e complexos padrões da formação do gelo.

A evolução do crescimento de um cristal de neve pode ser simulada como uma variante complexa do tipo "problema de Stefan", tendo em conta a curvatura e as singularidades da superfície do gelo através das equações que descrevem as leis físicas da solidificação da água por arrefecimento. Numa série de trabalhos recentes utilizando modelos do fluxo geométrico com funcionais de energia envolvendo a curvatura das superfícies dos cristais e condições cinéticas[31], uma equipa de matemáticos conseguiu simular os dois principais tipos de crescimento dos flocos de neve: o crescimento facetado, determinado por formas simples como triângulos e hexágonos, e o crescimento dendrítico, com formas ramificadas que se subdividem de forma anisotrópica (Figura 15).

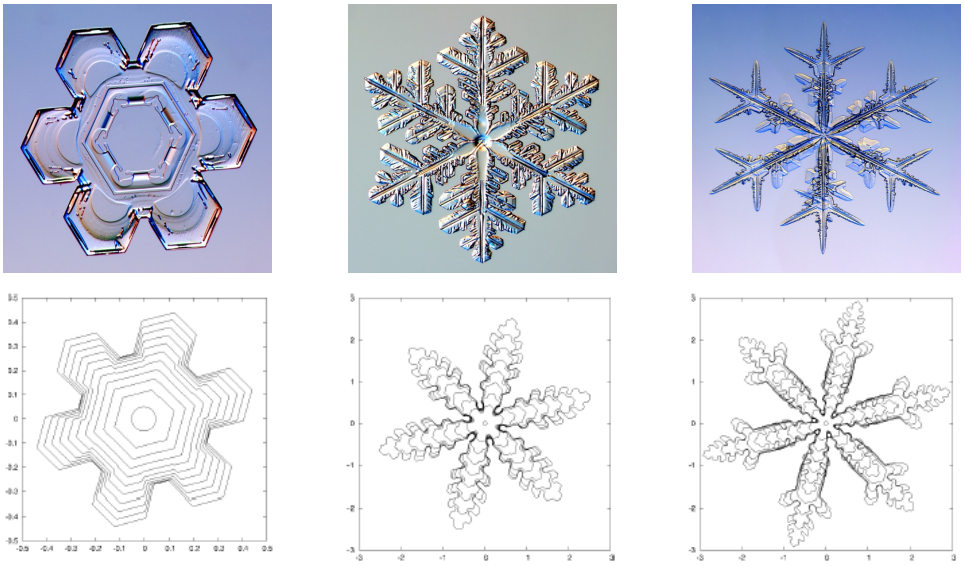


Figura 15: As simulações numéricas de Barrett-Garcke-Nürnberg[31], resolvendo problemas evolutivos com fronteira livre, a duas e a três dimensões, recriaram de modo notável configurações análogas a fotografias de cristais de neve reais[††]. Os métodos matemáticos atuais permitem obter as formas dos cristais de gelo que haviam sido classificados pelo físico japonês Ukichiro Nakaya, que também criou os primeiros flocos de neve artificiais em meados do século XX.

Se os cristais começaram a ser estudados no fim da Renascença e têm, hoje em dia, inúmeras aplicações na ciência dos materiais, na nanotecnologia, na biologia, na medicina e nas ciências farmacêuticas, foi durante o século XIX que os cientistas introduziram o conceito de simetria e usaram matemática para formalizarem a sua classificação. A utilização da noção de grupo em Geometria, por Camille Jordan em 1869, foi motivada e teve impacto na cristalografia, em particular nos trabalhos do matemático Artur Schönflies, em Göttingen, que conduziram à identificação dos 230 grupos de simetria espacial, concluída em 1891 em colaboração com o mineralogista russo Evgraf Fedorov que, em S. Petersburgo, havia iniciado de forma independente a catalogação completa destas simetrias, antecipando algumas das ideias equivalentes às do matemático alemão.

A primeira identificação dos grupos cristalográficos no plano, conduzindo aos 17 padrões, também foi efetuada por Fedorov em 1891, tendo o matemático húngaro George Polya redescoberto este resultado em 1924, como uma consequência de as isometrias do plano euclidiano se limitarem às transla-

[††] http://snowcrystals.com

ções, rotações, reflexões e reflexões deslizantes. A estas simetrias, que na linguagem topológica das orbivariedades (*orbifolds*) se representam, respetivamente, por O, n ($n = 2, 3, 4, 6$), $*$ e X, somam-se as 7 simetrias dos frisos, que resultam dos grupos lineares bidimensionais e também envolvem o símbolo ∞, completando o total das 24 possíveis simetrias do plano euclidiano (Figura 16).

17 TIPOS DE PADRÕES [menus que não envolvem o símbolo ∞]					7 TIPOS DE FRISOS [menus que envolvem o símbolo ∞]		
*632	632	*442	442	*333	*22∞	22∞	2*∞
333	*2222	2222	4*2	3*3	2*22	*∞∞	∞∞
22*	**	*X	XX	22X	O	∞*	∞X

Figura 16: O quadro dos 17 padrões e 7 frisos planos, na notação matemática, com as correspondentes imagens dos pavimentos da cidade de Lisboa cujas calçadas exibem as respetivas simetrias, segundo um baralho de cartas, publicado em 2014 pela associação portuguesa *Ludus,* onde se reconhecem 5 dos padrões então ainda inexistentes e substituídos pelas cartas com os sólidos platónicos. Na calçada do Mosteiro dos Jerónimos aparece o padrão XX, que pode ser também interpretado como repetições do friso ∞X através de uma leitura linear. O padrão $4*2$ foi identificado por Ana C. Silva no jardim Amália Rodrigues em Lisboa[35].

O quadro abstrato da topologia algébrica das orbivariedades permitiu, nos anos 1980's, dar uma explicação geométrica para a existência do número 24 como máximo das simetrias de grupo a duas dimensões, sendo apenas 17 os grupos cristalográficos, os quais têm as limitações de 230, de 4895 e 222097 grupos, respetivamente, para as dimensões 3, 4 e 5. O conceito de orbivariedade, um espaço topológico que generaliza o conceito de variedade (curvas e superfícies em várias dimensões) foi utilizado pelo matemático norte-americano William Thurston, medalhado *Fields* em 1982, na sua

célebre conjetura da geometrização, que estabeleceu que o conjunto das variedades tridimensionais admitem uma espécie de decomposição envolvendo apenas oito geometrias, generalizando o facto, conhecido desde o século XIX, da existência de três possíveis a duas dimensões (Figura 17 **(b)**), a geometria elítica, a parabólica (ou euclidiana) e a hiperbólica. A conjetura de Thurston foi demonstrada em 2003 pelo matemático russo Grigori Perelman, juntamente com a centenária conjetura de Poincaré, o que lhe mereceu a não aceite medalha *Fields* em 2006, atribuída pela União Internacional de Matemática.

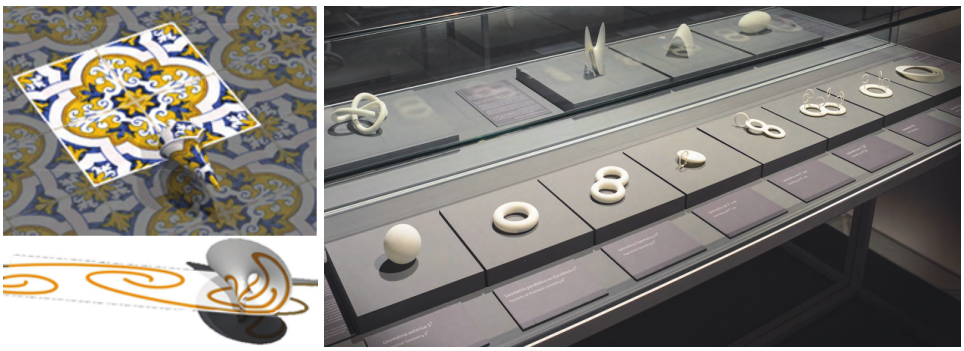


Figura 17: **(a)** Inspirada nas orbivariedades, a associação *Atractor* realizou em 2009 uma interpretação dinâmica e interativa dos 17 padrões e 7 frisos planos utilizando carimbos de simetria[32]. **(b)** Representação, com impressões 3-d, das três geometrias clássicas das variedades bidimensionais e das oito geometrias tridimensionais de Thurston da exposição *Formas & Fórmulas*[‡‡], realizada no Museu da Universidade de Lisboa em 2012.

A simetria não é importante só nas ciências e nas engenharias, pois esteve presente na civilização humana desde os seus primórdios, seja nas formas primitivas de arte de todos os povos, seja nas variadas culturas através da arquitetura e das artes decorativas, como por exemplo a arquitetura grega ou os mosaicos romanos da civilização clássica. Na Idade Média, a cultura islâmica criou formas geométricas com simetrias elaboradas, tendo os caleidoscópios e os mosaicos da Alhambra, em Granada, atingido uma criatividade sem precedentes nos séculos XIII e XIV. Só em 1987, o matemático espanhol José Maria Montesinos integrou a identificação completa dos 17 padrões cristalográficos dos ornamentos da Alhambra na literatura matemática[33], o que desencadeou alguma controvérsia, sobretudo respeitante ao padrão mais raro $*333$ (*p3m1* na notação cristalográfica), o qual, no entanto, existe abstraindo do estuque em causa os ornamentos florais.

[‡‡] http://formas-formulas.fc.ul.pt/

Uma outra utilização notável da simetria está patente na calçada portuguesa. Esta forma de arte pública decorativa é uma original síntese oitocentista dos antigos mosaicos e calçadas dos romanos, que teve origem em Lisboa, quando esta cidade aceitou, em 1848, o projeto de empedramento da praça do Rossio do tenente-general de engenharia Eusébio Furtado[34], com o padrão artístico que se chamou *Mar Largo* (Figura 18 **(a)**). A calçada do *Mar Largo*, que voltou a decorar aquela praça lisboeta em 1976, é uma homenagem à partida das naus da Lisboa renascentista, no arranque da expansão europeia que iniciou a globalização evocada por Camões na epopeia marítima dos portugueses a propósito do rei

> ***Manuel, que a Joane sucedeu***
> ***No Reino e nos altivos pensamentos,***
> ***Logo como tomou do Reino cargo,***
> ***Tomou mais a conquista do mar largo.*** (IV.66)

Esse motivo é uma das 20 simetrias, no total das 24 possíveis (Figura 16), que existem hoje em dia nas ruas e praças da capital portuguesa[35]. O *Mar Largo* esteve presente na Exposição Universal de Paris em 1900 e foi introduzido no Brasil no início do século XX, primeiro em Manaus, no Largo de S. Sebastião, e em seguida no Rio de Janeiro, integrando posteriormente o icónico calçadão de Copacabana e, hoje em dia, está reproduzido em muitos outros sítios espalhados pelo mundo lusófono e não só.

Figura 18: **(a)** Gravura antiga do Rossio do século XIX, em Lisboa, com a calçada "*Mar Largo*" inaugurada em 1849. **(b)** O conceito de um projeto de exposição que explora o encontro da simetria plana, através da calçada portuguesa, com a simetria espacial dos cristais, exemplificado com a associação do topázio ao padrão do "*Mar Largo*", cuja simetria tem uma assinatura $22*$, ou *p2mg* na notação cristalográfica, por ter duas rotações de 180° e uma reflexão de espelho.

Referências

[1] Euclides, *Os elementos*, Trad. Irineu Bicudo, Editora UNESP, São Paulo, 2009.

[2] P. Nunes, *Tratado da Sphera*, Biblioteca Nacional de Portugal, Lisboa, 1537, Reedição, com comentários, pela Academia de Ciências de Lisboa – Fundação Calouste Gulbenkian, Lisboa, 2002.

(http://purl.pt/14445)

[3] L. Pereira da Silva, *A astronomia dos Lusíadas*, Imprensa da Universidade de Coimbra, Coimbra, 1915.

(https://archive.org/details/astronomiadoslus00pere)

[4] Luis de Camões, *Os Lusíadas*, 4.ª ed., Instituto Camões, Lisboa, 2000.

(http://cvc.instituto-camoes.pt/conhecer/biblioteca-digital-camoes/literatura-1/182-os-lusiadas/file.html)

[5] T. Heath, *A History of Greek Mathematics: From Aristarchus to Diophantus*, Vol. II (1921), Dover, New York, 1981.

[6] J. F. Rodrigues, "Alguns aspetos matemáticos do planeta Terra", in *Matemática do Planeta Terra*, IST Press, Lisboa, 2013, Eds. F. P. Costa, J. Buescu, J. T. Pinto., pp. 3–28.

[7] D. Waters, "Portuguese nautical science and the origins of the scientific revolution", *Boletim da Academia Internacional da Cultura Portuguesa*, Vol. 2 (1966), pp. 165–191.

[8] P. Nunes, *Petri Nonni Salaciensis Opera*, Basileae, 1566, Biblioteca Nacional de Portugal, Reedição e tradução, com comentários, pela Academia de Ciências de Lisboa – Fundação Calouste Gulbenkian, Lisboa, 2008 e 2011.

(http://purl.pt/14447)

[9] J. Gaspar e H. Leitão, "How Mercator Did It in 1569: From Tables of Rhumbs to Cartographic Projection", *EMS Newsletter*, #99 (March 2016), pp. 44–49.

[10] F. Cajori, "On an Integration ante-dating the Integral Calculus", *Bibliotheca Mathematica*, Vol. 4 (1915), pp. 312–319.

[11] R. D'Hollander, "La théorie de la loxodromie de Pedro Nunes", *Proceedings of: International Conference Petri Nonii Salaciensis Opera*, Dep. Matemática, Faculdade de Ciências, Universidade de Lisboa, 2003, Eds. L. T. Campos, H. Leitão, J. F. Queiró, pp. 63–111.

[12] F. Gomes Teixeira, "Traité des courbes spéciales remarquables, planes et gauches", Tome II, *in Obras*, Vol. V, Coimbra, 1909.

[13] V. Frederick Rickey e Philip M. Tuchinsky, "An Application of Geography to Mathematics: History of the Integral of the Secant", *Mathematics Magazine*, Vol. 53, No. 3 (1980), pp. 162–166.

[14] E. Halley, "An easie demonstration of the analogy of the logarithmick tangents to the meridian line or sum of secants: with various methods for computing the same to the utmost exactness", *Philos. Trans., Roy. Soc. London*, Vol. 19 (1696), pp. 202–214.

[15] J. V. Pepper, "Harriot's calculation of the meridional parts as logarithmic tangents", *Arch. Hist. Exact Sci.*, Vol. 4 (1968), pp. 359–413.

[16] http://echo.mpiwg-berlin.mpg.de/content/scientific_revolution/harriot

[17] I. Newton, *O método das fluxões e das séries infinitas*, Associação dos Professores de Matemática/Ed.Prometeu, Lisboa, 2004.
(Original inglês: 1736, em
https://archive.org/details/methodfluxionsa00newtgoog).

[18] G. Leibniz, "Nova Methodus pro Maximis et Minimis", *Acta Eruditorum*, Vol. 3 (1684), tradução em inglês em *A Source Book in Mathematics, 1200-1800*, Ed. D.J.Struik, Princeton Univ Press, 1969, pp. 272–280.

[19] S. S. Demidov, "Création et développement de la théorie des équations différentielles aux dérivées partielles dans les travaux de J. d'Alembert", *Rev. Hist. Sci.*, Vol. XXXV (1982), pp. 3–42.

[20] V. A. Solonnikov, "On the stability of nonsymmetric equilibrium figures of a rotating viscous incompressible liquid", *Interfaces and Free Boundaries*, Vol. 6 (2004), pp. 461–492.

[21] J. Fourier, "Remarques générals sur la température du globe terrestre et des espaces planétaires", *Annales de Chimie et de Physique*, Vol. 27 (1824), pp. 136–167 (in Oeuvres, II, 97–125).

[22] J. I. Diaz (Ed.), "The mathematics of models for climatology and environment", *Proceedings of the NATO Advanced Institute, January 11-21, 1995*, Tenerife, Spain, Springer-Verlag Berlin, 1997.

[23] J. C. K. T. Jones, "Will climate change mathematics (?)", *IMA J. Appl. Math.*, Vol. 76, No. 3 (2011), pp. 353–370.

[24] D. Tarzia, "A Bibliography on moving-free boundary problems for the heat-diffusion equation. The Stefan and Related Problems", *MAT, Series A: Conferencias, seminarios y trabajos de matemática. Univ Austral, Rosario*, #2 (2000), pp. 1–297.

[25] J. F. Rodrigues e L. Santos, "Some Free Boundary Problems in Theoretical Glaciology", in *NATO ASI Series I: Global Environmental Change*, No. 48, Springer-Verlag, Heidelberg, 1997, pp. 337–365.

[26] M. A. Fontelos e A. I. Muñoz, "A free boundary problem in glaciology: The motion of grounding lines", *Interfaces and Free Boundaries*, Vol. 9 (2007), pp. 67–93.

[27] G. Jouvet et al., "Numerical simulation of Rhonegletscher from 1874 to 2100", *J. Comput. Phys.*, Vol. 228 (2009), pp. 6426–6439.

[28] C. Rousseau, "Four themes with potential examples of modules for a virtual exhibition on the Mathematics of Planet Earth", *Centro Internacional de Matemática Bulletin*, #30 (Jul 2011), pp. 31–32.

(http://www.cim.pt , e a sequência no portal https://imaginary.org/exhibition/mathematics-of-planet-earth)

[29] T. C. Hales, "Cannonballs and Honeycombs", *Notices of the AMS*, Vol. 47, No. 4 (April 2000), pp. 440–449.

[30] J. Milnor, "Hilbert's problem 18: on crystalographich groups, fundamental domains, and on sphere packing", *Proceedings of Symposia in Pure Mathematics*, Vol. 28, Part 2 (1976), pp. 491-506.

[31] J. W. Barrett, H. Garcke e R. Nürnberg, "Numerical computations of faceted pattern formation in snow crystal growth", *Physical Review E*, Vol. 86, 011604 (2012).

[32] M. Arala Chaves, "Atractor", in *Raising Public Awareness of Mathematics*, Springer-Verlag, Berlin Heidelberg, 2012, Eds. E. Behrends, N. Crato & J.F. Rodrigues, pp. 109–134.

[33] J. M. Montesinos, *Classical Tesselations and Three-Manifolds*, Springer, New-York, 1987.

[34] E. Bairrada, *Empedrados artísticos de Lisboa*, Câmara Municipal de Lisboa, 1985.

[35] A. C. Silva, *Calçadas de Portugal. Simetria passo a passo*, CTT Correios de Portugal, Lisboa, julho 2016.

[36] L. Bindi et al., "Natural quasicrystal with decagonal symmetry", *Scientific Reports*, Vol. 5, 9111 (2015).

(http://www.nature.com/articles/srep09111)

Nota sobre o Melhor Par Aproximante de Duas Variedades Lineares Enviesadas

M. A. Facas Vicente, *José Vitória*
Departamento de Matemática, Universidade de Coimbra
e-mail: vicente@mat.uc.pt

P. Saraiva
CMUC & CeBER, Universidade de Coimbra
e-mail: psaraiva@fe.uc.pt

P. D. Beites[1]
CMA & DM, Universidade da Beira Interior
e-mail: pbeites@ubi.pt

Armando Gonçalves
IPC, ESEC, MEM, Coimbra
e-mail: adsgoncalves@esec.pt

Resumo: Nesta nota obtém-se o melhor par aproximante de duas variedades lineares enviesadas após determinar a projeção de um ponto sobre uma variedade linear. Concretamente, calculam-se por duas vezes os vetores de norma mínima, adequados em cada caso, pertencentes a duas variedades lineares paralelas. Obtêm-se ainda expressões para as referidas projeções envolvendo a inversa de Moore-Penrose de cada uma das matrizes associadas às variedades lineares consideradas (na sua formulação matricial).

Abstract: The best approximation pair of two skew linear varieties is obtained after getting the projection of a point onto a linear variety. Concretely, we compute twice the minimum norm vectors of two, adequate each time, parallel linear varieties. In addition, we obtain expressions for the mentioned projections by means of the Moore-Penrose inverse of each of the matrices associated with the linear varieties considered (in its matrix version).

palavras-chave: variedades lineares enviesadas; melhor par aproximante; teoria da aproximação; vetor de norma mínima; inversa de Moore-Penrose.

keywords: skew linear varieties; best approximation pair; approximation theory; minimum norm vector; Moore-Penrose inverse.

[1] Financiada por Fundação para a Ciência e a Tecnologia (Portugal), projeto UID/MAT/00212/2013, e Ministerio de Economía y Competitividad (España), projeto MTM2013-45588-C3-1-P.

1 Introdução

A presente nota aborda dois problemas clássicos de Geometria Analítica Euclidiana n-dimensional – determinar a projeção ortogonal de um ponto sobre uma dada variedade linear e obter uma expressão para a distância entre duas variedades lineares enviesadas – de um modo que justifica revisitá-los. Com efeito, no primeiro resultado principal desta nota, recorremos aos vetores de norma mínima pertencentes a duas variedades lineares paralelas. O segundo problema principal diz respeito à obtenção do melhor par aproximante de pontos pertencentes a duas variedades lineares enviesadas, mediante a resolução de um sistema de equações lineares, construído por aplicação do primeiro resultado. Ao obtermos o melhor par aproximante, vamos mais longe do que o habitual, pois em [2], [4] e [7] apenas se focou o problema da distância entre duas variedade lineares. De facto, a julgar por estas referências, os pontos que materializam tal distância não parecem ter sido considerados.

O problema da determinação da (mínima) distância euclidiana entre duas variedades lineares foi também recentemente tratado em [6] e [5], aplicando, respetivamente, a Teoria de Gram e a inversa de Moore-Penrose de uma matriz particionada. Ainda que nestas referências se exiba o melhor par aproximante de pontos, as abordagens aí utilizadas são distintas da que se aplica na presente nota. Além disso, os resultados agora apresentados situam-se no contexto dos problemas de projeções sobre certos conjuntos convexos, baseados portanto em resultados sobre existência, unicidade e caracterização de soluções para problemas de melhor aproximação. Também no presente trabalho se recorre à inversa de Moore-Penrose, a qual constitui um excelente instrumento teórico (veja-se, *e.g.*, [1] e [11]). Contudo, é justo referir que, do ponto de vista numérico, tal ferramenta contém algumas fragilidades: a inversa de Moore-Penrose não é contínua e não é computacionalmente estável [10, pp. 423–424].

O presente trabalho está organizado do seguinte modo. Começamos por enunciar algumas definições, notações e resultados, coligidos na secção 2. Na secção 3, apresentamos os resultados sobre a projeção de um ponto sobre uma variedade linear. Na secção 4, recorrendo aos resultados da anterior secção, apresentamos um método original para obter o melhor par aproximante de duas variedades lineares enviesadas.

2 Preliminares

Ao longo da presente nota, $\mathbb{R}^n$ designa o espaço vetorial real n-dimensional euclidiano usual. O *produto interno* será denotado por $\bullet$ e define-se, como habitualmente, do seguinte modo: dados $\vec{p} = \begin{bmatrix} p_1 & p_2 & \cdots & p_n \end{bmatrix}^T$ e $\vec{q} = \begin{bmatrix} q_1 & q_2 & \cdots & q_n \end{bmatrix}^T \in \mathbb{R}^{n\times 1}$, onde T designa a transposta, tem-se

$$\vec{p} \bullet \vec{q} = \sum_{i=1}^{n} p_i q_i.$$

Recorde-se também que a *norma euclidiana* de $\vec{p}$ é dada por

$$\|\vec{p}\| = \sqrt{\vec{p} \bullet \vec{p}}.$$

Além disso, considerando o espaço afim associado a $\mathbb{R}^n$, a extremidade de cada vetor posicional (relativamente à origem das coordenadas) será identificada com o próprio vetor.

Dados um ponto $\vec{a}$ de $\mathbb{R}^n$ e um subespaço N de $\mathbb{R}^n$, de dimensão m, o conjunto

$$V_{\vec{a}} = \vec{a} + N$$

diz-se uma *variedade linear*, [9]. Uma variedade linear pode então ser entendida como o resultado da translação de um subespaço.

De entre as várias maneiras de representar uma variedade linear (ver [3], [8] e [9]), recorreremos ainda à *versão matricial*, a qual possibilita a utilização do método cartesiano. Assim, dada uma variedade linear $V_{\vec{a}} = \vec{a} + N$, para alguma matriz $A \in \mathbb{R}^{p\times n}$, com característica completa por linhas, e para algum $\vec{c} \in \mathbb{R}^p$ com $p \in \{1, \ldots, n\}$, tem-se

$$V_{\vec{a}} = \{\vec{x} \in \mathbb{R}^n : \ A\vec{x} = \vec{c}\}.$$

Relativamente à representação dada na definição, é importante reter que

$$N = \mathcal{N}(A) \text{ (i.e., o núcleo de } A)$$

e que p é a codimensão de N ($p \leq n$). Tendo isto em conta, doravante iremos identificar uma variedade linear $\mathcal{L}$ dos seguintes dois modos:

$$\mathcal{L} := \{\vec{x} \in \mathbb{R}^n : L\vec{x} = \vec{c}\} = \vec{l}_0 + \mathcal{N}(L),$$

apenas optando por uma delas se a outra se verificar desnecessária.

Sejam dados pontos $\vec{l}_0, \vec{m}_0 \in \mathbb{R}^n$ e consideremos duas variedade lineares

$$\mathcal{L} = \ \vec{l}_0 + \mathcal{N}(L) \text{ e } \mathcal{M} = \ \vec{m}_0 + \mathcal{N}(M).$$

Seguindo [4], as variedades $\mathcal{L}$ e $\mathcal{M}$ dizem-se *paralelas* se $\mathcal{N}(L) = \mathcal{N}(M)$.

Por outro lado, admitamos que as variedades $\mathcal{L}$ e $\mathcal{M}$ são tais que

$$\dim \ \mathcal{N}(L) + \dim \ \mathcal{N}(M) < n.$$

Nesse caso, tais variedades dizem-se *enviesadas* se $\mathcal{L} \cap \mathcal{M} = \emptyset$ e $\mathcal{N}(L) \cap \mathcal{N}(M) = \{\vec{0}\}$, [4].

A inversa de *Moore-Penrose*[2] de uma matriz $M \in \mathbb{R}^{m\times n}$ é a única matriz $M^\dagger \in \mathbb{R}^{n\times m}$ satisfazendo

$$MM^\dagger M = M,\ M^\dagger MM^\dagger = M^\dagger,\ (M^\dagger M)^T = M^\dagger M \text{ e } (MM^\dagger)^T = MM^\dagger.$$

Em particular, se M é uma matriz com característica completa por linhas, então

$$M^\dagger = M^T\left(MM^T\right)^{-1}.$$

No que se segue, assumiremos que todas as matrizes envolvidas têm característica completa por linhas.

No final desta nota, exibiremos o melhor par aproximante $\left(\vec{x^*}, \vec{y^*}\right) \in \mathcal{L}\times\mathcal{M}$ de duas variedades lineares enviesadas $\mathcal{L}$ e $\mathcal{M}$, querendo isto dizer que

$$\left\|\vec{x^*} - \vec{y^*}\right\| = d\left(\mathcal{L},\mathcal{M}\right),$$

onde $d\left(\mathcal{L},\mathcal{M}\right)$ designa a distância euclidiana entre as variedades $\mathcal{L}$ e $\mathcal{M}$.

3 Projeção de um Ponto sobre uma Variedade Linear

Considere uma variedade linear

$$\mathcal{L} = \vec{l_0} + \mathcal{N}(L) = \{\vec{x}\in\mathbb{R}^n : L\vec{x} = \vec{c}\},$$

onde $\vec{l_0}$ é um ponto fixo de $\mathbb{R}^n$. Procura-se determinar $\mathbb{P}_{\mathcal{L}}\left(\vec{m}\right)$, a projeção (ortogonal) de um dado ponto externo $\vec{m}\in\mathbb{R}^n$ sobre $\mathcal{L}$, isto é, pretende-se determinar o ponto $\vec{l}\in\mathcal{L}$ que minimize a distância $d\left(\vec{l},\vec{m}\right)$.

Para atingir o referido objetivo, propõe-se um método original que recorre duas vezes a um resultado que fornece o vetor de norma mínima de uma variedade linear com codimensão finita. Tal resultado pode ser consultado, *e.g.*, em [3, Theorem 9.26, p. 215], [8, Théorème 2.2.5, p. 45] e [9, Theorem 2, p. 51], assumindo a seguinte forma.

Teorema 3.1. *Considere uma variedade linear $\mathcal{L} = \vec{l_0} + \mathcal{N}(L)$ de $\mathbb{R}^n$. Uma condição necessária e suficiente para que $\vec{l}\in\mathcal{L}$ seja a melhor aproximação de $\vec{m}\notin\mathcal{L}$ em $\mathcal{L}$ é que $\vec{l}-\vec{m}$ seja ortogonal a $\mathcal{N}(L)$.*

Considerando esta caracterização do ponto que constitui a melhor aproximação de uma variedade linear, temos então o principal resultado desta secção.

[2] Na referência [1] é detalhada a teoria das inversas generalizadas de matrizes.

Teorema 3.2. *Sejam $\mathcal{L} = \vec{l}_0 + \mathcal{N}(L)$ uma variedade linear e $\vec{m} \notin \mathcal{L}$ um dado ponto fixo. A projeção $\mathbb{P}_{\mathcal{L}}(\vec{m})$ é dada por*

$$\mathbb{P}_{\mathcal{L}}(\vec{m}) = \vec{x}_0 + \vec{m} - \vec{x_0'}, \tag{1}$$

onde $\vec{x}_0$ e $\vec{x_0'}$ são, respetivamente, os vetores de norma mínima das variedades lineares paralelas $\mathcal{L}$ e $\mathcal{L}' = \vec{m} + \mathcal{N}(L)$.

Demonstração. Tendo em conta o Teorema 3.1, temos de provar que $\vec{m} - \mathbb{P}_{\mathcal{L}}(\vec{m})$ é ortogonal ao subespaço $\mathcal{N}(L)$.

De facto, tem-se

$$\mathbb{P}_{\mathcal{L}}(\vec{m}) - \vec{m} = \vec{x}_0 - \vec{x_0'}.$$

Mas $\vec{x}_0 - \vec{x_0'}$ é ortogonal a $\mathcal{N}(L)$ em virtude do facto de $\vec{x}_0$ e $\vec{x_0'}$ serem ambos ortogonais a $\mathcal{N}(L)$ (ver [9, Theorem 1, p. 64]). □

Teorema 3.3. *Considere uma variedade linear*

$$\mathcal{L} = \{\vec{x} \in \mathbb{R}^n : \ L\vec{x} = \vec{c}\}.$$

1. *O vetor de norma mínima de $\mathcal{L}$ é dado por $\vec{x}_0 = L^\dagger \vec{c}$;*
2. $\mathbb{P}_{\mathcal{N}(L)}(\vec{y}) = \left(I - L^\dagger L\right)\vec{y}$;
3. $\mathbb{P}_{\mathcal{L}}(\vec{y}) = \left(I - L^\dagger L\right)\vec{y} + L^\dagger \vec{c}$.

Demonstração. Relativamente a 1. e 2., consulte [10, pp. 423 e 434–435].

No que diz respeito a 3., seja $\vec{x}_0$ o vetor de norma mínima de $\mathcal{L}$. Por definição de projeção ortogonal, temos

$$\mathbb{P}_{\mathcal{L}}(\vec{y}) = \vec{x}_0 + \mathbb{P}_{\mathcal{N}(L)}(\vec{y} - \vec{x}_0),$$

de onde resulta

$$\mathbb{P}_{\mathcal{L}}(\vec{y}) = \left(I - \mathbb{P}_{\mathcal{N}(L)}\right)\vec{x}_0 + \mathbb{P}_{\mathcal{N}(L)}(\vec{y}). \tag{2}$$

Por 1. e 2. e pela definição de $L^\dagger$ tem-se:

$$\left(I - \mathbb{P}_{\mathcal{N}(L)}\right)\vec{x}_0 = \left[I - \left(I - L^\dagger L\right)\right]\vec{x}_0 = L^\dagger L\vec{x}_0 = L^\dagger L L^\dagger \vec{c} = L^\dagger \vec{c}.$$

A expressão em 3. obtém-se após substituição em (2) e aplicação de 2. □

Tendo em conta o resultado anterior, podemos agora exprimir o Teorema 3.2 à custa da inversa de Moore-Penrose aplicada à representação dos vetores de norma mínima das variedades lineares paralelas ali consideradas.

Corolário 3.4. *Nas condições do Teorema 3.2, sendo* $\mathcal{L} = \{\vec{x} \in \mathbb{R}^n : L\vec{x} = \vec{c}\}$ *a representação matricial da variedade* $\mathcal{L}$*, tem-se:*

$$\mathbb{P}_{\mathcal{L}}(\vec{m}) = \left(I - L^{\dagger}L\right)\vec{m} + L^{\dagger}\vec{c}. \tag{3}$$

Demonstração. De modo a exprimir os vetores de norma mínima em (1), vamos recorrer à notação matricial das variedades $\mathcal{L}$ e $\mathcal{L}'$. Assim, dado que tais variedades são paralelas, e tendo em conta a representação de $\mathcal{L}$, podemos assumir que

$$\mathcal{L}' = \{\vec{x} \in \mathbb{R}^n : L\vec{x} = \vec{d}\},$$

para determinados $\vec{d} \in \mathbb{R}^p$. Uma vez que $\mathcal{L}'$ passa por $\vec{m}$, tem-se $L\vec{m} = \vec{d}$. Deste modo, por aplicação de 1. do Teorema 3.3 tem-se:

$$\vec{x}_0 = L^{\dagger}\vec{c} \quad \text{e} \quad \vec{x_0'} = L^{\dagger}\vec{d} = L^{\dagger}L\vec{m}.$$

A expressão (3) obtém-se após substituição de $\vec{x}_0$ e $\vec{x_0'}$ em (1), seguida de simplificação. □

4 Melhor Par Aproximante

Nesta última secção, procuramos obter o mais curto segmento de reta

$$\left[\vec{x^*}\ \vec{y^*}\right]$$

ligando duas variedades lineares enviesadas $\mathcal{L}$ e $\mathcal{M}$.

Para tal, aplicamos duas vezes o Teorema 3.2, levando em conta que, no presente caso, os pontos externos a considerar são pontos genéricos das variedades lineares $\mathcal{L}$ e $\mathcal{M}$, respetivamente.

Teorema 4.1. *Considere duas variedades lineares enviesadas,* $\mathcal{L} = \vec{l}_0 + \mathcal{N}(L)$ *e* $\mathcal{M} = \vec{m}_0 + \mathcal{N}(M)$*. Então, o melhor par aproximante* $\left(\vec{x^*},\ \vec{y^*}\right) \in \mathcal{L} \times \mathcal{M}$ *das variedades lineares* $\mathcal{L}$ *e* $\mathcal{M}$ *é a solução do sistema de equações lineares*

$$\begin{cases} \mathbb{P}_{\mathcal{L}}(\vec{g}_{\mathcal{M}}) = \vec{g}_{\mathcal{L}} \\ \mathbb{P}_{\mathcal{M}}(\vec{g}_{\mathcal{L}}) = \vec{g}_{\mathcal{M}} \end{cases},$$

onde $\vec{g}_{\mathcal{L}}$ *e* $\vec{g}_{\mathcal{M}}$ *são pontos genéricos pertencentes às variedades lineares* $\mathcal{L}$ *e* $\mathcal{M}$*, respetivamente.*

Demonstração. Comece-se por aplicar duas vezes o Teorema 3.2, uma vez que o ponto genérico $\vec{g}_{\mathcal{L}}$ é externo a $\mathcal{M}$ e que o ponto genérico $\vec{g}_{\mathcal{M}}$ é externo a $\mathcal{L}$. Depois, basta ter em conta que o vetor $\overrightarrow{\vec{x^*}\vec{y^*}}$ — cujas extremidades são $\vec{x^*} \in \mathcal{L}$ e $\vec{y^*} \in \mathcal{M}$ — é simultaneamente ortogonal aos subespaços $\mathcal{N}(L)$ e $\mathcal{N}(M)$. □

Para terminar, vamos exibir o melhor par aproximante obtido no teorema anterior em que os pontos se exprimem à custa da inversa de Moore-Penrose.

Corolário 4.2. *Nas condições do Teorema 4.1, admita que as variedades lineares $\mathcal{L}$ e $\mathcal{M}$ são dadas na forma matricial do seguinte modo:*

$$\mathcal{L} = \{\vec{x} \in \mathbb{R}^n : L\vec{x} = \vec{c}\} \text{ e } \mathcal{M} = \{\vec{x} \in \mathbb{R}^n : M\vec{x} = \vec{d}\}.$$

Então o melhor par aproximante $\left(\vec{x^*},\ \vec{y^*}\right) \in \mathcal{L} \times \mathcal{M}$ é a solução do sistema

$$\begin{cases} \vec{x} = \left(I - L^\dagger L\right)\vec{y} + L^\dagger \vec{c} \\ \vec{y} = \left(I - M^\dagger M\right)\vec{x} + M^\dagger \vec{d}. \end{cases}$$

Demonstração. Sendo $\vec{x} \in \mathcal{L}$ e $\vec{y} \in \mathcal{M}$ tais que

$$\begin{cases} \vec{x} = \mathbb{P}_{\mathcal{L}}(\vec{y}) \\ \vec{y} = \mathbb{P}_{\mathcal{M}}(\vec{x}), \end{cases}$$

resulta de imediato por aplicação do Corolário 3.4 que

$$\begin{cases} \vec{x} = \left(I - L^\dagger L\right)\vec{y} + L^\dagger \vec{c} \\ \vec{y} = \left(I - M^\dagger M\right)\vec{x} + M^\dagger \vec{d}. \end{cases}$$

□

Agradecimentos

Os autores agradecem ao revisor, cujas importantes observações os levaram a refazer o artigo, tanto na forma como no conteúdo.

Referências

[1] A. Ben-Israel e T. N. E. Greville, *Generalized Inverses: Theory and Applications*, Springer-Verlag, New York, 2003.

[2] A. Dax, "The Distance between Two Convex Sets", *Linear Algebra and its Applications*, Vol. 416, No. 1 (2006), pp. 184–213.
DOI: https://doi.org/10.1016/j.laa.2006.03.022

[3] F. Deutsch, Best *Approximation in Inner Product Spaces*, CMS books in mathematics, 7, Springer-Verlag, New York, 2001.

[4] A. M. DuPré e S. Kass, "Distance and Parallelism between Flats in $\mathbb{R}^n$", *Linear Algebra and its Applications*, Vol. 171 (1992), pp. 99-107.
DOI: https://doi.org/10.1016/0024-3795(92)90252-6

[5] M. A. Facas Vicente, Armando Gonçalves e José Vitória, "Euclidean distance between two linear varieties", *Applied Mathematical Sciences*, Vol. 8, No. 21 (2014), pp. 1039-1043.
DOI: http://dx.doi.org/10.12988/ams.2014.311656

[6] Armando Gonçalves, M. A. Facas Vicente e José Vitória, "Optimal pair of two linear varieties", *Applied Mathematical Sciences*, Vol. 9, No. 12 (2015), pp. 593-596.
DOI: http://dx.doi.org/10.12988/ams.2015.410869

[7] J. Gross e G. Trenkler, "On the Least Squares Distances between Affine Subspaces", *Linear Algebra and its Applications*, Vols. 237/238 (1996), pp. 269-276. DOI: http://dx.doi.org/10.1016/0024-3795(95)00648-6

[8] P.-J. Laurent, *Approximation et Optimisation*, Hermann, Paris, 1972.

[9] D. G. Luenberger, *Optimization by Vector Space Methods*, J. Wiley, New York, 1969.

[10] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.

[11] G. Wang, Yimin Wei e Sanzheng Qiao, *Generalized Inverses: Theory and Computations*, Science Press, Beijing, New York, 2004.

Soluções interpolatórias de equações às diferenças lineares

Mário M. Graça
Departamento de Matemática
Instituto Superior Técnico, Universidade de Lisboa
e-mail: mgraca@math.tecnico.ulisboa.pt

Resumo: Dada uma equação às diferenças linear e homogénea, de coeficientes constantes, começamos por construir uma função contínua interpolatória da respectiva solução. Tal função é posteriormente usada para definir uma função paramétrica, com valores em $\mathbb{R}^2$, que designamos por 'retrato de fase'. Embora com algum abuso de linguagem, a definição que propomos para *retrato de fase de uma equação às diferenças* revela-se interessante para efeitos de estudo da dinâmica das soluções de equações às diferenças, tal como acontece com sistemas de equações diferenciais onde os denominados *retratos de fase* constituem uma técnica bem conhecida. Como ilustração, apresentam-se alguns exemplos de construção da referida função interpolatória bem como de retratos de fase associados a certas equações às diferenças de segunda ordem, em particular equações ligadas a sucessões do tipo Fibonacci.

Abstract: Given a linear and homogeneous difference equation, with constant coefficients, we begin by constructing a continuous function which is interpolatory of the difference equation solution. This function leads to a $\mathbb{R}^2$ valued parametric function which we call, with some language abuse, 'phase portrait' of the difference equation. The 'phase portrait' proves to be an interesting tool in order to understand the dynamics of the solutions of a difference equation, similarly to the so called *phase portrait* in the context of systems of ordinary differential equations. As an illustration we present some examples where the referred interpolatory function is considered as well as *phase portraits* of certain second order difference equations connected to some Fibonacci type sequences.

Palavras-chave: Equação às diferenças, interpolação, retrato de fase, sucessões de Fibonacci.

1 Introdução

Dados k valores iniciais $x_0, x_1, \ldots, x_{k-1}$, equações às diferenças (de ordem $k \geq 1$, lineares, homogéneas e de coeficientes constantes),

$$x_{n+k} = a_{k-1}\, x_{n+(k-1)} + a_{k-2}\, x_{n+(k-2)} + \ldots + a_1\, x_{n+1} + a_0\, x_n, \qquad (1)$$

onde $a_0 \neq 0$, $a_1, a_2, \ldots, a_{k-1} \in \mathbb{R}$, ocorrem como modelos evolutivos em inúmeras aplicações, em particular como versão discreta de certas equações diferenciais ordinárias com designação análoga. Uma vez que as soluções destas últimas são geralmente funções contínuas num certo domínio real, é nosso objectivo inicial construir uma função (pelo menos) contínua $g : \mathbb{R}_0^+ \mapsto \mathbb{R}$, que interpole os dados, isto é, tal que g seja interpoladora do conjunto infinito de pontos do plano $(0, x_0), (1, x_1), (2, x_2), \ldots$. A abordagem que propomos, embora elementar, constitui porventura uma ponte com o chamado 'método das funções geradoras' envolvendo funções e séries complexas ([3], Ch. 7.4).

O ponto de vista interpolatório (Secção 2) oferece vantagens, quer do ponto de vista teórico como computacional, sugerindo conexões entre temas tão diversos como sucessões recorrentes, sistemas lineares sobredeterminados, equações diferenciais, equações diofantinas, etc. Nomeadamente, ao associarmos a uma equação às diferenças uma função contínua g 'contendo' a solução do problema (1), somos naturalmente levados a considerar o respectivo 'retrato de fase', no sentido descrito adiante na Secção 3. O retrato de fase da sucessão $(x_n)_{n\geq 0}$, obtém-se simplesmente considerando a função paramétrica $H(t) = (g(t), g(t+1) - g(t))$, para $t \geq 0$ (Definição 3.1).

Em particular, o retrato de fase de sucessões do tipo Fibonacci (ver em [1] uma perspectiva histórica a respeito deste famoso matemático) evidencia geometricamente uma propriedade fundamental destas sucessões ligada à ocorrência do célebre 'número de ouro' (ver Proposição 3.1).

Como é bem sabido, para se resolver a equação (1), isto é, para se determinar uma fórmula explícita da sucessão $(x_n)_{n\geq 0}$, os modelos possíveis dessa solução dependem de as k raízes do polinómio característico associado à equação serem, respectivamente, simples ou múltiplas. Embora aqui apenas consideremos equações de segunda ordem, isto é, para $k = 2$ a abordagem que propomos é generalizável para equações às diferenças de ordem superior.

Tanto no caso de raízes simples como no caso de raízes múltiplas do referido polinómio característico, mostramos por indução matemática (Secção 2) que o sistema linear, traduzindo as condições interpolatórias, é um sistema de solução única (ver Proposições 2.1 e 2.2 e Exemplo 2.1, 2.2 e 2.3).

Na Secção 3 apresentamos um pequeno número de exemplos de construção do 'retrato de fase' de equações às diferenças de segunda ordem do tipo (1), em particular de um certo conjunto de sucessões de Fibonacci. O leitor poderá ensaiar outros exemplos à sua escolha modificando convenientemente o programa *Mathematica* [4] dado em Anexo.

2 Equações às diferenças de segunda ordem

A equação às diferenças de segunda ordem

$$x_{n+2} = a_1\, x_{n+1} + a_0\, x_n, \quad a_0 \neq 0, \quad n \geq 0, \tag{2}$$

com valores iniciais $x_0 = \alpha, \quad x_1 = \beta$, tem para equação característica

$$p(\lambda) = \lambda^2 - a_1\, \lambda - a_0 = 0\ . \tag{3}$$

Mostramos a seguir, respectivamente nos parágrafos 2.1 e 2.2, que o facto do polinómio p possuir ou não raízes *distintas* é a informação crucial que interessa na consideração dos dois tipos de soluções possíveis para a equação às diferenças (2). No caso das raízes serem distintas e positivas, a solução interpolatória g que obtemos é uma combinação linear de exponenciais de variável real. No caso de alguma das raízes ser negativa ou complexa (sendo, portanto, neste segundo caso, a outra raiz conjugada da primeira) a função g é igualmente combinação linear de funções exponenciais. No entanto, tais exponenciais tomam valores em $\mathbb{C}$ e iremos considerar apenas a parte real desses valores. Em ambos os casos, para efeitos computacionais, a solução real pretendida pode ser facilmente calculada recorrendo à função `Re[ ]`, disponível no sistema *Mathematica* [4].

Quanto às soluções da equação (2) temos o seguinte resultado ([2], Ch. 2.2):

Teorema 2.1. *Considere-se a equação às diferenças (2) e a equação característica associada (3). Sejam λ_1 e λ_2 as raízes do polinómio[1] p e x_n a solução geral de (2). Então:*

1. Se λ_1 e λ_2 são raízes reais e distintas,

$$x_n = c_1\, \lambda_1^n + c_2\, \lambda_2^n, \qquad c_1, c_2\ \in \mathbb{R}\ .$$

2. Se $\lambda_1 = \lambda_2 = \lambda$,

$$x_n = c_1\, \lambda^n + c_2\, n\, \lambda^n, \qquad c_1, c_2\ \in \mathbb{R}\ .$$

3. Se λ_1 e λ_2 são raízes complexas, $\lambda_1 = r\, e^{i\,\theta}$ e $\lambda_2 = r\, e^{-i\,\theta}$,

$$\begin{aligned} x_n &= c_1 \left(r\, e^{i\,\theta}\right)^n + c_2 \left(r\, e^{-i\,\theta}\right)^n, \qquad c_1, c_2\ \in \mathbb{C}, \\ &= r^n\ ((c_1 + c_2)\, \cos(n\,\theta) + i\,(c_1 - c_2)\, \sin(n\,\theta)) \\ &= r^n\ (C_1\, \cos(n\,\theta) + C_2\, \sin(n\,\theta)) \end{aligned}$$

onde $C_1 = c_1 + c_2$ e $C_2 = i\,(c_1 - c_2)$.

[1] Note-se que, dado que $a_0 \neq 0$, o polinómio p não tem raízes nulas.

Do Teorema 2.1, concluímos que, sendo λ_1 e λ_2 duas raízes *distintas* do polinómio p em (3), a função interpolatória real g será uma combinação linear de funções reais do seguinte tipo:
(i) caso $\lambda_1, \lambda_2 > 0$, funções $e^{\ln(\lambda_1)\,t}$ e $e^{\ln(\lambda_2)\,t}$;
(ii) caso $\lambda_1 < 0$ e $\lambda_2 > 0$, funções $\cos(\pi\,t)\,e^{\ln(|\lambda_1|)\,t}$ e $e^{\ln(\lambda_2)\,t}$;
(iii) caso $\lambda_1, \lambda_2 < 0$, funções $\cos(\pi\,t)\,e^{\ln(|\lambda_1|)\,t)}$ e $\cos(\pi\,t)\,e^{\ln(|\lambda_2|)\,t}$;
(iv) caso $\lambda_1 \in \mathbb{C}$ (λ_2 raiz conjugada), funções $\cos(n\,\theta)\,e^{\ln(r)\,t}$ e $\sin(n\,\theta)\,e^{\ln(r)\,t}$, onde $r = |\lambda_1|$ e $\theta = Arg(\lambda_1)$.

Sendo λ uma *raiz dupla* do polinómio p, a função interpolatória real g é combinação linear das funções $e^{\ln(\lambda)\,t}$ e $t\,e^{\ln(\lambda)\,t}$, no caso em que $\lambda > 0$, e das funções $\cos(\pi\,t)\,e^{\ln(|\lambda|)\,t}$ e $t\,\cos(\pi\,t)\,e^{\ln(|\lambda|)\,t}$, no caso em que $\lambda < 0$.

Os Exemplos 2.1, 2.2 e 2.3 adiante ilustram o cálculo de funcões interpolatórias g para algumas equações às diferenças de segunda ordem.

2.1 Raízes distintas

Para $\Delta = a_1^2 + 4\,a_0 > 0$, o polinómio p em (3) tem duas raízes distintas (reais ou complexas conjugadas), $\lambda_1, \lambda_2 \neq 0$, tais que

$$\lambda_1 = \frac{a_1 + \sqrt{\Delta}}{2}, \qquad \lambda_2 = \frac{a_1 - \sqrt{\Delta}}{2}\ . \tag{4}$$

Atendendo a (3) as raízes λ_1, λ_2 satisfazem as relações fundamentais

$$\begin{array}{l} \lambda_1^2 = a_1\,\lambda_1 + a_0 \\ \lambda_2^2 = a_1\,\lambda_2 + a_0\ . \end{array} \tag{5}$$

No caso em que $\lambda_1, \lambda_2 > 0$, mostramos adiante (ver Proposição 2.1) que a seguinte função contínua[2] $g : \mathbb{R} \mapsto \mathbb{R}$,

$$g(t) = c_1\,\lambda_1^t + c_2\,\lambda_2^t, \tag{6}$$

onde c_1, c_2 são constantes, é solução da equação às diferenças (2), no sentido de que existem constantes c_1 e c_2 para as quais é satisfeita a infinidade de condições interpolatórias, $g(0) = \alpha, g(1) = \beta$ e $g(j) = x_j$, para $j = 2, 3, \ldots$, isto é, a função g considerada interpola a seguinte tabela com um número infinito de nós inteiros não negativos,

t_i	0	1	2	3	...
x_i	x_0	x_1	x_2	x_3	...

[2] Neste caso a função $g \in C^\infty(\mathbb{R}_0^+)$. No entanto, para o efeito interpolatório basta-nos que a função seja contínua e, posteriormente , quando se tratar de 'retratos de fase', basta que a função g seja pelo menos de classe C^1.

A solução explícita da equação às diferenças (2) é dada por

$$x_n = g(n) = c_1\,\lambda_1^n + c_2\,\lambda_2^n, \qquad n = 0, 1, 2, \ldots \tag{7}$$

onde c_1 e c_2 são constantes (únicas), solução das equações $g(0) = x_0$ e $g(1) = x_1$.
Os casos em λ_1 ou λ_2 é um número negativo, ou quando λ_1 e λ_2 são complexos são também contemplados.

Observação : No caso particular de a_1, a_0 e x_0, x_1 serem inteiros, a sucessão $(x_k)_{k\geq 0}$ é constituída por números inteiros e, por conseguinte, o gráfico da função contínua g dada por (6), possui uma infinidade de pontos $P_k = (k, g(k))$, de coordenadas inteiras ('látice'), isto é, $P_k \in \mathbb{Z}^2$.

2.1.1 Solução na forma exponencial

Quando λ_1, λ_2 são números reais positivos, a função real g dada em (6) pode ser escrita na forma de combinação linear de funções exponenciais

$$g(t) = c_1\, e^{\ln(\lambda_1)\,t} + c_2\; e^{\ln(\lambda_2)\,t}, \qquad t \in \mathbb{R}_0^+, \tag{8}$$

Para contemplar os casos em que λ_1 ou λ_2 são números negativos, a função g deverá ser definida como

$$\begin{aligned} &g(t) = Re\,[c_1\, e^{\ln(\tilde{\lambda}_1)\,t} + c_2\; e^{\ln(\tilde{\lambda}_2)\,t}], \qquad t \in \mathbb{R}_0^+, \\ &\text{onde} \quad \tilde{\lambda}_i = \lambda_i, \text{ se } \lambda_i > 0, \qquad \tilde{\lambda}_i = |\lambda_i|\, e^{i\,\pi}, \text{ se } \lambda_i < 0\ . \end{aligned} \tag{9}$$

(Re designa a parte real de um número complexo). Repare-se que

$$e^{\ln(|\lambda_i| e^{i\,\pi})\,t} = e^{(\ln(|\lambda_i|) + i\,\pi)\,t} = e^{\ln(|\lambda_i|)\,t}\, e^{i\,\pi\,t} = e^{\ln(|\lambda_i|)\,t}\,(\cos(\pi\,t) + i\,\sin(\pi\,t))$$

e assim, $Re[e^{\ln(|\lambda_i| e^{i\,\pi})\,t}] = \cos(\pi\,t)\; e^{\ln(|\lambda_i|)\,t}$.

Um caso em que $\lambda_1 = 1 > 0$ e $\lambda_2 = -1 < 0$ é ilustrado no Exemplo 2.1. Para os dados iniciais considerados neste exemplo, a função

$$h(t) = -2\,\lambda_2^t = -2\,(-1)^t,$$

interpola os dados para $t = j$, com $j = 0, 1, 2, \ldots$, mas a função h não está definida quando t não é inteiro. A correspondente função contínua é obtida considerando a expressão (9). Assim, sem quebra de generalidade, nas Proposições 2.1 e 2.2 a seguir, assume-se que os números reais λ_1, λ_2 são distintos, não nulos, e positivos. Caso algum desses números seja negativo a função g a considerar deverá ter a forma da expressão (9).

Proposição 2.1. *Dada a equação às diferenças (2), cujo polinómio característico possui duas raízes reais distintas e não nulas, λ_1, λ_2, $\lambda_1 \neq \lambda_2$, a função contínua g dada em (6) (ou a sua versão (9)) satisfaz as condições interpolatórias*

$$g(j) = x_j, \quad j = 0, 1, 2, \ldots \ . \tag{10}$$

Demonstração. As constantes c_1 e c_2 podem calcular-se univocamente, porquanto as condições $g(0) = x_0$ e $g(1) = x_1$ equivalem à existência de solução do sistema linear

$$\begin{cases} c_1 + c_2 & = x_0 \\ \lambda_1\, c_1 + \lambda_2\, c_2 & = x_1 \ . \end{cases}$$

A solução do sistema existe e é única já que $\lambda_1 \neq \lambda_2$,

$$c_1 = \frac{x_0\, \lambda_2 - x_1}{\lambda_2 - \lambda_1}, \quad c_2 = \frac{x_1 - x_0\, \lambda_1}{\lambda_2 - \lambda_1} \ . \tag{11}$$

Vamos mostrar por indução que g satisfaz um número infinito de condições interpolatórias nos inteiros não negativos, isto é, que $g(x_j) = x_j$, qualquer que seja $j \geq 1$. Fixado j, tome-se para hipóteses de indução as igualdades

$$\begin{array}{lll} g(j-1) & = c_1\, \lambda_1^{j-1} + c_2\, \lambda_2^{j-1} & = x_{j-1} \\ g(j) & = c_1\, \lambda_1^{j} + c_2\, \lambda_2^{j} & = x_j \ . \end{array} \tag{12}$$

Prove-se que $g(j+1) = x_{j+1}$. Dado que

$$\begin{aligned} g(j+1) &= c_1\, \lambda_1^{j+1} + c_2\, \lambda_2^{j+1} \\ &= c_1\, \lambda_1^2\, \lambda_1^{j-1} + c_2\, \lambda_2^2\, \lambda_2^{j-1} \ , \end{aligned}$$

atendendo às relações fundamentais (5), obtém-se

$$\begin{aligned} g(j+1) &= c_1\, (a_1\, \lambda_1 + a_0)\, \lambda_1^{j-1} + c_2\, (a_1\, \lambda_2 + a_0)\, \lambda_2^{j-1} \\ &= a_0\, (c_1\, \lambda_1^{j-1} + c_2\, \lambda_2^{j-1}) + a_1 (c_1\, \lambda_1^{j} + c_2\, \lambda_2^{j}) \ . \end{aligned}$$

Levando em conta as hipóteses de indução e a definição recursiva da equação à diferenças considerada, resulta

$$g(j+1) = a_1\, x_j + a_0\, x_{j-1} = x_{j+1} \ . \qquad \square$$

Corolário 2.1. *A solução explícita da equação às diferenças (6) é a sucessão*

$$x_n = g(n), \qquad n = 0, 1, \ldots \tag{13}$$

O Corolário 2.1 resulta trivialmente de se substituir na função g a variável real t pela variável inteira $n \geq 0$.

Para o caso em que λ_1 e λ_2 são raízes complexas conjugadas, a função g deverá ser definida na forma

$$g(t) = Re\left[c_1\, e^{\ln(\lambda_1)\, t} + c_2\, e^{\ln(\lambda_2)\, t}\right], \qquad t \in \mathbb{R}_0^+ \ .$$

As constantes c_1 e c_2 são sempre complexas conjugadas e quando $t = n$, para $n = 0, 1, \ldots$, temos que $c_1\, e^{\ln(\lambda_1)\, t} + c_2\, e^{\ln(\lambda_2)\, t}$ assume um valor real (ver Teorema 2.1) e, portanto,

$$g(n) = c_1\, e^{\ln(\lambda_1)\, n} + c_2\, e^{\ln(\lambda_2)\, n} = c_1\, \lambda_1^n + c_2\, \lambda_2^n = x_n \ .$$

2.2 Raízes múltiplas

Dada a equação às diferenças (2), se z é uma raiz dupla do seu polinómio característico, $p(x) = x_2 - a_1\, x - a_0$, tem-se

$$x^2 - a_1\, x - a_0 = (x - z)^2 = x^2 - 2\, z\, x + z^2 \ .$$

Comparando os coeficientes, resulta

$$z^2 = -\, a_0, \quad \text{e} \quad z = \frac{a_1}{2} \ . \tag{14}$$

Note-se que terá de ser $a_1 \neq 0$.

Proposição 2.2. *Dada uma equação às diferenças (2), cujo polinómio característico possui uma raíz não nula z, de multiplicidade dois, a função contínua*

$$g(t) = c_1\, z^t + c_2\, t\, z^t = (c_1 + c_2\, t)\, z^t, \quad t \in \mathbb{R}_0^+ \tag{15}$$

satisfaz as condições interpolatórias

$$g(j) = x_j, \quad j = 0, 1, 2, \ldots \ . \tag{16}$$

Demonstração. As constantes c_1 e c_2 determinam-se univocamente considerando as condições interpolatórios $g(0) = x_0$ e $g(1) = x_1$. Obtém-se,

$$\begin{aligned} c_1 &= x_0 \\ c_2 &= \frac{x_1 - z\, x_0}{z} = \frac{x_1 - a_1/2\, x_0}{a_1/2} = \frac{2\, x_1 - a_1\, x_0}{a_1} \ . \end{aligned} \tag{17}$$

De modo semelhante ao que se fez na prova da Proposicão 2.1, para mostrar que tal função g satisfaz as condições interpolatórias para qualquer inteiro não negativo, considere-se agora para base de indução as igualdades

$$\begin{aligned} g(j-1) &= (c_1 + c_2\, (j-1))\; z^{j-1} &&= x_{j-1} \\ g(j) &= (c_1 + c_2\, j)\; z^j &&= x_j \ . \end{aligned} \tag{18}$$

Prove-se que $g(j+1) = x_{j+1}$, isto é,

$$g(j+1) = (c_1 + c_2\,(j+1))\ z^{j+1} = a_1\,x_j + a_0\,x_{j-1}\ .$$

Atendendo a que $a_1 = 2\,z$ e $a_0 = -z^2$ (ver (14)), resulta da segunda expressão em (18) que

$$a_1\,x_j = 2\,z\,(c_1 + c_2\,j)\,z^j,$$

e da primeira expressão em (18)

$$a_0\,x_{j-1} = -\,z^2\ (c_1 + c_2\,(j-1))\ z^{j-1}.$$

Então,

$$\begin{aligned} x_{j+1} &= a_1\,x_j + a_0\,x_{j-1} \\ &= 2\ (c_1 + c_2\,j)\ z^{j+1} - (c_1 + c_2\,(j-1))\ z^{j+1} \\ &= (c_1 + 2\,c_2\,j - c_2\,j + c_2)\ z^{j+1} \\ &= (c_1 + c_2\,(j+1))\ z^{j+1} = g(x_{j+1})\ . \end{aligned}$$

□

Corolário 2.2. *A solução explícita da equação às diferenças (6) é a sucessão*

$$x_n = g(n), \qquad n = 0, 1, \ldots \tag{19}$$

O Corolário 2.2 resulta de substituir a variável real t pelo inteiro $n \geq 0$.

2.3 Exemplos

Exemplo 2.1. (Duas raízes reais de sinal contrário)

A equação às diferenças

$$x_{n+2} = x_n, \quad n = 0, 1, \ldots, \quad \text{onde} \quad x_0 = -2, \quad x_1 = 2,$$

define recursivamente a sucessão periódica, de período dois,

$$-2, 2, -2, 2, \ldots\ .$$

A equação característica associada é da forma

$$\lambda^2 - 1 = 0, \quad \text{donde} \quad \lambda_1 = 1, \lambda_2 = -1\ .$$

Assim, a função modelo correspondente, para $j = 0, 1, 2, \ldots$, é da forma

$$g(t) = c_1 + c_2\,(-1)^j,$$

sendo que (c_1, c_2) satisfaz as condições interpolatórias $g(0) = -2$, $g(1) = 2$, isto é,

$$\begin{cases} c_1 + c_2 & = -2 \\ c_1 - c_2 = & = 2, \end{cases}$$

donde $c_1 = 0$ e $c_2 = -2$. Por conseguinte, a função contínua,

$$g(t) = -2\, Re\left[e^{(\ln(|-1|)+i\,\pi)\,t}\right] = -2\, Re\left[e^{i\,\pi\,t}\right] = -2\,\cos(\pi\,t), \quad t \in \mathbb{R}_0^+$$

é interpolatória dos dados já que $g(j) = x_j$, para $j \geq 0$. Na Figura 1 observa-se o gráfico da função g e os pontos, (j, x_j) que satisfazem a equação às diferenças dada, no intervalo $0 \leq t \leq 20$.

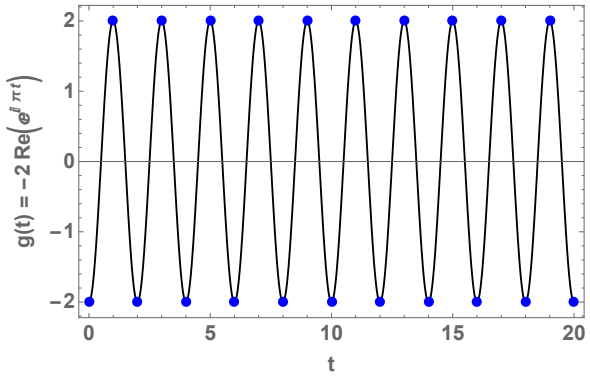


Fig. 1: $x_{n+2} = x_n, \quad n = 0, 1, 2, \ldots \quad x_0 = -2, x_1 = 2.$

Exemplo 2.2. (Duas raízes reais negativas)

Seja $x_{n+2} = -7x_{n+1} - 10x_n$, $n \geq 0$, $x_0 = 0$, $x_1 = 1$. Interessa-nos saber se existe

$$\lim_{k\to\infty} \frac{x_{k+1} - x_k}{x_k} \,. \tag{20}$$

O polinómio característico associado à equação às diferenças dada tem raízes negativas $\lambda_1 = -2$ e $\lambda_2 = -5$. Por conseguinte, a função interpoladora (com valores complexos), seja h, é da forma

$$\begin{aligned} h(t) \quad & = c_1\, e^{(\ln(2)+i\,\pi)\,t} + c_2\, e^{(\ln(5)+i\,\pi)\,t} = c_1\, e^{\ln(-2)\,t} + c_2\, e^{\ln(-5)\,t} \\ & = c_1\, e^{\ln(2)\,t}\,(\cos(\pi t) + i\,\sin(\pi\,t)) + c_2\, e^{\ln(5)\,t}\,(\cos(\pi t) + i\,\sin(\pi\,t)) \,. \end{aligned}$$

Assim, a função interpoladora real escreve-se como

$$g(t) = c_1\,\cos(\pi\,t)\, e^{\ln(2)\,t} + c_2\,\cos(\pi\,t)\, e^{\ln(5)\,t}, \quad t \geq 0 \,.$$

Levando em consideração as condições iniciais, $g(0) = 0$ e $g(1) = 1$, obtém-se $c_1 = 1/3$, $c_2 = -1/3$. Logo, a função interpoladora dos dados é definida como

$$g(t) = \frac{1}{3}\cos(\pi\, t)\,(2^t - 5^t), \quad t \geq 0\ .$$

Atendendo a que

$$g(t+1) - g(t) = \cos(\pi\, t)\,\frac{1}{3}\left(-3\,.\,2^t + 6\,.\,5^t\right),$$

resulta

$$Q(t) = \frac{g(t+1) - g(t)}{g(t)} = \frac{3\,(-2^t + 2\,.\,5^t)}{2^t - 5^t}\ .$$

Assim,

$$\lim_{t\to\infty} Q(t) = 3\,\lim_{t\to\infty} \frac{(-2/5)^t + 2}{(2/5)^t - 1} = -6\ .$$

Por conseguinte, o limite em (20) existe e tem o valor de -6.

Exemplo 2.3. (Duas raízes complexas conjugadas)

Para a equação

$$x_{n+2} = \frac{1}{2}\, x_{n+1} - \frac{1}{2}\, x_n, \quad \text{onde} \quad n \geq 0, \ \text{e} \ x_0 = -2, \ x_1 = -1, \tag{21}$$

o polinómio característico possui duas raízes complexas conjugadas (logo distintas)

$$\lambda_1 = \frac{1}{4}\left(1 + i\sqrt{7}\right), \quad \lambda_2 = \frac{1}{4}\left(1 - i\sqrt{7}\right) .$$

Denotando por h a função de variável real com valores complexos $h(t) = c_1\,\lambda_1^t + c_2\,\lambda_2^t$, as constantes c_1 e c_2 calculam-se como solução das equações $h(0) = x_0$ e $h(1) = x_1$. Obtém-se,

$$c_1 = \frac{1}{7}\left(-7 + i\sqrt{7}\right), \quad c_2 = \frac{1}{7}\left(-7 - i\sqrt{7}\right) .$$

Por conseguinte, a função $g(t) = Re\left(c_1\, e^{\ln(\lambda_1)\, t} + c_2\, e^{\ln(\lambda_2)\, t}\right)$, onde $t \in \mathbb{R}$, interpola os pontos (j, x_j), para $j \geq 0$, conforme é mostrado na Figura 2, na qual se considerou o intervalo $t \in [-3, 20]$.

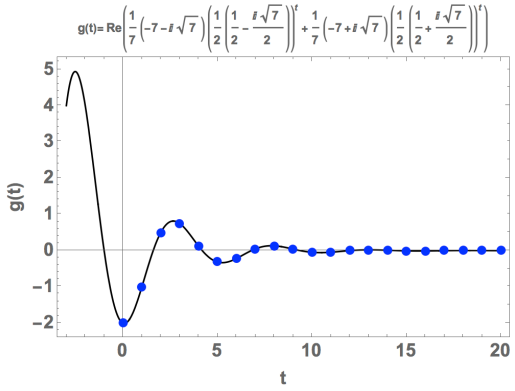


Fig. 2: $x_{n+2} = \frac{1}{2}\, x_{n+1} - \frac{1}{2}\, x_n, \quad n = 0, 1, 2, \ldots \quad x_0 = -2, x_1 = -1.$

Uma vez que $|\lambda_1| < 1$ e $|\lambda_2| < 1$, conclui-se imediatamente da expressão calculada para a função g que

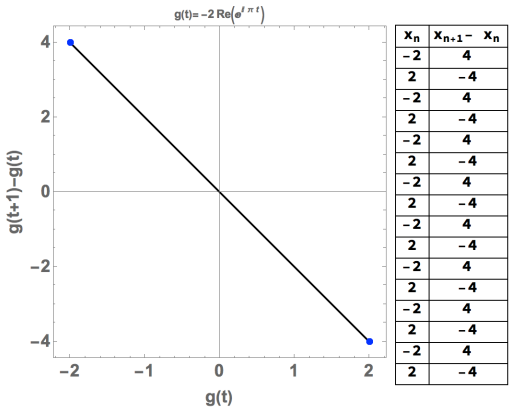
$$\lim_{t \to \infty} g(t) = 0 \ .$$

Dado que $g(j) = x_j$ para $j \geq 0$, podemos concluir igualmente que a sucessão considerada é tal que

$$\lim_{n\to\infty} x_n = 0,$$

conforme sugere a observação do gráfico apresentado na Figura 2.

3 Retratos de fase de equações às diferenças



x_n	$x_{n+1} - x_n$
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4
−2	4
2	−4

Fig. 3: $x_{n+2} = x_n, \quad n = 0, 1, 2, \ldots \quad x_0 = -2, x_1 = 2.$

As funções g anteriormente consideradas são particularmente úteis para desenhar o 'retrato de fase' associado à solução de uma dada equação às diferenças. Tais retratos permitem-nos observar a evolução da sucessão ao longo do 'tempo' t, o seu comportamento assimptótico (isto é para n suficientemente grande) – em particular quando desejamos estudar a dinâmica de sucessões recursivas em função dos valores iniciais $x_0 = \alpha$, $x_1 = \beta$ considerados, ou quando estamos interessados em observar o comportamento de classes de sucessões distintas em função dos parâmetros a_1 e a_0.

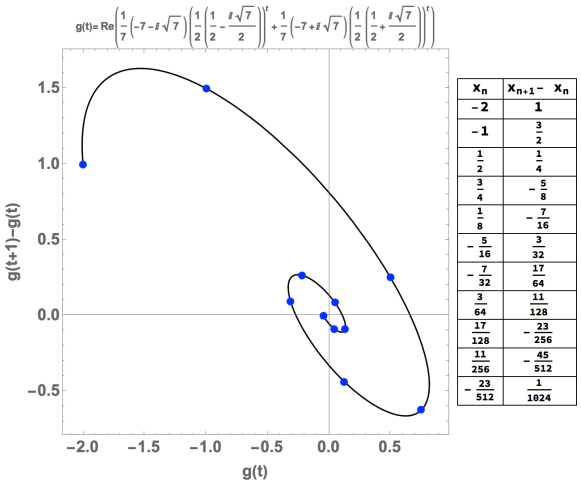
Dado que para quaisquer valores iniciais x_0 e x_1, a função interpolatória g é única, podemos definir a seguinte função H, a qual será designada por 'retrato de fase' associado à equação às diferenças (2):

Definição 3.1. *Designamos por 'retrato de fase' associado à equação às diferenças (2), a função $H : \mathbb{R}_0^+ \mapsto \mathbb{R}^2$, definida por*

$$H(t) = (g(t), g(t+1) - g(t)), \quad t \in \mathbb{R}_0^+ \ . \tag{22}$$

Da mesma forma que a cada termo x_n da sucessão $(x_n)_{n\geq 0}$ podemos associar um ponto $P_n = (x_n, x_{n+1} - x_n)$ do plano, atendendo a que g é interpolatória de $(x_n)_{n\geq 0}$, a função H satisfaz as igualdades

$$H(j) = (x_j, x_{j+1} - x_j) = P_j, \quad j = 0, 1, , 2, \ldots \tag{23}$$



x_n	$x_{n+1} - x_n$
-2	1
-1	$\frac{3}{2}$
$\frac{1}{2}$	$\frac{1}{4}$
$\frac{3}{4}$	$-\frac{5}{8}$
$\frac{1}{8}$	$-\frac{7}{16}$
$-\frac{5}{16}$	$\frac{3}{32}$
$-\frac{7}{32}$	$\frac{17}{64}$
$\frac{3}{64}$	$\frac{11}{128}$
$\frac{17}{128}$	$-\frac{23}{256}$
$\frac{11}{256}$	$-\frac{45}{512}$
$-\frac{23}{512}$	$\frac{1}{1024}$

Fig. 4: $x_{n+2} = \frac{1}{2}\, x_{n+1} - \frac{1}{2}\, x_n, \quad n = 0, 1, 2, \ldots \quad x_0 = -2, x_1 = -1.$

Assim, do ponto de vista cinemático, os pontos P_j do plano coordenado representam a 'posição' do termo x_j da solução da equação às diferenças, sendo a sua 'velocidade' representada pela segunda coordenada de P_j. Estas considerações de carácter geométrico justificam a designação aqui adoptada para *retrato de fase* da solução de uma equação às diferenças, representado pela função paramétrica (22).

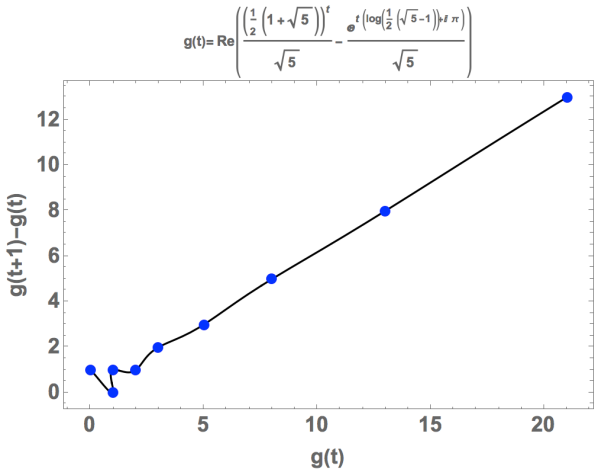
Na Figura 3 mostra-se o retrato de fase da sucessão considerada no Exemplo 2.1 e na Figura 4 é representado o retrato de fase correspondente ao Exemplo 2.3. Note-se que as abcissas e ordenadas figuradas neste caso são pontos de $\mathbb{Q}$, como se evidencia na tabela que acompanha a Figura 4.

O carácter periódico da solução $(x_n)_{n\geq 0}$ da equação às diferenças do primeiro exemplo é imediatamente aparente na Figura 3, enquanto a evolução 'espiral' da sucessão $(x_n)_{n\geq 0}$ do segundo exemplo é observável na Figura 4. O retrato de fase permite-nos prever a ocorrência do ponto $(0, 0)$ como estado limite do processo evolutivo modelado pela equação às diferenças (21) do Exemplo (2.3), confirmando o que se disse na parte final deste exemplo.

3.1 Sucessões tipo Fibonacci

A classe de sucessões de segunda ordem

$$x_{n+2} = x_n + x_{n-1}, \quad x_0 = \alpha \in \mathbb{Z}, \quad x_1 = \beta \in \mathbb{Z}, \tag{24}$$



x_n	$x_{n+1} - x_n$
0	1
1	0
1	1
2	1
3	2
5	3
8	5
13	8
21	13

Fig. 5: Retrato de fase da sucessão de Fibonacci.

goza de reputação universal, uma vez que para $\alpha = 0$, $\beta = 1$ a sucessão $(x_n)_{n\geq 0}$ é a famosa sucessão de Fibonacci [1]. Para qualquer elemento da classe o polinómio característico tem duas raízes reais distintas,

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \simeq 1.618 > 0, \quad \lambda_2 = \frac{1-\sqrt{5}}{2} \simeq -0.618 < 0 \ .$$

A maior raiz é o famoso 'número de ouro' $\Phi = (1+\sqrt{5})/2$. Na Figura 5 é mostrado o retrato de fase da sucessão de Fibonacci. A observação do respectivo gráfico, leva-nos a conjecturar que, assimptoticamente, o gráfico da função $g(t+1)-g(t)$ deverá conter pontos numa certa direcção invariante, ou seja, que existe $\lim_{t->\infty}(g(t+1)-g(t))/g(t)$. Atendendo aos valores da tabela da Figura 5, tal limite deverá ser aproximadamente $13/21 \simeq 0.62$. De facto, para uma infinidade de equações às diferenças do tipo (24) o referido limite existe e toma o valor $1/\Phi$, conforme previsto na seguinte proposição:

Proposição 3.1. *Sejam $\alpha \in \mathbb{Z}$, $\beta \in \mathbb{Z}$ valores iniciais da classe de equações às diferenças (24). Se existir*

$$\lim_{t\to\infty} \frac{g(t+1)-g(t)}{g(t)}, \tag{25}$$

tal limite é o número $1/\Phi$.

Demonstração. Sabemos que $g(t) = c_1\,\lambda_1^t + c_2\,\lambda_2^t$. Admitamos que $c_1 \neq 0$ e $c_1 + c_2\,(\lambda_2/\lambda_1)^t \neq 0$. Então,

$$\frac{g(t+1)-g(t)}{g(t)} = \frac{c_1\lambda_1^{t+1}+c_2\,\lambda_2^{t+1}-c_1\lambda_1^t-c_2\lambda_2^t}{c_1\,\lambda_1^t+c_2\,\lambda_2^t}$$
$$= \frac{c_1\lambda_1^t(\lambda_1-1)+c_2\lambda_2^t(\lambda_2-1)}{\lambda_1^t\,(c_1+c_2(\lambda_2/\lambda_1)^t)} = \frac{c_1(\lambda_1-1)+c_2(\lambda_2/\lambda_1)^t(\lambda_2-1)}{c_1+c_2(\lambda_2/\lambda_1)^t}\,.$$

Atendendo a que $|\lambda_2/\lambda_1|<1$, resulta

$$\lim_{t\to\infty}\frac{g(t+1)-g(t)}{g(t)} = \frac{c_1\,(\lambda_1-1)}{c_1} = \lambda_1-1 = \frac{1}{\Phi}\,. \qquad \square$$

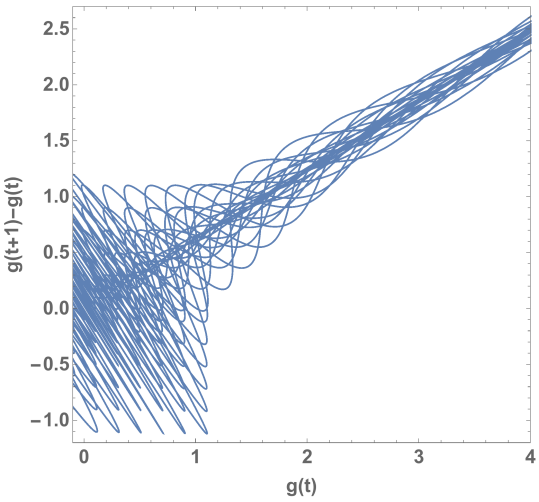


Fig. 6: Retrato de fase de sucessões de Fibonacci com x_0, x_1 assumindo valores com passo 0.1 .

Considerando o conjunto de valores iniciais (de passo 0.1), $x_0 = \alpha \in \{-1.1, -1.0, \ldots, 0.9, 1.0, 1.1\}$, e $x_1 = \beta \in \{-1.1, -1.0, \ldots, 0.9, 1.0, 1.1\}$, mostra-se na Figura 6 o retrato de fase das correspondentes sucessões $x_{n+2} = x_{n+1} + x_n$. A figura sugere a existência do limite do quociente $((g(t+1)-g(t))/g(t)$, de valor $1/\Phi \simeq 0.618$, correspondente ao declive de aproximadamente $35.4°$, que poderemos atribuir a um segmento de recta facilmente identificável na figura, cujo declive está de acordo com o enunciado na Proposição 3.1.

4 Anexo

A fim de que o leitor possa desenhar o retrato de fase para equações às

diferenças do tipo $x_{n+2} = a_1\, x_{n+1} + a_0\, x_n$, com valores iniciais $x_0 = \alpha$, $x_1 = \beta$, é incluído abaixo código *Mathematica*, cujos dados deverão ser modificados em função do problema a tratar. A nomenclatura do código é análoga à utilizada ao longo do texto e, por isso, não foram incluídos comentários adicionais.

```
ClearAll["Global`*"];
SetOptions[ParametricPlot, BaseStyle → {Bold, 18},
  Frame → True, ImageSize → 500];
tmin = 0;
tmax = 10;
cab = {"xn", "xn+1- xn"}; (* legenda para tabela *)
Manipulate[
 a1 = 1 / 2 ; a0 = -1 / 2;
 Δ = a1 ^ 2 + 4 a0;
 λ1 = (a1 + Sqrt[Δ]) / 2;
 λ2 = (a1 - Sqrt[Δ]) / 2;
 x[0] := α; x[1] := β;
 x[n_] := a1 x[n - 1] + a0 x[n - 2];
 h[t_] := c1 λ1 ^ t + c2 λ2 ^ t;
 {c11, c22} = {c1, c2} /. Solve[{h[0] == α, h[1] == β}, {c1, c2}][[1]];
 g[t_] := Re[c11 Exp[Log[λ1] t] + c22 Exp[Log[λ2] t]];
 tab = Table[{x[n], x[n + 1] - x[n]}, {n, 0, tmax}];
 {ParametricPlot[{g[t], g[t + 1] - g[t]}, {t, tmin, tmax},
    PlotStyle → {Black},
    Frame → True,
    AspectRatio → 1,
    PlotRange → Automatic,
    FrameLabel → {"g(t)", "g(t+1)-g(t)"},
    PlotLabel → Style["g(t)= " <> ToString[g[t] // TraditionalForm], 12],
    Epilog → {PointSize[0.02], Blue, Point[tab]}], " ",
   Style[Grid[Prepend[tab, cab], Frame → All], Bold, 16],
   Grid[{{" c1=", c11, " c2=", c22},
     {{"λ1=", λ1, N[λ1]} // Column,
      {"λ2=", λ2, N[λ2]} // Column,
      {"x[n]= " <> ToString[a1 // TraditionalForm] <> "   x[n-1]
" <> ToString[a0 // TraditionalForm] <> "  x[n-2]"} // Row
     }}]} // Row,
 {α, {-2, -1, 0, 2}}, {β, {-1, 0, 1, 2}}]
```

Fig. 7: Código para representação do retrato de fase $H(t) = (g(t), g(t+1) - g(t))$.

Referências

[1] K. Devlin, *Finding Fibonacci: The Quest to Rediscover The Forgotten Mathematical Genius Who Changed the World*, Princeton University Press, 2017.

[2] S. Elaydi, *An Introduction to Difference Equations*, Springer, third ed., 2005.

[3] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 1, John Wiley and Sons, New York, 1974.

[4] S. Wolfram, *The Mathematica Book*, Wolfram Media, fifth ed., 2003.

Em busca da unidade: conexões de Galois e inversões de Möbius-Rota[1]

Jorge Picado
CMUC, DMat
Universidade de Coimbra
e-mail: picado@mat.uc.pt

Pedro M. Silva
Dep. de Física
Universidade de Coimbra
e-mail: pmsilva@student.fisica.uc.pt

Resumo: Este artigo está organizado em duas partes distintas, desenvolvidas em paralelo, onde ilustramos a utilidade das conexões de Galois na teoria dos números (primeira parte) e das inversões de Möbius-Rota na combinatória (segunda parte). Estas ferramentas permitem abordar problemas aparentemente difíceis, transformando-os, com um simples processo de inversão, em problemas equivalentes mais simples. O fio condutor comum é a visão conceptual da teoria dos reticulados.

Abstract: This paper is organized in two distinct but parallel parts. Our goal is to illustrate the parallel between the usefulness of Galois connections (quasi-inversions) in number theory and of Möbius-Rota inversions in enumerative combinatorics. These tools allow to address apparently hard problems in an illuminating unifying way, by (quasi-)inverting them into much simpler equivalent problems. The common setting is the conceptual point of view of lattice theory.

palavras-chave: Sequências complementares, desencontros, princípio da inclusão-exclusão, conexão de Galois, inversão de Möbius, álgebra de incidência de Rota.

keywords: Complementary sequences, derangements, inclusion-exclusion principle, Galois connection, Möbius inversion, Rota incidence algebra.

> «At a time when mathematical fashion despises generality (seen as gratuitous "generalities", i.e. vacuities) I affirm the principal force in all my work has been the quest for the "general." In truth I prefer to accent "unity" rather than "generality." But for me these are two aspects of one quest. Unity represents the profound aspect, and generality the superficial aspect.»
>
> — Alexandre Grothendieck
> (Récoltes et Semailles, 1986)

[1] Trabalho realizado no âmbito do programa *Novos Talentos em Matemática* da Fundação Calouste Gulbenkian.

1 Um problema elementar de números

«Quem? O infinito?
Diz-lhe que entre.
Faz bem ao infinito
estar entre gente.»

— ALEXANDRE O'NEILL
(De Porta em Porta, 1960)

Denotemos por $\mathbb{N}_0$ o conjunto dos números naturais (onde incluímos o número 0) e consideremos duas funções bem conhecidas da teoria dos números: seja $f(n)$ o n-ésimo número primo e seja $g(n)$ o número de números primos que não excedem n. Será conveniente considerarmos a sucessão f dos números primos a começar em $n = 0$: assumimos que $f(0) = 0$. Claro que $g(0) = 0$. Calculemos agora os primeiros valores das sucessões definidas pelas somas $f(n) + n$ e $g(n) + n + 1$ ($n \in \mathbb{N}_0$):

n	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
$f(n)$	0	2	3	5	7	11	13	17	19	23	29	
$f(n)+n$	0	3	5	8	11	16	19	24	27	32	39	
$g(n)$	0	0	1	2	2	3	3	4	4	4	4	
$g(n)+n+1$	1	2	4	6	7	9	10	12	13	14	15	

Observemos melhor os números nessas duas linhas:

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, \ldots$$

Surpreendente, não? Será que o 17 e o 18 aparecerão em seguida na linha correspondente a $g(n) + n + 1$? (o 16 e o 19 já estão na linha de $f(n) + n$.) Isto é, será que estas sequências infinitas de naturais são *complementares*, ou seja, não têm elementos comuns e em conjunto esgotam todos os naturais? Desafiamos o caro leitor a tentar demonstrar tal facto. Não parece ser um exercício fácil, pois não? De facto, está longe de ser óbvio!

E mais: este facto não tem nada a ver com as propriedades dos números primos! Por exemplo, se $f(n)$ for agora o n-ésimo quadrado perfeito (convencionamos mais uma vez $f(0) = 0$) e $g(n)$ for o número de quadrados perfeitos que não excedem n, obtemos a tabela

n	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
$f(n)$	0	0	1	4	9	16	25	36	49	64	81	
$f(n)+n$	0	1	3	7	13	21	31	43	57	73	91	
$g(n)$	1	2	2	2	3	3	3	3	3	4	4	
$g(n)+n+1$	2	4	5	6	8	9	10	11	12	14	15	

E poderíamos continuar com outros exemplos. Mais geralmente, para uma qualquer propriedade P, se $f(n)$ é o n-ésimo número P, $f\colon \mathbb{N}_0 \to \mathbb{N}_0$ uma sucessão crescente, com $f(0)=0$, tendendo para ∞ com n, e $g(n)$ é o número de números P que não excedem n, o problema resume-se a:

Problema 1. *Será verdade que as sequências*

$$F=\{f(n)+n \mid n\in\mathbb{N}_0\} \quad e \quad G=\{g(n)+n+1 \mid n\in\mathbb{N}_0\}$$

são complementares?

Este problema foi originalmente resolvido por J. Lambek e L. Moser em 1954 [10], tendo resposta positiva, como veremos mais adiante. O padrão comum a este e outros problemas análogos, envolvendo funções bem conhecidas da aritmética e da teoria dos números, são as *conexões de Galois* [9]. Isso mesmo, o tipo de conexão entre os corpos intermédios de uma extensão de corpos e os subgrupos do correspondente grupo de automorfismos, base da teoria de Galois moderna reformulada por E. Artin.

Este artigo é constituído por duas partes aparentemente distintas.

Na primeira parte, abrangendo as primeiras cinco secções, veremos como as conexões de Galois [6, 8], combinando grande clareza estrutural e facilidade computacional, resolvem o Problema 1 de uma maneira surpreendentemente simples e elegante. Observaremos como, olhadas como uma generalização de pares de bijecções mutuamente inversas (mais especificamente, pares de funções *quase inversas*), permitem transferir informação do lado onde esta é mais completa para o lado oposto. Aproveitaremos ainda para mostrar a sua utilidade no ensino dos fundamentos da teoria dos conjuntos na possível unificação e sistematização de muitas propriedades .

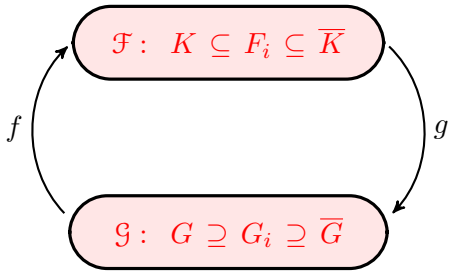
Na segunda parte do artigo (secções 6-10) passamos dos números para a combinatória. O objectivo é apresentar, numa abordagem em tudo paralela à primeira parte, uma ferramenta da combinatória enumerativa que desempenha um papel análogo ao das quase-inversões de Galois descrito na primeira parte: as *inversões de Möbius-Rota* [16].

2 Conexões de Galois

«Adjunctions arise everywhere»

— SAUNDERS MAC LANE
(Categories for the Working Mathematician, 1971)

A famosa solução do problema da resolubilidade algébrica de uma equação polinomial baseia-se na correspondência, descoberta por Galois (1811-1832) em 1830, para uma dada extensão de corpos $K \subseteq \overline{K}$, entre a colecção $\mathcal{F}$ dos subcorpos de $\overline{K}$ contendo K e a colecção $\mathcal{G}$ de todos os subgrupos do grupo de automorfismos de $\overline{K}$ que deixam K invariante (o chamado *grupo de Galois* dessa extensão):



A moderna *teoria de Galois* baseia-se nesta correspondência e no facto essencial de que as funções f e g são, em geral, *quase inversas* uma da outra:

$$\forall G \in \mathcal{G},\ \forall F \in \mathcal{F}\ (f(G) \subseteq F \ \Leftrightarrow \ G \supseteq g(F)).$$

Esta correspondência é um exemplo daquilo a que hoje se chama uma conexão de Galois (dual ou contravariante): Birkhoff, em 1940, associou uma conexão deste tipo a qualquer relação binária (a que chamou *polaridade* [2]) e, em 1944, Ore [14] generalizou o conceito a quaisquer conjuntos parcialmente ordenados. Invertendo a ordem num dos lados chegamos à definição (covariante) moderna de conexão de Galois [2, p. 124]:

Definição 2.1. Sejam $(A, \leq)$ e $(B, \leq)$ dois conjuntos parcialmente ordenados. Um par de funções

$$(A, \leq) \underset{g}{\overset{f}{\rightleftarrows}} (B, \leq)$$

diz-se uma *conexão* (ou *adjunção*[2]) *de Galois* quando

$$\forall a \in A,\ \forall b \in B\ (f(a) \leq b \ \Leftrightarrow\ a \leq g(b)).$$

A função f chama-se o *adjunto à esquerda* enquanto g é o *adjunto à direita*. Abrevia-se tudo escrevendo simplesmente $f \dashv g$.

Note que, no caso em que as relações de ordem em A e B são a relação de igualdade, $f \dashv g$ significa simplesmente que f e g são um par de bijeccções mutuamente inversas.

É agora claro que em todos os exemplos da secção anterior temos precisamente uma conexão de Galois $f \dashv g$ em $(\mathbb{N}_0, \leq)$.

Note que um adjunto à esquerda (resp. à direita) de uma dada função monótona pode não existir, mas caso exista é necessariamente único: se $f_i(a) \leq b \Leftrightarrow a \leq g(b)$, $i = 1, 2$, então, evidentemente, $f_1(a) \leq b \Leftrightarrow f_2(a) \leq b$.

Proposição 2.2. *Sejam* $f\colon A \to B$ *e* $g\colon B \to A$.

(1) *Se* $f \dashv g$ *então:*

- (i) f *e* g *são funções quase-inversas, isto é,* $fg \leq \mathrm{id}$ *e* $\mathrm{id} \leq gf$.
- (ii) f *e* g *são monótonas.*[3]
- (iii) $fgf = f$ *e* $gfg = g$.
- (iv) f *e* g *definem bijecções entre* $f[A]$ *e* $g[B]$*, inversas uma da outra.*
- (v) f *preserva supremos*[4]*,* g *preserva ínfimos,*

$$f(a) = \inf\{b \in B \mid a \leq g(b)\} \quad e \quad g(b) = \sup\{a \in A \mid f(a) \leq b\}.$$

(2) *Reciprocamente, se* f *e* g *são monótonas, então:*

- (i) *Se, para cada* $b \in B$*, existe o supremo de* $\{a \in A \mid f(a) \leq b\}$ *em* A*, e* f *preserva estes supremos, então* f *possui um (único) adjunto à direita dado por esta fórmula.*

[2] Terminologia influenciada pela teoria das categorias, uma vez que este conceito é um exemplo do conceito mais geral de *adjunção* entre duas categorias, aplicado a conjuntos parcialmente ordenados (considerados como categorias magras, de modo *standard*). Segundo este ponto de vista, toda a teoria de reticulados pode ser vista como teoria das categorias em categorias magras, precisamente $(0,1)$-categorias (ver [*nLab HomePage*, `ncatlab.org/nlab/show/partial+order`]).

[3] As condições (1:i) e (1:ii), em conjunto, caracterizam a adjunção $f \dashv g$: se $f(a) \leq b$ então $gf(a) \leq g(b)$ pelo que $a \leq gf(a) \leq g(b)$; analogamente, se $a \leq g(b)$, então $f(a) \leq fg(b) \leq b$.

[4] Dizemos que f *preserva supremos* quando preserva todos os supremos que existem em A; analogamente para os ínfimos.

(ii) *Se, para cada $a \in A$, existe o ínfimo de $\{b \in B \mid a \leq g(b)\}$ em B, e g preserva estes ínfimos, então g possui um (único) adjunto à esquerda dado por esta fórmula.*

Demonstração. (1) (i) Como $f(a) \leq f(a)$, então $a \leq gf(a)$ para qualquer $a \in A$. Analogamente, $fg(b) \leq b$ para qualquer $b \in B$.
(ii) Se $a \leq a'$ em A, então $a \leq gf(a')$, isto é, $f(a) \leq f(a')$. Analogamente para g.
(iii) A desigualdade $fgf \leq f$ decorre imediatamente da desigualdade $fg \leq \mathrm{id}$, enquanto $\mathrm{id} \leq gf$ e o facto de f ser monótona implicam $f \leq fgf$. De modo análogo, $gfg = g$.
(iv) É consequência imediata da alínea anterior.
(v) Sejam $S \subseteq A$, $x = \sup S$. Temos que mostrar que $f(x)$ é o supremo de $\{f(s) \mid s \in S\}$ em B:

- $f(x) \geq f(s)$ para qualquer $s \in S$ pois f é monótona.
- Se algum $b \in B$ satisfaz $b \geq f(s)$ para todo o $s \in S$ então, pela adjunção, $g(b) \geq s$ para todo o $s \in S$ e, portanto, $g(b) \geq x$. Pela adjunção, isto significa que $b \geq f(x)$.

De modo análogo, pode provar-se que g preserva ínfimos.

Finalmente, mostremos que $g(b) = \sup\{a \in A \mid f(a) \leq b\}$ (o resultado dual para f segue de modo semelhante):

- Denotemos o conjunto $\{a \in A \mid f(a) \leq b\}$ por S. Claramente $g(b) \in S$ pois $fg(b) \leq b$.
- Por outro lado, $g(b) \geq s$ para qualquer $s \in S$, uma vez que, por definição de S, $b \geq f(s)$ para qualquer $s \in S$.

(2) (i) Consideremos a função $h\colon B \to A$ definida por

$$h(b) = \sup\{a \in A \mid f(a) \leq b\}.$$

Trata-se de um adjunto à direita de f: se $f(a) \leq b$, evidentemente $a \leq h(b)$; reciprocamente, se $a \leq h(b)$, então, como f é monótona e preserva supremos, $f(a) \leq \sup\{f(a') \mid a' \in A, f(a') \leq b\} \leq b$.
(ii) Por dualidade. □

Corolário 2.3. *Sejam $f\colon A \to B$ e $g\colon B \to A$ funções monótonas.*

(1) *f é um adjunto à esquerda se e só se preserva supremos e para cada $b \in B$ existe o supremo de $\{a \in A \mid f(a) \leq b\}$ em A.*

(2) *g é um adjunto à direita se e só se preserva ínfimos e para cada $a \in A$ existe o ínfimo de $\{b \in B \mid a \leq g(b)\}$ em B.* □

Portanto, quando A e B são reticulados completos,

(1) f é um adjunto à esquerda se e só se preserva supremos, e

(2) g é um adjunto à direita se e só se preserva ínfimos.

3 Resolvendo o Problema com conexões de Galois

> «The structural contribution of Galois was not so much to do with fields and groups but with their relationship.»
> — Øystein Ore
> (Galois connexions, 1944)

Analisemos agora a situação que nos interessa, do Problema 1:

$$A = B = (\mathbb{N}_0, \leq), \text{ com a ordem usual } \leq .$$

Trata-se de um conjunto totalmente ordenado, onde todo o subconjunto não vazio tem um ínfimo – o seu mínimo –, mas só os subconjuntos *finitos* têm supremo – o máximo do conjunto –; em particular, o máximo do conjunto vazio é o zero.

Lema 3.1. *Seja $f\colon \mathbb{N}_0 \to \mathbb{N}_0$. Então:*

(1) *$f[\mathbb{N}_0]$ é um conjunto infinito se e só se $\{m \in \mathbb{N}_0 \mid f(m) \geq n\} \neq \emptyset$ para qualquer $n \in \mathbb{N}_0$.*

(2) *Se f é monótona então as seguintes afirmações também são equivalentes:*

   (i) *$f[\mathbb{N}_0]$ é um conjunto infinito.*

   (ii) *$f(n) \to \infty$ quando $n \to \infty$ (isto é, $\forall m \in \mathbb{N}_0\ \exists N \in \mathbb{N}_0\colon f(n) \geq m$ para qualquer $n \geq N$).*

   (iii) *$\{n \in \mathbb{N}_0 \mid f(n) \leq m\}$ é finito para qualquer $m \in \mathbb{N}_0$.*

Demonstração. (1) Suponhamos que $f[\mathbb{N}_0]$ é infinito. O caso $n = 0$ é óbvio: $\{m \in \mathbb{N}_0 \mid f(m) \geq 0\} = \mathbb{N}_0$. Os restantes casos ($n \in \mathbb{N}$) também são evidentes: se $\{m \in \mathbb{N}_0 \mid f(m) \geq n\}$ fosse vazio, teríamos $f[\mathbb{N}_0] \subseteq [0, n-1]$, um absurdo.

Reciprocamente, se $f[\mathbb{N}_0]$ fosse finito, igual a, digamos, $\{y_1 < y_2 < \cdots < y_k\} \subseteq \mathbb{N}_0$, para $n = y_k + 1$ existiria, por hipótese, $m \in \mathbb{N}_0$ tal que $f(m) \geq n > y_k$, pelo que $f(m)$ não pertenceria a $f[\mathbb{N}_0]$, um absurdo.

(2) (i)⇒(ii): Para cada $m \in \mathbb{N}_0$, como $f[\mathbb{N}_0]$ é infinito existe N tal que $f(N) \geq m$. Pela monotonia de f, $f(n) \geq f(N) \geq m$ para qualquer $n \geq N$. A implicação recíproca (ii)⇒(i) é óbvia.

(ii)⇒(iii) Seja $m \in \mathbb{N}_0$. Se $\{n \in \mathbb{N}_0 \mid f(n) \leq m\}$ fosse infinito, teríamos uma sucessão

$$n_1 < n_2 < \cdots < n_k < \cdots \quad (k \in \mathbb{N})$$

tal que $f(n_i) \leq m$. Como f é monótona, isto implicaria $f(n) \leq m$ para todo o $n \in \mathbb{N}_0$, o que contraria a hipótese.

(iii)⇒(ii) Seja $m \in \mathbb{N}_0$. Como $A_m := \{n \in \mathbb{N}_0 \mid f(n) \leq m\}$ é finito então existe $N \in \mathbb{N}_0$ tal que $N \notin A_m$, ou seja, $f(N) > m$. □

O Corolário 2.3 reduz-se agora a:

Corolário 3.2. *Sejam $f, g\colon \mathbb{N}_0 \to \mathbb{N}_0$ funções monótonas.*

(1) *f é um adjunto à esquerda se e só se $f(0) = 0$ e $f[\mathbb{N}_0]$ é infinito. Nesse caso, o seu adjunto direito é dado pela fórmula $g(m) = \max\{n \in \mathbb{N}_0 \mid f(n) \leq m\}$.*

(2) *g é um adjunto à direita se e só se $g[\mathbb{N}_0]$ é infinito. Nesse caso, o seu adjunto direito é dado pela fórmula $f(n) = \min\{m \in \mathbb{N}_0 \mid n \leq g(m)\}$.*

Demonstração. (1) Como $(\mathbb{N}_0, \leq)$ é totalmente ordenado, qualquer função monótona $f\colon \mathbb{N}_0 \to \mathbb{N}_0$ preserva supremos binários. Portanto, preserva todos os supremos que existem em $\mathbb{N}_0$, isto é, os supremos finitos, se e só se $f(0) = 0$. Por outro lado, para cada $m \in \mathbb{N}_0$, o conjunto $\{n \in \mathbb{N}_0 \mid f(n) \leq m\}$ tem supremo se e só se é finito. Pelo Lema 3.1(2), a finitude daquele conjunto é equivalente à condição '$f[\mathbb{N}_0]$ é infinito'.

(2) Em primeiro lugar, qualquer função monótona $g\colon \mathbb{N}_0 \to \mathbb{N}_0$ preserva ínfimos de subconjuntos não vazios e, portanto, preserva todos os ínfimos que existem em $\mathbb{N}_0$. Além disso, para cada $n \in \mathbb{N}_0$, o conjunto $\{m \in \mathbb{N}_0 \mid n \leq g(m)\}$ tem ínfimo se e só se não é vazio, ou seja, se e só se a imagem $g[\mathbb{N}_0]$ é infinita (Lema 3.1(1)). □

Daqui decorre, sem dificuldade, o resultado de Lambek-Moser [10, 9] que confirma que o Problema 1 (pág. 3) tem, de facto, resposta positiva. O

Teorema de Lambek-Moser[5] é universal, no sentido em que descreve qualquer partição dos naturais em dois subconjuntos infinitos, em termos de conexões de Galois em $(\mathbb{N}_0, \leq)$:

Teorema 3.3. (1) *Sejam F e G subconjuntos infinitos complementares de $\mathbb{N}_0$ com $0 \in F$. Sendo $F(n)$ o $(n+1)$-ésimo elemento de F e $G(m)$ o $(m+1)$-ésimo elemento de G, as funções $f, g\colon \mathbb{N}_0 \to \mathbb{N}_0$ definidas por*

$$f(n) = F(n) - n, \quad g(m) = G(m) - m - 1$$

determinam uma conexão de Galois $f \dashv g$.

(2) *Seja $f \dashv g$ uma conexão de Galois entre $(\mathbb{N}_0, \leq)$ e ele próprio. Os conjuntos*

$$F = \{f(n) + n \mid n \in \mathbb{N}_0\} \quad e \quad G = \{g(n) + n + 1 \mid n \in \mathbb{N}_0\}$$

são infinitos e formam uma partição de $\mathbb{N}_0$ (com $0 \in F$).

Demonstração. (1) Teremos que mostrar que $f(n) \leq m$ se e só se $n \leq g(m)$, isto é,

$$F(n) \leq n + m \quad \Leftrightarrow \quad n + m + 1 \leq G(m).$$

'$\Rightarrow$': Por absurdo: Se $F(n) \leq n+m$ e $G(m) \leq n+m$, então $F(0), \ldots, F(n)$ e $G(0), \ldots, G(m)$ seriam $m+n+2$ números naturais distintos $\leq m+n$.

'$\Leftarrow$': Por absurdo: Se $G(m) \not\leq n+m$ e $F(n) \not\leq n+m$, então F teria no máximo n elementos $\leq n+m$ e G teria no máximo m elementos $\leq n+m$, pelo que a sua união teria no máximo $n+m$ elementos $\leq n+m$, contradizendo a existência de $n+m+1$ naturais $\leq n+m$.

(2) O Corolário 3.2 garante que o subconjunto F é infinito e contém o número 0. O complementar $G' = \mathbb{N}_0 \setminus F$ também é infinito: se não fosse, existiria $n_0 \in \mathbb{N}_0$ tal que $n \in F$ para qualquer $n \geq n_0$, pelo que teríamos $F(n) = F(n_0) + (n - n_0)$, isto é, $f(n) + n = f(n_0) + n_0 + n - n_0$, ou seja, $f(n) = f(n_0)$ para qualquer $n \geq n_0$, contradizendo o facto de que $f[\mathbb{N}_0]$ é infinito (Corolário 3.2).

Assim, por (1), f teria um adjunto à direita g' definido por

$$g'(m) = G'(m) - m - 1$$

onde $G'(m)$ é o $(m+1)$-ésimo elemento de G'. Mas, como vimos na secção anterior, os adjuntos são únicos, logo $g' = g$ e, consequentemente,

$$G'(m) = g(m) + m + 1 = G(m).$$

Portanto, G é o complemento (infinito) de F em $\mathbb{N}_0$. □

[5] O resultado análogo a este para o conjunto parcialmente ordenado $(\mathbb{Z}, \leq)$ encontra-se em [9].

4 Ilustrando as potencialidades do método

> «In their most general sense, adjunctions and/or Galois connections make it possible to relate two "worlds" of (more or less mathematical) objects with each other in order to gain information about one world by passing to the other, perhaps better known world.»
>
> — MARCEL ERNÉ
> (Capítulo 1 de [6], 2004)

Exemplos 4.1. (1) A vantagem que podemos tirar de uma conexão de Galois[6], tal como o próprio Galois fez, é concluir factos novos num dos lados da correspondência (no problema de Galois, no lado das extensões de corpos) a partir de factos mais facilmente prováveis no lado oposto. Por exemplo, a fórmula de cálculo para o n-ésimo termo da sucessão em $\mathbb{N}_0$ dos *quadrados perfeitos* é óbvia: $(n-1)^2$. Mas já não parece tão fácil determinar uma fórmula geral para o n-ésimo número que não é quadrado perfeito. Vejamos como com uma simples conexão de Galois podemos descobrir esta fórmula a partir da primeira.

Seja F a sucessão dos quadrados perfeitos em $\mathbb{N}_0$ e G o respectivo complemento:

n	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
$F\colon f(n)+n$	0	1	4	9	16	25	36	49	64	81	100	
$f(n)$	0	0	2	6	12	20	30	42	56	72	90	
$G\colon g(n)+n+1$	2	3	5	6	7	8	10	11	12	13	14	
$g(n)$	1	1	2	2	2	2	3	3	3	3	3	

O problema resume-se à determinação de uma fórmula para $G(n-1) = g(n-1)+n$. Para isso basta aplicar o Teorema 3.3:

Primeiro,

$$f(n) = n^2 - n, \quad g(n) = G(n) - n - 1.$$

Da adjunção $f \dashv g$ sabemos então que

$$g(n) = \max\{m \in \mathbb{N}_0 \mid f(m) \leq n\} = \max\{m \in \mathbb{N}_0 \mid m^2 - m \leq n\}$$

[6] Para mais informação sobre conexões de Galois consulte, por exemplo, [6, 8, 15].

e, como

$$m^2 - m \leq n \Leftrightarrow m^2 - m + \frac{1}{4} < n + 1$$
$$\Leftrightarrow \left(m - \frac{1}{2}\right)^2 < n + 1 \Leftrightarrow m - \frac{1}{2} < \sqrt{n+1},$$

então $g(n) = \lfloor \sqrt{n+1} + \frac{1}{2} \rfloor = [\sqrt{n+1}]$, onde $\lfloor x \rfloor$ denota a parte inteira do real x (mais adiante usaremos também a notação $\lceil x \rceil$ para referir o menor inteiro que não é menor que x) e $[x]$ designa o inteiro mais próximo de x (onde, no caso de x ser a metade de um inteiro ímpar y, escolhemos $[x] = \frac{y-1}{2}$). Concluindo,

$$G(n-1) = n + [\sqrt{n}]$$

é o n-ésimo número em $\mathbb{N}_0$ que não é quadrado perfeito.

(2) Sejam p e q números irracionais positivos tais que $\frac{1}{p} + \frac{1}{q} = 1$. A função f definida por $f(n) = \lfloor (p-1)n \rfloor$ é um adjunto à esquerda da função g dada por $g(m) = \lfloor (q-1)(m+1) \rfloor$. Portanto,

$$F(n) = f(n) + n = \lfloor pn \rfloor \text{ e } G(m) = g(m) + m + 1 = \lfloor q(m+1) \rfloor,\ n, m \in \mathbb{N}_0,$$

enumeram uma partição de $\mathbb{N}_0$ em subconjuntos infinitos. Esquecendo $F(0) = 0$, temos então que as sequências

$$A = \{\lfloor p \rfloor, \lfloor 2p \rfloor, \lfloor 3p \rfloor, \ldots\} \quad \text{e} \quad B = \{\lfloor q \rfloor, \lfloor 2q \rfloor, \lfloor 3q \rfloor, \ldots\}$$

constituem uma partição de $\mathbb{N}$. Trata-se de um resultado de Beatty [1] com 90 anos; por isso, as sequências em A e B são chamadas *sequências de Beatty*. Por exemplo, $p = \sqrt{2}$ gera a sequência de Beatty

$$A = \{1, 2, 4, 5, 7, 8, 9, 11, 12, 14, 15, 16, 18, 19, 21, \ldots\}.$$

A grande surpresa que o resultado de Beatty nos proporciona é que o complemento de A em $\mathbb{N}$,

$$\mathbb{N} \setminus A = \{3, 6, 10, 13, 17, 20, 23, 27, 30, 34, 37, 40, 44, \ldots\},$$

é também uma sequência de Beatty, gerada por $q = \frac{p}{p-1}$:

$$\mathbb{N} \setminus A = B = \left\{ \left\lfloor \frac{n\sqrt{2}}{\sqrt{2}-1} \right\rfloor : n \in \mathbb{N} \right\}.$$

Para mais exemplos elementares em $\mathbb{N}$ (ordenado pela relação de divisibilidade) e em $\mathbb{Z}$ (com a ordem usual) consulte [9].

5 Mais sobre conexões de Galois

> «Since the beginning of the century, computational procedures have become so complicated that any progress by those means has become impossible, without the elegance which modern mathematicians have brought to bear on their research, and by means of which the spirit comprehends quickly and in one step a great many computations.»
>
> — ÉVARISTE GALOIS
> (Do prefácio do seu último manuscrito, 1832)

Talvez valha a pena atentarmos em mais alguns exemplos elementares de conexões de Galois [8, 9, 15] que revelam algum do seu potencial o ensino dos fundamentos da teoria dos conjuntos.

Exemplos 5.1. (1) Suponhamos que uma função $f\colon \mathbb{N} \to \mathbb{N}$ pode ser estendida a uma função real crescente $\widetilde{f}$ no intervalo $\langle 1, +\infty)$ e seja ϕ a sua inversa. É evidente que $\lceil \phi(-) \rceil$ é um adjunto à esquerda de f e $\lfloor \phi(-) \rfloor$ é um adjunto à direita de f (consequência do facto óbvio de que $\lceil \phi(m) \rceil \leq n$ sse $\phi(m) \leq n$, e $\lfloor \phi(m) \rfloor \geq n$ sse $\phi(m) \geq n$).

Logo, por exemplo, $\lceil \log_2 \rceil$ e $\lfloor \log_2 \rfloor$ são os adjuntos à esquerda e à direita da exponencial $n \mapsto 2^n$.

(2) Sejam X e Y conjuntos arbitrários e $f\colon X \to Y$ uma função arbitrária. Uma vez que, para quaisquer $A \subseteq X$ e $B \subseteq Y$,

$$f[A] \subseteq B \text{ se e só se } A \subseteq f^{-1}[B], \tag{5.1.1}$$

as funções

$$f[-]\colon \mathcal{P}(X) \to \mathcal{P}(Y) \quad \text{e} \quad f^{-1}[-]\colon \mathcal{P}(Y) \to \mathcal{P}(X)$$

são adjuntas, $f[-]$ à esquerda e $f^{-1}[-]$ à direita. Isto significa que $f[-]$ preserva supremos enquanto $f^{-1}[-]$ preserva ínfimos. Daí as fórmulas básicas da teoria dos conjuntos

$$f[\bigcup_i A_i] = \bigcup_i f[A_i] \quad \text{e} \quad f^{-1}[\bigcap_i B_i] = \bigcap_i f^{-1}[B_i].$$

Compare agora a Proposição 2.2 com as fórmulas *standard* da teoria dos conjuntos

$$f[f^{-1}[B]] \subseteq B \quad \text{e} \quad A \subseteq f^{-1}[f[A]],$$
$$f[f^{-1}[f[A]]] = f[A] \quad \text{e} \quad f^{-1}[f[f^{-1}[B]]] = f^{-1}[B].$$

Mas $f^{-1}[-]$ também tem um adjunto à direita:

$$f^{-1}[B] \subseteq A \text{ se e só se } B \subseteq Y \setminus f[X \setminus A]. \tag{5.1.2}$$

(De facto, se $b \in B$ e, por absurdo, b pertencesse a $f[X \setminus A]$, teríamos $b = f(a')$ para algum $a' \in X \setminus A$, donde $a' \in f^{-1}[B] \subseteq A$, uma contradição; reciprocamente, se $x \in f^{-1}[B]$ então $f(x) \in B \subseteq Y \setminus f[X \setminus A]$, isto é, $f(x) \notin f[X \setminus A]$, pelo que $x \in A$.)

Daqui decorrem as propriedades bem conhecidas da teoria dos conjuntos

$$f^{-1}[\bigcup_i B_i] = \bigcup_i f^{-1}[B_i], \quad B \subseteq Y \setminus f[X \setminus f^{-1}[B]],$$
$$f^{-1}[Y \setminus f[X \setminus A]] \subseteq A, \quad \text{etc.}$$

Em geral, no entanto, a função imagem $f[-]$ não tem nenhum adjunto à esquerda, uma vez que, ao contrário das pré-imagens, as imagens não preservam ínfimos.

De facto, se $f[-]$ preserva ínfimos, f é necessariamente uma bijecção: se não fosse injectiva, existiriam $a \neq b$ em X tais que $f(a) = f(b) = y$, o que implicaria $f[A \cap B] = f[\emptyset] = \emptyset$, para $A := \{a\}$ e $B := \{b\}$, enquanto $y \in f[A] \cap f[B]$; por outro lado, a sobrejectividade de f decorre simplesmente da preservação de ínfimos para famílias *vazias*: em $\mathcal{P}(X)$ (resp. $\mathcal{P}(Y)$) esse ínfimo é precisamente X (resp. Y), pelo que $f[X] = Y$.

Claro que no caso em que f é uma bijecção, com função inversa $g\colon Y \to X$, a equivalência (5.1.1) aplicada à inversa g diz-nos que $g[B] \subseteq A$ se e só se $B \subseteq g^{-1}[A]$, isto é, $f^{-1}[B] \subseteq A$ se e só se $B \subseteq f[A]$, ou seja, neste caso especial tem-se mesmo $f^{-1}[-] \dashv f[-]$.

(3) Sejam X, Y espaços topológicos e $f\colon X \to Y$ uma função contínua. A adjunção (5.1.2) pode ser modificada numa adjunção entre as duas topologias:

$$f^{-1}[B] \subseteq A \text{ se e só se } B \subseteq \operatorname{int}(Y \setminus f[X \setminus A]) \tag{5.1.3}$$

para quaisquer abertos A e B de X e Y, respectivamente. Contudo, a adjunção (5.1.1) não tem correspondente neste contexto, uma vez que, como é fácil de verificar, para funções contínuas e conjuntos abertos, a função imagem não preserva uniões enquanto a pré-imagem preserva uniões mas, em geral, não preserva ínfimos (note que, no caso infinito, estes não coincidem, em geral, com as intersecções mas sim com o interior das intersecções).

6 Um problema elementar de combinatória

> «God created infinity, and man, unable to understand infinity, had to invent finite sets.»

— Gian-Carlo Rota
('Combinatorics', em: Discrete Thoughts, 1969)

«A vida é a arte do encontro
Embora haja tanto desencontro pela vida»

— Vinícius de Moraes
(Samba da benção, 1967)

Como qualquer estudante rapidamente se apercebe, o estudo da combinatória enumerativa torna-se um desafio mais sério a partir do momento em que começamos a impor restrições nalgumas posições das configurações em análise. Por exemplo, no chamado 'jogo dos pares' (de casino):

Problema 2. *As 52 cartas de um baralho são dispostas sequencialmente, com o seu valor à vista. O 'croupier' dispõe então as cartas de um segundo baralho, uma a uma, por cima das primeiras. Ganha-se o jogo caso nenhuma carta do segundo baralho coincida com a carta do primeiro baralho com quem emparelha. Qual é a probabilidade de vitória?*

Uma das fórmulas mais úteis na resolução destes problemas é o chamado Princípio da Inclusão-Exclusão, também conhecido por *fórmula do crivo* ou *fórmula de da Silva-Sylvester*[7]. Este princípio estende a conjuntos arbitrários o princípio óbvio, habitualmente apelidado de *Princípio da Adição*, de que o cardinal da união de conjuntos disjuntos (dois a dois) é a soma dos cardinais de cada um desses conjuntos.

Por exemplo, no caso de três conjuntos A, B e C não necessariamente disjuntos, se somarmos os elementos em A, B e C (Fig. 1)

[7] O Princípio da Inclusão-Exclusão foi publicado pela primeira vez em 1854, num artigo de Daniel da Silva, e redescoberto mais tarde, em 1883, por Sylvester. Por isso, a fórmula do crivo e suas similares são, por vezes, apelidadas de fórmulas de da Silva ou de Sylvester. Realçamos o facto de Daniel da Silva, na opinião de Gomes Teixeira o mais notável matemático português do séc. XIX, ter sido estudante da Universidade de Coimbra; transcrevemos de [J. Silva Oliveira, *Daniel Augusto da Silva*, Boletim da SPM 2 (1979) 3-15]: «Daniel da Silva (1814-1878) foi, além de matemático eminente do seu tempo, oficial da Armada e professor da Escola Naval. Como estudante frequentou primeiro a Academia Real de Marinha e prosseguiu depois os seus estudos na Universidade de Coimbra onde se licenciou em Matemática e acabou por se doutorar.»

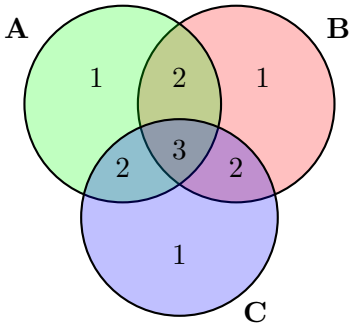


Figura 1: $|A| + |B| + |C|$.

estaremos a contar, uma vez cada um, os elementos de $A \setminus (B \cup C)$, os de $B \setminus (A \cup C)$ e os de $C \setminus (A \cup B)$, mas estaremos a contar por duas vezes os elementos de $(A \cap B) \setminus C$, $(A \cap C) \setminus B$ e $(B \cap C) \setminus A$, e, pior ainda, estaremos a contar por três vezes os elementos da intersecção $A \cap B \cap C$. Podemos começar por descontar os primeiros, subtraindo $|A \cap B|$, $|A \cap C|$ e $|B \cap C|$:

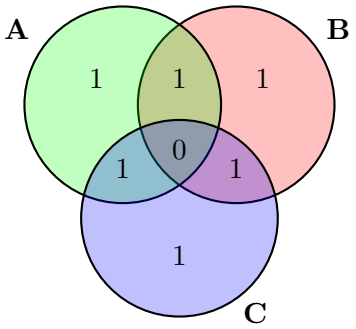


Figura 2: $|A| + |B| + |C| - (|A \cap B| + |A \cap C| + |B \cap C|)$.

Mas agora acabámos por descontar os elementos da intersecção $A \cap B \cap C$ mais do que devíamos (o zero na Fig. 2 indica que os elementos dessa região ainda não foram considerados para a contagem dos elementos de $A \cup B \cup C$), tendo que os repor novamente, para que a contagem fique finalmente certa:

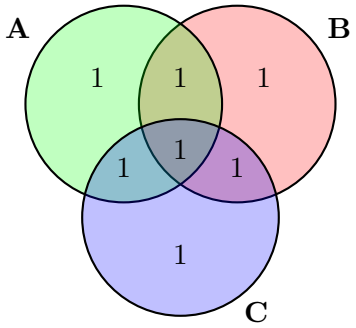


Figura 3: $|A| + |B| + |C| - (|A \cap B| + |A \cap C| + |B \cap C|) + |A \cap B \cap C|$.

No caso geral de n conjuntos $A_1, A_2, \ldots, A_n$, esta fórmula estende-se facilmente a (ver, por exemplo, [3] para uma demostração):

Proposição 6.1. [*Princípio da Inclusão-Exclusão*] *Para cada* $I \subseteq \{1, 2, \ldots, n\}$ *seja* $n(I) = |\cap_{i \in I} A_i|$. *O cardinal da união* $A_1 \cup A_2 \cup \cdots \cup A_n$ *é dado pela fórmula*

$$\sum_{|I|=1} n(I) - \sum_{|I|=2} n(I) + \sum_{|I|=3} n(I) - \cdots + (-1)^{n+1} \sum_{|I|=n} n(I).$$

O problema do jogo dos pares é um caso particular do bem conhecido **problema dos desencontros**:

Problema 2B. *Uma permutação* $a_{j_1} a_{j_2} \ldots a_{j_n}$ *de* $S = \{a_1, a_2, \ldots, a_n\}$ *diz-se um desencontro de* S *caso* $j_k \neq k$ *para qualquer* $k \in \{1, 2, \ldots, n\}$. *Quantas das* $n!$ *permutações de* S *são desencontros?*

De facto, denotando por D_n o número de desencontros de um conjunto com n elementos, a probabilidade de vitória no jogo dos pares é evidentemente igual a

$$\frac{D_{52}}{52!}.$$

Deixamos agora ao cuidado do leitor o exercício (recorrente em qualquer curso de Matemática Discreta) de verificar, com a ajuda do Princípio da Inclusão-Exclusão, que, para cada $n \in \mathbb{N}$,

$$D_n = \sum_{r=0}^{n} (-1)^{n-r} \binom{n}{r} r! = n! \left(\frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} \right). \qquad (6.1.1)$$

n	2	3	4	5	6	7	8	9	10	11
D_n	1	2	9	44	265	1854	14833	133496	1334961	14684570

Aqui, o segredo de aplicação do princípio de da Silva reside na constatação de que, não sendo fácil determinar directamente o número D_n, é, por outro lado, imediato o cálculo, para cada k, do número de permutações $a_{j_1}a_{j_2}\ldots a_{j_n}$ de S tais que $j_k = k$ (portanto aquelas em que a_k está na sua posição original), bem como o cálculo para cada par k, l, do número de permutações nas quais os elementos a_k e a_l estão nas suas posições primitivas k e l, etc.:

$$(n-1)! \text{ no primeiro caso,} \quad (n-2)! \text{ no segundo caso, etc.}$$

7 Inversões de Möbius

> «The apex of mathematical achievement occurs when two or more fields which were thought to be entirely unrelated turn out to be closely intertwined. Mathematicians have never decided whether they should feel excited or upset by such events.»
>
> — GIAN-CARLO ROTA
> ('A Mathematician's Gossip', em: Indiscrete Thoughts, 1997)

Como veremos mais adiante, o Princípio da Inclusão-Exclusão pode obter-se como exemplo de aplicação do processo de inversão de Möbius num conjunto parcialmente ordenado. A inversão de Möbius [13], introduzida originalmente em 1832 por August Ferdinand Möbius (1790-1868) no contexto da teoria dos números, pode ser descrita, nos seus aspectos básicos, do seguinte modo. Uma *função aritmética* é uma função $\mathbb{N} \to \mathbb{R}$ (ou, mais geralmente, $\mathbb{N} \to \mathbb{C}$, mas aqui consideraremos apenas funções reais). Qualquer par de funções aritméticas f, g tem um produto *de convolução* (de Dirichlet) $f * g$, definido por

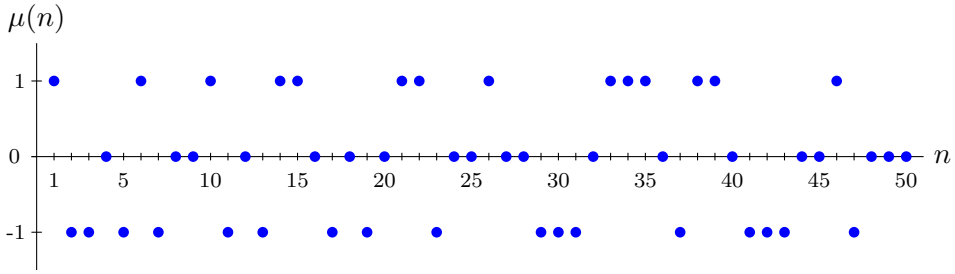
$$(f*g)(n) = \sum_{k,m:\ km=n} f(k)\, g(m) = \sum_{d:\ d|n} f(\frac{n}{d})\, g(d).$$

Este produto tem uma identidade: a *função unidade* δ definida por $\delta(1) = 1$ e $\delta(n) = 0$ para $n \geq 2$. A função constante $\zeta = (1, 1, \ldots)$ tem um inverso: a *função de Möbius* μ, muito usada em teoria dos números[8], definida para

[8] A função μ aparece já implicitamente nos trabalhos de Euler (em 1748) mas foi Möbius o primeiro a investigar de modo sistemático as suas propriedades (em 1832).

cada natural n, com factorização prima $p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$ $(n_i \geq 1)$, por

$$\mu(n) \equiv \begin{cases} 1 & \text{se } n = 1 \\ (-1)^k & \text{se } n_1 = n_2 = \cdots = n_k = 1 \\ 0 & \text{se } n_i \geq 2 \text{ para algum } i. \end{cases}$$



Note que

$$\sum_{d:\, d|n} \mu(d) = \begin{cases} 1 & \text{se } n = 1, \\ 0 & \text{senão.} \end{cases} \tag{7.0.1}$$

De facto, para qualquer natural $n > 1$ com factorização prima $p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$ $(n_i \in \mathbb{N})$, tem-se

$$\begin{aligned} \sum_{d:\, d|n} \mu(d) &= \mu(1) + \sum_{i=1}^{k} \mu(p_i) + \sum_{1\leq i_1<i_2\leq k} \mu(p_{i_1}p_{i_2}) + \cdots + \mu(p_1p_2\cdots p_k) \\ &= 1 + \binom{k}{1}(-1) + \binom{k}{2}(-1)^2 + \cdots + \binom{k}{k}(-1)^k \\ &= (1 + (-1))^k = 0. \end{aligned}$$

É este facto que torna μ tão relevante na teoria das funções aritméticas e está na base da fórmula seguinte da *inversão de Möbius*:

Teorema 7.1. *Se $f, g\colon \mathbb{N} \to \mathbb{R}$ são funções aritméticas tais que*

$$f(n) = \sum_{d:\, d|n} g(d) \quad \textit{para qualquer } n \in \mathbb{N}, \tag{7.1.1}$$

então podemos recuperar g com a identidade

$$g(n) = \sum_{d:\, d|n} f(\frac{n}{d})\, \mu(d) \quad \textit{para qualquer } n \in \mathbb{N}. \tag{7.1.2}$$

Demonstração. Este resultado é trivial quando formulado na linguagem do *anel de Dirichlet* (o conjunto das funções aritméticas com a soma de funções usual, ponto a ponto, e o produto de Dirichlet): a identidade (7.1.1) significa simplesmente $f = g * \zeta$ pelo que, imediatamente, $g = f * \zeta^{-1} = f * \mu$, isto é, (7.1.2). □

Exemplos 7.2. (1) Seja φ a função *tociente* de Euler, definida por

$$\varphi(n) = |\{k \in \{1, 2, \ldots, n\} \colon \text{mdc}(k, n) = 1\}|.$$

A fórmula da inversão de Möbius (7.1.2), aplicada à identidade bem conhecida de Gauss

$$n = \sum_{d \colon d|n} \varphi(d)$$

dá imediatamente

$$\varphi(n) = \sum_{d \colon d|n} \frac{n}{d}\, \mu(d) = n \sum_{d \colon d|n} \frac{\mu(d)}{d}. \tag{7.2.1}$$

Assim, por exemplo,

$$\begin{aligned} \varphi(36) = \varphi(2^2 \times 3^2) &= 36 \left(\frac{\mu(1)}{1} + \frac{\mu(2)}{2} + \frac{\mu(3)}{3} + \frac{\mu(2 \times 3)}{2 \times 3} \right) \\ &= 36 \left(1 - \frac{1}{2} - \frac{1}{3} + \frac{1}{6} \right) = 12. \end{aligned}$$

(2) Vejamos agora um exemplo um pouco mais complicado: determinemos o número de elementos do conjunto $C(n)$ de sequências *circulares* de zeros e uns de comprimento n (ou seja, sequências $a_1 a_2 \ldots a_n \in \{0, 1\}^n$ onde quaisquer duas sequências, uma das quais se obtém da outra por rotação, são consideradas iguais). Para isso, seja $\overline{P}(n)$ o conjunto de sequências circulares de comprimento n que não são periódicas (por exemplo, 010011 não é periódica, enquanto 010010 o é). Dada uma sequência arbitrária $S \in C(n)$, das duas uma: ou S não é periódica, isto é, $S \in \overline{P}(n)$, ou S é periódica, de período $d \mid n$. No segundo caso,

$$S = a_1 a_2 \ldots a_d a_1 a_2 \ldots a_d \cdots a_1 a_2 \ldots a_d,$$

pelo que podemos supor $S = a_1 a_2 \ldots a_d \in \overline{P}(d)$. Portanto,

$$|C(n)| = \sum_{d \colon d|n} |\overline{P}(d)|.$$

O problema resume-se então à determinação do número $|\overline{P}(d)|$. Mas como cada sequência em $\overline{P}(d)$ é igual às suas d permutações cíclicas (obtidas por rotação), tem-se

$$\sum_{d\colon\, d|n} d\,|\overline{P}(d)| = 2^n.$$

Logo, pelo Teorema 7.1,

$$n\,|\overline{P}(n)| = \sum_{d\colon\, d|n} 2^{\frac{n}{d}}\,\mu(d) = \sum_{d\colon\, d|n} 2^d\,\mu(\frac{n}{d})$$

e, consequentemente,

$$\begin{aligned}|C(n)| &= \sum_{d\colon\, d|n} |\overline{P}(d)| = \sum_{d\colon\, d|n} \frac{1}{d} \sum_{k\colon\, k|d} 2^k\,\mu(\frac{d}{k}) \\ &= \sum_{d\colon\, d|n} \sum_{k\colon\, k|d} \frac{2^k}{k\,\frac{d}{k}}\,\mu(\frac{d}{k}) = \sum_{k\colon\, k|n} \frac{2^k}{k} \sum_{l\colon\, l|\frac{n}{k}} \frac{\mu(l)}{l}.\end{aligned}$$

Finalmente, por (7.2.1),

$$|C(n)| = \frac{1}{n} \sum_{k\colon\, k|n} \varphi(\frac{n}{k})\,2^k. \tag{7.2.2}$$

8 Resolvendo o Problema 2 com inversões de Möbius-Rota

> «We often hear that mathematics consists mainly of 'proving theorems'. Is a writer's job mainly that of 'writing sentences'?»
>
> — GIAN-CARLO ROTA
>
> (Prefácio a [P. Davis e R. Hersh, *The Mathematical Experience*], 1981)

Em 1964, num artigo revolucionário [16], Gian-Carlo Rota (1932-1999) generalizou a inversão de Möbius a quaisquer conjuntos parcialmente ordenados *localmente finitos*, com o intuito de a tornar útil também na combinatória (e na teoria dos grupos). Estes conjuntos localmente finitos são aqueles $(P, \leq)$ nos quais qualquer intervalo

$$[x, y] := \{z \in P \mid x \leq z \leq y\}$$

é finito.

O par $(\mathbb{N}_0, \leq)$ é um exemplo de conjunto parcialmente ordenado localmente finito; outro é o par $(\mathbb{N}, |)$ dos inteiros positivos com a relação de divisibilidade.

Definição 8.1. Seja $(P, \leq)$ um conjunto parcialmente ordenado localmente finito e denotemos por $\text{int}(P)$ o respectivo conjunto de intervalos. A *álgebra de incidência* [16, 7], $\mathcal{I}(P)$, é o conjunto das funções

$$f \colon \text{int}(P) \to \mathbb{R}.$$

Abreviaremos $f([x, y])$ por $f(x, y)$. Se convencionarmos que $f(x, y) = 0$ sempre que $x \not\leq y$, podemos considerar cada $f \in \mathcal{I}(P)$ como uma função $P \times P \to \mathbb{R}$.

Qualquer álgebra de incidência $\mathcal{I}(P)$ é um espaço vectorial real com as operações de adição e multiplicação escalar definidas ponto a ponto. Com o produto de convolução

$$(f * g)(x, y) = \sum_{z \in [x,y]} f(x, z)g(z, y)$$

torna-se uma álgebra associativa. A identidade de $\mathcal{I}(P)$ é a função de Kronecker

$$\delta_P(x, y) \equiv \begin{cases} 1 & \text{se } x = y \\ 0 & \text{caso contrário} \end{cases}$$

e $f \in \mathcal{I}(P)$ é invertível se e só se $f(x, x) \neq 0$ para qualquer x. A função de Möbius μ_P é definida recursivamente sobre o comprimento dos intervalos:

(M1) $\mu_P(x, x) = 1$ para qualquer $x \in P$.

(M2) Se $x \not\leq y$, então $\mu_P(x, y) = 0$.

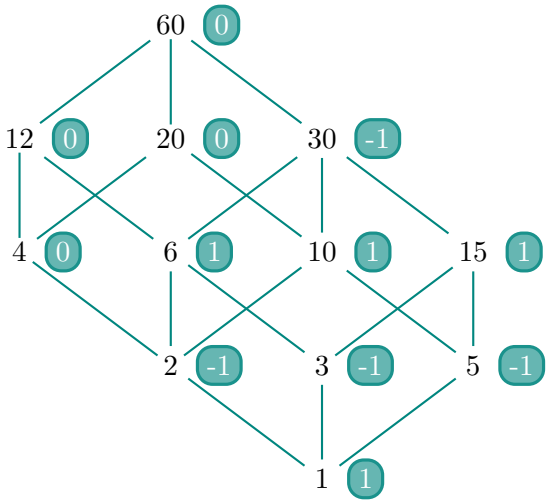
(M3) Se $x < y$, então $\mu_P(x, y) = -\sum_{z \colon x \leq z < y} \mu_P(x, z)$.

É surpreendente o dom da ubiquidade, na combinatória, da função de Möbius de um conjunto parcialmente ordenado.

Exemplos 8.2. (1) Calculemos alguns valores particulares de $\mu_P(1, n)$, no caso

$$(P, \leq) := (\mathbb{N}, |),$$

enumerados nos círculos da figura seguinte:



Estes valores correspondem precisamente à função de Möbius clássica:

$$\mu_P(1,n) = \mu(n) \text{ para qualquer } n \in \mathbb{N}.$$

De facto, $\mu_P(1,1) = 1 = \mu(1)$; para $n > 1$, supondo, por hipótese de indução, que $\mu_P(1,d) = \mu(d)$ para qualquer $d < n$ obtemos, usando (M3) e (7.0.1):

$$\mu_P(1,n) = -\sum_{d:\ 1\leq d<n} \mu_P(1,d) = -\sum_{d|n,\, d\neq n} \mu(d) = \mu(n).$$

Observe também como a função de Kronecker corresponde à função δ clássica: $\delta_P(1,n) = \delta(n)$ para qualquer natural n.

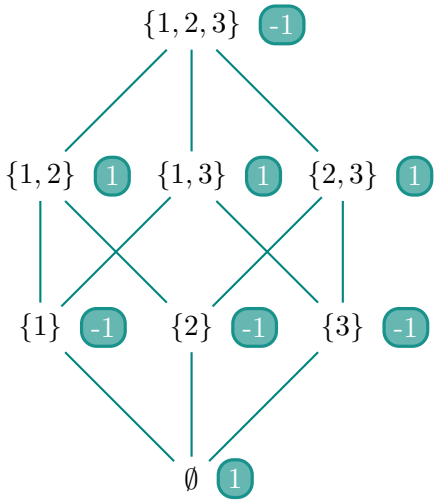
(2) No caso da álgebra de Boole

$$(P,\leq) = (\mathcal{P}(X_n),\subseteq)$$

sobre $X_n = \{1,2,\ldots,n\}$, para cada par $S \subseteq T$ de elementos de P,

$$\mu_P(S,T) = (-1)^{|T|-|S|}. \tag{8.2.1}$$

A figura seguinte enumera os diferentes valores de $\mu_P(\emptyset,T)$ no caso $n = 3$:



A prova de (8.2.1) segue por indução[9] sobre $|T| - |S|$:

Se $S = T$ então, por (M1), $\mu_P(S,T) = 1$ e (8.2.1) confirma-se; se $S \neq T$ e $p = |T \smallsetminus S| = |T| - |S|$, por (M3) e pela hipótese de indução obtemos

$$\mu_P(S,T) = -\sum_{R:\ S\subseteq R\subset T} \mu_P(S,R) = -\sum_{R:\ S\subseteq R\subset T} (-1)^{|R|-|S|} = -\sum_{k=0}^{p-1} (-1)^k \binom{p}{k}.$$

Como $0 = (1-1)^p = \sum_{k=0}^{p} (-1)^k \binom{p}{k}$, então, finalmente,

$$\mu_P(S,T) = (-1)^p \binom{p}{p} = (-1)^p = (-1)^{|T|-|S|}.$$

Rota mostrou que a função zeta[10] definida por $\zeta_P(x,y) = 1$ para quaisquer $x \leq y$ é a inversa de μ_P. Portanto, mais uma vez,

$$f = g * \zeta_P \quad \Rightarrow \quad g = f * \mu_P,$$

[9] Outra maneira de concluir isto é observar que: (1) P é isomorfo a $\mathcal{P}(\{1\})^n = \mathbf{2}^n$; (2) $\mathbf{2}$ pode ser visto como o intervalo $[0,1] \subseteq (\mathbb{N}, \leq)$; (3) portanto, a função de Möbius de $\mathbf{2}$ é precisamente $(-1)^n$, $n \in \{0,1\}$; (4) a função de Möbius do produto directo de dois conjuntos parcialmente ordenados P_1, P_2 é o produto das funções de Möbius de P_1 e P_2, ou seja, é dada por $\mu_{P_1\times P_2}((x,y),(x',y')) = \mu_{P_1}(x,x')\mu_{P_2}(y,y')$ para quaisquer $(x,y) \leq (x',y')$ em $P_1 \times P_2$ (regra do produto, uma das ferramentas mais úteis para calcular funções de Möbius em conjuntos parcialmente ordenados [16]).

[10] Sim, ζ como na função de Riemann, e não é por acaso!

Para mais informação ver as primeiras postagens em [*The n-Category Café*, `golem.ph.utexas.edu/category/2011/05/mbius_inversion_for_categories.html`].

Sobre a motivação e desenvolvimentos das ideias de Rota na transferência de conceitos da teoria dos números para a combinatória, ver [4].

ou seja:

Teorema 8.3 (Inversão de Möbius-Rota)**.** *Sejam* $f, g \in \mathcal{I}(P)$*. Se* $f(x,y) = \sum_{z\in[x,y]} g(x,z)$*, então* $g(x,y) = \sum_{z\in[x,y]} f(x,z)\,\mu_P(z,y)$*.* □

Quando P possui primeiro elemento 0, a restrição de $f\colon P\times P\to\mathbb{R}$ a $\{0\}\times P$, que continuaremos a designar por f, pode ser vista como uma função $f\colon P\to\mathbb{R}$, $f(y) = f(0,y)$. O Teorema anterior, no caso particular $x = 0$, garante então o seguinte:

Corolário 8.4. *Sejam* $f, g\colon P\to\mathbb{R}$*. Se* $f(y) = \sum_{z\le y} g(z)$*, então* $g(y) = \sum_{z\le y} f(z)\,\mu_P(z,y)$*.* □

De modo perfeitamente análogo ao que fizemos com as quase-inversões de Galois no Problema 1, a inversão de Möbius-Rota (Corolário 8.4) resolve o problema dos desencontros (Problema 2) de modo trivial a partir da correspondente identidade inversa, que é óbvia. Com efeito, denotemos por $\mathrm{Per}(X_n)$ o conjunto das $n!$ permutações dos elementos de $X_n = \{1,2,\ldots,n\}$. Basta então tomar para P a álgebra de Boole $(\mathcal{P}(X_n),\subseteq)$ do Exemplo 8.2(2). Se considerarmos, para cada $S\in P$, o conjunto $\mathrm{Des}(S)$ das permutações $(p_1,p_2,\ldots,p_n)$ de X_n nas quais $p_i\neq i$ para todo o $i\in S$, é evidente que

$$\mathrm{Per}(X_n) = \bigcup_{S\subseteq X_n} \mathrm{Des}(S) \quad \text{(união disjunta)}$$

pelo que

$$|\mathrm{Per}(X_n)| = \sum_{S\subseteq X_n} |\mathrm{Des}(S)|\,.$$

Logo, pela inversão de Möbius-Rota, tomando para $f, g\colon P\to\mathbb{R}$ as funções $S\mapsto|\mathrm{Per}(S)|$ e $S\mapsto|\mathrm{Des}(S)|$, respectivamente, obtemos

$$\begin{aligned}
|\mathrm{Des}(X_n)| &= \sum_{S\subseteq X_n} |\mathrm{Per}(S)|\;\mu_P(S,X_n) = \sum_{S\subseteq X_n} (-1)^{n-|S|}\;|\mathrm{Per}(S)| \\
&= \cancel{n!} - \binom{n}{n-1}\cancel{(n-1)!} + \binom{n}{n-2}(n-2)! - \cdots + (-1)^n\,0! \\
&= \frac{n!}{2!} - \frac{n!}{3!} + \frac{n!}{4!} - \cdots + (-1)^n\,\frac{n!}{n!},
\end{aligned}$$

precisamente a fórmula (6.1.1).

O paralelismo com as conexões de Galois é evidente: resolver problemas complicados, por inversão, a partir de identidades mais óbvias no lado oposto.

Para terminar esta secção, vejamos como se pode obter o Princípio da Inclusão-Exclusão como caso particular de aplicação da inversão de Möbius-Rota. Sejam $A_1, A_2, \ldots, A_n$ subconjuntos de um conjunto finito X. Designaremos por $\overline{A_i}$, $i = 1, 2, \ldots, n$, o complementar de A_i em X. Consideremos mais uma vez $P = (\mathcal{P}(X_n), \subseteq)$, e a função $f\colon P \to \mathbb{R}$ definida por

$$f(I) = \left|\{x \in X \mid x \in \overline{A_i} \text{ sse } i \in I\}\right| = \left|\bigcap_{i\in I} \overline{A_i} \cap \bigcap_{i\in X_n \setminus I} A_i\right|.$$

Claro que $f(X_n) = \left|\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}\right|$. Além disso, seja $g\colon P \to \mathbb{R}$ a função definida por $g(I) = \sum_{J\subseteq I} f(J)$. Não é difícil provar que

$$g(I) = |\bigcap_{i\in X_n \setminus I} A_i|.$$

Aplicando a inversão de Möbius-Rota, podemos então concluir que

$$f(I) = \sum_{J\subseteq I} g(J)\, \mu_P(J, I) = \sum_{J\subseteq I} (-1)^{|I|-|J|} g(J).$$

Em particular,

$$\begin{aligned}\left|\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}\right| = f(X_n) &= \sum_{J\subseteq X_n} (-1)^{n-|J|}\, g(J) \\ &= \sum_{J\subseteq X_n} (-1)^{n-|J|}\, |\bigcap_{j\in X_n\setminus J} A_j| = \sum_{I\subseteq X_n} (-1)^{|I|}\, |\bigcap_{i\in I} A_i|.\end{aligned}$$

Concluindo, $|A_1 \cup A_2 \cup \cdots \cup A_n|$ é igual a

$$\begin{aligned}|X| - \left|\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}\right| &= |X| - \left(\sum_{I\subseteq X_n} (-1)^{|I|}\, |\bigcap_{i\in I} A_i|\right) \\ &= |X| - \left(|X| + \sum_{\varnothing\neq I\subseteq X_n} (-1)^{|I|}\, |\bigcap_{i\in I} A_i|\right) \\ &= \sum_{\varnothing\neq I\subseteq X_n} (-1)^{|I|+1}\, |\bigcap_{i\in I} A_i|,\end{aligned}$$

que é precisamente a fórmula do Princípio da Inclusão-Exclusão em 6.1.

9 Ilustrando as potencialidades do método

> «Mathematics is the study of analogies between analogies. All science is. Scientists want to show that things that don't look alike are really the same. That is one of their innermost Freudian motivations. In fact, that is what we mean by understanding.»
> — GIAN-CARLO ROTA
> ('A Mathematician's Gossip', em: Indiscrete Thoughts, 1997)

Exemplos 9.1. (1) De modo análogo, o número de funções sobrejectivas de um conjunto X num conjunto Y pode ser calculado observando primeiro que

$$\mathrm{Func}(X,Y) = \bigcup_{S\subseteq Y} \mathrm{Sobrej}(X,S) \quad \text{(união disjunta)}$$

(uma vez que cada função $f\colon X \to Y$ pode ser identificada pela função sobrejectiva $f\colon X \to f[X]$). Portanto,

$$|\mathrm{Func}(X,Y)| = \sum_{S\subseteq Y} |\mathrm{Sobrej}(X,S)|\,.$$

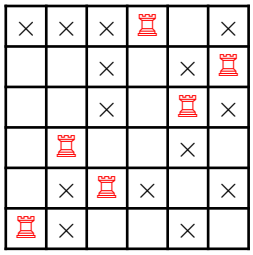
Pela inversão de Möbius-Rota, considerando $f, g\colon \mathcal{P}(Y) \to \mathbb{R}$ dadas respectivamente por $S \mapsto |\mathrm{Func}(X,S)|$ e $S \mapsto |\mathrm{Sobrej}(X,S)|$, obtemos

$$|\mathrm{Sobrej}(X,Y)| = \sum_{S\subseteq Y} (-1)^{|Y|-|S|}\ |\mathrm{Func}(X,S)|$$

pelo que, se X tem m elementos e Y tem n elementos ($m \geq n$), então

$$|\mathrm{Sobrej}(X,Y)| = \sum_{r=0}^{n} (-1)^{n-r} \binom{n}{r} r^m. \tag{9.1.1}$$

(2) Apliquemos agora a inversão de Möbius-Rota ao problema, mais complicado, do cálculo do número de maneiras de dispor n torres num tabuleiro de xadrez $n \times n$ com *posições proibidas*, de modo a não se atacarem mutuamente (isto é, de modo a que nenhum par de torres esteja numa linha ou coluna comuns). Por exemplo, no caso $n = 6$, com posições proibidas marcadas com $\times$, uma dessas configurações é a seguinte:



Representemos o tabuleiro com posições proibidas pela matriz binária (0: posição proibida)

$$T = [t_{ij}] = \begin{bmatrix} 0 & 0 & 0 & \mathbf{1} & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & \mathbf{1} \\ 1 & 1 & 0 & 1 & \mathbf{1} & 0 \\ 1 & \mathbf{1} & 1 & 1 & 0 & 1 \\ 1 & 0 & \mathbf{1} & 0 & 1 & 0 \\ \mathbf{1} & 0 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

É claro que colocar seis torres no tabuleiro, sem se atacarem, corresponde a uma colecção de seis 1's em T (marcados a carregado na matriz) com a seguinte propriedade: cada linha e cada coluna contém exactamente um desses 1's. Podemos ainda representar esta colecção pela sequência (permutação) dos números das colunas onde estão esses 1's (começando de cima para baixo, a partir da linha 1): $(4, 6, 5, 2, 3, 1)$. Em geral, n torres colocadas num tabuleiro $n \times n$ correspondem a uma permutação σ (no grupo simétrico S_n de todas as permutações de X_n) com $t_{i\,\sigma(i)} = 1$ $(i = 1, 2, \ldots, n)$ e, portanto,

$$\prod_{i=1}^{n} t_{i\,\sigma(i)} = t_{1\,\sigma(1)} t_{2\,\sigma(2)} \cdots t_{n\,\sigma(n)} = 1.$$

Como este produto só poderá ser, além de 1, igual a 0 (precisamente quando pelo menos uma das torres for colocada numa posição proibida), torna-se evidente que o número que queremos calcular é precisamente o chamado *permanente* da matriz T, ou seja, a soma

$$\sum_{\sigma \in S_n} \prod_{i=1}^{n} t_{i\,\sigma(i)}. \tag{9.1.2}$$

Consideremos novamente $P = (\mathcal{P}(X_n), \subseteq)$. Cada subconjunto S de cardinalidade k de X_n corresponde a uma escolha de k colunas de T. Seja $\mathrm{Func}(X_n, S)$ o conjunto de todas as funções $\sigma\colon X_n \to S$ e seja $\mathrm{Sobrej}(X_n, S)$

o respectivo subconjunto de funções sobrejectivas. Como já observámos no exemplo anterior, $\mathrm{Func}(X_n, S)$ é a união disjunta $\bigcup_{R\subseteq S} \mathrm{Sobrej}(X_n, R)$.

Consideremos agora a função $f\colon \mathcal{P}(X_n) \to \mathbb{R}$ definida por

$$f(S) = \sum_{\sigma\in \mathrm{Sobrej}(X_n,S)} \prod_{i=1}^{n} t_{i\,\sigma(i)}.$$

Note que $f(\varnothing) = 0$ e que $f(X_n)$ é a soma (9.1.2), uma vez que qualquer função sobrejectiva $\sigma\colon X_n \to X_n$ é uma bijecção.

Finalmente, definindo

$$g(S) = \sum_{R\subseteq S} f(R) \quad (S \in \mathcal{P}(X_n)),$$

a inversão de Möbius-Rota diz-nos que

$$f(X_n) = \sum_{S\subseteq X_n} (-1)^{n-|S|} g(S).$$

Mas, contrariamente a $f(X_n)$ (mais geralmente $f(S)$), não é difícil calcular o valor de $g(S)$ directamente, pelo que conseguiremos ter assim uma forma de cálculo para $f(X_n)$. De facto, não é difícil concluir que

$$g(S) = \sum_{\theta\in \mathrm{Func}(X_n,S)} t_{1\,\theta(1)}\; t_{2\,\theta(2)} \cdots t_{n\,\theta(n)} \quad (S \in \mathcal{P}(X_n))$$

é igual a

$$\Big(\sum_{j\in S} t_{1j}\Big) \cdot \Big(\sum_{j\in S} t_{2j}\Big) \cdots \Big(\sum_{j\in S} t_{nj}\Big).$$

Logo

$$f(X_n) = \sum_{S\subseteq X_n} (-1)^{n-|S|} \prod_{i=1}^{n}\Big(\sum_{j\in S} t_{ij}\Big). \tag{9.1.3}$$

Temos assim uma fórmula de cálculo do número de maneiras de dispor n torres num tabuleiro de xadrez $n \times n$ com *posições proibidas*, de modo a não se atacarem mutuamente, de fácil implementação computacional: escolher um conjunto S de colunas, calcular a soma dos elementos de cada linha nessas colunas, multiplicar estas somas todas, justapor-lhe o sinal apropriado, e somar os resultados sobre todas as escolhas de S (o número de parcelas é igual a 2^n mas algumas são nulas, precisamente as correspondentes a conjuntos S para os quais a matriz tem uma linha que só tem zeros nas colunas de S). Deixamos a cargo do leitor o cálculo (um pouco fastidioso...) desse número no tabuleiro do exemplo inicial com matriz associada T.

10 Mais sobre funções de Möbius-Rota

> «We tend to think of generating functions as related to combinatorics, and Dirichlet series as related to number theory. But this is because combinatorists prefer adding finite sets, while number theorists get more excited about multiplying them. (Primes and all that.)»
>
> — John Baez
> (The n-Category Café, Maio de 2011)

Terminamos o artigo com o retorno às conexões de Galois num resultado de Rota [16] que mostra como se relacionam as funções de Möbius disponíveis em cada um dos lados da conexão.

Teorema 10.1. *Sejam P e Q conjuntos parcialmente ordenados finitos, P com primeiro e último elementos 0 e 1 (diferentes) e Q com último elemento 1. Sejam μ_P e μ_Q as respectivas funções de Möbius. Se $f\colon P \to Q$ e $g\colon Q \to P$ constituírem uma conexão de Galois $f \dashv g$ tal que*

(G1) *$f(a) = 1$ se e só se $a = 1$,*

(G2) *$g(1) = 1$,*

então

$$\mu_P(0,1) = \sum_{y<1} \mu_Q(y,1)\,\zeta(g(y),0) = \sum_{y\colon g(y)=0} \mu_Q(y,1).$$

Demonstração. Como $f(a) \leq b$ se e só se $a \leq g(b)$, então, para cada $b \in Q$,

$$\sum_{y\colon y\leq b} \delta(f(a),y) = \zeta_Q(a,g(b)). \tag{10.1.1}$$

Aplicando a (10.1.1) a inversão de Möbius-Rota relativamente a Q obtemos

$$\delta(f(a),1) = \sum_{y\leq 1} \mu_Q(y,1)\,\zeta(a,g(y)). \tag{10.1.2}$$

$\delta(f(a),1)$ toma o valor 1 se e só se $f(a) = 1$, isto é, $a = 1$, por (G1). Para os restantes valores de a, $\delta(f(a),1) = 0$. Assim, $\delta(f(a),1) = 1-\zeta(a,1)+\delta(a,1)$. Denotando a função de incidência $\zeta - \delta$ por n, temos $\delta(f(a),1) = 1 - n(a,1)$ e a identidade (10.1.2) pode então ser reescrita como

$$1 - n(a,1) = \zeta(a,g(1)) + \sum_{y<1} \mu_Q(y,1)\,\zeta(a,g(y)).$$

Mas a condição (G2) implica $\zeta(a, g(1)) = \zeta(a, 1) = 1$ para qualquer $a \in P$. Portanto,

$$-n(a, 1) = \sum_{y<1} \mu_Q(y, 1)\, \zeta(a, g(y)).$$

Agora, como $\zeta = \delta + n$, temos $\delta - \mu * n = \delta - \mu * (\zeta - \delta) = \delta - \delta + \mu * \delta = \mu$. Logo

$$\mu_P(0, 1) = - \sum_{0 \le a \le 1} \mu_P(0, a)\, n(a, 1) = \sum_{0 \le a \le 1} \sum_{y<1} \mu_Q(y, 1) \mu_P(0, a) \zeta(a, g(y)).$$

Finalmente, trocando a ordem dos somatórios,

$$\mu_P(0, 1) = \sum_{y<1} \mu_Q(y, 1) \sum_{0 \le a \le 1} \mu_P(0, a) \zeta(a, g(y))$$

e o somatório mais à direita é igual a $(\mu_P * \zeta)(0, g(y)) = \delta(0, g(y)) = \zeta(g(y), 0)$. □

Com este resultado, fixando um dos conjuntos parcialmente ordenados, e variando o outro entre ordens parciais nos quais a função de Möbius é bem conhecida, podemos obter informação sobre a função de Möbius no conjunto previamente fixado.

Comentários finais. Do mesmo modo que uma conexão de Galois é um exemplo muito particular de um conceito fundamental da moderna teoria das categorias – o conceito de adjunção –, base de muitos teoremas importantes que relacionam áreas distintas da matemática, as inversões de Möbius também podem ser formuladas categorialmente em contextos mais gerais que os clássicos (o conjunto parcialmente ordenado dos inteiros positivos, ordenado pela relação de divisibilidade, no caso de Möbius, e qualquer conjunto parcialmente ordenado *localmente finito*, na generalização de Rota). As ideias de Rota criaram as condições para essas recentes extensões (a categorias com alguma condição de *finitude*). Isto foi feito essencialmente de dois modos distintos, por Content, Lemay e Leroux [5] e, mais recentemente, de um modo mais geral, por Leinster [11, 12]. A segunda abordagem faz parte da teoria da característica de Euler de uma categoria [11], que coincide com a característica de Euler topológica quando esta existe (mas é também válida em situações diversas nas quais esta não existe). Curiosamente, neste contexto categorial é possível generalizar ainda mais o princípio da inclusão-exclusão a fórmulas sobre cardinais de colimites de conjuntos!

> «Go to the roots, of these calculations! Group the operations. Classify them according to their complexities rather than their appearances! This, I believe, is the mission of future mathematicians. This is the road on which I am embarking in this work.»
>
> — Évariste Galois
> (Do prefácio do seu último manuscrito, 1832)

Referências

[1] S. Beatty, Problem 3173, *Amer. Math. Monthly* 33 (1926) 159.

[2] G. Birkhoff, *Lattice Theory*, Amer. Math. Soc. Colloq. Publ. 25 (1967).

[3] R. Brualdi, *Introductory Combinatorics*, 5^a edição, Prentice Hall (2010).

[4] T. Y. Chow, The combinatorics behind number-theoretic sieves, *Adv. Math.* 138 (1998) 293-305.

[5] M. Content, F. Lemay e P. Leroux, Catégories de Möbius et fonctorialités: Un cadre général pour l'inversion de Möbius, *Journal of Combinatorial Theory Series A* 28 (1980) 169-190.

[6] K. Denecke, M. Erné e S. L. Wismath, *Galois Connections and Applications*, Mathematics and its Applications, vol. 565, Kluwer, Dordrecht (2004).

[7] P. Doubilet, G.-C. Rota e R. P. Stanley, On the foundations of combinatorial theory VI: The idea of generating function, *Berkeley Symp. on Math. Statist. and Prob.*, vol. 2, pp. 267-318, Univ. of Calif. Press (1972).

[8] M. Erné, J. Koslowski, A. Melton e G. Strecker, A primer on Galois connections, *Papers on general topology and applications* (Madison, WI, 1991), pp. 103–125, Ann. New York Acad. Sci., 704 (1993).

[9] J. Lambek, Some Galois connections in elementary number theory, *J. Number Theory* 47 (1994) 371–377.

[10] J. Lambek e L. Moser, Inverse and complementary sequences of natural numbers, *Amer. Math. Monthly* 61 (1954) 454-458.

[11] T. Leinster, The Euler characteristic of a category, *Documenta Math.* 13 (2008) 21-49.

[12] T. Leinster, Notions of Möbius inversion, *Bulletin of the Belgian Mathematical Society* 19 (2012) 911-935.

[13] A. F. Möbius, Über eine besondere Art von Umkehrung der Reihen, *J. reine angew. Math.* 9 (1832) 105-123.

[14] O. Ore, Galois connexions, *Trans. Amer. Math. Soc.* 55 (1944) 493-513.

[15] J. Picado e A. Pultr, *Frames and locales: Topology without points*, Frontiers in Mathematics, vol. 28, Springer, Basel (2012).

[16] G.-C. Rota, On the foundations of combinatorial theory I: Theory of Möbius functions, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 2 (1964) 340-368.

INS AND OUTS OF INCLUSION–EXCLUSION

R.E. Hartwig, Min Kang
North Carolina State University
e-mail: hartwig@math.ncsu.edu
kang@math.ncsu.edu

Abstract Inclusion-Exclusion identities and inequalities are obtained for valuations. Applications to cardinality, probability, max/min and least common multiples are presented.

keywords: Inclusion-Exclusion principle; Boole's inequality; Bonferroni inequality.

1 Introduction

In this treatise we shall examine the identities and inequalities associated with a valuation on a set $(S, +, \cdot)$. For finite sets these give the "inclusion-exclusion" (InEx) formulae and inequalities, while for probability the former yields the Poincaré's formula.

The basic valuation formula can directly be applied to such non-negative functions as content or measure (volume, area, length, weight, size, probability), as well as dimension, min/max and gcd/lcm. In the former case it is also possible to use indicator functions followed by the taking of a suitable linear functional, such as expected value, which will get us back to probability. When f is multiplicative, an easier and non-inductive way of proving the valuation formula is by using the symmetric functions of the roots $f(a_i)$ of a suitable polynomial.

Throughout the paper, $(S, +, \cdot)$ is a set S with two binary operations "$+$" and "$\cdot$", which are commutative and associative, and we further assume multiplicative idempontency $a \cdot a = a$ for all $a \in S$ and the distributive law $a \cdot (b + c) = a \cdot b + a \cdot c$, for all $a, b, c \in S$. The target space is a set $(T, \oplus, \otimes)$ with binary operations $\oplus$ and $\otimes$.

In general, we do not assume idempotency for addition, but when we do, we will clearly state that $a + a = a$ for all $a \in S$ as well. For instance, $(S, +, \cdot)$ can be a commutative distributive complemented lattice with $a \cdot b = a \wedge b = glb(a, b)$ and $a + b = a \vee b = lub(a, b)$. The most important of these is the power set $(\mathcal{P}(X), \cup, \cap)$.

Definition 1.1. *A function* $f:(S,+,\cdot)\to(T,\oplus,\otimes)$ *is called an* α-valuation *if*

$$f(a+b)\oplus[\alpha\otimes f(a\cdot b)]=f(a)\oplus f(b), \tag{1}$$

where $\alpha\in T$ *and* $a,b\in S$.

We shall examine the interplay between an α-valuation f and the operations of $+$ and $(\cdot)$ defined on S, and $\oplus$ and $\otimes$ defined on T. When there is no risk of ambiguity we shall as always write this as $f(a+b)+\alpha f(ab)=f(a)+f(b)$, in which the "multiplicative dot" has been dropped.

When T admits an additive inverse, we may rewrite this as $f(a+b)=f(a)+f(b)-\alpha f(ab)$. Needless to say when α is absent or $\alpha=1$, we have a more symmetric unit-valuation. When this is the case, we do not need to define multiplication on T.

We shall primarily be interested in the case where (T,$\oplus,\otimes$) is $\mathbb{R}^+$, and $\alpha=1+\varepsilon\geq 1$ (see [2]).

We shall examine additive as well as multiplicative results for α-valuations. Non-zero values of ε are used for example in the **exclusive-or** case (addition on a powerset), where $\alpha=2$ and $f(a+a)=f(\emptyset)=0$. It should be noted that $f(a+a)=(2-\alpha)f(a)$ and hence we have $f(a+a)=f(a)$ iff $\alpha=1$.

2 Additive Results for α-valuations

We shall first need several definitions and notations dealing with triangular "slices".

For a given set $S\subseteq\mathbb{N}$, with $\#(S)=M$, we introduce the associated collection of lists.

Definition 2.1. *For* $k\leq M$,

$$V_k^S=\{(i_1,\ldots,i_k);i_1<i_2<\cdots<i_k,i_r\in S,\ \forall r=1,\ldots,k\}.$$

We may alternatively think of this as the collection of all $\binom{M}{k}$ combinations of the M objects in S, taken k at a time. For example $V_2^{2,3,4}=\{(2,3),(2,4),(3,4)\}$. Initially we shall focus on $S=\{1,2,\ldots,n\}$, and shorten $V_k^{1,2,\ldots,n}$ to $U_k^{(n)}$. In particular, $U_n^{(n)}=\{(1,2,\ldots,n)\}$ is made of a single string.

Now let $A=\{x_1,\ldots,x_n\}$ be a collection of symbols. Then for $a\in A$ and $V\subseteq A^k$ we define $(V,a)=\{(x_1,\ldots,x_k,a);(x_1,\ldots,x_k)\in V\}$. We now have

Lemma 2.1. $U_k^{(n+1)} = U_k^{(n)} \cup (U_{k-1}^{(n)}, n+1)$

Proof. This is nothing but a partitioning of $U_k^{(n+1)}$ into terms that do or do not contain the highest index $n+1$. □

Next, we introduce a collection of functions $g_k : A^k \to (T, \oplus)$, $k = 1, \dots, n$, where T is a suitable set with addition $\oplus$. For each $k = 1, 2, \dots, n$ we now define

Definition 2.2. $\bigoplus_{U_k^{(n)}} g_k(x_{i_1}, \dots, x_{i_k}) = \bigoplus_{1 \le i_1 < i_2 < \cdots < i_k \le n} g_k(x_{i_1}, x_{i_2}, \dots, x_{i_k})$, *where the addition is in* T.

We may now state:

Corollary 2.1.

(i) $$\bigoplus_{U_k^{(n+1)}} g_k(x_{i_1}, \dots, x_{i_k}) = \bigoplus_{U_k^{(n)}} g_k(x_{i_1}, \dots, x_{i_k}) \oplus \bigoplus_{U_{k-1}^{(n)}} g_k(x_{i_1}, \dots, x_{i_{k-1}}, n+1).$$

(ii) $$\bigoplus_{k=1}^{n+1} \bigoplus_{U_k^{(n+1)}} g_k(x_{i_1}, \dots, x_{i_k})$$
$$= \bigoplus_{k=1}^{n} \bigoplus_{U_k^{(n)}} g_k(x_{i_1}, \dots, x_{i_k}) \oplus \bigoplus_{k=1}^{n+1} \bigoplus_{U_{k-1}^{(n)}} g_k(x_{i_1}, \dots, x_{i_{k-1}}, n+1).$$

Note that in the second summation the term with $k = n+1$ is absent as $U_{n+1}^{(n)} = \emptyset$.

As a special case, we choose $g_k(a_1, \dots, a_k) = (-\alpha)^k f(a_1 \cdots a_k)$, where f is an α-valuation from $(S, +, \cdot)$ to $T = \mathbb{R}^+$ evaluated at the **product** of a_i. We further let $\mathbf{a} = (a_1, a_2, \dots)$ be a sequence of elements from S and for any $b \in S$ we set b$\mathbf{a}$= $(ba_1, ba_2, \dots)$.

For convenience we now also define the following "symmetric functions"

$$\sigma_k^{(n)}(\mathbf{a}, f) = \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} f(a_{i_1} a_{i_2} \dots a_{i_k}).$$

Note that $\sigma_n^{(n)}(\mathbf{a}, f) = f(a_1 \cdots a_n)$. When there is no risk of confusion we shall drop the brackets in the superscript and use σ_k^n instead of $\sigma_k^{(n)}$.

Again, if f is an α-valuation from $(S, +, \cdot)$ to $\mathbb{R}^+$ then we may recast Corollary (2.1) in terms of f as

Corollary 2.2.

$$(i)\quad \sigma_k^{n+1}(\boldsymbol{a}, f) = \sigma_k^n(\boldsymbol{a}, f) + \sigma_{k-1}^n(b\boldsymbol{a}, f),$$

where $k = 1, 2, \ldots, n+1$ *and* $b = a_{n+1}$.

$$(ii)\quad \sum_{k=1}^{n+1}(-\alpha)^{k-1}\sigma_k^{n+1}(\boldsymbol{a}, f) = \sum_{k=1}^{n}(-\alpha)^{k-1}\sigma_k^n(\boldsymbol{a}, f) + \sum_{k=1}^{n+1}(-\alpha)^{k-1}\sigma_{k-1}^n(b\boldsymbol{a}, f)\,.$$

For consistency, we let $\sigma_0^n(b\mathbf{a}, f) = f(b)$ for $k = 1$ in part (i) of Corollary (2.1) and also note that $\sigma_k^n(\mathbf{a}, f)$ is undefined for $k > n$.

We shall primarily be interested in the special sums,

$$S_n = S_n(\mathbf{a}, f) = f(\sum_{i=1}^{n} a_i),$$

where $\mathbf{a} = (a_1, a_2, \ldots)$. Using (1), we observe that

Lemma 2.2. *If* $b = a_{n+1}$, *then*

$$S_{n+1}(\boldsymbol{a}) + \alpha S_n(b\boldsymbol{a}) = S_n(\boldsymbol{a}) + f(b) \tag{2}$$

Proof. $f(\sum_{i=1}^{n} a_i + b) + \alpha f[(\sum_{i=1}^{n} a_i)b] = f(\sum_{i=1}^{n} a_i) + f(b).$ □

Our aim is to solve the recurrence (2) for S_n. When T does not admit an additive inverse, we have to separate the even and odd values of n. We shall first introduce the following summations.

Definition 2.3. *(i)* $\lambda_k^n = \sum_{i=1}^{k} \alpha^{2i-2}\sigma_{2i-1}^n, \quad 2k-1 \le n,$

(ii) $\mu_k^n = \sum_{i=1}^{k} \alpha^{2i-1}\sigma_{2i}^n, \quad 2k \le n.$

In this we dropped the superscript braces for convenience. For example,

$$\lambda_1^n = \sigma_1^n, \quad \lambda_2^n = \sigma_1^n + \alpha^2\sigma_3^n, \quad \lambda_3^n = \sigma_1^n + \alpha^2\sigma_3^n + \alpha^4\sigma_5^n$$

and

$$\mu_1^n = \alpha\sigma_2^n, \quad \mu_2^n = \alpha\sigma_2^n + \alpha^3\sigma_4^n, \quad \mu_3^n = \alpha\sigma_2^n + \alpha^3\sigma_4^n + \alpha^5\sigma_6^n.$$

Here μ_0^n is not defined, or it can be taken as the additive identity in T, if it exists. Before we can use induction we need the following identities.

Lemma 2.3. *Let* $b = a_{n+1}$, *then the following identities among* $\{\lambda_r^n\}$ *and* $\{\mu_r^n\}$ *hold.*

(i) $\lambda_r^{n+1}(\boldsymbol{a}) = \lambda_r^n(\boldsymbol{a}) + \alpha\mu_{r-1}^n(b\boldsymbol{a}) + f(b)$.

(ii) $\lambda_{r+1}^{n+1}(\boldsymbol{a}) = \lambda_r^n(\boldsymbol{a}) + \alpha\mu_r^n(b\boldsymbol{a}) + f(b) + \alpha^{2r}\sigma_{2r+1}^n(\boldsymbol{a})$.

(iii) $\mu_r^{n+1}(\boldsymbol{a}) = \mu_r^n(\boldsymbol{a}) + \alpha\lambda_r^n(b\boldsymbol{a})$.

(iv) $\mu_r^{n+1}(\boldsymbol{a}) + \alpha^{2r+1}\sigma_{2r+1}^n(b\boldsymbol{a}) = \mu_r^n(\boldsymbol{a}) + \alpha\lambda_{r+1}^n(b\boldsymbol{a})$.

Proof. Note that $b = a_{n+1}$ throughout.

(i) $\lambda_r^n(\mathbf{a})+\alpha\mu_{r-1}^n(b\mathbf{a})+f(b) = \sum_{k=1}^{r}\alpha^{2k-2}\sigma_{2k-1}^n(\mathbf{a})+\alpha\{\sum_{k=1}^{r-1}\alpha^{2k-1}\sigma_{2k}^n(b\mathbf{a})\}+ f(b) = [\sigma_1^n(\mathbf{a}) + f(b)] + \sum_{k=2}^{r}\alpha^{2k-2}\sigma_{2k-1}^n(\mathbf{a}) + \sum_{k=1}^{r-1}\alpha^{2k}\sigma_{2k}^n(b\mathbf{a}) = \sigma_1^{n+1}(\mathbf{a}) + \sum_{t=1}^{r-1}\alpha^{2t}\sigma_{2t+1}^n(\mathbf{a}) + \sum_{k=1}^{r-1}\alpha^{2k}\alpha_{2k}^n(b\mathbf{a}) = \sigma_1^{n+1}(\mathbf{a}) + \sum_{t=1}^{r-1}\alpha^{2t}[\sigma_{2t+1}^n(\mathbf{a}) + \sigma_{2t}^n(b\mathbf{a})] = \sum_{k=0}^{r-1}\alpha^{2k}\sigma_{2k+1}^{n+1}(\mathbf{a}) = \sum_{s=1}^{r}\alpha^{2s-2}\sigma_{2s-1}^{n+1}(\mathbf{a}) = \lambda_r^{n+1}(\mathbf{a})$.

(ii) $\lambda_r^n(\mathbf{a}) + \alpha\mu_r^n(b\mathbf{a}) + f(b) + \alpha^{2r}\sigma_{2r+1}^n(\mathbf{a}) = \lambda_r^n(\mathbf{a}) + \alpha\mu_{r-1}^n(b\mathbf{a}) + f(b) + \alpha^{2r}\sigma_{2r+1}^n(\mathbf{a})+\alpha^{2r}\sigma_{2r}^n(b\mathbf{a}) = \lambda_r^{n+1}(\mathbf{a})+\alpha^{2r}\sigma_{2r}^n(b\mathbf{a})+\alpha^{2r}\sigma_{2r+1}^n(\mathbf{a}) = \lambda_r^{n+1}(\mathbf{a})+ \alpha^{2r}\sigma_{2r+1}^{n+1}(\mathbf{a}) = \lambda_{r+1}^{n+1}(\mathbf{a})$, which proves (ii).

(iii) The right hand side equals $\sum_{k=1}^{r}\alpha^{2k-1}\sigma_{2k}^n(\mathbf{a})+\alpha\sum_{k=1}^{r}\alpha^{2k-2}\sigma_{2k-1}^n(b\mathbf{a}) = \sum_{k=1}^{r}\alpha^{2k-1}[\sigma_{2k}^n(\mathbf{a}) + \sigma_{2k-1}^n(b\mathbf{a})] = \sum_{k=1}^{r}\alpha^{2k-1}\sigma_{2k}^{n+1}(\mathbf{a}) = \mu_r^{n+1}(\mathbf{a})$.

(iv) From the right hand side, we have $\sum_{k=1}^{r}\alpha^{2k-1}\sigma_{2k}^n(\mathbf{a}) + \alpha\sum_{k=1}^{r+1}\alpha^{2k-2}\sigma_{2k-1}^n(b\mathbf{a}) = \sum_{k=1}^{r}\alpha^{2k-1}[\sigma_{2k}^n(\mathbf{a}) + \sigma_{2k-1}^n(b\mathbf{a})] + \alpha^{2r+1}\sigma_{2r+1}^n(b\mathbf{a})$, which is the left hand side. □

It should be noted that when $2r \geq n$, then σ_{2r+1}^n is absent. Moreover by part (iii) and (iv) in Lemma 2.3, we also have

Corollary 2.3. *(i)* $\mu_m^{2m+2}(\boldsymbol{a}) = \mu_m^{2m+1}(\boldsymbol{a}) + \alpha\lambda_m^{2m+1}(b\boldsymbol{a})$

(ii) $\mu_m^{2m+2}(\boldsymbol{a}) + \alpha^{2m+1}\sigma_{2m+1}^{2m+1}(b\boldsymbol{a}) = \mu_m^{2m+1}(\boldsymbol{a}) + \alpha\lambda_{m+1}^{2m+1}(b\boldsymbol{a})$.

We are now ready for the following identities on $S_n(\mathbf{a})$.

Theorem 2.1.

$$(i)\ S_{2m}(\boldsymbol{a}) + \mu_m^{2m}(\boldsymbol{a}) = \lambda_m^{2m}(\boldsymbol{a}) \quad \textit{for all } m \in \mathbb{N}\,, \tag{3}$$

$$(ii)\ S_{2m+1}(\boldsymbol{a}) + \mu_m^{2m+1}(\boldsymbol{a}) = \lambda_{m+1}^{2m+1}(\boldsymbol{a}) \quad \textit{for all } m \in \mathbb{N}\cup\{0\}\,. \tag{4}$$

Proof. (a) Both results are clearly true for initial values of m (in other words, $m = 1$ for part (i) and $m = 0$ for part (ii)). So let us assume that the result is true for $n = 2m$, and then show it for $n = 2m + 1$. Consider (2) with $n = 2m$ and add $\mu_m^{2m}(\mathbf{a})$ and $\alpha\mu_m^{2m}(b\mathbf{a})$ to both sides. This gives

$$S_{2m+1}(\mathbf{a})+\alpha[S_{2m}(b\mathbf{a})+\mu_m^{2m}(b\mathbf{a})]+\mu_m^{2m}(\mathbf{a}) = S_{2m}(\mathbf{a})+\mu_m^{2m}(\mathbf{a})+f(b)+\alpha\mu_m^{2m}(b\mathbf{a})$$

which by induction hypothesis reduces to

$$S_{2m+1}(\mathbf{a}) + \alpha\lambda_m^{2m}(b\mathbf{a}) + \mu_m^{2m}(\mathbf{a}) = \lambda_m^{2m}(\mathbf{a}) + f(b) + \alpha\mu_m^{2m}(b\mathbf{a}).$$

This in turn can be reduced by part (ii) and (iii) in Lemma (2.3) to

$$S_{2m+1}(\mathbf{a}) + \mu_m^{2m+1}(\mathbf{a}) = \lambda_{m+1}^{2m+1}(\mathbf{a}) - \alpha^{2m}\sigma_{2m+1}^{2m}(\mathbf{a}) = \lambda_{m+1}^{2m+1}(\mathbf{a}).$$

(b) Next we assume that the result holds for $n = 2m + 1$ and we will show it holds for $n = 2m + 2$. Consider (2) with $n = 2m + 1$ and $b = a_{2m+2}$, and add $\mu_m^{2m+1}(\mathbf{a})$ and $\alpha\mu_m^{2m+1}(b\mathbf{a})$ to both sides. This gives

$$\begin{aligned} & S_{2m+2}(\mathbf{a}) + \alpha[S_{2m+1}(b\mathbf{a}) + \mu_m^{2m+1}(b\mathbf{a})] + \mu_m^{2m+1}(\mathbf{a}) = \\ = \quad & S_{2m+1}(\mathbf{a}) + \mu_m^{2m+1}(\mathbf{a}) + f(b) + \alpha\mu_m^{2m+1}(b\mathbf{a}) \end{aligned}$$

which by induction hypothesis reduces to

$$S_{2m+2}(\mathbf{a}) + \alpha\lambda_{m+1}^{2m+1}(b\mathbf{a}) + \mu_m^{2m+1}(\mathbf{a}) = \lambda_{m+1}^{2m+1}(\mathbf{a}) + \alpha\mu_m^{2m+1}(b\mathbf{a}) + f(b)\,.$$

On account of Lemma (2.3) part (iv) and Corollary (2.3), we see that

$$S_{2m+2}(\mathbf{a}) + \mu_m^{2m+2}(\mathbf{a}) + \alpha^{2m+1}\sigma_{2m+1}^{2m+1}(b\mathbf{a}) = \lambda_{m+1}^{2m+2}(\mathbf{a}).$$

Lastly, note that $\sigma_{2m+1}^{2m+1}(b\mathbf{a}) = \sigma_{2m+2}^{2m+2}(\mathbf{a})$ with $b = a_{2m+2}$, and so we reach

$$S_{2m+2}(\mathbf{a}) + \mu_{m+1}^{2m+2}(\mathbf{a}) = \lambda_{m+1}^{2m+2}(\mathbf{a}).$$ □

If T admits additive inverses, we have the following valuation formulæ.

Corollary 2.4. (i) $S_n(\boldsymbol{a}) = f(\sum_{i=1}^{n} a_i) = \sum_{k=1}^{n}(-\alpha)^{k-1}\sigma_k^n(\boldsymbol{a}, f)$ (5)

$$(ii) \quad S_{2m}(\boldsymbol{a}) = \lambda_m^{2m}(\boldsymbol{a}) - \mu_m^{2m}(\boldsymbol{a})$$

$$(iii) \quad S_{2m+1}(\boldsymbol{a}) = \lambda_{m+1}^{2m+1}(\boldsymbol{a}) - \mu_m^{2m+1}(\boldsymbol{a})$$

Proof. We may list the first few sums in which $\sigma_k^n = \sigma_k^n(\mathbf{a})$:

$$\begin{array}{lcll} S_2 & + & \alpha\sigma_2^2 & = \sigma_1^2 \\ S_3 & + & (\alpha\sigma_2^3) & = (\sigma_1^3 + \alpha^2\sigma_3^3) \\ S_4 & + & (\alpha\sigma_2^4 + \alpha^3\sigma_4^4) & = (\sigma_1^4 + \alpha^2\sigma_3^4) \\ S_5 & + & (\alpha\sigma_2^5 + \alpha^3\sigma_4^5) & = (\sigma_1^5 + \alpha^2\sigma_3^5 + \alpha^4\sigma_5^5) \end{array}$$

$$\begin{array}{rcl} S_6 + (\alpha\sigma_2^6 + \alpha^3\sigma_4^6 + \alpha^5\sigma_6^6) &=& (\sigma_1^6 + \alpha^2\sigma_3^6 + \alpha^4\sigma_5^6) \\ S_7 + (\alpha\sigma_2^7 + \alpha^3\sigma_4^7 + \alpha^5\sigma_6^7) &=& (\sigma_1^7 + \alpha^2\sigma_3^7 + \alpha^4\sigma_5^7 + \alpha^6\sigma_7^7)\,. \end{array}$$

More generally,

$$S_{2m} + (\alpha\sigma_2^{2m} + \alpha^3\sigma_4^{2m} + \cdots + \alpha^{2m-1}\sigma_{2m}^{2m}) = \sigma_1^{2m} + \alpha^2\sigma_3^{2m} + \cdots + \alpha^{2m-2}\sigma_{2m-1}^{2m}$$

and

$$S_{2m+1} + (\alpha\sigma_2^{2m+1} + \alpha^3\sigma_4^{2m+1} + \cdots + \alpha^{2m-1}\sigma_{2m}^{2m+1}) \\ = \sigma_1^{2m+1} + \alpha^2\sigma_3^{2m+1} + \cdots + \alpha^{2m}\sigma_{2m+1}^{2m+1}\,. \qquad \square$$

Examples If an additive inverse exists, we may write

$$\text{(i)}\ f(a+b+c) = f(a)+f(b)+f(c) - \alpha[f(ab)+f(ac)+f(bc)]+\alpha^2 f(abc). \quad (6)$$

$$\text{(ii)}\ \begin{array}{c} f(a+a) = (2-\alpha)f(a) \\ f(a+b+b) = f(a) + (2-\alpha)f(b) - (2-\alpha)\alpha f(ab). \end{array} \quad (7)$$

3 The valuation inequalities

Throughout let f be an α-valuation from $(S, +, \cdot)$ to $\mathbb{R}^+$. We begin by considering the matrices $M = (m_{ij})$ and $N = (n_{ij})$ with

$$m_{ij} = \lambda_i^{2m} - \mu_j^{2m}, \qquad \text{for} \quad 1 \le i, j \le m$$

$$n_{ij} = \lambda_i^{2m+1} - \mu_j^{2m+1}, \qquad \text{for} \quad 1 \le i \le m+1, 1 \le j \le m.$$

Then

$$m_{ij} \le m_{i+1,j}, \quad m_{i,j+1} \le m_{ij} \quad \textit{and} \quad n_{ij} \ge n_{i,j+1}, \quad n_{i+1,j} \ge n_{ij}.$$

In other words, the elements in the matrix M as well as N **increase** from right to left and top to bottom. We shall next use induction to prove the valuation inequalities. We first assume the validity for even n and prove it for odd n, and then reverse the parity.

Theorem 3.1. *(a) For $n = 2m$ with $m \in \mathbb{N}$ and any $k \ge 1$,*

$$(a_1) S_{2m} \le \sigma_1^{2m}(\boldsymbol{a}) - \alpha\sigma_2^{2m}(\boldsymbol{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m}(\boldsymbol{a}) = \lambda_{k+1}^{2m} - \mu_k^{2m},\ \textit{if}\ k+1 \le m, \quad (8)$$

$$(a_2) S_{2m} \ge \sigma_1^{2m}(\boldsymbol{a}) - \alpha\sigma_2^{2m}(\boldsymbol{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m}(\boldsymbol{a}) = \lambda_k^{2m} - \mu_k^{2m},\ \textit{if}\ k \le m. \quad (9)$$

(b) For $n = 2m+1$ with $m \in \mathbb{N} \cup \{0\}$ and any $k \geq 1$,

(b_1) $S_{2m+1} \leq \sigma_1^{2m+1}(\boldsymbol{a}) - \alpha\sigma_2^{2m+1}(\boldsymbol{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m+1}(\boldsymbol{a}) = \lambda_{k+1}^{2m+1} - \mu_k^{2m+1}$, *if $k \leq m$,*

(b_2) $S_{2m+1} \geq \sigma_1^{2m+1}(\boldsymbol{a}) - \alpha\sigma_2^{2m+1}(\boldsymbol{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m+1}(\boldsymbol{a}) = \lambda_k^{2m+1} - \mu_k^{2m+1}$, *if $k \leq m$.*

Proof. (a) The inequalities clearly hold for the initial values of m (in other words, $m = 1$ for part (a) and $m = 0$ for part (b)), so let us assume they both hold for all values of $r \leq n$ and that $b = a_{n+1}$. Recall that

$$S_{n+1} = f(\sum_{i=1}^{n} a_i + b) = f(\sum_{i=1}^{n} a_i) + f(b) - \alpha f[\sum_{i=1}^{n}(a_i b)].$$

By induction hypothesis, in the first summation on the right hand side, we may apply the inequality from (8) with $2k+1$ terms and sequence $\mathbf{a}$, while in the second summation we apply (9) with 2k terms and sequence b$\mathbf{a}$, respectively. This gives

$S_{2m+1} \leq [\sigma_1^{2m}(\mathbf{a}) - \alpha\sigma_2^{2m}(\mathbf{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m}(\mathbf{a})] + f(b) - \alpha[\sigma_1^{2m}(b\mathbf{a}) - \alpha\sigma_2^{2m}(b\mathbf{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m}(b\mathbf{a})]$

$= [f(b) + \sigma_1^{2m}(\mathbf{a})] - \alpha[\sigma_2^{2m}(\mathbf{a}) + \sigma_1^{2m}(b\mathbf{a})] + \cdots + \alpha^{2k}[\sigma_{2k+1}^{2m}(\mathbf{a}) + \sigma_{2k}^{2m}(b\mathbf{a})]$

$= \sigma_1^{2m+1}(\mathbf{a}) - \alpha\sigma_2^{2m+1}(\mathbf{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m+1}(\mathbf{a})$.

On the other hand, if we apply (9) in the first summation on the right hand side, with $2k$ terms and sequence $\mathbf{a}$, and apply (8) with $2k-1$ terms and sequence $b\mathbf{a}$, then we obtain

$S_{2m+1} \geq [\sigma_1^{2m}(\mathbf{a}) - \alpha\sigma_2^{2m}(\mathbf{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m}(\mathbf{a})] + f(b) - \alpha[\sigma_1^{2m}(b\mathbf{a}) - \alpha\sigma_2^{2m}(b\mathbf{a}) + \cdots + \alpha^{2k-2}\sigma_{2k-1}^{2m}(b\mathbf{a})]$

$= [f(b) + \sigma_1^{2m}(\mathbf{a})] - \alpha[\sigma_2^{2m}(\mathbf{a}) + \sigma_1^{2m}(b\mathbf{a})] + \cdots - \alpha^{2k-1}[\sigma_{2k}^{2m}(\mathbf{a}) + \sigma_{2k-1}^{2m}(b\mathbf{a})]$

$= \sigma_1^{2m+1}(\mathbf{a}) - \alpha\sigma_2^{2m+1}(\mathbf{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m+1}(\mathbf{a})$.

Next, we assume that (b1) and (b2) hold for $r \leq 2m+1$ and use (a1) and (a2) in the α-evaluation formula for $2m+2$,

$$S_{2m+2}(\mathbf{a}) = S_{2m+1}(\mathbf{a}) + f(b) - \alpha S_{2m+1}(b\mathbf{a})$$

to give

$S_{2m+2} \leq [\sigma_1^{2m+1}(\mathbf{a}) - \alpha\sigma_2^{2m+1}(\mathbf{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m+1}(\mathbf{a})] + f(b) - \alpha[\sigma_1^{2m+1}(b\mathbf{a}) - \alpha\sigma_2^{2m+1}(b\mathbf{a}) + \cdots - \alpha^{2k-1}\sigma_{2k}^{2m+1}(b\mathbf{a})] = [f(b) + \sigma_1^{2m+1}(\mathbf{a})] - \alpha[\sigma_2^{2m+1}(\mathbf{a}) + \sigma_1^{2m+1}(b\mathbf{a})] + \cdots + \alpha^{2k}[\sigma_{2k+1}^{2m+1}(\mathbf{a}) + \sigma_{2k}^{2m+1}(b\mathbf{a})] = \sigma_1^{2m+2}(\mathbf{a}) - \alpha\sigma_2^{2m+2}(\mathbf{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m+2}(\mathbf{a})$.

Likewise, noting that $k+1 \le m+1$, we reach

$$S_{2m+2} \ge [\sigma_1^{2m+1}(\mathbf{a}) - \alpha\sigma_2^{2m+1}(\mathbf{a}) + \cdots - \alpha^{2k+1}\sigma_{2k+2}^{2m+1}(\mathbf{a})] + f(b) - \alpha[\sigma_1^{2m+1}(b\mathbf{a}) - \alpha\sigma_2^{2m+1}(b\mathbf{a}) + \cdots + \alpha^{2k}\sigma_{2k+1}^{2m+1}(b\mathbf{a})] = [f(b) + \sigma_1^{2m+1}(\mathbf{a})] - \alpha[\sigma_2^{2m+1}(\mathbf{a}) + \sigma_1^{2m+1}(b\mathbf{a})] + \cdots - \alpha^{2k+1}[\sigma_{2k+2}^{2m+1}(\mathbf{a}) + \sigma_{2k+1}^{2m+1}(b\mathbf{a})] = \sigma_1^{2m+2}(\mathbf{a}) - \alpha\sigma_2^{2m+2}(\mathbf{a}) + \cdots - \alpha^{2k+1}\sigma_{2k+2}^{2m+2}(\mathbf{a}),$$

which completes the proof. □

The InEx inequalities suggest that there should be inequalities relating the symmetric functions. We shall now show this, for the case where $\alpha f(ab) \le f(a)$ and $k+1 \le n \le 2k+1$. We first need some basic facts about the binomial sets. We shall denote the existence of an injective map from $V_{k+1}^{1,\dots,n}$ into $V_k^{1,..n}$ by $V_{k+1}^{1,\dots,n} \hookrightarrow V_k^{1,\dots,n}$.

Lemma 3.1. *If $k+1 \le n \le 2k+1$ then we can find an injection (one-to-one map) from $V_{k+1}^{1,\dots,n}$ into $V_k^{1,..n}$.*

Proof. The proof follows by induction on n, and is very similar to that of Theorem (3.1) in that we have to separate even and odd values of n. When $n=3$, the only possible values for k are $k=1$ or 2. The result is now easily seen for these cases because $\binom{3}{2} = \binom{3}{1}$ and the injection is supplied by taking the "complement" in $\{1,2,3\}$. On the other hand, because $\binom{3}{3} = 1$, we can drop any one of the 3 digits in $V_3^{1,2,3}$=$\{(1,2,3)\}$, to obtain a unique image in $V_2^{1,2,3}$. Let us now assume that $V_{k+1}^{1,2,\dots,2m} \hookrightarrow V_k^{1,2,\dots,2m}$ for all k such that $k+1 \le 2m \le 2k+1$, in other words, $k = m, m+1, \dots, 2m-1$. Now we wish to show that $V_{k+1}^{1,2,\dots,2m+1} \hookrightarrow V_k^{1,2,\dots,2m+1}$ again for all k such that $k+1 \le 2m+1 \le 2k+1$, in other words, $k = m, \dots, 2m$. We observe that $V_k^{1,2,\dots,2m+1}$ can be written as a disjoint union, by separating the terms that start with a digit 1 from those that start with a digit 2 etc. Indeed, through this natural decomposition, we have a bijection between $V_{m+1}^{1,\dots,2m+1}$ and $\hat{V}_m^{2,\dots,2m+1} \cup \hat{V}_m^{3,\dots,2m+1} \cup \cdots \cup \hat{V}_m^{m+2,\dots,2m+1}$ where $\hat{V}_m^{a+1,\dots,b} = \{(a, x_1, \dots, x_m) \,|\, a+1 \le x_1 < \cdots < x_m \le b\} \subset V_{m+1}^{1,\dots,b}$. Obviously there is a natural bijection between $\hat{V}_m^{a+1,\dots,b}$ and $V_m^{a+1,\dots,b}$ by dropping the first coordinate of the vectors. Hence we note that the number of list in $\hat{V}_m^{2,\dots,2m+1}$ equals $\binom{2m}{m}$, while the number in $\hat{V}_m^{3,\dots,2m+1}$ is $\binom{2m-1}{m}$ etc. By the hypothesis we can find injections from each of the $V_m^{r,r+1,\dots,2m+1}$ into $V_{m-1}^{r,r+1,\dots,2m+1}$ for all $r = 2, \dots, m+2$. That is, we have, by the induction hypothesis,

$$V_m^{2,3,\dots,2m+1} \hookrightarrow V_{m-1}^{2,3,\dots,2m+1},$$

$$V_m^{3,4,\dots,2m+1} \hookrightarrow V_{m-1}^{3,4,\dots,2m+1},$$

$$\vdots$$

$$V_m^{m+2,\ldots,2m+1} \hookrightarrow V_{m-1}^{m+2,\ldots,2m+1}.$$

Combining these and the natural bijections between $\hat{V}_n^{a+1,\ldots,b}$ and $V_n^{a+1,\ldots,b}$, we have an injection from $V_{m+1}^{1,2,\ldots,2m+1}$ into $\hat{V}_{m-1}^{2,3,\ldots,2m+1} \cup \hat{V}_{m-1}^{3,4,\ldots,2m+1} \cup \cdots \cup \hat{V}_{m-1}^{m+2,\ldots,2m+1}$ which is a subset of $V_m^{1,2,\ldots,2m+1}$.

It similarly follows that $V_{m+2}^{1,2,\ldots,2m+1} \hookrightarrow V_{m+1}^{1,2,\ldots,2m+1}$ etc. Combining these and the natural bijections between $\hat{V}_n^{a+1,\ldots,b}$ and $V_n^{a+1,\ldots,b}$, the end result is that $V_{k+1}^{1,2,\ldots,2m+1} \hookrightarrow V_k^{1,2,\ldots,2m+1}$ for $k = m, \ldots, 2m$. Hence this shows that if the result holds for even n then it also holds for the next odd n.

Let us now turn to the converse and assume that

$$V_{k+1}^{1,2,\ldots,2m-1} \hookrightarrow V_k^{1,2,\ldots,2m-1} \text{ for } k \geq m-1.$$

We wish to show that

$$V_{m+1}^{1,2,\ldots,2m} \hookrightarrow V_m^{1,2,\ldots,2m}, \quad V_{m+2}^{1,2,\ldots,2m} \hookrightarrow V_{m+1}^{1,2,\ldots,2m} \quad , \ldots, V_{2m}^{1,2,\ldots,2m} \hookrightarrow V_{2m-1}^{1,2,\ldots,2m}.$$

We focus on the first term with $k = m$, and again write it as a disjoint union of subsets, as

$$V_{m+1}^{1,2,\ldots,2m} = \hat{V}_m^{2,\ldots,2m} \cup \hat{V}_m^{3,\ldots,2m} \cup \cdots \cup \hat{V}_m^{m+1,\ldots,2m}.$$

It again follows by the induction hypothesis that

$V_m^{2,\ldots,2m} \hookrightarrow V_{m-1}^{2,\ldots,2m}$, $V_m^{3,\ldots,2m} \hookrightarrow V_{m-1}^{3,\ldots,2m}$,..., and $V_m^{m+1,\ldots,2m} \hookrightarrow V_{m-1}^{m+1,\ldots,2m}$.

The end result is that $V_{m+1}^{1,2,\ldots,2m} \hookrightarrow V_m^{1,2,\ldots,2m}$. It similarly follows for the other pieces, giving the desired injection

$$V_{k+1}^{1,2,\ldots,2m} \hookrightarrow V_k^{1,2,\ldots,2m}$$

for all $k = m, \ldots, 2m-1$, i.e. for all k such that $k+1 \leq 2m \leq 2k+1$. □

We can now capitalize on the existence of the injection to derive the following inequalities for the "symmetric" functions. In the following Corollary, we do not need to assume the existence of the additive inverse in T.

Corollary 3.1. *Suppose that $\alpha f(ab) \leq f(a)$. Then for $k+1 \leq n \leq 2k+1$,*

$$\sigma_k^n(\boldsymbol{a}) \geq \alpha \sigma_{k+1}^n(\boldsymbol{a}) \tag{10}$$

Proof. We first observe that $\#(\sigma_k^n(\mathbf{a})) = \binom{n}{k}$. Now because of the injection, we

can associate to each term $f(a_{i_1}a_{i_2}\dots a_{i_{k+1}})$ in $\sigma^n_{k+1}(\mathbf{a})$ a distinct term $f(a_{j_1}\cdots a_{j_k})$ in $\sigma^n_k(\mathbf{a})$ obtained from the former by deleting **exactly one** of the a_{i_r}. Since $\alpha f(ab)\le f(a)$, we see that $\alpha f(a_{i_1}a_{i_2}\cdots a_{i_{k+1}})\le f(a_{j_1}\cdots a_{j_k})$ and hence that $\alpha\sigma^n_{k+1}(\mathbf{a})\le\sigma^n_k(\mathbf{a})$. □

Remark For the case where $\alpha f(ab)\le f(a)$, half of the valuation inequalities follow from the local fact that $\sigma^n_k(\mathbf{a})\ge\alpha\sigma^n_{k+1}(\mathbf{a})$.

4 Additive valuation formula for multiplicative valuations with $f(ab)=f(a)f(b)$.

(a) When f is "multiplicative" i.e. $f(ab)=f(a)f(b)$, then the valuation formula can be simplified. In this case the symmetric functions simplify to

$$\sigma_k=\sigma_k(\mathbf{a},f)=\sum_{1\le i_1<i_2<\cdots<i_k\le n} f(a_{i_1})f(a_{i_2})\cdots f(a_{i_k}).$$

Then

$$\begin{aligned} f(\sum_{i=1}^n a_i) &= \sigma_1-\alpha\sigma_2+\alpha^2\sigma_3-\dots(-1)^{n-1}\alpha^{n-1}\sigma_n \\ &= \tfrac{1}{\alpha}[1-\prod_{i=1}^n(1-\alpha f(a_i))]. \end{aligned} \tag{11}$$

A particularly important example is that of the **indicator function** $\chi_A(s)$ of a set A. We shall examine these functions shortly in detail.

(b) On the other hand, if we have for each a in S a "complement" a' in S such that
(i) $(a+b)'=a'b'$ and (ii) $f(a')=1-\alpha f(a)$, then we can obtain (11) *directly* via

$$\begin{aligned} f(\sum_{i=1}^n a_i) &= \tfrac{1}{\alpha}[1-f[(\sum_{i=1}^n a_i)']]=\tfrac{1}{\alpha}[1-f[\prod_{i=1}^n a_i']] \\ &= \tfrac{1}{\alpha}[1-\prod_{i=1}^n f(a_i')]=\tfrac{1}{\alpha}[1-\prod_{i=1}^n[1-\alpha f(a_i)]] \end{aligned} \tag{12}$$

(c) For the case when $\alpha=1$, it suffices to have a "complement" a' such that

$$f(a)=f(ab')+f(ab)\,.$$

4.1 Examples of Valuation equalities

Let us now turn to the various applications of the valuation formula.

Indeed, most of these deal with a collection of subsets of set X as the poset $(\mathcal{P}(X),\subseteq)$ and the lattice $(S,+,\cdot)=(\mathcal{P}(X),\cup,\cap)$, where in addition, f is some

type of "content/size" such as volume, area, length or cardinality. We shall denote the collection of all **finite** subsets of set X, by $F(X)$.

Example 4.1. *Let $(S,+,\cdot) = (F(X),\cup,\cap)$ with $f(A) = \#(A)$, the cardinality of A. Then*

$$\#(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \#(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k})\,.$$

Example 4.2. *Let $I(\mathbb{R})$ be an algebra or σ-algebra of intervals, with Lebesgue measure as length. Suppose S is a the collection of unions of intervals of $\mathbb{R}$, and let $f(A)$ be the length of $\ell(A)$ of A, so that $(S,+,\cdot) = (I(\mathbb{R}),\cup,\cap)$. Then*

$$\ell(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} \ell(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k})\,.$$

Example 4.3. *$(S,+,\cdot) = (\mathcal{F},\cup,\cap)$, where $\mathcal{F}\subset\mathcal{P}(\Omega)$ is a σ-field on Ω and $(\Omega,\mathcal{F},\mathbb{P})$ is a probability triple. We set $f(A) = P(A)$, the probability of A. Then P is a valuation and hence we have the Poincare formula:*

$$P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1\le i_1<i_2<\cdots<i_k\le n} P(A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k})\,.$$

Needless to say, no indicator functions, expected values or integrals were needed to derive this.

Example 4.4. *Let $(S,+,\cdot) = (\mathbb{R}^+,\max,\min)$ and let f be the identity map. (This is not non-negative, so we have at first to restrict S to $\mathbb{R}^+$.)*

Now for two real numbers a and b, it is easily seen that $\max(a,b) = a + b - \min(a,b)$ and $\min\{a,\max(b,c)\} = \max\{\min(a,b),\min(a,c)\}$, so that the identity map is a valuation. This gives

$$\begin{aligned}\max\{a_1,\ldots,a_n\} &= (a_1+\cdots+a_n) - \sum_{i<j}\min\{a_i,a_j\} + \\ &+ \sum_{i<j<k}\min\{a_i,a_j,a_k\} - \cdots + (-1)^{n-1}\min\{a_1,a_2,\ldots,a_n\}.\end{aligned}$$

Example 4.5. *Let S be a collection of all the subsets of X with $(S,+,\cdot) = (\mathcal{P}(X),\cup,\cap)$. For any prime p, and any valuation f with $\alpha = 1$, we have the multiplicative formula,*

$$p^{f(\cup_{i=1}^n A_i)} = \frac{\prod_{i=1}^n p^{f(A_i)} \cdot \prod_{1\le i<j<k\le n} p^{f(A_i\cap A_j\cap A_k)} \cdots \prod_{1\le i_1<\cdots<i_m\le n} p^{f(A_{i_1}\cap\cdots\cap A_{i_m})} \cdot}{\prod_{i<j} p^{f(A_i\cap A_j)} \cdots \prod_{1\le i_1<\cdots<i_\ell\le n} p^{f(A_{i_1}\cap\cdots\cap A_{i_\ell})}},$$

where $m := \max\{k \le n \,|\, k \text{ is odd}\}$ *and* $\ell := \max\{k \le n \,|\, k \text{ is even}\}$.

Example 4.6. *$S = \mathbb{N}$ and $a \le b$ if and only if $a \mid b$.*
Any positive integer can be expressed as a unique product of prime powers $a = p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}$, with $p_1 < p_2 < \dots$. If b is likewise expanded as $b = p_1^{t_1} p_2^{t_2} \dots p_r^{t_r}$, then clearly $a \mid b$ iff $k_i \le t_i$ for all i. Moreover $(a,b) = gcd(a,b) = \prod p_i^{min\{k_i,t_i\}}$ and $[a,b] = lcm(a,b) = \prod p_i^{max\{k_i,t_i\}}$. Since min/max obey the InEx law on $\mathbb{N}$, we may conclude that the gcd and lcm satisfy the multiplicative version *of the InEx law. In other words,*

$$[a_1,\dots,a_n] = \frac{\prod_{i=1}^n a_i \cdot \prod_{1\le i_1<i_2<i_3\le n} (a_{i_1},a_{i_2},a_{i_3}) \cdots \prod_{1\le i_1<\cdots<i_m\le n} (a_{i_1},\dots,a_{i_m})}{\prod_{1\le i_1<i_2\le n} (a_{i_1},a_{i_2}) \cdots \prod_{1\le i_1<\cdots<i_\ell\le n} (a_{i_1},\dots,a_{i_\ell})}, \tag{13}$$

where $m := \max\{k \le n \,|\, k \text{ is odd}\}$ and $\ell := \max\{k \le n \,|\, k \text{ is even}\}$. For example,

$$\begin{aligned} [a,b,c] &= \tfrac{abc(a,b,c)}{(a,b)(a,c)(b,c)} \quad \text{and} \\ [a,b,c,d] &= \tfrac{abcd(a,b,c)(a,b,d)(a,c,d)(b,c,d)}{(a,b)(a,c)(a,d)(b,c)(b,d)(c,d)(a,b,c,d)} \end{aligned} \tag{14}$$

This is a "multiplicative version" of the Inclusion Exclusion rule for an evaluation map f with $\alpha = 1$ such as

$$f(a+b+c) = f(a) + f(b) + f(c) - [f(ab) + f(ac) + f(bc)] + f(abc).$$

Example 4.7. *Let S be the collection of all subspaces of a vector space U, with $V + W$ being the vector space direct sum and $V \cdot W = V \cap W$. Then $f(V) = \dim(V)$ is a valuation on S with $\alpha = 1$, and*

$$dim(\sum_{k=1}^n V_k) = \sum_{k=1}^n (-1)^{k-1} \cdot \sum_{1\le i_1<i_2<\cdots<i_k\le n} \dim(V_{i_1} \cap V_{i_2} \cap \cdots \cap V_{i_k})$$

Example 4.8. *$S = \mathbb{Z}_2^n = \{\boldsymbol{x} = (x_1,\dots,x_n) \,|\, x_i \in \mathbb{Z}_2, \forall i = 1,\dots,n\}$ with coordinate-wise addition and scalar multiplication. These operations are commutative, multiplication is idempotent and distributes over addition. The Hamming metric $h(\boldsymbol{a})$ counts the number of ones in the vector $\boldsymbol{a}$. It is a valuation that satisfies the "cosine rule"*

$$h(\boldsymbol{a}+\boldsymbol{b}) = h(\boldsymbol{a}) + h(\boldsymbol{b}) - 2h(\boldsymbol{a}.\boldsymbol{b}), \tag{15}$$

which is the InEx equation with $\alpha = 2$.

Example 4.9. *$S = \mathbb{B}^n = \{\boldsymbol{x}; (x_1, \ldots, x_n) \,|\, x_i \in \mathbb{B}\,, \forall i = 1, \ldots, n\}$ with the Boolean scalar operations $x + y = x \vee y$ and $x \cdot y = x \wedge y$. The vector operations are again defined coordinate-wise. The Hamming metric is again a valuation, but this satisfies In-Ex equation with $\alpha = 1$, i.e.*

$$h(\boldsymbol{a} + \boldsymbol{b}) = h(\boldsymbol{a}) + h(\boldsymbol{b}) - h(\boldsymbol{a}.\boldsymbol{b}) \tag{16}$$

5 Examples of Valuation inequalities

Example 5.1. *$(S, +, \cdot) = (\mathcal{F}, \cup, \cap)$, where $\mathcal{F} \subset \mathcal{P}(\Omega)$ is a σ-field on Ω and $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability triple such that $f(A) = P(A)$, the probability of A. We have*

(i) $$P(\bigcup_{k=1}^{n} A_k) \leq \sum_{k=1}^{n} P(A_i)\,, \quad \textit{(Boole's law)}$$

(ii) $$P(\bigcup_{k=1}^{n} A_k) \geq \sum_{k=1}^{n} P(A_i) - \sum_{1 \leq i_1 < i_2 \leq n} P(A_{i_1} \cap A_{i_2})\,.$$

Example 5.2. *Next let $(S, +, \cdot) = (\mathbb{R}, \max, \min)$ and let f be the identity map (extended from $\mathbb{R}^+$). We have*

$$\begin{aligned}
&\max\{a_1, \ldots, a_n\} \leq (a_1 + \cdots + a_n)\,,\\
&\max\{a_1, \ldots, a_n\} \geq (a_1 + \cdots + a_n) - \sum_{i<j} \min\{a_i, a_j\}\,,\\
&\max\{a_1, \ldots, a_n\} \leq (a_1 + \cdots + a_n) - \sum_{i<j} \min\{a_i, a_j\} + \sum_{i<j<k} \min\{a_i, a_j, a_k\}\,.
\end{aligned}$$

Example 5.3. *Applying the above to prime powers we arrive at*

$$\begin{aligned}
&[a, b, c, d] \leq abcd\,,\\
&[a, b, c, d] \geq \tfrac{abcd}{(a,b)(a,c)(a,d)(b,c)(b,d)(c,d)}\,,\\
&[a, b, c, d] \leq \tfrac{abcd}{(a,b)(a,c)(a,d)(b,c)(b,d)(c,d)}(a, b, c)(a, b, d)(a, c, d)(b, c, d),
\end{aligned}$$

where we actually attain equality in the last inequality above.

6 Indicator functions

Many of the applications of the "inclusion-exclusion" formula using sets can be derived using the indicator functions (also called characteristic functions). The main purpose of introducing these functions is to convert manipulations involving sets into manipulations involving functions. i.e. turn a Boolean algebra into a Boolean ring. Let us now recap some of its properties.

Given a set S and a subset $A \subset S$, the indicator function of A, $\chi_A : S \to \{0,1\}$ is defined by

$$\chi_A(s) = \begin{cases} 1 & if s \in A \\ 0 & if\ s \notin A \end{cases}.$$

Closely related is the function

$$\phi_A(s) = 1 - 2\chi_A(s) = \begin{cases} 1 & if s \in A \\ -1 & if\ s \notin A \end{cases}.$$

Again, for subsets $A, B, C, \ldots$ of S and $\cup, \cap$ and $\oplus$ stand for union, intersection and the operation, "exclusive or (XOR)", i.e. $A \oplus B = (A \cap B^c) \cup (B \cap A^c)$ and $A \oplus A = (A \cap A^c) \cup (A \cap A^c) = \emptyset$.

For real valued functions on S we define $f \leq g$ iff $f(s) \leq g(s)$ for all $s \in S$. We shall also abbreviate χ_{A_i} to χ_i, when necessary. Some of the valuation maps associated with indicator functions are demonstrated below.

1. $\chi_{A\cup B} = \chi_A + \chi_B - \chi_{A\cap B} \quad \phi_{A\cup B} = \phi_A + \phi_B - \phi_{A\cap B}$ (1-valuations)
2. $\chi_{A\oplus B} = \chi_A + \chi_B - 2\chi_A\chi_B$ (2-valuation), $\quad \phi_{(A\oplus B)} = \phi_A.\phi_B$
3. χ provides a commutative valuation on $(\mathcal{P}(X), \cup, \cap)$ with $\alpha = 1$ and as such the derivation of (12) holds. It is traditional to derive this as

$$\chi_{\cup A_i} = 1 - \chi_{\cap A_i^c} = 1 - \prod_{i=1}^{n}(1 - \chi_{A_i}) = \sigma_1 - \sigma_2 + \cdots + (-1)^{n-1}\sigma_n,$$

which rewrites as

$$\chi_{\cup A_i} = \sum_1^n \chi_i - \sum_{i<j} \chi_i\chi_j + \sum_{i<j<k} \chi_i\chi_j\chi_k + \cdots + (-1)^{n-1}\chi_1\chi_2\cdots\chi_n\,. \tag{17}$$

4. Since χ is a 2-valuation with $\alpha = 2$ on $(\mathcal{P}(X), \oplus, \cap)$ with $\oplus$ = XOR, we have from (11)

$$\chi_{\oplus_{i=1}^n A_i} = \sigma_1 - 2\sigma_2 + 4\sigma_3 + \cdots + (-1)^{n-1}2^{n-1}\sigma_n.$$

5. $\sum\limits_{k=1}^{n} \prod\limits_{i\neq k}(1 - \chi_i) = \sum\limits_{k=1}^{n} \prod\limits_{i\neq k} \chi_{A_i^c} = \sum\limits_{k=1}^{n} \chi_{(\bigcap\limits_{i\neq k} A_i^c)} = \sum\limits_{k=1}^{n} \chi_{(\bigcup\limits_{i\neq k} A_i)^c} = n - \sum\limits_{k=1}^{n} \chi_{\bigcup\limits_{i\neq k} A_i}$,
which parallels Lagrange interpolation.

6. Since indicator functions are multiplicative and act on sets, the valuation inequalities can also be derived using combinatorics. This is instructive in its own right. The valuation inequalities for indicator functions take the following forms,

   (i) $\chi_{\cup A_i} \leq \sum_1^n \chi_i$

   (ii) $\chi_{\cup A_i} \geq \sum_1^n \chi_i - \sum_{i<j} \chi_i\chi_j$

   (iii) $\chi_{\cup A_i} \leq \sum_1^n \chi_i - \sum_{i<j} \chi_i\chi_j + \sum_{i<j<k} \chi_i\chi_j\chi_k$

   (iv) $\chi_{\cup A_i} \geq \sum_{k=1}^r (-1)^{k-1} \cdot \sum_{1\leq i_1<\cdots<i_k\leq n} \chi_{i_1}\chi_{i_2}\cdots\chi_{i_k}$ when r is even

   (v) $\chi_{\cup A_i} \leq \sum_{k=1}^r (-1)^{k-1} \cdot \sum_{1\leq i_1<\cdots<i_k\leq n} \chi_{i_1}\chi_{i_2}\cdots\chi_{i_k}$ when r is odd.

   To prove these using sets, it suffices to show the inequality at any point s in the union. Now the union $\cup A_i$ is made up of 2^n disjoint subsets (called atoms or petals) of the form $B_i = A_1 \cap A_2 \cdots \cap A_i \cap A_{i+1}^c \cdots \cap A_n^c$ etc. and it suffices to show that the inequalities hold for for some fixed s $\in B_i$. For the remaining subsets the result follows by symmetry. As an example we consider the case where $r = 3$.

   If $s \in B_1 = A_1 \cap A_2 \cdots \cap A_i \cap A_{i+1}^c \cdots \cap A_n^c$ then $s \in A_1 \cap A_2 \cdots \cap A_i$ and the left hand side which equals $\chi_{\cup A_i}(s)$ yields the value 1. On the other hand the right hand side gives the value: $\binom{i}{1} - \binom{i}{2} + \binom{i}{3}$. The result now follows from the binomial identity

$$\sum_{k=0}^{r} (-1)^k \binom{i}{k} = \begin{cases} 0 & if\ r = i \\ \binom{i-1}{r}(-1)^r & if\ r < i. \end{cases} \tag{18}$$

   This yields the desired inequalities:
   (i) $\binom{n}{0} \geq \binom{n}{1} - \binom{n}{2} + \cdots + (-1)^{r-1}\binom{n}{r}$ when r is even and
   (ii) $\binom{n}{0} \leq \binom{n}{1} - \binom{n}{2} + \ldots + (-1)^{r-1}\binom{n}{r}$ when r is odd.

7 Multiplicative Formulae for α-valuations

Let f be an α-valuation on $(S, +, \cdot)$. We begin by noting that if $\alpha = 1 + \varepsilon$ where $\varepsilon \geq 0$, then

$$f(a+b)f(a\cdot b) = f(a)f(b) - [f(a) - f(a\cdot b)][f(b) - f(a\cdot b)] - \varepsilon[f(a\cdot b)]^2 \tag{19}$$

and also

$$\alpha.f(a+b)f(a\cdot b) = f(a)f(b) - [f(a) - \alpha f(a\cdot b)][f(b) - \alpha f(a\cdot b)]. \quad (20)$$

These immediately imply the following In-Ex inequalities.

Corollary 7.1. *(I) If* $f(a\cdot b) \le f(a)$ *then*

$$f(a+b)f(a\cdot b) \le f(a)f(b) - \varepsilon[f(a\cdot b)]^2 \le f(a)f(b)\,. \quad (21)$$

If $\alpha = 1$ *then*

$$f(a+b)f(a\cdot b) = f(a)f(b) \text{ iff } f(a) = f(a\cdot b) = f(b).$$

If $\alpha > 1$ *then*

$$f(a+b)f(a\cdot b) = f(a)f(b) \text{ iff } f(a) = f(a\cdot b) = f(b) = 0.$$

(II) If $\alpha f(a\cdot b) \le f(a)$ *with* $\alpha > 1$*, then*

$$\alpha f(a+b)f(a\cdot b) = f(a)f(b) \text{ iff } f(a) = \alpha f(a\cdot b) = f(b)\,.$$

Corollary 7.2. *(i) If* $f(a\cdot b) \le f(a)$*, then*

$$f(a+b)f(a\cdot b) \le f(a)f(b) - \varepsilon[f(a\cdot b)]^2 \le f(a)f(b)\,. \quad (22)$$

(ii) If $\alpha f(a\cdot b) \le f(a)$ *then*

$$f(a+b)f(a\cdot b) \le \frac{1}{\alpha} f(a)f(b)\,. \quad (23)$$

Needless to say, when $\alpha \ge 1$, the assumption that $\alpha f(a\cdot b) \le f(a)$ is much stronger than that of $f(a\cdot b) \le f(a)$. In particular, for XOR operation, this will generally NOT be true.

Examples with $\alpha = 1$ **and** $\varepsilon = 0$

(i) If $(S, +, \cdot) = (F(X), \cup, \cap)$ and $f(\cdot) = \#(\cdot)$, then we have

$$\#(A\cup B).\#(A\cap B) = \#(A).\#(B) - \#(A^c\cap B)\#(A\cap B^c) \le \#(A).\#(B) \quad (24)$$

(ii) If $S = \mathcal{F}$, a σ-field in $\mathcal{P}(X)$, $(+, \cdot) = (\cup, \cap)$ and $f(\cdot) = P(\cdot)$ a probability measure, then

$$P(A\cup B)P(A\cap B) = P(A)P(B) - P(A^c\cap B)P(A\cap B^c) \le P(A)P(B) \quad (25)$$

(iii) If $S = \mathbb{R}, (+, \cdot) = (\max, \min)$ and $f =$ identity, then

$$\max(a, b) \cdot \min(a, b)) = a \cdot b - (a - \min\{a, b\})(b - \min\{a, b\}) = ab \quad (26)$$

The latter follows from the fact that $a - \min\{a, b\} = \max\{a - b, 0\}$, therefore we get to $[a - min\{a, b\}][b - min\{a, b\}] = 0$.

(iv) Applying (iii) to integer prime powers we arrive at

$$[a, b](a, b) = a \cdot b.$$

As such we actually have equality in the latter two "inequalities".

(v) Let S be the collection of all subspaces of a vector space U, with $V + W$ being the vector space direct sum and $V \cdot W = V \cap W$. Then $f(V) = dim(V)$ is a valuation on S, and

$$\begin{aligned} dim(V_1+V_2)dim(V_1\cap V_2) &= dim(V_1)dim(V_2) - [dim(V_1) - dim(V_1 \cap V_2)] \\ &\quad \times [dim(V_1) - dim(V_1 \cap V_2)] \\ &\leq dim(V_1)dim(V_2). \end{aligned}$$

8 InEx equalities and inequalities for $n \leq 3$.

Consider InEx equation for three elements

$$f(a + b + c) = f(a) + f(b) + f(c) - \alpha[f(ab) + f(ac) + f(bc)] + \alpha^2 f(abc).$$

There are numerous equalities that follow from it, and some of these are non-trivial. Also the difference between the cases where $\alpha = 2$ (XOR operation) and $\alpha \neq 2$ becomes striking.

8.1 InEx equalities

We first present some of necessary equalities that must hold as a consequence of InEx for an α-valuation with $\alpha \geq 1$.

1. $f(ab + a) = f(a) - (\alpha - 1)f(ab) \geq f(a)$,
   which we refer to as a "weak version" of InEx. This in turn implies
2. $f(a + a) = (2 - \alpha)f(a)$ and thus $f(a + a) = f(a)$ iff $\alpha = 1$.
   For the XOR operation, this means that $f(a + a) = 0$.
3. $f(a + b + b) = f(a) + (2 - \alpha)f(b) - \alpha(2 - \alpha)f(ab)$.

4. We also have

$$\begin{aligned} f(ab+ac) &= f(ab+ac+bc)+\alpha(2-\alpha)t-f(bc) \\ &= f(ab+c)+f(ac)-f(c)\,. \end{aligned} \tag{27}$$

where $t=f(abc)$. Now if $\alpha f(ab)\le f(b)$ for all a and b then $\alpha t\le f(ab)$ as well as $\alpha t\le f(ac)$. We then have $f(ab+ac)=f(ab)+f(ac)-\alpha t\ge \alpha t+\alpha t-\alpha t=\alpha t$. In particular, for $\alpha\ge 1$,

$$f(abc)\le \alpha f(abc)\le f(ab+ac).$$

5. The following identities follow from InEx and can actually be used to characterize it for an α-valuation.

$$\begin{aligned} f(a)-f(ab+ac) &= f(a+c)-f(ab+c)+(\alpha-1)f(ac) \\ &= f(a)+f(c)-f(ac)-f(ab+c). \end{aligned} \tag{28}$$

Lemma 8.1. *The following are equivalent for $\alpha\ge 1$ and $n=3$:*

(i) *InEx identity*
(ii) $f(a)-f(ab+ac)=f(a+c)-f(ab+c)+(\alpha-1)f(ac)$ *and* $f(ab+a)=(1-\alpha)f(ab)+f(a)$

(29)

Proof. The necessity follows from the above equalities (28). For sufficiency, suppose (29) holds, and set $c=b$. Then we get $f(a)-f(ab+ab)=f(a+b)-f(ab+b)+(\alpha-1)f(ab)$ in which we substitute the weak InEx identity $f(ab+b)=(1-\alpha)f(ab)+f(b)$ to give the desired result. □

8.2 InEx Inequalities

We next come to some of the inequalities spawned by InEx. First, we note that $f(a+b)-f(a)=f(b)-\alpha f(ab)$, with $\alpha\ge 1$, ensures that

$$\alpha f(ab)\le f(b)\Leftrightarrow f(a)\le f(a+b). \tag{30}$$

A useful consequence is

Corollary 8.1.

$$t=f(abc)\le \alpha f(abc)\le f(ab)\le f(ab+ac). \tag{31}$$

We also have

Lemma 8.2. *Let $\alpha f(ab) \leq f(a)$ with $\alpha \geq 1$. Then all the following statements hold.*

(i) $f(ab) \leq f(a)$ for all a and b,

(ii) $f(a(b+c)) \leq f(a)$ for all a,b and c,

(iii) $f(ab+c) \leq f(b) + f(c) - f(bc)$ for all a,b and c,

(iv) $f(ab+c) \leq f(b+c) + \varepsilon f(bc)$ for all a,b and c,

(v) $f(ab) - \alpha f(abc) \leq f(b) - f(bc)$ for all a,b and c (monotonicity inequality)

Proof. (i) holds and also clearly (i) implies (ii). The equivalence of (ii), (iii) and (iv) follows at once from the identity:

$$0 \leq f(b) - f(ab+bc) = f(b+c) - f(ab+c) + \varepsilon f(bc) = f(b) + f(c) - f(bc) - f(ab+c).$$

The equivalence of (ii) and (v) follows from the identity

$$0 \leq f(b) - f(ab+bc) = f(b) - f(ab) - f(bc) + \alpha f(abc).$$

□

For $\alpha = 1$, we may combine Lemma 8.2 (v) with the inequality (30). In addition, the monotone inequality corresponds to $q \leq q + r$ in the following Venn diagram.

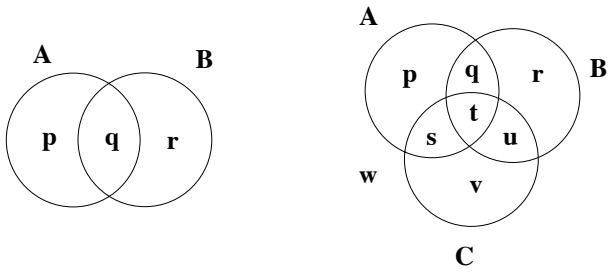


Figure 1: Three sets

Furthermore, it is clear from (30) that

$$0 \leq \alpha^2 f(abc) \leq \alpha f(ab) \leq f(a)$$

and

$$3\alpha^2 f(abc) \leq \alpha[f(ab) + f(ac) + f(bc)] \leq [f(a) + f(b) + f(c)]\,.$$

8.3 Generalizations of the InEx Inequality to 3-d.

Let us now examine possible generalization of the 2-d InEx inequality

$$f(a+b)f(ab) \leq f(a)f(b)\,, \tag{32}$$

to the sum $f(a+b+c)$, assuming that $\alpha = 1$ and $f(ab) \leq f(a)$.

There are numerous "sharp" generalizations, that all reduce to (32), when two of the variables are set equal. We shall present three of these under the assumption that the addition is also idempotent.

Our first method is to replace b by $b+c$ in the inequality (32). This gives

$$f(a+b+c)f[a(b+c)] \leq f(a)f(b+c), \tag{33}$$

which we combine with

$$f(abc) \leq f(ab) \leq f(ab+ac) \tag{34}$$

to arrive at

$$f(a+b+c)f(abc) \leq f(a+b+c)f[a(b+c)] \leq f(a)f(b+c)\,. \tag{35}$$

Rotating the variables we may conclude that

$$f(a+b+c)f(abc) \leq \min\{f(a)f(b+c), f(b)f(a+c), f(c)f(a+b)\}. \tag{36}$$

This result is sharp, in the sense that if we set $b=c$, this reduces to

$$f(a+b)f(ab) \leq \min\{f(a)f(b), f(b)f(a+b), f(b)f(a+b)\} = f(a)f(b)\,,$$

as $f(a) \leq f(a+b)$.

Alternatively, we may multiply (35) by f(bc) and use InEx identity for $n=2$, to arrive at

$$f(a+b+c)f(abc)f(bc) \leq f(a)f(b+c)f(bc) \leq f(a)f(b)f(c)\,. \tag{37}$$

Rotating the variables then yields

$$f(a+b+c)f(abc)\max\{f(ab), f(ac), f(bc)\} \leq f(a)f(b)f(c)\,. \tag{38}$$

Setting b = c, shows that $f(a+b)f(ab)\max\{f(ab), f(ab), f(b)\} \leq f(a)f(b)f(b)$, which reduces to (32). Based on the inequalities that we have seen for the additive case, it would be natural to expect that for a

sub-multiplicative valuation, with $f(ab) \leq f(a)$ and one may expect the following to hold

$$f(a+b+c)f(ab)f(ac)f(bc) \leq f(a)f(b)f(c)f(abc). \tag{39}$$

However, numerical tests prove that this is not true, even for probability measures. In fact, the inequality (39)

$$\begin{aligned}&(t+p+q+r+s+u+v)(t+q)(t+s)(t+u) \leq \\ &\leq t(t+p+q+s)(t+q+r+u)(t+s+u+v)\end{aligned} \tag{40}$$

is violated when we set $p=q=r=s=u=v=\frac{1}{10}$ and $t=\frac{1}{100}$ for instance.

However, a slight perturbation from this does hold as below.

$$\begin{aligned}&(t+p+q+r+s+u+v)[(t+q)(t+s)(t+u)-qsu] \leq \\ &\leq t[(t+p+q+s)(t+q+r+u)(t+s+u+v)+qsu].\end{aligned} \tag{41}$$

We may re-express this as

$$f(a+b+c)[f(ab)f(ac)f(bc)-qsu] \leq f(abc)[f(a)f(b)f(c)+qsu], \tag{42}$$

where $q=f(ab)-f(abc), s=f(ac)-f(abc)$ and $u=f(bc)-f(abc)$. If we set $b=c$, then $q=0$, and we are back to $n=2$ case.

Note that the outer layer of "petals" are p,r and v while the inner layer is made up of q,s and u. The sum $ab+ac+bc$ corresponds to the inner "flower" of the Venn diagram, made up of the petals q,s,u and t. The proof of the inequality (42) will be given later when we actually clarify when the equality holds.

Examples

(i) For a probability measure, we may state:

$$P(A \cup B)P(A \cap B) \leq P(A)P(B) \tag{43}$$

with the equality holding if and only if $P(A \setminus B)=0=P(B \setminus A)$. Also we have

$$P(A\cup B\cup C)P(A\cap B\cap C) \leq \min\{P(A)P(B\cup C), P(B)P(A\cup B), P(C)P(A\cup B)\} \tag{44}$$

as well as

$$P(A\cup B\cup C)P(A\cap B\cap C)\max\{P(A\cap B), P(A\cap C), P(B\cap C)\} \leq P(A)P(B)P(C). \tag{45}$$

(ii) For a probability measure, the inequality (42) becomes

$$\begin{aligned} P(A\cup B\cup C)\Big[&P(A\cap B)P(A\cap C)P(B\cap C)\\ &-\Big[[P(A\cap B)-P(A\cap B\cap C)][[P(A\cap C)-P(A\cap B\cap C)]\\ &\quad\times[P(B\cap C)-P(A\cap B\cap C)]\Big]\Big] \le P(A\cap B\cap C)\Big[P(A)P(B)P(C)\\ &+[P(A\cap B)-P(A\cup B\cup C)][P(A\cap C)-P(A\cap B\cap C)]\\ &\quad\times[P(B\cap C)-P(A\cap B\cap C)]\Big], \end{aligned} \tag{46}$$

which reduces to (43) when $B=C$.

To see when we actually achieve equality in (42), we start by expressing both sides as polynomials in t. We let

$$\begin{aligned} (t+p+q+r+s+u+v)&[(t+q)(t+s)(t+u)-qsu]\\ &= t^4+e_3t^3+e_2t^2+e_1t+e_0-f(a+b+c)\cdot qsu \end{aligned}$$

and set

$$t[(t+p+q+s)(t+q+r+u)(t+s+u+v)+qsu] = t^4+f_3t^3+f_2t^2+f_1t+t.qsu\,.$$

For the two sides to be equal we must have

$$(f_3-e_3)t^3+(f_2-e_2)t^2+(f_1-e_1)t+qsu.t+f(a+b+c).qsu-e_0=0. \tag{47}$$

Now note that $f_3=e_3$ while $f_2-e_2=(pu+qv+rs)+(pr+pv+rv)$, $e_0=[f(a+b+c)-t](qsu)$ and $f_1-e_1=\lambda-2qsu$, where

$$\lambda=(pqv+pvu)+(prs+pru)+(rsv+rvq)+(q^2v+u^2p+s^2r)+prv.$$

Substituting these gives

$$\begin{aligned} &[(pu+qv+rs)+(pr+pv+rv)]t^2+(\lambda-2qsu)t+(qsu)t+\\ &+f(a+b+c)qsu-[f(a+b+c)-t]qsu=0 \end{aligned}$$

and consequently

$$t[t(f_2-e_2)+\lambda]=0.$$

We can now conclude that the inequality (42) must hold since the difference between the two sides in (41) is given by $t[t(f_2-e_2)+\lambda]$, in which each term is non-negative.

We close by examining the equality case. Indeed, since all terms are non-negative, they must vanish and we have the following two cases.

Case (i) $t = 0$ or Case (ii) $t \neq 0$. In the latter case we must have $f_2 = e_2$ and $\lambda = 0$.

The equality $f_2 = e_2$ ensures that

$$pu = 0, qv = 0, rs = 0, pr = 0, pv = 0, rv = 0. \tag{48}$$

The first three contain "cross products" between inner and outer petals while the latter three involve only the three outer petals. These conditions in turn imply that $\lambda = 0$.

Let us close with some relevant comments and open questions.

9 Remarks and open Questions

1. We have seen a partial parallel between the additive and multiplicative inequalities for valuations. It would be of interest to find more general multiplicative inequalitiesfor valuations.
2. In how far does convexity play a role? We have met the consequences of composing an evaluation map with the exponential functions. If would be interesting to explore the composition of an evaluation map with other convex or concave functions.
3. Investigations into **sub**-valuation for which $f(a+b) \leq f(a) + f(b) - \alpha f(a \cdot b)$ would be of interest. The catch, however, is that inequalities cannot be lined up in this case.
4. Do analogous inequalities exist for multinomial coefficients?
5. The derivation of the valuation inequalities may be done by replacing the assumption that an additive inverse exists, by the assumption that a partial order exists.

References

[1] G.Pólya and G. Szegö, *Aufgabe und Lehrsätze aus der Analysis*, Springer Berlin, 1964, 3rd ed, pp. 119-121.

[2] D.E. Rutherford, *Introduction to Lattice Theory*, Hafner, New York, 1965, p. 20.

[3] G. Birkhoff, *Lattice theory*, vol 25, 3rd ed, Providence RI, AMS series.

Construções impossíveis com régua e compasso

José Carlos Santos
Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
e-mail: jcsantos@fc.up.pt

Resumo: Será visto como lidar com problemas de construções com régua e compasso que levem a equações algébricas de grau 4. Mais precisamente, é dado um critério que permite decidir se um tal problema tem ou não solução.

Abstract: It will be seen how to deal with constructions that can be carried out with compass and straightedge when they lead to quartic equations. More precisely, it will be given a criterion that allows to determine whether or not such a problem is solvable.

palavras-chave: Régua e compasso. Geometria euclidiana

keywords: Compass and straightedge. Euclidian Geometry.

1 Introdução

A maneira usual de demonstrar que certos problemas geométricos não têm solução usando somente régua e compasso baseia-se no facto, demonstrado por Pierre-Laurent Wantzel (veja-se [13]), de que o uso daqueles instrumentos leva sempre a números algébricos cujo grau é necessariamente uma potência de 2. Este facto é empregue em [5, § 4.2] e em [12, ch. 7] a fim de demonstrar que há problemas geométricos que não podem ser resolvidos usando somente a régua e compasso, pois dão origem ou a números algébricos cujo grau *não é* uma potência de 2 (duplicação do cubo e trissecção do ângulo; veja-se também [9] para este último problema) ou a um número não algébrico (quadratura do círculo).

Esta abordagem às soluções de problemas geométricos recorrendo somente a régua e compasso leva naturalmente à seguinte questão: há problemas geométricos que *não* possam ser resolvidos usando somente aqueles instrumentos mas que, no entanto, dêem origem a um número algébrico cujo grau *é* uma potência de 2? Posto de outro modo: haverá números algébricos não construtíveis cujo grau seja uma potência de 2? A resposta é afirmativa. Neste artigo será visto um critério que permite, dado um número algébrico de grau 4, determinar se é ou não construtível. Como ponto de partida,

será visto um problema que não pode ser resolvido usando somente régua e compasso, embora leve a um número algébrico de grau 4.

Em [8] podem ser vistos resultados semelhantes àqueles que são expostos aqui, bem como alguns problemas adicionais.

2 O problema

Considere-se o seguinte problema: dados um rectângulo r e um comprimento c, é possível construir um rectângulo r' inscrito no rectângulo r tal que um dos seus lados tenha comprimento c (veja-se a figura 1)? Como será visto, este problema não pode, em geral, ser resolvido usando somente régua e compasso, embora leve a um número algébrico de grau 4. Será também visto como resolver este problema por meio de intersecção de cónicas. Naturalmente, qualquer número algébrico de grau menor ou igual a 4 pode ser obtido como a abcissa do ponto de instersecção de duas cónicas mas, como será visto, a solução do problema em questão via cónicas é particularmente simples.

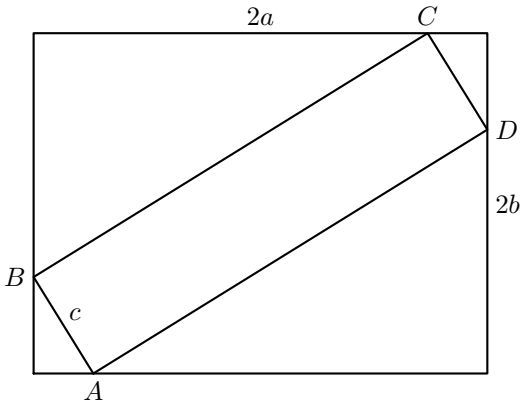


Figura 1: O problema

3 Solução via cónicas

Podemos, sem perda de generalidade, supor que as coordenadas dos vértices de r, relativamente a algum referencial ortonormado, são $(\pm a, \pm b)$, para dois números a e b tais que $a \geq b > 0$. Uma vez que c é um comprimento, $c > 0$ e, naturalmente, para que o problema tenha solução, c terá que ser

menor do que a diagonal do rectângulo r, ou seja, $c < 2\sqrt{a^2+b^2}$. Vamos procurar uma solução com um vértice A no lado de baixo de r, um vértice B no lado esquerdo de r e tal que $\overline{AB} = c$. Então, para algum $x \in [-a, a]$ e para algum $y \in [-b, b]$, $A = (x, -b)$ e $B = (-a, y)$. Afirmar que $\overline{AB} = c$ equivale a afirmar que $(x+a)^2 + (y+b)^2 = c^2$. Seja C (respectivamente D) a reflexão de A (respectivamente B) no centro de r. Então $[ABCD]$ é um paralelogramo e é então necessário determinar quando é que é um rectângulo. Não é difícil ver que isto ocorre quando e só quando A e B são equidistantes do centro de r. Por outras palavras, $[ABCD]$ é um rectângulo se e só se $x^2 + b^2 = y^2 + a^2$, o que é o mesmo que afirmar que $x^2 - y^2 = a^2 - b^2$. Assim sendo, afirmar que $[ABCD]$ é um rectângulo tal que $\overline{AB} = c$ equivale a afirmar que

$$\begin{cases} (x+a)^2 + (y+b)^2 = c^2 \\ x^2 - y^2 = a^2 - b^2 \\ -a \leq x \leq a \\ -b \leq y \leq b. \end{cases} \tag{1}$$

Geometricamente, afirmar que (x_0, y_0) é solução de ambas as equações do sistema (1) significa que o ponto (x_0, y_0) está na intersecção da hipérbole $x^2 - y^2 = a^2 - b^2$ (que, de facto, só é uma hipérbole quando $a > b$; quando $a = b$ é a reunião de duas rectas concorrentes) com a circunferência $(x + a)^2 + (y + b)^2 = c^2$. Trata-se da circunferência de raio c centrada no vértice do canto inferior esquerdo de r, enquanto que a hipérbole em questão é a única hipérbole que passa pelos vértices de r e cujas assímptotas são as rectas de declives ± 1 que passam pelo centro de r. Logo, se a hipérbole e a circunferência se intersectam num ponto P e se

- A é o ponto do lado de baixo de r mais próximo de P;
- B é o ponto do lado esquerdo de r mais próximo de P;
- C é a reflexão de A no centro de r;
- D é a reflexão de B no centro de r,

então $[ABCD]$ é uma solução do problema (veja-se a figura 2) e qualquer solução do problema pode ser obtida por este processo.

A segunda equação do sistema (1) também pode ser obtida por outro processo. O paralelogram $[ABCD]$ é um rectângulo se e só se os segmentos de recta $[AB]$ and $[AD]$ forem perpendiculares e, uma vez que $\overrightarrow{AB} = (-x -$

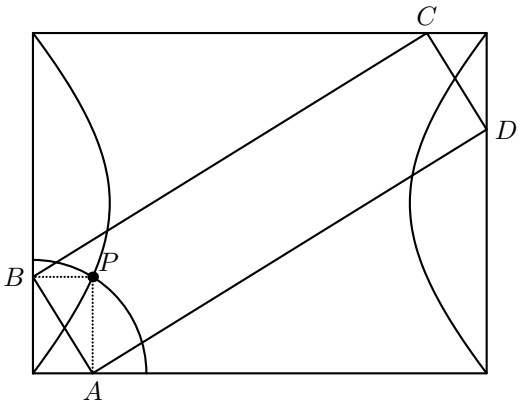


Figura 2: Solução por meio de intersecção de cónicas

$a, y+b)$ e que $\overrightarrow{AD} = (-x+a, -y+b)$,

$$\begin{aligned}\overrightarrow{AB} \perp \overrightarrow{AD} &\iff (-x-a, y+b).(-x+a, -y+b) = 0\\ &\iff x^2 - y^2 = a^2 - b^2.\end{aligned}$$

A demonstração anterior emprega coordenadas, mas a descrição do método em si envolve somente Geometria sintética. Seria interessante encontrar uma demonstração que também recorresse somente a Geometria sintética.

Observe-se que tanto as equações como as desigualdades do sistema (1) são homogéneas (tanto relativamente às variáveis quanto aos parâmetros) e que, portanto, se (x_0, y_0) for uma solução do sistema e se $\lambda > 0$, então $(\lambda x_0, \lambda y_0)$ é uma solução do sistema obtido de (1) ao substituir-se a, b e c por λa, λb e λc respectivamente.

4 Régua e compasso

Vamos agora provar que o problema anterior não pode ser resolvido usando somente régua e compasso. De facto, vai ser demonstrado um critério para determinar quando é que um número algébrico de grau 4 é construtível e esse critério será então aplicado a este problema, bem como a outros.

Preliminares

Seja $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in \mathbb{Q}[x]$ um polinómio mónico de grau n. Se $a_{n-1} = 0$, diremos que $p(x)$ é um *polinómio reduzido*. Há

somente um polinómio reduzido da forma $p(x+\lambda)$, que é aquele para o qual $\lambda = -a_{n-1}/n$; o polinómio $p(x - a_{n-1}/n)$ designa-se por *forma reduzida* de $p(x)$.

Diz-se que um número complexo α é um número algébrico se for raiz de algum polinómio não nulo com coeficientes racionais. O menor grau de um tal polinómio designa-se por *grau* de α e existe um e um só polinómio mónico com coeficientes racionais cujo grau é o grau de α do qual α é raiz, que é o *polinómio minimal* de α, o qual divide qualquer outro polinómio com coeficientes racionais do qual α seja raiz (veja-se [5, § 2.11] ou [10, § 5.6.2]).

Dado um polinómio mónico de grau 4, $P(x) = x^4 + px^3 + qx^2 + rx + s$, com raízes r_1, r_2, r_3 e r_4, não é complicado provar que o polinómio mónico de grau 3 cujas raízes são $(r_1+r_2)(r_3+r_4)$, $(r_1+r_3)(r_2+r_4)$ e $(r_1+r_4)(r_2+r_3)$ é o polinómio $x^3 - 2qx^2 + (q^2 + pr - 4s)x - pqr + r^2 + p^2 s$, o qual se designa por *cúbica resolvente* de $P(x)$. Se $P(x)$ for reduzido (ou seja, se $p = 0$), um cálculo simples (veja-se [12, § 1.4]) mostra que, para qualquer número complexo u, u é raiz da cúbica resolvente de $P(x)$ se e só se

$$P(x) = \left(x^2 + \sqrt{-u}x + \frac{q}{2} - \frac{u}{2} - \sqrt{-u}\right)\left(x^2 - \sqrt{-u}x + \frac{q}{2} - \frac{u}{2} + \sqrt{-u}\right). \tag{2}$$

A fim de provar que certos polinómios $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ com coeficientes inteiros são irredutíveis em $\mathbb{Q}[x]$, iremos recorrer ao *critério de Eisenstein*: se existir um primo p tal que $p \mid a_i$ para cada $i \in \{0, 1, \ldots, n-1\}$ mas tal que $p \nmid a_n$ e que $p^2 \nmid a_0$, então $P(x)$ é irredutível em $\mathbb{Q}[x]$ (veja-se [12, § 3.4]). Este critério será aplicado somente a polinómios mónicos, caso em que a condição $p \nmid a_n$ se verifica automaticamente. Convém observar que o polinómio minimal de um número algébrico é sempre irredutível.

Um número complexo z é *construtível* caso possa ser construído a partir de 0 e de 1 usando somente régua e compasso. Por exemplo, os números $\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$ são construtíveis, pois são os pontos de intersecção das circunferências de raio 1 centradas em 0 e em 1. Naturalmente, todos os números racionais são construtíveis.

Caso $u_1, u_2, \ldots, u_n$ sejam números complexos, seja $\mathbb{Q}(u_1, u_2, \ldots, u_n)$ o mais pequeno subcorpo de $\mathbb{C}$ ao qual pertencem.

Teorema 1 *Um número complexo z é construtível se e só se pertencer a algum subcorpo de $\mathbb{C}$ da forma $\mathbb{Q}(u_1, u_2, \ldots, u_n)$, onde ${u_1}^2 \in \mathbb{Q}$ e, para cada $j \in \{2, 3, \ldots, n\}$, ${u_j}^2 \in \mathbb{Q}(u_1, u_2, \ldots, u_{j-1})$.*

Uma generalização deste teorema pode ser vista em [5, § 4.2].

Os números construtíveis formam um corpo, como pode ser visto em [5, § 4.2] ou em [12, § 7.2]. De facto, só iremos precisar de saber que a soma e o produto de dois números construtíveis também são números construtíveis.

Critério de impossibilidade

Se (x_0, y_0) for solução do sistema (1), resulta da segunda equação que $x_0 = \pm\sqrt{{y_0}^2 + a^2 - b^2}$. Se isto for empregue para eliminar x da primeira equação, obtém-se que y_0 é raiz do polinómio

$$4y^4 + 8by^3 + 4(a^2 + b^2 - c^2)y^2 + 4b(2a^2 - c^2)y + 4a^2(b^2 - c^2) + c^4. \quad (3)$$

Suponha-se que $a = 4$, $b = 3$ e $c = 2$ (que são os valores empregues na criação das figuras 1 e 2). Após divisão por 4, o polinómio (3) fica

$$p(y) = y^4 + 6y^3 + 21y^2 + 84y + 84. \quad (4)$$

Resulta do critério de Eisenstein (com $p = 3$) que $p(y)$ é irredutível em $\mathbb{Q}[y]$ e que, portanto, as suas raízes são números algébricos de grau 4.

O critério atrás mencionado que permite determinar se um número algébrico de grau 4 é ou não construtível é o seguinte:

Teorema 2 *Seja α um número algébrico de grau 4 e seja $p(x) \in \mathbb{Q}[x]$ o polinómio minimal de α. Então α é construtível se e só se a cúbica resolvente de $p(x)$ tiver alguma raiz racional.*

Um cálculo simples revela que a cúbica resolvente do polinómio (4) é $q(x) = x^3 - 42x^2 + 609x - 504$. Resulta do critério de Eisenstein (com $p = 7$) que este polinómio é irredutível em $\mathbb{Q}[x]$, pelo que não tem raízes racionais. Também se poderia chegar à mesma conclusão aplicando o teorema das raízes racionais (veja-se [7, § 4.3] ou [10, § 4.4.1]). Com efeito, resulta deste teorema que as raízes racionais de $q(x)$ são necessariamente números inteiros. Mas, uma vez que $q(0) < 0$, que $q(1) > 0$ e que $(\forall x \in \mathbb{R}) : q'(x) = 3(x - 14)^2 + 21 > 0$, a única raiz de $q(x)$ está em $]0, 1[$ e, portanto, não é inteira.

Vejamos como demonstrar o teorema 2. Seja α um número algébrico de grau 4, seja $p(x)$ o seu polinómio minimal e suponha-se que a cúbica resolvente $q(x)$ de $p(x)$ tem alguma raiz racional r; quer-se provar que α é construtível. Se β, γ e δ forem as restantes raízes de $p(x)$, está-se a supor que um dos números $(\alpha+\beta)(\gamma+\delta)$, $(\alpha+\gamma)(\beta+\delta)$ e $(\alpha+\delta)(\beta+\gamma)$ (ou seja, uma das raízes de $q(x)$) é racional. Pode-se, sem perda de generalidade,

supor que é o primeiro destes. Seja $p^*(x)$ a forma reduzida de $p(x)$. Então $p^*(x)$ é da forma $p(x+\lambda)$, para algum número racional λ, pelo que as raízes de $p^*(x)$ são $\alpha-\lambda$, $\beta-\lambda$, $\gamma-\lambda$ e $\delta-\lambda$. Mas então

$$(\alpha-\lambda+\beta-\lambda)(\gamma-\lambda+\delta-\lambda)$$

é uma raiz da cúbica resolvente $q^*(x)$ de $p^*(x)$. Acontece que esta raiz é igual a

$$(\alpha+\beta)(\gamma+\delta)+4\lambda^2-2\lambda(\alpha+\beta+\gamma+\delta),$$

que é um número racional ($\alpha+\beta+\gamma+\delta$ é racional por ser o simétrico do coeficiente de x^3 em $p(x)$). Sendo assim, resulta da relação (2) que $\alpha-\lambda$ é raiz de um polinómio quadrático em que cada coeficiente é racional ou raiz quadrada de um número racional, pelo que $\alpha-\lambda$ é construtível e, portanto, α é construtível.

Suponha-se agora que α é construtível e suponha-se também que β, γ e δ são construtíveis. Uma vez que a soma e o produto de números construtíveis são novamente números construtíveis, as raízes de $q(x)$ são construtíveis. Mas então $q(x)$ é redutível, pois se assim não fosse as suas raízes seriam números algébricos de grau 3, que não é uma potência de 2. Sendo redutível e de grau 3, $q(x)$ tem necessariamente uma raiz racional.

Assim sendo, a fim de terminar a demonstração do teorema 2, basta provar que se α for construtível, então as restantes raízes de $p(x)$ também o são. Provemos então que β é construtível. Para tal, vamos começar por provar que existe um isomorfismo ι de $\mathbb{Q}(\alpha)$ sobre $\mathbb{Q}(\beta)$ tal que $\iota(\alpha)=\beta$. Cada elemento de $\mathbb{Q}(\alpha)$ é da forma $q(\alpha)$, para algum polinómio $q(x)\in\mathbb{Q}[x]$. Além disso, se $r(x)\in\mathbb{Q}[x]$ for tal que $q(\alpha)=r(\alpha)$, então α é raiz de $q(x)-r(x)$ e, portanto, $p(x)\mid q(x)-r(x)$; em particular, β também é raiz de $q(x)-r(x)$, pelo que $q(\beta)=r(\beta)$. Logo, faz sentido definir

$$\begin{array}{rccc}\iota\colon & \mathbb{Q}(\alpha) & \mapsto & \mathbb{Q}(\beta)\\ & q(\alpha) & \mapsto & q(\beta),\end{array}$$

que é um isomorfismo tal que $\iota(\alpha)=\beta$.

Como α é construtível, sabe-se, pelo teorema 1, que $\alpha\in\mathbb{Q}(u_1,u_2,\ldots,u_n)$, onde ${u_1}^2\in\mathbb{Q}$ e onde, para cada $j\in\{2,3,\ldots,n\}$, ${u_j}^2\in\mathbb{Q}(u_1,u_2,\ldots,u_{j-1})$. Observe-se que $\mathbb{Q}(u_1,u_2,\ldots,u_n)\supset\mathbb{Q}(\alpha)$, uma vez que $\mathbb{Q}(u_1,u_2,\ldots,u_n)\supset\mathbb{Q}$ e que $\alpha\in\mathbb{Q}(u_1,u_2,\ldots,u_n)$. Assim sendo, se se provar que é possível prolongar ι a um homomorfismo de corpos (que também será representado por ι) de domínio $\mathbb{Q}(u_1,u_2,\ldots,u_n)$, poder-se-á deduzir do teorema 1 que β é construtível pois, se se definir $v_k=\iota(u_k)$ para

cada $k \in \{1, 2, \ldots, n\}$, é claro que $\beta \in \mathbb{Q}(v_1, v_2, \ldots, v_n)$, que ${v_1}^2 \in \mathbb{Q}$ e que, para cada $j \in \{2, 3, \ldots, n\}$, ${v_j}^2 \in \mathbb{Q}(v_1, v_2, \ldots, v_{j-1})$.

Comecemos por provar que ι se pode prolongar a $\mathbb{Q}(u_1)$. Caso $u_1 \in \mathbb{Q}(\alpha)$, então $\iota(u_1)$ já se encontra definido e, portanto, o domíno de ι já contém $\mathbb{Q}(u_1)$. Caso contrário, uma vez que ${u_1}^2 \in \mathbb{Q} \subset \mathbb{Q}(\alpha)$, $\iota({u_1}^2)$ já se encontra definido e basta então definir $\iota(u_1)$ como sendo uma das raízes quadradas de $\iota({u_1}^2)$. Visto que ι já está definido em $\mathbb{Q}$, isto permite prolongar ι a $\mathbb{Q}(u_1)$. Pode-se agora prolongar ι a $\mathbb{Q}(u_1, u_2)$ pelo mesmo processo: caso $u_2 \in \mathbb{Q}(\alpha)$, não há nada a fazer. Caso contrário, ${u_2}^2 \in \mathbb{Q}(u_1)$, onde ι já se encontra definido. Então define-se $\iota(u_2)$ como sendo uma das raízes quadradas de $\iota({u_1}^2)$. Prosseguindo deste modo, prolonga-se ι a $\mathbb{Q}(u_1, u_2, \ldots, u_n)$, como se pretendia. Isto conclui a demonstração do teorema 2.

Em particular está agora provado que o problema que consiste em inscrever um rectângulo de comprimento dado num rectângulo dado não tem, em geral, solução com régua e compasso. Naturalmente, para certos valores concretos de a, b e c uma tal solução existe. Por exemplo, quando $a = b$ (ou seja, quando o rectângulo dado é um quadrado), o problema *pode* ser resolvido usando apenas régua e compasso, seja qual for o comprimento c (desde que seja menor do que a diagonal do quadrado, naturalmente), pois neste caso, como já foi referido, não temos uma hipérbole, mas sim uma cónica degenerada, a qual consiste nas duas rectas definidas pelos dois pares de vértices opostos do quadrado, ou seja, são as rectas suporte das diagonais do quadrado. É claro geometricamente que o problema tem uma e uma só solução (com, como atrás, uma extremidade de um dos lados com comprimento c no lado de baixo do quadrado e a outra extremidade do lado esquerdo) quando $0 < c < \sqrt{2}a$ e que há exactamente três soluções (nas mesmas condições) quando $\sqrt{2}a \leq c < 2\sqrt{2}a$. A figura 3 mostra duas destras três soluções quando $c = \frac{5}{3}a$; a terceira pode ser obtida a partir da da direita por reflexão em qualquer das diagonais do quadrado. Qualquer duas duas últimas soluções forma uma imagem que surge numa das mais conhecidas demonstrações do teorema de Pitágoras.

Que o problema tem sempre solução quando $a = b$ também resulta do facto de, neste caso, a expressão polinomial (3) ser

$$4y^4 + 8ay^3 + 4(2a^2 - c^2)y^2 + 4a(2a^2 - c^2)y + 4a^2(a^2 - c^2) + c^4.$$

Acontece que este polinómio é redutível em $\mathbb{Q}[x]$, sejam quais for os valores de a e de c (desde que sejam racionais), pois é igual a

$$\left(2y^2 + 4ay + 2a^2 - c^2\right)\left(2y^2 + 2a^2 - c^2\right).$$

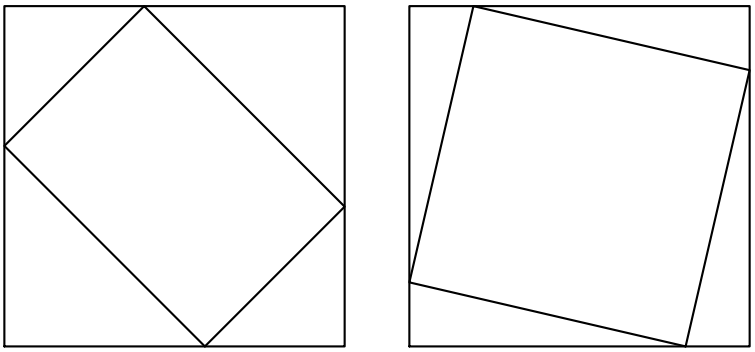


Figura 3: Soluções quando o rectângulo dado é um quadrado

Outra situação na qual o problema pode ser resolvido usando somente régua e compasso é no caso em que $c = 2a$ ou que $c = 2b$ pois, em qualquer dos casos, o rectângulo que se pretende construir coincide com o rectângulo dado.

5 Número de soluções

Resulta imediatamente da figura 2 que o problema tem sempre solução quando $0 < c \leq 2b$, pois nesse caso a circunferência de raio r centrada no canto inferior esquerdo do rectângulo dado intersecta o ramo da esquerda da hipérbole. E também tem alguma solução quando c for maior ou igual à distância $d(a, b)$ do canto inferior esquerdo do rectângulo ao ramo da direita da hipérbole (supondo, é claro que $c < 2\sqrt{a^2 + b^2}$). Mas caso aconteça que

$$2b < d(a, b), \tag{5}$$

então o problema não tem solução quando $2b < c < d(a, b)$. A fim de calcular o valor de $d(a, b)$, o método dos multiplicadores de Lagrange pode ser empregue para determinar o ponto (x, y) do ramo da direita da hipérbole que fica mais próximo do vértice inferior esquerdo de r. Mas é mais simples (embora leve essencialmente aos mesmos cálculos) ver que o vector que vai de $(-a, -b)$ a (x, y) tem necessariamente que ser paralelo ao gradiente da função $h\colon \mathbb{R}^2 \longrightarrow \mathbb{R}$ definida por $h(x, y) = x^2 - y^2$; veja-se a figura 4.

Está-se então interessado em encontrar o ponto $(x, y) \in \mathbb{R}^2$ que pertence ao ramo da direita da hipérbole $x^2 - y^2 = a^2 - b^2$ tal que $(x + a, y + b)$ seja múltiplo de $(2x, -2y)(= \nabla h(x, y))$. Por outras palavras, quer-se resolver o

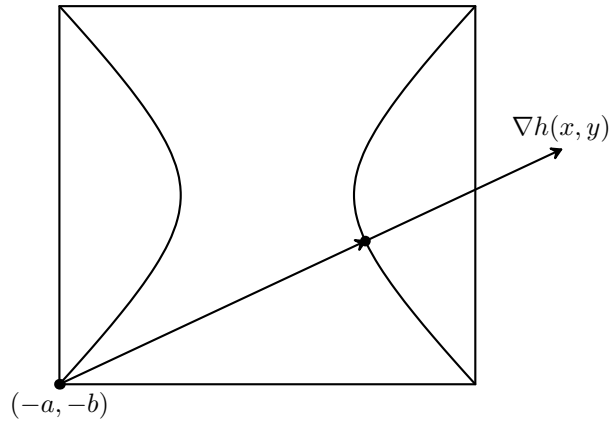


Figura 4: Ponto mais próximo do canto inferior esquerdo

sistema

$$\begin{cases} x+a = 2\lambda x \\ y+b = -2\lambda y \\ x^2-y^2 = a^2-b^2 \\ x > 0. \end{cases}$$

Uma vez que é claro que não existe nenhuma solução do sistema para a qual se tenha $2\lambda = \pm 1$, as duas primeiras equações podem ser substituídas por $x = -\frac{a}{1-2\lambda}$ e por $y = -\frac{b}{1+2\lambda}$ respectivamente. Então a terceira equação passa a ser

$$-4\frac{4(a^2-b^2)\lambda^4 - 3(a^2-b^2)\lambda^2 - (a^2+b^2)\lambda}{(1-2\lambda)^2(1+2\lambda)^2} = 0.$$

Uma solução desta equação é $\lambda = 0$, mas esta solução é irrelevante neste contexto, pois significa tomar-se $(x,y) = (-a,-b)$, o qual pertence ao ramo da esquerda da hipérbole. A outra solução pode ser obtida resolvendo a equação

$$\lambda^3 - \frac{3}{4}\lambda - \frac{a^2+b^2}{4(a^2-b^2)} = 0,$$

a qual pode ser resolvido recorrendo à fórmula de Cardano. (É claro que isto não faz sentido quando $a = b$, mas já se lidou com este caso.) A solução é

$$\lambda = \frac{1}{2}\left(\sqrt[3]{\frac{a-b}{a+b}} + \sqrt[3]{\frac{a+b}{a-b}}\right) = \frac{1}{2}\left(\sqrt[3]{\frac{\frac{a}{b}-1}{\frac{a}{b}+1}} + \sqrt[3]{\frac{\frac{a}{b}+1}{\frac{a}{b}-1}}\right) \qquad (6)$$

e, portanto, o ponto do ramo da direita da hipérbole mais próximo de $(-a,-b)$ é $(-\frac{a}{1-2\lambda},-\frac{b}{1+2\lambda})$, com o valor de λ dado por (6). Assim sendo,

$$\begin{aligned} d(a,b) &= \text{distância de } \left(-\frac{a}{1-2\lambda},\frac{b}{1+2\lambda}\right) \text{ a } (-a,-b) \\ &= 2\lambda\sqrt{\left(\frac{a}{1-2\lambda}\right)^2+\left(\frac{b}{1+2\lambda}\right)^2} \end{aligned}$$

e a desigualdade (5) equivale a

$$1 \leq \lambda\sqrt{\left(\frac{a/b}{1-2\lambda}\right)^2+\frac{1}{(1+2\lambda)^2}}.$$

O membro da direita desta desigualdade depende somente do quociente a/b e um cálculo numérico revela que a desigualdade tem lugar quando a/b for menor que um número μ cujo valor é aproximadamente 1,18. Consequentemente, o problema pode ter até três soluções quando $a < \mu b$. De facto, se isto se verificar, o problema tem exactamente três soluções quando $d(a,b) < c < 2b$; para um exemplo desta situação (obtido com $a = 1{,}1 \times b$ e $c = 1{,}9 \times b$), veja-se a figura 5.

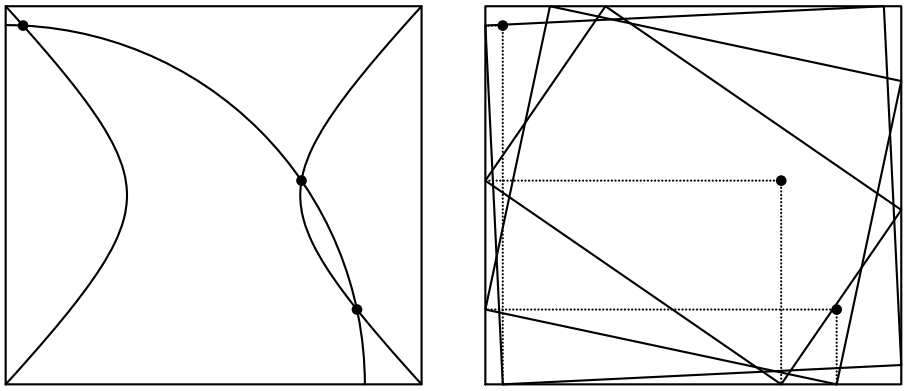


Figura 5: Existência de três soluções

Por outro lado, se $a > \mu b$, então, embora o problema tenha solução quando c está próximo de 0 ou de $2\sqrt{a^2+b^2}$, não tem solução quando $2b < c < d(a,b)$. Se, por exemplo, $a = 4$ e $b = 3$ (que, como já foi afirmado, são os valores usados para o rectângulo dado no caso das figuras 1 e 2), então $d(a,b) \simeq 7\frac{1}{9}$ e, portanto, o problema não tem solução quando $c = 7$, ou seja, quando o comprimento c for a média aritmética dos comprimentos de dois lados adjacentes do rectângulo dado.

É interessante observar que o problema aqui estudado de determinar a distância de um ponto a uma hipérbole também não pode, em geral, ser resolvido usando somente régua e compasso; veja-se [1].

6 Outros problemas

Vejamos o resultado de aplicar o teorema 2 a dois problemas clássicos.

Construção do pentágono regular

O primeiro destes é o de construir um pentágono regular. Mais precisamente, quer-se construir o pentágono de centro 0 do qual 1 é um dos vértices (aqui, estão a encarar-se os números 0 e 1 como sendo números complexos). Então os restantes vértices do pentágono são as raízes do polinómio x^5-1 distintas de 1. Visto que $x^5-1=(x-1)(x^4+x^3+x^2+x+1)$, as raízes em questão são as raízes do polinómio $q(x)=x^4+x^3+x^2+x+1$, o qual é irredutível (basta aplicar o critério de Eisenstein a $q(x+1)$ com $p=5$). Logo, as raízes de $q(x)$ são números algébricos de grau 4 e $q(x)$ é o polinómio minimal de cada uma delas. A cúbica resolvente de $q(x)$ é x^3-2x^2-2x+1, da qual -1 é uma raiz. Logo, o pentágono regular em questão pode ser construído com régua e compasso. Naturalmente, que é possível construir um pentágono regular dados o centro e um dos vértices é algo que já se sabe desde o tempo de Euclides; veja-se [4, Livro IV, proposição 11].

O problema de Alhazan

Hasan Ibn al-Haytham (c. 965–c. 1040), mais conhecido no Ocidente por Alhazen, propôs o seguinte problema no seu tratado de Óptica (veja-se [11] para mais detalhes): dados dois pontos A e B de um círculo de centro C, determinar um ponto P da circunferência do círculo tal que a bissectriz do ângulo $A\hat{P}B$ passe por C. Isto também pode ser visto como um problema de Óptica (o que é natural, dada a sua origem). Basta encarar a circunferência como um espelho circular e o problema passa então a ser o de determinar um ponto P desse espelho tal que um sinal luminoso emitido de A, ao ser reflectido em P passe por B. Este problema não pode, em geral, ser resolvido usando somente régua e compasso, como é demonstrado em [3], embora possa ser resolvido através da intersecção da circunferência com uma cónica, como foi descoberto por Christiaan Huygens (veja-se [11]).

Suponha-se que a circunferência em questão é a circunferência de centro $(0,0)$ e raio 1 e que os pontos A e B são $(1/6, 1/6)$ e $(-1/2, 1/2)$, respec-

tivamente. Pode-se provar que, neste caso, a abcissa das soluções P do problema são as raízes reais do polinómio $p(x) = x^4 - 2x^3 + 4x^2 + 2x - 1$, o qual é irredutível (veja-se [6, § 2]; a irredutibilidade de $p(x)$ também pode ser demonstrada recorrendo ao algoritmo descrito em [2]). Portanto, a abcissa de P é um número algébrico de grau 4. A cúbica resolvente de $p(x)$ é $x^3 - 8x^2 + 16x + 16$. Pelo teorema das raízes racionais, as suas raízes racionais só podem ser os divisores de 16, mas verifica-se facilmente que nenhum daqueles números é raiz deste polinómio. Isto mostra que o problema de Alhazen não pode ser resolvido usando somente régua e compasso. No caso concreto anterior, $p(x)$ tem duas raízes reais e, consquentemente, o problema tem duas soluções P e P^*, que podem ser vistas na figura 6, onde C é o centro da circunferência (note-se que, ao contrário do que possa parecer, os pontos P, B e P^* não são colineares).

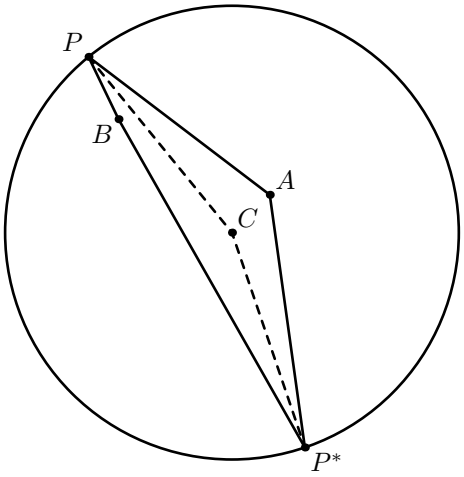


Figura 6: Soluções do problema de Alhazen

Referências

[1] H. Azad e A. Laradji, Some impossible constructions in elementary geometry, *Math. Gaz.*, **88** (2004), pp. 548–551.

[2] Gary Brookfield, Factoring quartic polynomials: A lost art, *Math. Mag.*, **80**(1) (2007), pp. 67–70. URL: http://www.jstor.org/stable/27642994

[3] Jack M. Elkin, A deceptively easy problem, *Math. Teacher*, **58** (1965), pp. 193–198. URL: http://www.jstor.org/stable/27968003

[4] Euclides, *Os Elementos*, Editora Unesp, São Paulo, 2009.

[5] Nathan Jacobson, *Basic Algebra I*, 2ª edição, W. H. Freeman and Company, New York, 1985.

[6] Peter M. Neumann, Reflections on reflection in a spherical mirror, *Amer. Math. Monthly*, **105**(6) (1998), pp. 523–528. URL: http://www.jstor.org/stable/2589403

[7] Ivan Niven, *Numbers: Rational and irrational*, The L. W. Singer Company, 1961.

[8] Russell A. Gordon e José Carlos Santos, An interesting construction problem, *Amer. Math. Monthly*, **125**(3) (2018), pp. 207–221

[9] José Carlos Santos, Another approach to the trisection problem, *Math. Gaz.*, **90** (2006), pp. 280–284.

[10] José Carlos Santos, *Números*, U. Porto Edições, Porto, 2014.

[11] John D. Smith, The Remarkable Ibn al-Haytham, *Math. Gaz.*, **76** (1992), pp. 189–198.

[12] Ian Stewart, *Galois Theory*, 4ª edição, Chapman & Hall, Boca Raton, 2015.

[13] Pierre-Laurent Wantzel, Recherches sur les moyens de reconnaître si un Problème de Géométrie peut se résoudre avec la règle et le compas, *J. Math. Pures Appl.*, **1**(2) (1837), pp. 366–372. URL: http://visualiseur.bnf.fr/ConsulterElementNum?O=NUMM-16381&Deb=374&Fin=380&E=PDF

Aproximações de π pelos métodos de Newton e de Wegstein

Mário M. Graça
Departamento de Matemática
Instituto Superior Técnico, Universidade de Lisboa
e-mail: mgraca@math.tecnico.ulisboa.pt

Resumo: A partir da função modelo $f(x) = \tan(x/4) - 1$, são construídos métodos iterativos de Newton e de Wegstein a fim de produzir novas fórmulas aproximantes do número π. Provamos que, para a função modelo, a iteradora de Wegstein possui ordem cúbica de convergência, no intervalo $[3, 4]$. Apresentam-se resultados numéricos mostrando que a auto composição de iteradoras de Newton/Wegstein pode ser usada recursivamente para obter aproximações de π com milhares de dígitos decimais correctos.

Abstract: Starting from the function model $f(x) = \tan(x/4) - 1$, are built iterative methods to produce approximating new formulae of the number π. We prove that, for the model function, the Wegstein iteration function has cubic convergence in the interval $[3, 4]$. Numerical results are presented showing that the auto composition of Newton/Wegstein iteration functions can be used recursively in order to obtain approximations of π having thousands of correct decimal digits.

palavras-chave: Método de Wegstein; método de Newton; ordem de convergência; iteradoras auto compostas.

keywords: Wegstein method; Newton method; convergence order; self composition.

1 Introdução

Um processo iterativo de ponto fixo do tipo $x_{k+1} = g(x_k)$ é frequentemente considerado não satisfatório caso (i) não convirja; (ii) seja de convergência lenta; ou (iii) quando a partir dele se pretenda construir um outro de convergência mais rápida. Um processo pensado para os casos (ii) ou (iii), ou seja, para acelerar a convergência de outro mais lento, é usado nesta nota a fim de aproximar o número π. Trata-se do método de Wegstein [11], o qual consiste essencialmente na aplicação de extrapolação linear de Aitken [2] sobre duas iteradas consecutivas da função g.

O método de Wegstein, além de não usar derivadas (tal como o método da secante com o qual tem semelhanças), oferece ainda a vantagem de permitir calcular aproximações de um determinado ponto fixo da função iteradora g, no caso em que tal ponto fixo é repulsor. Tal significa que o método de Wegstein pode ser útil em todos os casos (i) a (iii) anteriormente referidos. Além da sua simplicidade, este método é facilmente generalizável a funções iteradoras definidas em $\mathbb{R}^n$, com $n \geq 2$, o que explica a sua utilidade na aproximação de soluções de sistemas lineares e não lineares com várias variáveis [10], [1], [9].

Designando por W a função iteradora de Wegstein, esta depende do ponto x onde actua bem como da função iteradora de base adoptada, isto é, $W = W(g, x)$. Como função iteradora g, escolhemos a função de Newton, tendo em vista ilustrarmos a construção do respectivo método de Wegstein e responder à seguinte questão: no que respeita a aproximar o número π, o método de Wegstein é mais interessante do que o método de Newton, quando este é aplicado a uma certa função f que admita π como um seu zero?

A fim de respondermos a tal questão, calculámos algumas fórmulas aproximantes do número π, comparando as que resultam do método de Newton com fórmulas aproximantes obtidas pelo método de Wegstein. Em ambos os casos essas fórmulas serão depois modificadas, a fim de se obter aproximações de π com milhares de dígitos correctos, até ao limite de cálculo numérico permitido pelo sistema *Mathematica* quando usado num computador pessoal[1]. A limitação mais importante refere-se à precisão utilizada nos cálculos. Uma precisão superior a 10^6 dígitos decimais implica tempos de execução não aceitáveis, pelo que essa precisão é aqui considerada como o máximo atingível no sistema computacional que usámos.

Na Secção 2 começamos por descrever o método de Wegstein. Depois mostramos que a função W possui ordem de convergência superior à do método de Newton, quando este método é aplicado a funções reais f satisfazendo certas propriedades.

Em particular, na Secção 3, adoptamos como modelo a função $f(x) = \tan(x/4) - 1$, para a qual $z = \pi$ é um seu zero (simples), e se nos afigura a função real mais simples e natural quando se trata de aproximar esse famoso número. Para esta função mostramos que o método de Wegstein possui convergência cúbica, por exemplo, no intervalo $[3, 4]$.

A função de Newton correspondente, bem como a respectiva função iteradora de Wegstein, são depois utilizadas para a obtenção de fórmulas aproxi-

[1] No presente caso um portátil MacBook Pro, 2 GHz Intel Core i7.

mantes de π, fórmulas essas que aparentemente não se encontram arroladas na imensa literatura existente dedicada ao número π. Em particular, em [3], [4] , [6], [7] encontram-se descritas várias abordagens e fórmulas usadas ao longo dos tempos para aproximar esse número extraordinário. Métodos de segunda, terceira e quarta ordens de convergência são descritos detalhadamente, por exemplo, em [5] e [8]. Tais métodos, todavia, não são métodos de ponto fixo pelo que não pertencem à classe de métodos que aqui propomos. A comparação de alguns desses métodos com os processos de Newton/Wegstein ilustrados neste trabalho será efectuada noutra ocasião.

O sistema *Mathematica* permite-nos simplificar convenientemente as expressões simbólicas das iteradoras de Newton e Wegstein associadas à função $f(x) = \tan(x/4) - 1$. Consequentemente, surge naturalmente a ideia de obter métodos de Newton/Wegstein de ordem de convergência superior, mediante auto composição dessas duas iteradoras de base, tal como descrito na Secção 3.1. Uma vez que tais fórmulas obedecem a um padrão recursivo, que descrevemos, isso leva-nos naturalmente à construção de novas fórmulas, numericamente estáveis, a partir das quais se podem calcular aproximações de π de alta precisão, com milhares de casas decimais correctas. Algumas dessas fórmulas são apresentadas nos Parágrafos 3.1.1 (para a iteradora auto composta de Newton) e 3.1.2 (para a iteradora auto composta de Wegstein).

A estabilidade numérica das fórmulas que considerámos é inferida por meio de experimentação, iniciando os processos em causa respectivamente no ponto $x = 3$ e no ponto 'perturbado' $x = 3.14$, pontos esses tomados como aproximações iniciais do número π. Os resultados são automaticamente controlados mediante estimativas de erro inerentes a métodos de convergência supralinear, como é o presente caso. Tais estimativas são depois confirmadas mediante recurso aos valores de π que o sistema *Mathematica* fornece com precisão arbitrária (até ao limite de memória disponível). A demonstração formal da estabilidade numérica das fórmulas que utilizámos está fora do âmbito deste trabalho, pelo não será aqui considerada.

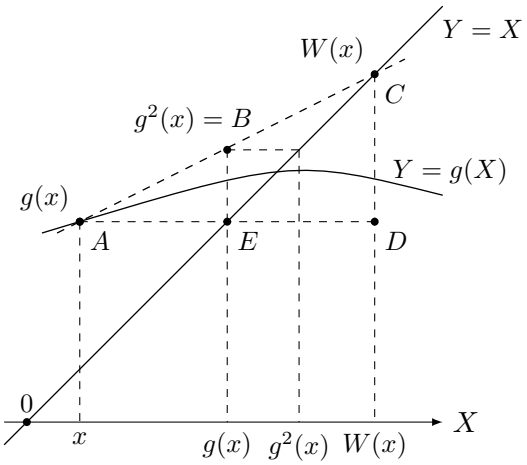
A fim de que o leitor possa fazer as suas próprias experiências, em Anexo é apresentado código para as iteradoras de Newton/Wegstein e respectivas auto composições.

2 O método de Wegstein

Dada uma função iteradora g, definida num domínio $D \subset \mathbb{R}$, e continuamente diferenciável até uma ordem conveniente, comecemos por deduzir expressões para a função iteradora de Wegstein correspondente.

Na figura a seguir é mostrada a geometria envolvida. A notação $g^2(x)$ refere-se à 2-auto composição de g, isto é, à função composta $g^2(x) = g(g(x))$. Em geral a j-auto composição de uma função g, será denotada por $g^j(x) = g \circ g \circ \cdots \circ g \quad (j \geq 2 \text{ vezes})$.

A função iteradora de Wegstein aplica o ponto x no ponto $W(x)$. Dado que o declive do segmento de recta que une os pontos $A = (x, g(x))$ e $B = (g(x), g^2(x))$ figurados é



$$m(x) = \frac{g^2(x) - g(x)}{g(x) - x}, \quad \text{onde} \quad g(x) \neq x, \tag{1}$$

o ponto C, localizado sobre a recta $Y = X$, de coordenadas $(W(x), W(x))$, obtém-se por extrapolação linear a partir dos pontos A e B. Atendendo à semelhança dos triângulos $\triangle ACD$ e $\triangle ABE$, resulta

$$m(x) = \frac{W(x) - g(x)}{W(x) - x}, \quad \text{onde} \quad W(x) \neq x \ . \tag{2}$$

Assim,

$$(m(x) - 1)\, W(x) = m(x)\, x - g(x) \ . \tag{3}$$

Admitindo que $m(x) \neq 1$, obtém-se

$$W(x) = \frac{1}{1 - m(x)}\, g(x) + \frac{m(x)}{m(x) - 1}\, x \ . \tag{4}$$

Fazendo

$$r(x) = \frac{m(x)}{1 - m(x)} \quad \Longrightarrow \quad 1 - r(x) = \frac{1}{1 - m(x)}, \tag{5}$$

a expressão (4) pode escrever-se na forma

$$W(x) = (1 - r(x))\, g(x) + r(x)\, x \ . \tag{6}$$

No caso de $r(x)$ ser função constante, a expressão (6) mostra-nos que $W(x)$ resulta de uma combinação linear convexa dos pontos x e $g(x)$.

Proposição 2.1. *Seja z um ponto fixo hiperbólico de g, isto é, $g(z) = z$ e $g'(z) \neq 1$.*
(i) A função 'declive' m definida em (1) pode prolongar-se por continuidade ao ponto fixo, isto é, a função

$$m(x) = \begin{cases} \dfrac{g^2(x) - g(x)}{g(x) - x}, & \text{se } x \neq z \\ g'(z), & \text{se } x = z \end{cases}$$

é contínua em z.
(ii) Um ponto fixo z de g é ponto fixo superatractor da função iteradora de Wegstein, ou seja,

$$\begin{aligned} W(z) &= z \\ W'(z) &= 0 \\ W''(z) &= \frac{2\, m'(z) - g''(z)}{g'(z) - 1} \ . \end{aligned} \tag{7}$$

Note-se que em (7), o símbolo $m'(z)$ deve ser entendido como $m'(z) = \lim_{h\to 0} \left(m(z+h) - m(z)\right)/h$.

Corolário 2.1. *Se*

$$m'(z) = \frac{1}{2}\, g''(z) \tag{8}$$

o método de Wegstein tem ordem de convergência 3 (pelo menos).

Demonstração. (i) A regra de l'Hôpital pode ser aplicada à expressão (1), obtendo-se

$$\begin{aligned} \lim_{x\to z} m(x) &= \lim_{x\to z} \frac{g'(g(x))\, g'(x) - g'(x)}{g'(x) - 1} \\ &= \lim_{x\to z} \frac{\left(g'(g(x)) - 1\right)\, g'(x)}{g'(x) - 1} = g'(z) \ . \end{aligned}$$

(ii) De (3), e atendendo a que $g(z) = z$, $m(z) = g'(z)$, vem

$$\left(g'(z) - 1\right)\, W(z) = g'(z)\, z - z = \left(g'(z) - 1\right)\, z \ .$$

Como $g'(z) \neq 1$, resulta que $W(z) = z$, isto é, z é ponto fixo da função W. A partir da igualdade (3), mediante derivação sucessiva resulta

$$m'(x)\,W(x) + (m(x)-1)\;W'(x) = m'(x)\,x + m(x) - g'(x) \tag{9}$$

e

$$\begin{aligned} m''(x)\,W(x) + 2\,m'(x)\,W'(x) + (m(x)-1)\;m''(x) \\ = m''(x) + 2\,m'(x) - g''(x)\;. \end{aligned} \tag{10}$$

Dado que $g(z) = z$ e $m(z) = g'(z)$, de (9) obtém-se

$$m'(z)\,z + (g'(z)-1)\;W'(z) = m'(z)\,z\;.$$

Como por hipótese, $g'(z) \neq 1$, é válida a segunda igualdade em (7). De (10), para $x = z$, vem

$$m''(z)\,z + (g'(z)-1)\;w''(z) = m''(z)\,z + 2\,m'(z) - g''(z),$$

isto é,

$$(g'(z)-1)\;w''(z) = 2\,m'(z) - g''(z),$$

sendo portanto válida a terceira igualdade em (7).

O Corolário deve-se ao facto de na última igualdade em (7) se ter $W''(z) = 0$, pelo que o método é (pelo menos) de terceira ordem de convergência. □

3 Iteradoras de Newton e de Wegstein como aproximantes de π

Adoptamos para equação modelo

$$f(x) = \tan\left(\frac{x}{4}\right) - 1 = 0\;. \tag{11}$$

A função f possui obviamente um só zero (simples), $z = \pi$, no intervalo $I = [3, 4]$. A iteradora de Newton correspondente tem a forma

$$g(x) = x + 2\left(1 + \cos\left(\frac{x}{2}\right) - \sin\left(\frac{x}{2}\right)\right) \tag{12}$$

donde,

$$g(\pi) = \pi \quad \text{e} \quad g'(\pi) = 0\;.$$

Assumindo continuidade para as funções g e g', é fácil concluir a partir das duas últimas igualdades que o método de Newton, se convergente, converge supralinearmente para π, ou seja, $\lim_{k\to\infty} |\pi - g(x_k)|/|\pi - x| = 0$, quando x suficientemente próximo de π. Levando em conta que a segunda derivada $g^{(2)}(\pi) = 1/2 \neq 0$, a convergência é de segunda ordem.

Pode mostrar-se que são válidas as condições do teorema do ponto fixo para a função g, no intervalo $I = [3, 4]$. Assim, o método de Newton converge quadraticamente para π, qualquer que seja o ponto inicial x_0 que se escolha no intervalo I para iniciar o processo iterativo $x_{k+1} = g(x_k), \quad k = 0, 1, \ldots$.
Uma vez que

$$m(x) = \frac{1 + \cos(u) - \sin(u)}{u(x) - x/2},$$

onde

$$u(x) = 1 + x/2 + \cos(x/2) - \sin(x/2),$$

resulta

$$\lim_{x\to\pi} m(x) = 0 = g'(\pi) \quad e \quad \lim_{x\to\pi} m'(x) = 1/4 \ .$$

Assim, $m'(\pi) = 1/2\, g''(\pi)$, isto é, é válida a igualdade (8). Por conseguinte, atendendo ao Corolário 2.1, o método de Wegstein é pelo menos de terceira ordem de convergência. Directamente, ou recorrendo a computação simbólica, como por exemplo a proporcionada pelo sistema *Mathematica* [12], pode concluir-se que $W'''(\pi) = -3/8 \neq 0$, pelo que o método de Wegstein associado à função (11) é de ordem cúbica de convergência. Tal significa que, neste caso, tanto a iteradora de Newton como a iteradora de Wegstein nos permitem construir processos iterativos que aproximam supralinearmente o número π, circunstância que nos habilita a estimar convenientemente o erro das respectivas iteradas, conforme se refere no Parágrafo 3.1.1 adiante.

3.1 Fórmulas aproximantes de π de ordem de convergência arbitrária

Tendo em conta a expressão (12) da função iteradora de Newton para a função f considerada, as fórmulas para a j-auto composição da função de Newton, $g^j(x) = g \circ g \circ \cdots \circ g$ ($j \geq 2$ vezes), podem ser obtidas recursivamente por aplicação sucessiva de uma função muito simples, abaixo denotada por h. A função iteradora correspondente possui ordem de convergência $p = 2^j$, pelo que o conjunto de iteradoras g^j, para $j \geq 1$ representa funções aproximantes de π, com ordem tão elevada quanto se queira.
Denotando por u a função

$$u(x) = 1 + x/2 + \cos(x/2) - \sin(x/2), \tag{13}$$

a função iteradora de Newton escreve-se como

$$g(x) = 2\, u(x), \tag{14}$$

exprimindo o facto de que quando $g(x)$ é aproximante de π, então $u(x)$ é aproximante de $\pi/2$, e reciprocamente.

Atendendo a que

$$g^2(x) = g(g(x)) = 2\,u\,(2\,u(x))$$
$$g^3(x) = g(g^2(x)) = 2\,u\,(2\,u(2\,u(x)))\,,$$

definindo uma função auxiliar h,

$$h(x) = 2\,u(x), \tag{15}$$

resulta que a função iteradora composta g^j, pode ser obtida recursivamente uma vez definida a função h dada em (15).

Assumindo já em memória a função f, dada em (11), segue-se pseudocódigo para as funções u, em (13), e para a função auto composta associada g^j, com $j \geq 1$. No pseudocódigo a seguir, o símbolo $\tilde{x} \leftarrow x$ significa que o valor em memória x é atribuído à variável $\tilde{x}$.

Função u

Input:

x;
$prec$; (* precisão a adoptar na aritmética de computador *)

$\tilde{x} \leftarrow x$ (* x representado com precisão $prec$ *);
$x_1 \leftarrow \tilde{x}/2$;

Output:

$$1 + x_1 + \cos(x_1) - \sin(x_1)\ .$$

Função g^j

Input:

x;
$prec$ (* precisão a adoptar na aritmética de computador *);
j (* $j \geq 1$, ordem da auto composição de g *);

$\tilde{x} \leftarrow x$ (* $\tilde{x}$ representa x com precisão $prec$ *);
$x_0 \leftarrow \tilde{x}$ (* x_0 é valor inicial *);

Repetir j vezes:
$y \leftarrow 2\,u(x_0)$;
$x_0 \leftarrow y$;

Output:

y .

A introdução da função h, dada em (15), representada pela variável y no pseudocódigo anterior, bem como o recurso a comandos recursivos como `Nest` do sistema *Mathematica*, em conjugação com programação dinâmica, permitem-nos calcular rapidamente expressões simbólicas de g^j, para j elevado, dentro das possibilidades de memória disponível. De outra forma, as expressões de g^j, calculadas directamente por auto composição a partir de (12), tornar-se-íam de tal modo complexas que o seu tempo de cálculo e ocupação de memória seriam proibitivos.

No Anexo 5.1 encontra-se código *Mathematica* para as funções g, g^j e W, W^j, traduzindo a recursividade que anteriormente descrevemos.

3.1.1 Fórmulas aproximantes de π pelo método de Newton

Passemos a ilustrar o processo recursivo descrito no Parágrafo 3.1, no caso particular da função g ser a função iteradora de Newton. Após $g(x) = 2\,u(x)$, a primeira função aproximante de Newton do número π, expressa em termos da função u dada em (13), é a seguinte:

$$\begin{aligned} g^2(x) &= 4 + x + 2\,\cos(x/2) + 2\,\cos(u(x)) - 2\,\sin(x/2) - 2\sin(u(x)) \\ &= 2\,\left[1 + u(x) + \cos(u(x)) - \sin(u(x))\right]\ . \end{aligned} \tag{16}$$

Denotando por v a função

$$v(x) = 1 + u(x) + \cos(u(x)) - \sin(u(x)), \tag{17}$$

a composta de Newton g^3 pode escrever-se como

$$g^3(x) = 2\,\left[1 + v(x) + \cos(v(x)) - \sin(v(x))\right]\ . \tag{18}$$

A função anterior é a mesma que se obtém auto compondo três vezes a função dada em (12), do que resulta explicitamente a seguinte fórmula aproximante de π:

$$\begin{aligned} g^3(x) &= 6 + x - 2\sin(x/2) + 2\cos(x/2) \\ &\quad +2\cos(x/2 - \sin(x/2) + \cos(x/2) + 1) \\ &\quad +2\cos(x/2 - \sin(x/2) + \cos(x/2) \\ &\quad +\cos(x/2 - \sin(x/2) + \cos(x/2) + 1) \\ &\quad -\sin(x/2 - \sin(x/2) + \cos(x/2) + 1) + 2) \\ &\quad -2\sin(x/2 - \sin(x/2) + \cos(x/2) + 1) \\ &\quad -2\sin(x/2 - \sin(x/2) + \cos(x/2) \\ &\quad +\cos(x/2 - \sin(x/2) + \cos(x/2) + 1) \\ &\quad -\sin(x/2 - \sin(x/2) + \cos(x/2) + 1) + 2)\ . \end{aligned}$$

Dada a complexidade da fórmula anterior e a sua potencial instabilidade numérica, ela não será usada directamente. Para obtenção de uma fórmula equivalente, será adoptado o esquema recursivo anteriormente descrito o qual, como se disse, será útil para calcular as funções auto compostas, g^j, para $j \geq 2$, algumas das quais são apresentadas a seguir.

Uma vez que as fórmulas associadas a $g^j(x)$ se destinam a aproximar $\pi \in [3,4]$, em particular começaremos por utilizar o ponto inicial $x = 3$, registando os resultados numéricos que resultem de um pequeno número de iterações de $g^j(3)$.
Por exemplo, é mostrado na Tabela 1, o número α de dígitos decimais correctos do valor obtido mediante auto composição de cada uma das funções g^j, com $1 \leq j \leq 4$, e anotada a respectiva ordem de convergência na última coluna tabelada.
Denotando por w o valor da quarta iterada de g^j, $w = g_4^j(3)$, o número α pode ser calculado pela expressão $\alpha = 1 + \lfloor -\log_{10}(|g^j(w) - w|)\rfloor$, uma vez que sendo g^j função iteradora de convergência supralinear, é bem conhecido que o erro absoluto de w é (muito bem) estimado por $|g^j(w) - w|$. Na expressão de α anterior $\lfloor \cdots \rfloor$ denota a função parte inteira superior.

Os cálculos foram efectuados começando por fixar a precisão do ponto inicial, $x = 3$, mediante a instrução `SetPrecision[3,10^5]`. A precisão adoptada foi aqui fixada em 10^5 porquanto o número final calculado, $g_4^4(3)$, possui um número de casas decimais correctas dessa ordem de grandeza, pelo que não terá interesse considerar uma precisão menor do que a prefixada. Relativamente à precisão a adoptar no cálculo de sucessivas iteradoras auto compostas, o preço a pagar quanto ao tempo de execução é decisivo. Por exemplo, a Tabela 1 foi calculada em 3.9 seg, enquanto a Tabela 2 demorou 85.7 seg.

Os resultados numéricos, ao passarmos do ponto inicial $x = 3$ ao ponto $x = 3.14$, sugerem estabilidade das fórmulas utilizadas, uma vez efectuado um ajustamento apropriado da precisão de cálculo. Assim, tomando para valor inicial $x = 3.14$ e fazendo `SetPrecision[3.14,10^6]`, mostra-se na segunda coluna da Tabela 2 o número de dígitos decimais correctos dos valores calculados. Uma vez que, tal como era de esperar, se observa um incremento do número de dígitos correctos calculados para as aproximações de π, por iteração de cada uma das funções g^j consideradas, concluímos que as fórmulas usadas para tais funções são estáveis para os valores iniciais usados.
O leitor poderá testar as funções iteradoras propostas escolhendo outros pontos iniciais no intervalo $[3,4]$.

j	α	ordem
1	23	2
2	373	4
3	5 965	8
4	95 434	16

Tabela 1: α é o número de dígitos decimais correctos da quarta iterada $g_4^j(3)$.

j	α	ordem
1	54	2
2	870	4
3	13 926	8
4	222 822	16

Tabela 2: α é o número de dígitos decimais correctos da quarta iterada $g_4^j(3.14)$.

Mostra-se a seguir (parcialmente) o número $w = g_4^4(3.14)$ calculado aquando da Tabela 2, apresentando apenas os seus primeiros 20 dígitos iniciais e 20 dígitos finais (dos 222 822 dígitos correctos do respectivo valor aproximado de π):

$$w = 3.1415926535897932384 \cdots 74986013697679794276 \ .$$

Uma vez que o sistema *Mathematica* nos dá o número π com precisão arbitrária, pode verificar-se que

$$\pi - w \simeq -2.8 \times 10^{-222\,822},$$

confirmando que o valor calculado por meio da função iteradora em causa possui, de facto, 222 822 dígitos correctos.

3.1.2 Fórmulas aproximantes de π pelo método de Wegstein

A função auxiliar u, dada em (13), $u(x) = 1 + x/2 + \cos(x(2) - \sin(x/2)$, é agora utilizada a fim de se obter uma versão numericamente estável da iteradora de Wegstein e respectivas auto compostas.

Efectuando simplificações das expressões que resultam das fórmulas (1) a (6), quando se considera a função f dada em (11), a função W respectiva pode escrever-se na forma

$$W(x) = \frac{p(x)}{q(x)} \tag{19}$$

onde

j	α	ordem
1	39	3
2	1 066	9
3	28 790	27

Tabela 3: α é o número de dígitos decimais correctos da terceira iterada $W_3^j(3)$.

j	α	ordem
1	92	3
2	2 479	9
3	66 924	27

Tabela 4: α é o número de dígitos decimais correctos da terceira iterada $W_3^j(3.14)$.

$p(x) = 4 + (x+4)\,(u(x) - (1 + x/2)) - 2\,\sin(x) + x\,(\sin(u(x)) - \cos(u(x)))$
e $q(x) = u(x) - (1 + x/2) + \sin(u(x)) - \cos(u(x))$.

Códigos *Mathematica* para definir esta função, bem como para as funções compostas W^j que dela resultam, é dado no Anexo 5.2, pressupondo que os cálculos são efectuados fixando previamente uma certa precisão `prec`.

Na segunda coluna da Tabela 3 apresenta-se o número α de dígitos correctos obtidos, partindo de $x = 3$, sendo `prec` $= 10^5$. Na Tabela 4 são mostrados os resultados para $x = 3.14$ e `prec` $= 10^6$. Em ambos os casos não se efectuaram mais do que três iterações pois, caso contrário, seria necessário aumentar a precisão dos cálculos para `prec` $> 10^6$, o que é inaceitável em termos do tempo de execução.
Quanto à função de Wegstein (19), inspecionando a segunda coluna das Tabelas 3 e 4, confirma-se que na passagem de W^j a W^{j+1}, para $1 \leq j \leq 3$, o número de dígitos correctos é aproximadamente multiplicado por $27 = 3^3$, dado que a iteradora W de base é de ordem 3 de convergência, conforme provado na Proposição 2.1. Há portanto melhoria dos resultados numéricos relativamente à iteradora de Newton, como se esperava.

4 Conclusões

Para responder à questão de saber se o método de Wegstein é vantajoso relativamente ao método de Newton, mostrámos que, de facto, o método de

Wegstein pode ser usado para melhorar os resultados do método de Newton aplicado à função $f(x) = \tan(x/4) - 1$, quando o objectivo consiste em obter valores aproximados de alta precisão do número π, mediante fórmulas estáveis das funções iteradoras de Newton/Wegstein auto compostas. O resultado teórico expresso na Proposição 2.1 reflecte-se nos resultados numéricos obtidos nos parágrafos 3.1.1 e 3.1.2, onde se confirma a convergência cúbica da função iteradora de Wegstein.

Evidentemente, poderíamos também considerar composições entre outras funções iteradoras g, e as correspondente iteradoras de Newton/Wegstein, e analisar as fórmulas aproximantes de π daí resultantes. Todavia, nesta nota, cingimo-nos apenas à auto composição de cada uma dessas duas funções iteradoras de base, por ser o processo mais óbvio de se obter funções iteradoras de ordem arbitrária. Além disso, no caso da função f adoptada, foi possível construir um processo recursivo para as nossas iteradoras auto compostas, tal como como descrito no Parágrafo 3.1 e seguintes.

5 Anexo (código *Mathematica*)

5.1 Para iteradoras de Newton

```
f[x_] := Tan[x/4] - 1;
u[x_] := u[x] = (xx = SetPrecision[x, prec]; x1 = xx/2;
      1 + x1 + Cos[x1] - Sin[x1]);
              (* iteradoras auto compostas: *)
g[j_, x_] := g[j, x] = (xx = SetPrecision[x, prec];
     Nest[2 u[#] &, xx, j]);

(*  teste :  *)
niter = 4;
prec = 10^5;
jmax = 4;
Timing[
 tab = Table[xx = SetPrecision[3, prec];
 v = Nest[g[k, #] &, xx, niter]; {k, Short[v, 0.6],
    1 + Floor[-Log10[Abs[g[k, v] - v]]]}, {k, 1, jmax}];
 Grid[tab, Frame -> All]]
```

5.2 Para iteradoras de Wegstein

(* iteradora de Wegstein *)

```
W[x_] := W[x] = Block[{xx, a, u, am, v, dif, p, q},
    xx = SetPrecision[x, prec];
    a = xx/2; am = 1 + a;
    u = am + Cos[a] - Sin[a];
```

```
    v = u - am;
    dif = Sin[u] - Cos[u];
    p = 4 + (xx + 4) v - 2 Sin[xx] + xx dif; (* numerador *)
    q = v + dif;  (* denominador *)
    p/q];
W[1, x_] := W[x];  (* iteradoras auto compostas *)
W[k_, x_] := W[k, x] = W[W[k - 1, x]];
```

Referências

[1] A. Arman, "Acceleration algorithms for process design simulations", Msc. Thesis, Oklahoma State University, 1986.

[2] A. C. Aitken, "On Bernoulli's numerical solution of algebraic equations", *Proc. Roy. Soc. Edinburgh*, Vol. 46 (1925), pp. 289–305.

[3] D. H. Bailey, "A short history of π formulas", 2016. Disponível em http://www.davidhbailey.com/dhbpapers/pi-history.pdf

[4] D. H. Bailey, "Finding new mathematical identities via numerical computations", *ACM SIGNUM*, Vol. 33, No. 1 (1998), pp. 17–22.

[5] D. H. Bailey, "The computation of π to $29,360,000$ decimal digits using Borwein's quartically convergent algorithm", *Math. Comp.*, Vol. 50, No. 181 (1988), pp. 283–296.

[6] J. M. Borwein, S. T. Chapman, "I prefer Pi: A brief history and anthology of articles in the American Mathematical Monthly", *Amer. Math. Monthly*, Vol. 121, No. 1 (2015).

[7] J. M. Borwein, "The life of Pi: from Archimedes to ENIAC and beyond", (2012). Disponível em http://www.cs.utsa.edu/~wagner/pi/LifeOfPi2.pdf

[8] J. M. Borwein, P. B. Borwein, "An explicit cubic iteration for π", *BIT*, Vol. 26 (1986), pp. 123–126.

[9] F. C. Knopf, Modeling, *Analysis and Optimization of Process and Energy Systems*, John Wiley and Sons, New York, 2012.

[10] C. H. Gutzler, "An iterative method of Wegstein for solving simultaneous nonlinear equations", Msc. Thesis, Oregon State College, 1959. Disponível em http://ir.library.oregonstate.edu/xmlui/bitstream/handle/1957/50411/GutzlerCharlesH1959.pdf?sequence=3

[11] J. H. Wegstein, "Accelerating convergence of iterative processes", *Commun. ACM*, Vol. 1 (1958), pp. 9–13.

[12] S. Wolfram, *The Mathematica Book*, Wolfram Media, fifth ed., 2003.

O método de Diofanto e a curva $a^b = b^a$

Maria Pires de Carvalho[1], *Alberto Cavaleiro Pacheco*[2]
Centro de Matemática da Universidade do Porto
Faculdade de Ciências da Universidade do Porto
e-mail: mpcarval@fc.up.pt; up201605820@fc.up.pt

Resumo: A curva descrita pelos pares (a, b) no plano tais que $a, b > 0$, $a \neq b$ e $a^b = b^a$ é parametrizável pelos declives $0 < t \neq 1$ do feixe de rectas de equações cartesianas $y = tx$, o que permite descrever facilmente os pontos do traço desta curva cujas coordenadas são ambas números algébricos, ou inteiros algébricos ou racionais. Em particular, uma tal parametrização fornece um método simples de encontrar coordenadas de pontos desta curva que são números transcendentes.

Abstract: The curve described by the pairs (a, b) in the plane satisfying $a, b > 0$, $a \neq b$ and $a^b = b^a$ is parameterizable by the slopes $0 < t \neq 1$ of the straight lines whose Cartesian equations are given by $y = tx$. This information allows us to easily detect those pairs whose coordinates are algebraic over the field of the rational numbers, or are algebraic integers, or else are both rational. In particular, such a parametrization provides a simple criterium to find coordinates of points of this curve which are transcendental numbers.

palavras-chave: Número algébrico sobre $\mathbb{Q}$; inteiro algébrico.

keywords: Algebraic number over $\mathbb{Q}$; algebraic integer.

1 A equação $a^b = b^a$, para $a, b > 0$

Para que valores reais de $a, b > 0$ se tem $a^b = b^a$? A igualdade é óbvia quando $a = b$. Para valores distintos de a e b, se reescrevermos a equação $a^b = b^a$ como $a^{\frac{1}{a}} = b^{\frac{1}{b}}$ e esboçarmos o gráfico da função $x > 0 \mapsto f(x) = x^{\frac{1}{x}}$, ilustrado na Figura 1, teremos uma ideia aproximada dos valores da imagem de f que são obtidos mais do que uma vez (e, nesse caso, exactamente duas vezes).

[1] MC tem sido financiada pelo CMUP (UID/MAT/00144/2013), que é suportado financeiramente pela FCT com fundos nacionais (MEC) e europeus, através dos programas FEDER e no âmbito do acordo PT2020. Os autores agradecem ao revisor os comentários e sugestões.

[2] Este artigo foi escrito no âmbito do *Programa Novos Talentos em Matemática*, da Fundação Calouste Gulbenkian.

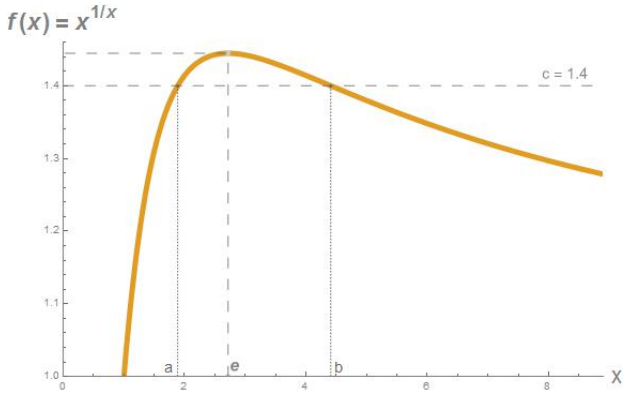


Fig. 1: Gráfico da função f.

Esta função tem um máximo global $e^{\frac{1}{e}}$, atingido apenas em $x = e$, e é estritamente crescente em $]0, e[$ e estritamente decrescente em $]e, +\infty[$. Além disso,

$$\lim_{x\to 0^+} f(x) = 0^+ \quad \text{e} \quad \lim_{x\to+\infty} f(x) = 1^+.$$

Logo, cada recta horizontal $y = c$ com $1 < c < e^{\frac{1}{e}}$ (e só para estes valores de c) intersecta o gráfico de f em dois pontos cujas abcissas determinam dois reais positivos distintos a e b tais que $a^{\frac{1}{a}} = b^{\frac{1}{b}}$, ou seja, $a^b = b^a$. Mais precisamente, dado $c \in]1, e^{\frac{1}{e}}[$, existe um e um só $b > e$ tal que $f(b) = b^{\frac{1}{b}} = c$; se agora resolvermos a equação $a^{\frac{1}{a}} = c$ com a incógnita $a \in]1, e[$, determinamos o outro valor a do domínio de f tal que $f(a) = f(b) = c$.

Para descrever o lugar geométrico de tais pares (a, b) com $a \neq b$, usemos o feixe de rectas $y = tx$ com declive $t \in \mathbb{R}^+\setminus\{1\}$ que, como um radar, permite detectá-los no 1º quadrante de $\mathbb{R}^2$ (veja-se em [2] outra instância em que este método, atribuído a Diofanto, é usado). Para cada t, determinamos a intersecção das condições $y = tx$ e $x^y = y^x$ resolvendo em conjunto as equações

$$\frac{y}{x} = t \quad \text{e} \quad x^{\frac{y}{x}} = y.$$

Obtemos então $x^t = tx$, logo $x = t^{\frac{1}{t-1}}$; e, como $y = tx = x^t$, tem-se $y = t^{\frac{t}{t-1}}$. As soluções descrevem a curva α em $\mathbb{R}^+ \times \mathbb{R}^+$, cujo traço está esboçado na Figura 2, parametrizada por

$$\alpha\colon t \in \mathbb{R}^+ \setminus \{1\} \quad \mapsto \quad \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right).$$

Observe-se que o traço de α, que designaremos por $\mathcal{T}_\alpha$, é simétrico relativamente à bissectriz do 1º quadrante (a semirecta de equação $y = x$, $x \geq 0$) uma vez que, se fixarmos $t > 0$ e o par correspondente $\alpha(t) = (a_t, b_t) = \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right)$ de α, então $\frac{1}{t}$ determina o ponto $\alpha(\frac{1}{t}) = (b_t, a_t)$ da mesma curva. Além disso, $\alpha(t)$ converge para (e, e) quando t tende para 1. Note-se ainda que as coordenadas dos pontos desta curva são ambas estritamente maiores do que 1, propriedade que resulta de só surgirem tais pares com abcissas no subconjunto $]1, +\infty[$ do domínio da função f.

Se a $\mathcal{T}_\alpha$ juntarmos a semirecta $\{(a, a) : a \in \mathbb{R}^+\}$, obtemos o conjunto $\mathfrak{X}$ de todos os pares $(a, b) \in \mathbb{R}^+ \times \mathbb{R}^+$ tais que $a^b = b^a$. Os conjuntos $\mathcal{T}_\alpha$ e $\mathfrak{X}$ estão representados na Figura 2. (Esta figura consta do artigo [1], onde se analisaram as soluções reais e complexas da equação $a^b = b^a$.) Os dois

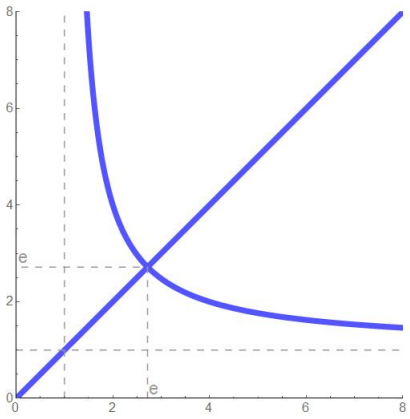


Fig. 2: Os conjuntos $\mathcal{T}_\alpha$ e $\mathfrak{X}$.

ramos de $\mathfrak{X}$ intersectam-se precisamente no ponto (e, e) e dividem $\mathbb{R}^+ \times \mathbb{R}^+$ em quatro regiões em cada uma das quais o sinal da diferença $a^b - b^a$, que varia continuamente com a e b, se mantém constante. Em particular, como a recta vertical $x = e$ só intersecta as regiões em que este sinal é positivo (uma vez que são gráficos de funções, $\mathcal{T}_\alpha$ e $\mathfrak{X}$ intersectam cada recta vertical quando muito uma vez), concluímos que, para todos os valores de $0 < x \neq e$, se tem $e^x > x^e$ (por exemplo, $e^\pi > \pi^e$).

Designemos por $\mathcal{A}$ o conjunto de números reais algébricos sobre $\mathbb{Q}$ e por $\mathcal{A}_\mathbb{Z}$ o seu subconjunto dos inteiros algébricos (definições na Secção 2). O objectivo deste texto é o de localizar pontos de $\mathcal{T}_\alpha$ cujas coordenadas sejam ambas racionais, ou ambas números algébricos. Recorde-se que, uma vez que o ponto do traço de α determinado por $\frac{1}{t}$ é (b_t, a_t), bastará analisar

estas coordenadas para $t \in]0,1[$. A tabela seguinte resume o que de mais relevante aqui se provará sobre as coordenadas dos pontos (a_t, b_t) de $\mathcal{T}_\alpha$ para valores de $t \in]0,+\infty[\setminus\{1\}$:

1. (a_t, b_t) tem ambas as coordenadas algébricas se e só se $t \in \mathbb{Q}^+ \setminus \{1\}$.
2. (a_t, b_t) tem ambas as coordenadas racionais se e só se $t = \frac{n}{n+1}$ ou $t = \frac{n+1}{n}$ para algum $n \in \mathbb{N}$.
3. (a_t, b_t) tem ambas as coordenadas inteiras algébricas se e só se $t = 1/n$ ou $t = n$ para algum $n \in \mathbb{N} \setminus \{1\}$.

$0 < t < 1$	Domínio a que pertence (a_t, b_t)	a_t	b_t
$\frac{n}{m}$ $(n, m \in \mathbb{N})$	$\mathcal{A} \times \mathcal{A}$	$\left(\frac{m}{n}\right)^{\frac{m}{m-n}}$	$\left(\frac{m}{n}\right)^{\frac{n}{m-n}}$
$\frac{n}{n+1}$ $(n \in \mathbb{N})$	$\mathbb{Q} \times \mathbb{Q}$	$\left(\frac{n+1}{n}\right)^{n+1}$	$\left(\frac{n+1}{n}\right)^{n}$
$\frac{1}{m}$ $(m \in \mathbb{N},\ m > 1)$	$\mathcal{A}_{\mathbb{Z}} \times \mathcal{A}_{\mathbb{Z}}$	$\sqrt[m-1]{m^m}$	$\sqrt[m-1]{m}$
$\frac{1}{2}$	$\mathbb{N} \times \mathbb{N}$	4	2

Em particular, resulta deste estudo que, quando a é um natural maior do que 1 e b é um real positivo tal que $a^b = b^a$, então $a = b$ ou $\{a, b\} = \{2, 4\}$ ou b é um número transcendente.

2 O corpo $\mathcal{A}$ dos números algébricos

Um número real diz-se algébrico sobre $\mathbb{Q}$ (ou, simplesmente, *algébrico*) se for zero de um polinómio não nulo com coeficientes racionais. Por exemplo, os números racionais são algébricos (pois a fracção irredutível $\frac{p}{q}$, com $p \in \mathbb{Z}$ e $q \in \mathbb{N}$, anula o polinómio $qx - p$); $\sqrt{2}$ é algébrico uma vez que é raiz de $x^2 - 2$; e $\sqrt{2} + \sqrt{3}$ também é algébrico já que é zero do polinómio $x^4 - 10x^2 + 1$:

$$\begin{aligned} \left(\sqrt{2} + \sqrt{3}\right)^2 &= 5 + 2\sqrt{6} \\ \left([\sqrt{2} + \sqrt{3}]^2 - 5\right)^2 &= 24 \\ \left(\sqrt{2} + \sqrt{3}\right)^4 - 10\left(\sqrt{2} + \sqrt{3}\right)^2 + 1 &= 0. \end{aligned}$$

O *grau* de um número algébrico z é o menor natural n tal que z é zero de um polinómio não nulo com coeficientes racionais e grau n. Por exemplo, os racionais são os algébricos de grau 1; $\sqrt{2}$ tem grau 2; $\sqrt{2}+\sqrt{3}$ é de grau 4.

Parece difícil demonstrá-lo, por a noção de número algébrico depender de propriedades da família dos polinómios, mas o conjunto dos números algébricos sobre $\mathbb{Q}$ com a soma e produto usuais forma um corpo, que designaremos por $\mathcal{A}$, que até é fechado para a operação de exponenciação por potências racionais. Demonstrações detalhadas destas e de outras propriedades de $\mathcal{A}$ podem ser lidas em [12] ou [14]. Resumiremos aqui o essencial deste argumento. Consideremos $z, w \in \mathcal{A}$, e sejam

$$\mathcal{P}\colon x \in \mathbb{R} \mapsto c_n x^n + \cdots + c_1 x + c_0 \quad \text{e} \quad \mathcal{Q}\colon x \in \mathbb{R} \mapsto d_m x^m + \cdots + d_1 x + d_0$$

dois polinómios não nulos com coeficientes racionais que se anulam, respectivamente, em z e em w. Então:

1. Se $\mathcal{P}^-$ é o polinómio definido por $x \in \mathbb{R} \mapsto \mathcal{P}(-x)$, então P^- tem coeficientes racionais (os mesmos de $\mathcal{P}$, a menos de sinal nas parcelas de grau ímpar) que se anula em $-z$. Assim sendo, $-z$ é algébrico.

2. Se $z \neq 0$ e $\mathcal{P}$ tem grau n, com coeficiente $c_j \in \mathbb{Q}$ de x^j para cada $j \in \{0, 1, \cdots, n\}$, seja $\mathcal{P}^{\div}$ o polinómio

$$x \in \mathbb{R} \mapsto \mathcal{P}^{\div}(x) = \begin{cases} x^n\, P(\frac{1}{x}) & \text{se } x \neq 0 \\ c_n & \text{se } x = 0 \end{cases}$$

Então, $\mathcal{P}^{\div}$ tem coeficientes racionais (os mesmos de $\mathcal{P}$, mas c_j é agora coeficiente da potência x^{n-j}, para $0 \leq j \leq n$) que se anula em $\frac{1}{z}$. O que confirma que $\frac{1}{z}$ é algébrico.

3. Sejam $k \in \mathbb{N}$, $k \geq 2$, e $\mathcal{P}^{\wedge}$ o polinómio $x \in \mathbb{R} \mapsto \mathcal{P}(x^k)$. Então, $\mathcal{P}^{\wedge}$ é também um polinómio de coeficientes racionais (os de $\mathcal{P}$) que se anula em $\sqrt[k]{z}$. Logo $\sqrt[k]{z}$ é algébrico. Note-se que esta raiz é um número real para todo o natural ímpar k e, se k é par, quando $z > 0$.

4. Suponhamos que os graus dos polinómios $\mathcal{P}$ e $\mathcal{Q}$ são n e m, respectivamente. Então, existem racionais $(c_i)_{i\in\{0,1,\cdots,n\}}$ e $(d_j)_{j\in\{0,1,\cdots,m\}}$ tais que

$$z^n = c_{n-1}z^{n-1} + \cdots + c_1 z + c_0 \quad \text{e} \quad w^m = d_{m-1}w^{m-1} + \cdots + d_1 w + d_0. \quad (1)$$

Multiplicando a primeira igualdade de (1) por z, concluimos que z^{n+1} é também combinação linear finita sobre o corpo $\mathbb{Q}$ de $\{1, z, \ldots, z^{n-1}\}$. Por

indução finita, deduzimos que, para todo o $k \in \mathbb{N}$, a potência z^k é combinação linear finita sobre $\mathbb{Q}$ de $1, z, \ldots, z^{n-1}$. Analogamente se conclui que w^k é combinação linear finita sobre o corpo $\mathbb{Q}$ de $\{1, w, \ldots, w^{m-1}\}$, para todo o $k \in \mathbb{N}$. Mais geralmente, todos os elementos do conjunto $\{1, z+w, (z+w)^2, \ldots, (z+w)^{nm}\}$ são combinações lineares finitas sobre $\mathbb{Q}$ das nm parcelas $z^i\, w^j$, com $i \in \{0, 1, \ldots, n-1\}$ e $j \in \{0, 1, \ldots, m-1\}$. Desta forma, temos $mn+1$ vetores num subespaço vectorial sobre o corpo $\mathbb{Q}$ gerado por mn vetores. E, portanto, esses $mn+1$ vetores são linearmente dependentes sobre $\mathbb{Q}$. Ou seja, existem racionais $e_0, \ldots, e_{mn}$ não todos nulos tais que $e_0 + e_1(z+w) + \cdots + e_{mn}(z+w)^{mn} = 0$. O que prova que $z+w$ é algébrico.

Analogamente, as potências $\{1, zw, (zw)^2, \ldots, (zw)^{nm}\}$ são combinações lineares finitas sobre $\mathbb{Q}$ das nm parcelas $z^i\, w^j$, para $i \in \{0, 1, \ldots, n-1\}$ e $j \in \{0, 1, \ldots, m-1\}$. Consequentemente, estes $mn+1$ vetores são linearmente dependentes sobre $\mathbb{Q}$. Isto é, existem racionais $E_0, \ldots, E_{mn}$ não todos nulos tais que $E_0 + E_1\, zw + \cdots + E_{mn}\, (zw)^{mn} = 0$. Logo zw é algébrico.

2.1 O anel $\mathcal{A}_{\mathbb{Z}}$ dos inteiros algébricos

Um número real diz-se um inteiro algébrico sobre $\mathbb{Q}$ (ou, simplesmente, *um inteiro algébrico*) quando é algébrico e zero de um polinómio mónico com coeficientes inteiros. Por exemplo, cada número inteiro z é inteiro algébrico (por ser zero do polinómio $x \in \mathbb{R} \mapsto x - z$); $\sqrt{3}$ também, já que anula o polinómio $x^2 - 3$; e $\sqrt[3]{2} + 1$, que anula o polinómio $(x-1)^3 - 2 = x^3 - 3x^2 + 3x - 3$, também é inteiro algébrico. O conjunto dos inteiros algébricos, que designaremos por $\mathcal{A}_{\mathbb{Z}}$, forma um anel com a soma e produto usuais, que também é fechado para a operação de exponenciação por potências racionais.

Relativamente aos argumentos de (1) e (3) da subsecção anterior, temos apenas de observar que, se $\mathcal{P}$ é mónico e tem coeficientes inteiros, então o mesmo vale para os polinómios $\mathcal{P}^-$ e $\mathcal{P}^\wedge$. Quanto a $\mathcal{P}^\div$, o argumento de (2) não funciona, a não ser que c_0 seja ± 1, pois c_0 é o coeficiente de x^n em $\mathcal{P}^\div$. Por exemplo, apesar de 2 ser inteiro algébrico, o racional $\frac{1}{2}$ não o é porque os zeros racionais de qualquer polinómio mónico com coeficientes inteiros são números inteiros (a propósito, veja-se o início da Secção 6).

Finalmente, o argumento de (4) também não se apropria pois $\mathbb{Z}$ não é um corpo. Contudo, é ainda verdade que, se z e w são inteiros algébricos, então zw e $z+w$ também o são. Sejam $\mathcal{P}\colon x \in \mathbb{R} \mapsto x^n + \cdots + c_1 x + c_0$ e $\mathcal{Q}\colon x \in \mathbb{R} \mapsto x^m + \cdots + d_1 x + d_0$ dois polinómios não nulos com coeficientes inteiros que se anulam, respectivamente, em z e em w. As funções $\mathcal{P}$ e $\mathcal{Q}$ são os polinómios característicos de duas matrizes quadradas A e B com

entradas inteiras, nomeadamente

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & c_0 \\ 1 & 0 & 0 & \cdots & 0 & c_1 \\ 0 & 1 & 0 & \cdots & 0 & c_2 \\ \vdots & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 1 & c_{n-1} \end{pmatrix} \quad \text{e} \quad B = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & d_0 \\ 1 & 0 & 0 & \cdots & 0 & d_1 \\ 0 & 1 & 0 & \cdots & 0 & d_2 \\ \vdots & & & & & \vdots \\ 0 & 0 & \cdots & 0 & 1 & d_{m-1} \end{pmatrix}.$$

Formemos o produto de Kronecker[3] entre as matrizes A e B; obtemos uma matriz quadrada $A \otimes B$, de dimensões $mn \times mn$, cujo conjunto de valores próprios contém zw. Além disso, como a matriz $A \otimes B$ tem entradas inteiras, o seu polinómio característico é não nulo, mónico e tem coeficientes inteiros, o que prova que zw é inteiro algébrico. Analogamente, se considerarmos a soma de Kronecker $A \otimes Id_{m\times m} + Id_{n\times n} \otimes B$, onde $Id_{k\times k}$ designa a matriz identidade com k linhas, que tem entradas inteiras e $z+w$ como valor próprio, concluimos que $z + w$ é inteiro algébrico.

Por exemplo, consideremos $z = \sqrt{2}$, $w = \sqrt[3]{5}$, $\mathcal{P}(x) = x^2 - 2$ e $\mathcal{Q}(x) = x^3 - 5$. Então

$$A = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \quad \text{e} \quad B = \begin{pmatrix} 0 & 0 & 5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

e o produto zw é valor próprio da matriz

$$A \otimes B = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

O polinómio característico de $A \otimes B$ é $x^6 - 200$, cujas raízes reais são precisamente $\pm\sqrt{2}\,\sqrt[3]{5}$. No Capítulo VI de [14] pode ler-se um outro argumento,

[3] O produto de Knonecker de duas matrizes $C = (c_{ij})_{i=1,\cdots,m;\ j=1,\cdots,n}$ e $D = (d_{k\ell})_{k=1,\cdots,p;\ j=1,\cdots,q}$, com dimensões $m \times n$ e $p \times q$, é a matriz $C \otimes D$ que tem mp linhas e nq colunas e cujas entradas são dadas por $e_{\theta\eta} = c_{ij}\, d_{k\ell}$, onde $\theta = p(i-1) + k$ e $\eta = q(j-1) + \ell$. Por exemplo, o produto de Kronecker entre uma matriz C quadrada 2×2 e uma matriz D que seja 3×2 é a matriz 6×4 representada por

$$C \otimes D = \begin{pmatrix} c_{11}D & c_{12}D \\ c_{21}D & c_{22}D \end{pmatrix}.$$

usando funções simétricas, para demonstrar que $\mathcal{A}_{\mathbb{Z}}$, com a soma e produto usuais em $\mathbb{R}$, é um anel.

2.2 $\mathbb{R} \setminus \mathcal{A}$

O conjunto dos números algébricos sobre $\mathbb{Q}$ é infinito (pois contém $\mathbb{Q}$) e numerável, uma vez que os seus elementos são os da união numerável, indexada por $k \in \mathbb{N}$ e por $(c_0, c_1 \cdots, c_k) \in \mathbb{Q}^k \times (\mathbb{Q} \setminus \{0\})$, dos conjuntos (finitos) dos zeros dos polinómios de grau k com coeficientes $c_0, c_1 \cdots, c_k$. Como a união de dois conjuntos numeráveis é numerável e $\mathbb{R}$ não é numerável (pode ler-se o argumento de Cantor que prova esta última afirmação no primeiro capítulo de [13]; uma outra demonstração consta de [3]), o complementar de $\mathcal{A}$ em $\mathbb{R}$, formado pelos números *transcendentes*, é não vazio. Não é fácil exibir exemplos concretos de números transcendentes, e é, em geral, bastante difícil provar que um número em particular é transcendente. Na verdade, ainda há números reais sobre os quais não se sabe se são irracionais, quanto mais se são transcendentes. A dificuldade está na definição deste tipo de números: para se provar que um número é algébrico basta encontrar um polinómio não nulo de coeficientes racionais de que o número seja raiz; pelo contrário, para se garantir que um número é transcendente há que confirmar que ele não anula nenhum polinómio não nulo de coeficientes racionais, e esta família é vasta. Por isso, são úteis os critérios expeditos de verificação da algebricidade de um número, alguns dos quais mencionaremos brevemente de seguida.

Liouville apresentou em [10] uma outra prova da existência de números transcendentes e o seu argumento permite construir exemplos destes números. Liouville demonstrou que, para qualquer número real algébrico z de grau $n > 1$, existe um natural M tal que, para todos os inteiros p e $q > 0$, se tem

$$\left| z - \frac{p}{q} \right| > \frac{1}{M\, q^n}.$$

Daqui resulta que, se um real z for irracional e, para todo o natural m, existirem inteiros p e $q > 1$ tais que $\left| z - \frac{p}{q} \right| < \frac{1}{q^m}$, então z é transcendente. Tais z's muito bem aproximados por racionais chamam-se *números de Liouville*. De facto, suponhamos que um tal z é algébrico de grau n (sendo $n > 1$ por z ser irracional). Então existiria um natural M tal que $\left| z - \frac{p}{q} \right| > \frac{1}{M\, q^n}$ para todos os inteiros p e $q > 0$. Tome-se, porém, um natural k tal que

$$2^k \geq 2^n M$$

que existe porque $\mathbb{N}$ não é majorado, e verifica $k \geq n$ por M ser um natural.

Sendo z um número de Liouville, existem inteiros p e $q > 1$ tais que

$$\left| z - \frac{p}{q} \right| < \frac{1}{q^k}$$

e, portanto, devemos ter $\frac{1}{q^k} > \frac{1}{M\, q^n}$, ou seja,

$$M > q^{k-n} \geq 2^{k-n} \geq M.$$

Esta contradição indica que números como os de Liouville não podem ser algébricos.

Com este critério, Liouville gerou o primeiro número transcendente conhecido: $L = \sum_{j=1}^{+\infty} \frac{1}{10^{j!}}$. Este número é irracional, uma vez que tem uma dízima infinita não periódica, e, dado um natural m, se considerarmos os inteiros $q = 10^{m!}$ e $p = q \sum_{j=1}^{m} \frac{1}{10^{j!}}$, obtemos

$$\begin{aligned} \left| L - \frac{p}{q} \right| &= \left| \sum_{j=1}^{+\infty} \frac{1}{10^{j!}} - \sum_{j=1}^{m} \frac{1}{10^{j!}} \right| = \sum_{j=m+1}^{+\infty} \frac{1}{10^{j!}} < \frac{9}{10^{(m+1)!}} \sum_{j=0}^{+\infty} \frac{1}{10^{j}} \\ &= \frac{10}{10^{(m+1)!}} < \frac{10^{m!}}{10^{(m+1)!}} = \frac{1}{10^{m(m)!}} = \frac{1}{q^m}. \end{aligned}$$

E, portanto, L é um número de Liouville, logo transcendente.

O conjunto $\mathcal{L}$ de números de Liouville é residual (isto é, intersecção numerável de abertos densos) de $\mathbb{R}$, logo é denso em $\mathbb{R}$. Contudo, tem medida de Lebesgue nula: dado $\varepsilon > 0$, existe uma sucessão de intervalos $(I_n)_{n\in\mathbb{N}}$ tal que $\sum_{n=1}^{+\infty} |I_n| < \varepsilon$ e

$$\mathcal{L} \subset \bigcup_{n\in\mathbb{N}} I_n.$$

Pode ler-se mais informação sobre $\mathcal{L}$ nos capítulos 3 e 7 de [12].

Usando outro tipo de argumento, em 1873 Hermite [7] provou que e é transcendente e, pouco depois, Lindemann [9] e Weierstrass [16] generalizaram esse método, demonstrando que π é transcendente, assim encerrando o celebrado problema sobre a impossibilidade de construir, com régua não graduada e compasso, um quadrado com área igual à de um círculo. Por estas referências, ficamos a saber que são também transcendentes os números

- e^x, $\operatorname{sen}(x)$, $\cos(x)$, $\operatorname{tg}(x)$, para todo o número algébrico $x \neq 0$;

- $\log(x)$, $\arcsen(x)$, $\arccos(x)$, $\arctan(x)$, para todo o número algébrico $x \notin \{0,1\}$ do domínio destas funções.

Estes dois resultados são casos particulares de um teorema posterior, provado em 1934 por A. O. Gelfond [6] e, independentemente, através de uma solução mais elementar, em 1935 por Th. Schneider [15]. Na referência [12] pode ler-se uma demonstração primorosa do que é agora conhecido como Teorema de Gelfond-Schneider. Estes autores resolveram o sétimo problema proposto por Hilbert [8] e criaram um gerador simples de números transcendentes. Ora acontece que os números associados à equação $a^b = b^a$ se encaixam muito bem nesse teorema porque ele garante que, *se* a, b *são números algébricos,* $a \notin \{0,1\}$ *e* $b \in \mathbb{R} \setminus \mathbb{Q}$*, então* a^b *é transcendente.* Este resultado prova que, por exemplo, são transcendentes os números $2^{\sqrt{2}}$, $\log_{10}(2)$, e^{π}; e os que analisaremos neste texto.

Todavia, sobram ainda reais a que nenhum destes teoremas, ou outros, se aplica, como 2^e ou $\pi + e$. Neste contexto, é curioso referir um resultado de Mahler (veja-se [11] ou [4]) que prova, em particular, que se f é a função $x \in \mathbb{R} \mapsto f(x) = \frac{x(x+1)}{2}$, então é transcendente o número cuja dízima é

$$0.f(1)\, f(2) \cdots f(n) \cdots = 0.1361015 \cdots .$$

3 A equação $n^x = x^n$, para $n \in \mathbb{N} \setminus \{1\}$ e $x \in \mathbb{R}$

Fixemos um natural $n > 1$. Na Secção 1 verificámos que a equação $n^x = x^n$ tem duas soluções positivas, nomeadamente $x_1(n) = n$ e $x_2(n) = t^{\frac{t}{t-1}}$, sendo t o único real de $]0,1[\cup \{2\}$ tal que $n = t^{\frac{1}{t-1}}$ (cf. Figura 3). Em particular, $x_2(2) = 4$, $x_2(4) = 2$ e $x_2(n) \in\ ]1, e[$ se $n \geq 3$, sendo $\lim_{n\to+\infty} x_2(n) = 1^+$. Mas, ao contrário do que acontece com a equação geral $a^b = b^a$ em que, para nos mantermos no domínio dos reais, precisamos que a e b sejam positivos, a equação $n^x = x^n$ também é válida para $x \in\ ]-\infty, 0[$ e, dependendo da paridade de n, pode ter soluções neste intervalo. A Figura 4 mostra a existência de uma tal solução negativa quando $n = 2$; na mesma figura à direita podemos notar por que não existem soluções negativas quando $n = 3$. Mais geralmente, quando n é ímpar, a igualdade $n^x = x^n$ não é verificada por nenhum real negativo uma vez que, se $x < 0$, temos $x^n < 0$ e $n^x = e^{x \log n} > 0$.

Suponhamos agora que n é par e consideremos as funções

$$\mathcal{F}\colon x \in \mathbb{R} \mapsto n^x \quad \text{e} \quad \mathcal{G}\colon x \in \mathbb{R} \mapsto x^n.$$

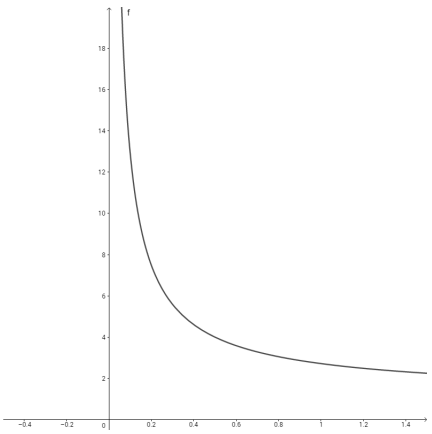


Fig. 3: Gráfico da função $t \mapsto t^{\frac{1}{t-1}}$.

Então, por n ser par, temos

$$\begin{aligned}(\mathcal{F}-\mathcal{G})(-1) &= \frac{1}{n}-(-1)^n = \frac{1}{n}-1<0\\ (\mathcal{F}-\mathcal{G})(0) &= 1-0>0\end{aligned}$$

e, portanto, como $\mathcal{F}-\mathcal{G}$ é contínua, pelo Teorema do Valor Intermédio existe um zero da função $\mathcal{F}-\mathcal{G}$ em $]-1,0[$; ou seja, existe um real negativo $x_3(n) \in]-1,0[$ tal que $n^{x_3(n)} = (x_3(n))^n$. Por outro lado, a derivada de $\mathcal{F}-\mathcal{G}$ é dada por

$$x \in \mathbb{R} \quad \mapsto \quad (\log\, n)\, n^x - n\, x^{n-1}$$

função que, como $n-1$ é natural ímpar, é positiva no intervalo $]-\infty,0[$. Consequentemente, a restrição a $]-\infty,0[$ da função $\mathcal{F}-\mathcal{G}$ é injectiva e, por isso, o zero $x_3(n)$ em $]-\infty,0[$ de $\mathcal{F}-\mathcal{G}$ é único.

Observe-se, finalmente, que para n par se tem $(x_3(n))^n = (-x_3(n))^n$ e

$$n^{x_3(n)} = (x_3(n))^n \quad \Leftrightarrow \quad \frac{\log\, n}{n} = \frac{\log\,(-x_3(n))}{x_3(n)}.$$

Logo, como $\lim_{n\to+\infty} \frac{\log\, n}{n} = 0$ e $x_3(n) \in]-1,0[$ para todo o n par, devemos ter

$$\lim_{n \text{ par}\, \to\, +\infty} x_3(n) = -1.$$

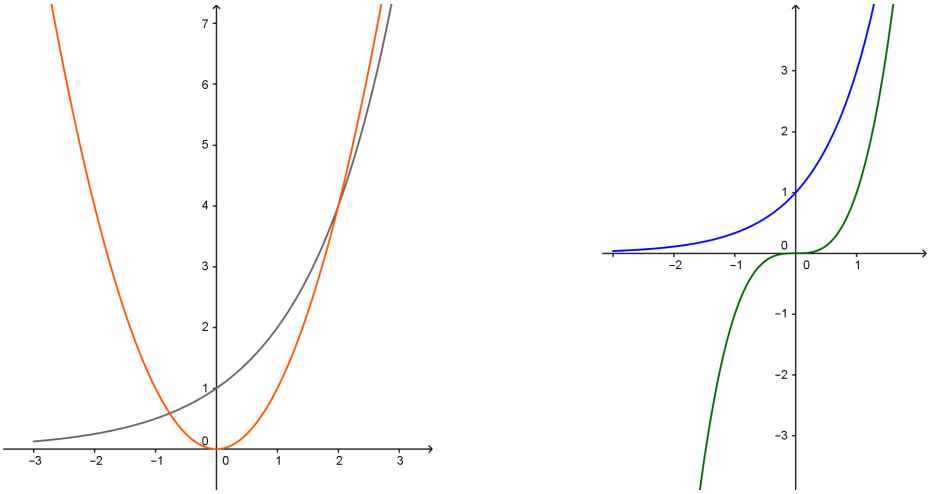


Fig. 4: Gráficos das funções $x \in \mathbb{R} \mapsto x^2$ e $x \in \mathbb{R} \mapsto 2^x$; e, à direita, $x \in \mathbb{R} \mapsto x^3$ e $x \in \mathbb{R} \mapsto 3^x$.

3.1 Irracionalidade de $x_2(n)_{n \neq 1,\, 2,\, 4}$ e de $x_3(n)_{n \text{ par}}$

Fixemos um natural $n > 1$. No que se segue, verificaremos que $x_2(n)$ é irracional quando $n \neq 2, 4$ e que, para n par, $x_3(n)$ é sempre irracional.

Comecemos por analisar o caso de $x_2(n)$. Recorde-se que, para $n \geq 3$, $x_2(n)$ está em $]1, e[$; em particular, $x_2(n)$ é um inteiro se e só se $x_2(n) = 2$, o que é equivalente a $n = 4$. Suponhamos que para algum valor de $n \neq 4$ se tem $x_2(n)$ racional. Se representarmos $x_2(n)$ pela sua fracção irredutível $\frac{p}{q}$, onde $p, q \in \mathbb{N}$ são tais que $mdc\,(p, q) = 1$, a igualdade $n^{\frac{p}{q}} = (\frac{p}{q})^n$ é equivalente a $n^p\, q^{nq} = p^{nq}$, o que implica que $q = 1$ por q ser primo com p. Temos então de resolver em $\mathcal{Z} = \{z \in \mathbb{N} \colon z > 1\}$ a equação $n^p = p^n$. Ora, é uma consequência simples do facto de a equação $n^x = x^n$ só ter duas soluções, nomeadamente $x_1(n) = n$ e $x_2(n) \in\,]1, e[$ se $n \geq 3$, que as únicas soluções $n, p \in \mathcal{Z}$ desta equação são (n, n), $(2, 4)$ e $(4, 2)$. Logo, se $n \neq 2, 4$, o número $x_2(n)$ é irracional.

Quanto a $x_3(n)$, que só existe quando n é par e está no intervalo $]-1, 0[$, se, para algum valor de n, $x_3(n)$ fosse uma fracção irredutível $-\frac{p}{q}$, onde $p, q \in \mathbb{N}$ são tais que $mdc\,(p, q) = 1$, então, como no caso anterior, concluiríamos que $p = 1$ (logo $q > 1$ pois $x_3(n) > -1$) e que q teria de ser solução da equação $n = q^{nq}$. Contudo, para todo o par de números naturais $n, q > 1$, tem-se $q^n > n$. De facto, é fácil provar por indução finita que, para todo o

natural $n > 1$ se tem $2^n > n$: $2^2 > 2$; e se, para um natural $k > 1$ fixado, se tem $2^k > k$, então, como $k > 1$, $2^{k+1} = 2^k \times 2 > 2k > k+1$. Logo, para todo o natural $q > 1$ e todo o natural $n > 1$, temos $q^n \geq 2^n > n$. Assim sendo, a equação $n = q^{nq}$ não tem soluções em $\mathcal{Z}$ e, portanto, $x_3(n)$ é irracional.

3.2 Transcendência de $x_2(n)_{n \neq 1,2,4}$ e de $x_3(n)_n$ par

Já sabemos que, para cada natural $n > 1$, a coordenada $x_2(n)$ do ponto $(n, x_2(n))$ do traço de α é irracional quando $n \neq 2, 4$ e que, para n par, $x_3(n)$ é sempre irracional. Mas podemos afirmar mais: estes números são transcendentes. Estudemos $x_2(n)$, adaptando-se facilmente o argumento para $x_3(n)$. Suponhamos, pelo contrário, que, para algum natural $1 < n \neq 2, 4$, o número $x_2(n)$ é algébrico. Então, como vimos na Secção 2, a potência $(x_2(n))^n$ também é um número algébrico. Contudo, $n \neq 0, 1$, $x_2(n)$ é irracional (veja-se a Secção 3.1) e, por hipótese, $x_2(n)$ é algébrico. Logo, pelo Teorema de Gelfond-Schneider, $n^{x_2(n)}$ é transcendente, o que não é compatível com a igualdade $(x_2(n))^n = n^{x_2(n)}$. Esta contradição indica que $x_2(n)$ não pode ser algébrico.

4 $\mathcal{T}_\alpha \cap (\mathbb{Q} \times \mathbb{Q})$

Já sabemos que no traço $\mathcal{T}_\alpha$ da curva α não há pontos (n, y) com $n \in \mathbb{N}$ e $y \in \mathbb{Q}$, com excepção de $(2, 4)$ e $(4, 2)$. E quanto a pontos com ambas as coordenadas racionais? Suponhamos que um tal ponto da intersecção do traço de α com $\mathbb{Q}^+ \times \mathbb{Q}^+$ é (a, b). Como vimos na Secção 1, existe $t \in \mathbb{R}^+ \setminus \{1\}$ tal que $a = a_t = t^{\frac{1}{t-1}}$ e $b = b_t = t^{\frac{t}{t-1}}$. Ora, sendo a e b racionais e distintos, o declive t da recta que une $(0, 0)$ e (a, b) está em $\mathbb{Q}^+ \setminus \{1\}$. Pela simetria do conjunto $\mathcal{T}_\alpha$ relativamente à recta $y = x$, basta prosseguir a análise no caso em que $a > e$, sendo que então se tem $1 < b < e$ e $0 < t < 1$.

Consideremos a fracção irredutível de t, digamos $\frac{n}{m}$, onde $n, m \in \mathbb{N}$, $n < m$ e $mdc\,(n, m) = 1$. Então

$$a = \left(\frac{m}{n}\right)^{\frac{m}{m-n}} \qquad \text{e} \qquad b = \left(\frac{m}{n}\right)^{\frac{n}{m-n}}. \tag{2}$$

Comecemos por notar que, se $m - n$ dividir n e m, então os expoentes de $\frac{m}{n}$ em a e em b são inteiros e, portanto, a e b são racionais. Ora, como n e m são primos entre si, para $m - n$ dividir ambos os naturais n e m devemos ter $m - n = 1$. Ou seja, para todos os valores racionais de $t = \frac{n}{n+1}$, os pontos $(a, b) = \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right)$ do traço de α têm ambas as coordenadas racionais. Por exemplo,

$n \in \mathbb{N}$	$t = \frac{n}{n+1}$	$a_t = \left(\frac{n}{n+1}\right)^{n+1}$	$b_t = \left(\frac{n}{n+1}\right)^{n}$
1	$\frac{1}{2}$	4	2
2	$\frac{2}{3}$	$\frac{27}{8}$	$\frac{9}{4}$
3	$\frac{3}{4}$	$\frac{254}{81}$	$\frac{64}{27}$

Haverá outros pares $(a, b) \in \mathcal{T}_\alpha$ com ambas as coordenadas racionais? Voltemos à expressão (2) de a e b determinados por $t = \frac{n}{m}$, para $n, m \in \mathbb{N}$, $n < m$ e $mdc\,(n, m) = 1$. Note-se que, como $b = ta$ e t é racional, basta determinar para que valores de n e m pode a ser racional. Estando a em $\mathbb{Q}$, existem $r, s \in \mathbb{N}$ tais que $r > s$, $mdc\,(r, s) = 1$ e $a = \frac{r}{s}$. Então, elevando a primeira igualdade de (2) ao expoente $m - n$, resulta que

$$\left(\frac{m}{n}\right)^m = \left(\frac{r}{s}\right)^{m-n}.$$

Observe-se que, como n e m são primos entre si, tem-se $mdc\,(m^m, n^m) = 1$: de facto, se p fosse um divisor primo de m^m e de n^m, então dividiria n e m; mas estes naturais são, por hipótese, primos entre si. Analogamente se conclui que $mdc\,(r^{m-n}, s^{m-n}) = 1$. Então as duas fracções

$$\frac{m^m}{n^m} \quad \text{e} \quad \frac{r^{m-n}}{s^{m-n}}$$

são escritas em fracção irredutível do mesmo racional. Consequentemente, devemos ter

$$m^m = r^{m-n} \quad \text{e} \quad n^m = s^{m-n}.$$

Mas então existem naturais u, v tais que

$$m = u^{m-n} \quad \text{e} \quad n = v^{m-n}$$

sendo $u > v$ porque $n < m$. Vejamos como encontrar tais u e v.

Como $m > 1$, para determinarmos u basta considerar a factorização de m em primos. Seja p um primo que divide m e surge na factorização de m com potência $\beta \in \mathbb{N}$. Então, como $m^m = r^{m-n}$, o primo p tem também de

dividir r, ocorrendo na factorização de r com uma potência máxima $\gamma \in \mathbb{N}$. Como a factorização em primos é única, concluimos que

$$\beta\, m = \gamma\,(m-n).$$

Ora m e $m-n$ são primos entre si e, portanto, esta igualdade implica que $m-n$ divide β; isto é, existe $d \in \mathbb{N}$ tal que $\beta = d\,(m-n)$. E, portanto, todos os primos da factorização de m ocorrem em m com potência que é divisível por $m-n$. Então, se $m = p_1^{\beta_1} \cdots p_k^{\beta_k}$ é a factorização de m em primos distintos e $\beta_i = d_i\,(m-n)$ para todo o $i \in \{1, \cdots, k\}$, tem-se

$$m = \left(p_1^{\beta_1} \cdots p_k^{\beta_k}\right) = p_1^{d_1\,(m-n)} \cdots p_k^{d_k\,(m-n)} = \left(p_1^{d_1} \cdots p_k^{d_k}\right)^{m-n}$$

e $u = p_1^{d_1} \cdots p_k^{d_k}$. Um argumento análogo permite concluir que ou $n = 1$, caso em que $v = 1$, ou $n > 1$ e se constrói v como u para m. Daqui resulta que

$$m - n = u^{m-n} - v^{m-n}$$

onde, recorde-se, $u > v$ e $m - n \geq 1$. Contudo, se $m - n \geq 2$, então

$$\begin{aligned} m-n &= u^{m-n} - v^{m-n} \\ &= (u-v)\left(u^{m-n-1} + u^{m-n-2}\,v + \cdots + u\,v^{m-n-2} + v^{m-n-1}\right) \\ &> m-n \end{aligned}$$

uma vez que na soma de naturais $u^{m-n-1} + u^{m-n-2}\,v + \cdots + u\,v^{m-n-2} + v^{m-n-1}$ há $m-n$ parcelas de números naturais das quais pelo menos uma, u^{m-n-1}, é estritamente maior do que 1. Consequentemente, devemos ter $m - n = 1$, ou seja,

$$t = \frac{n}{n+1}, \quad a = \left(\frac{n+1}{n}\right)^{n+1} \quad \text{e} \quad b = \left(\frac{n+1}{n}\right)^{n}.$$

Note-se ainda que $\lim_{n\to+\infty} t = 1$ e $\lim_{n\to+\infty} a = \lim_{n\to+\infty} b = e$.

5 $\mathcal{T}_\alpha \cap (\mathcal{A} \times \mathcal{A})$

Dadas as propriedades que referimos na Secção 2 sobre o corpo dos números algébricos, sabemos que, para cada número racional $t \in \mathbb{R} \setminus \{1\}$, o par (a, b) tal que $a = t^{\frac{1}{t-1}}$ e $b = t^{\frac{t}{t-1}}$ tem ambas as coordenadas algébricas. Reciprocamente, suponhamos que o par $(a, b) = \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right)$ de $\mathcal{T}_\alpha$ tem ambas as coordenadas algébricas. Então devemos ter $t \in \mathbb{Q}$, caso contrário, pelo Teorema de Gelfond-Schneider, $b = a^t$ seria transcendente.

6 $\mathcal{T}_\alpha \cap (\mathcal{A}_\mathbb{Z} \times \mathcal{A}_\mathbb{Z})$

Comecemos por notar que $\mathcal{A}_\mathbb{Z} \cap \mathbb{Q} = \mathbb{Z}$, uma vez que os zeros racionais de um polinómio mónico $x \in \mathbb{R} \mapsto x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$ com coeficientes $c_i \in \mathbb{Z}$, para $i = 0, 1, \ldots, n$, são fracções irredutíveis $\frac{r}{s}$ tais que $r \in \mathbb{Z}$, $s \in \mathbb{N}$, $mdc\,(r, s) = 1$, s divide o coeficiente 1 de x^n e r divide c_0; e, portanto, $s = 1$ e $\frac{r}{s} = r$ é um inteiro.

Consideremos um ponto $(a, b) = \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right)$ de $\mathcal{T}_\alpha$, onde $t \in]0, 1[$, e suponhamos que as suas coordenadas são inteiros algébricos (o argumento é análogo para $t \in]1, +\infty[$). Então, como vimos na Subsecção 5, o declive t é racional, digamos, $t = \frac{n}{m}$ com $n, m \in \mathbb{N}$, $n < m$ e $mdc\,(n, m) = 1$. Observe-se que, como a é inteiro algébrico, a^{m-n} também é; além disso, a^{m-n} é racional porque

$$a^{m-n} = \left(t^{\frac{1}{t-1}}\right)^{m-n} = \left(\frac{m}{n}\right)^m.$$

Logo a^{m-n} é inteiro, e portanto $n = 1$. Consequentemente, $t = \frac{1}{m}$ para algum natural $m \geq 2$.

Reciprocamente, quando $t = \frac{1}{m}$ para algum natural $m \geq 2$, o par $(a, b) = \left(t^{\frac{1}{t-1}}, t^{\frac{t}{t-1}}\right)$ tem coordenadas que são inteiros algébricos. Efectivamente, neste caso,

$$a = \sqrt[m-1]{m^m} \quad \text{e} \quad b = \sqrt[m-1]{m}$$

e estes números são, respectivamente, zeros dos polinómios mónicos de coeficientes inteiros

$$\mathcal{P}(x) = x^{m-1} - m^m \quad \text{e} \quad \mathcal{Q}(x) = x^{m-1} - m.$$

Por exemplo,

m	t	$a_t = t^{\frac{1}{t-1}}$	$b_t = t^{\frac{t}{t-1}}$	$\mathcal{P}$	$\mathcal{Q}$
2	$\frac{1}{2}$	4	2	$x - 4$	$x - 2$
3	$\frac{1}{3}$	$3\sqrt{3}$	$\sqrt{3}$	$x^2 - 27$	$x^2 - 3$
4	$\frac{1}{4}$	$4\sqrt[3]{4}$	$\sqrt[3]{4}$	$x^3 - 256$	$x^3 - 4$

Referências

[1] Associação Atractor, *Dinâmica de uma família de exponenciais*, Gazeta de Matemática 181 (2017) 3–7.

[2] M. Carvalho, *O método das cordas*, Bol. Soc. Port. Mat. Número especial do tricentenário de Leonard Euler (2008) 61–72.

[3] M. Carvalho, S. Cavadas, *Jogando no limite*, Bol. Soc. Port. Mat. 69 (2013) 1–19.

[4] J. H. Cadwell, *Topics in Recreational Mathematics*, Cambridge University Press, 1980.

[5] L. Euler, *De formulis exponentialibus replicatis*, Acta Academiae Scientarum Imperialis Petropolitinae 1 (1778) 38–60.

[6] A. O. Gelfond, *Sur le septieme problème de D. Hilbert*, Doklady Akad. Nauk. 2 (1934) 1–6.

[7] C. Hermite, *Sur la fonction exponentielle*, C. R. Acad. Sci. Paris 77 (1873) 18–24.

[8] D. Hilbert, *Sur les problèmes futures des mathématiques*, C. R. Deuxième Congrès International des Mathématiciens, Paris, 1900, 58–114.

[9] F. Lindemann, *Ueber die Zahl p*, Math. Annalen 20 (1882) 213–225.

[10] J. Liouville, *Sur des classes très-étendues de quantités dont la valeur n'est ni rationnelle ni même réductible à des irrationnelles algébriques*, C. R. Acad. Sci. Paris 18 (1844) 883–885; J. Math. Pures Appl. 16:1 (1851) 133-142.

[11] K. Mahler, *Lectures on Transcendental Numbers*, LNM 546, Springer New York, 1976.

[12] I. Niven, *Irrational Numbers*, Wiley New York, 2012.

[13] J. C. Oxtoby, *Measure and Category*, Springer New York, 1980.

[14] H. Pollard, H. G. Diamond, *The Theory of Algebraic Numbers*, Carus Mathematical Monographs 9, MAA, 1975.

[15] Th. Schneider, *Transzendenzuntersuchungen periodischer Funktionen I*, Journal fur Mathematik 172 (1935) 65–69.

[16] K. Weierstrass, *Math. Werke II* (1895) 341–362.

Invariants and TQFT's for cut cellular surfaces from finite 2-groups

Diogo Bragança
CENTRA & Physics Department
Instituto Superior Técnico, Universidade de Lisboa
e-mail: diogo.braganca@tecnico.ulisboa.pt

Roger Picken
CAMGSD & Mathematics Department
Instituto Superior Técnico, Universidade de Lisboa
e-mail: roger.picken@tecnico.ulisboa.pt

Resumo: Nesta breve continuação de um artigo anterior, relembramos a noção de uma superfície celular recortada (CCS), sendo uma superfície com fronteira, que é cortada de uma maneira especificada para ser representada no plano, e é composta de 0-, 1 - e 2-células. Obtemos invariantes de CCS sob transformações semelhantes às de Pachner na estrutura celular, contando as colorações das 1- e 2-células com elementos de um 2-grupo finito, sujeito a uma condição de "planicidade falsa" para cada 2-célula. Essas invariantes, que estendem as invariantes de Yetter para esta classe de superfícies, também são descritas numa configuração de TQFT. Um resultado do artigo anterior relativo à fração de comutação de um grupo é generalizado para o contexto de 2-grupos.

Abstract: In this brief sequel to a previous article, we recall the notion of a cut cellular surface (CCS), being a surface with boundary, which is cut in a specified way to be represented in the plane, and is composed of 0-, 1- and 2-cells. We obtain invariants of CCS's under Pachner-like moves on the cellular structure, by counting colorings of the 1- and 2-cells with elements of a finite 2-group, subject to a "fake flatness" condition for each 2-cell. These invariants, which extend Yetter's invariants to this class of surfaces, are also described in a TQFT setting. A result from the previous article concerning the commuting fraction of a group is generalized to the 2-group context.

palavras-chave: Superfície celular recortada; TQFT; grupo finito; módulo cruzado; 2-grupo; fração de comutação.

keywords: Cut cellular surface; TQFT; finite group; crossed module; 2-group; commuting fraction.

1 Introduction

In our previous work [3] we studied invariants of a class of surfaces with boundary, obtained by counting certain G-colorings of the 1-cells of the surface, where G is a finite group. We called these surfaces *cut cellular surfaces* (CCS's), since they come equipped with a planar representation which arises from cutting the surface along some 1-cells to get a simply-connected planar region made up of 2-cells bounded by a circuit of 1- and 0-cells (see Section 2 for the full definition). Such surfaces include triangulated surfaces, but allow for a considerably more economical description in terms of the number of cells needed. For instance, a triangulation of the 2-sphere S requires at least four 2-cells, six 1-cells and four 0-cells, whereas its minimal representation as a CCS has just one 2-cell, one 1-cell and two 0-cells (see Section 2).

The invariants that we studied in [3] involved counting the number of so-called flat G-colorings of the 1-cells of the surface, i.e. assignments of elements of G to the 1-cells of the surface, such that, taking into account the orientation of the 1-cells, their product around the boundary of each 2-cell equals 1_G, the identity element of G. For a triangulated surface without boundary, these invariants coincide with the Dijkgraaf-Witten invariants of the surface [4]. They are invariant under simple moves on the cellular structure, namely subdividing or combining 1-cells and subdividing or combining 2-cells. We showed that these two types of moves generate the well-known Pachner moves for triangulated surfaces. The invariants also behave well under gluing of surfaces along shared boundary components and we showed that they give rise to a topological quantum field theory (TQFT). See the second section of our previous article [3] for an introduction to the notion of TQFT.

The number of flat G-colorings for minimal CCS representations of some elementary surfaces, like the sphere, cylinder, pants surface and torus, has a group-theoretical significance, e.g. for the torus this is the number of commuting pairs of elements of G. Using topological arguments we were able to derive some group-theoretical properties, such as:

Proposition 1.1 *The number of conjugacy classes of G is equal to the commuting fraction of G times the order of G.*

We recall that the commuting fraction of G is defined to be the number of commuting pairs of elements of G divided by the overall number of pairs.

The constructions in [3] were intended to pave the way for an analogous approach using finite 2-groups, which is the subject of the present article.

In Section 3, we recall the definition of a finite 2-group $\mathcal{G}$, also known as a finite crossed module. It consists of two finite groups G and H, a group homomorphism from H to G, and a left action of G on H by automorphisms, subject to two conditions.

We then define invariants of CCS's (Def. 3.4) which involve counting the number of $\mathcal{G}$-colorings of the surface, i.e. assignments of elements of G to the 1-cells and elements of H to the 2-cells. These assignments, are subject to a "fake flatness" condition, which reduces to the flatness condition when the group H is trivial. We prove several properties of the expressions of Definition 3.4, in particular that they are invariant under the aforementioned two types of move on the cellular structure. For triangulated surfaces these invariants correspond to Yetter's invariants [10, 5]. In section 4 we calculate the invariant for some elementary examples.

In Section 5, we describe how the invariant behaves when gluing two CCS's together along a common boundary component, and use this to get a TQFT for these surfaces. We focus on properties of the invariant for the cylinder, and in Proposition 5.9 we obtain a generalization of Proposition 1.1 in the 2-group context. Finally, in the conclusions of section 6, we comment on some features of the TQFT and give an interpretation for the invariants in terms of the notion of groupoid cardinality.

To make this article self-contained, we have repeated some material from [3]. We invite the reader to consult this previous article for fuller details concerning a number of points.

To conclude this introduction we will say a brief word about notation. When we wish to describe a linear map $Z : V \to W$ in concrete terms, we may introduce a basis $\{e_i\}_{i=1,\ldots,n}$ of V and a basis $\{f_j\}_{j=1,\ldots,m}$ of W. Then Z is represented by an $m \times n$ matrix $[c_{ji}]$, where

$$Z(e_i) = \sum_{j=1}^{m} c_{ji} f_j.$$

We will be using the suggestive physicists' notation for the matrix elements c_{ji}, namely:

$$c_{ji} = \langle f_j \,|Z|\, e_i \rangle \,.$$

2 Cut cellular surfaces

We will be considering surfaces with boundary, which are cut in a specified way to be represented in the plane (like the well-known rectangle with opposite edges identified representing the torus), and which are composed of 0-, 1- and 2-cells, generalizing the familiar notion of a triangulated surface.

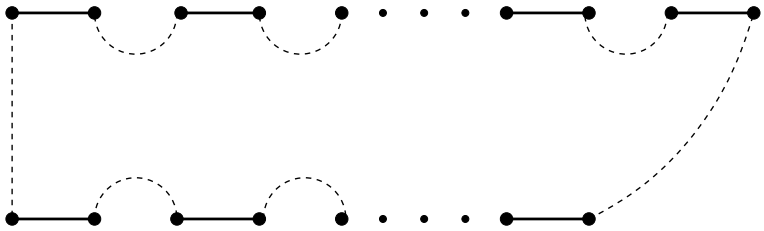


Figure 1: General appearance of a cut cellular surface (CCS)

Definition 2.1 *A cut cellular surface (CCS) is an orientable 2-manifold M with boundary, endowed with a finite cell-structure, such that*

a) Each boundary component of M consists of a single 0-cell and a single 1-cell.

b) M has a specified planar representation, obtained by cutting M along 1-cells in such a way as to obtain a simply connected region in the plane. The cut 1-cells are labeled and given an orientation to make explicit how they are identified in M.

c) The planar representation has the schematic structure shown in Fig. 1: the boundary components, represented by solid lines, lie either along the bottom or the top edge of the planar representation. Those along the bottom edge are called "in" boundary components, those along the top edge are called "out" boundary components. When there are no "in/out" boundary components, the bottom/top edge contains a single 0-cell. The dotted lines on the left and right, and the dotted lines between boundary components along the bottom and top edge, each represent one or more cut 1-cells, separated by 0-cells when there are more than one of them.

d) The simply connected planar region is made up of one or more 2-cells, separated by 1-cells and 0-cells when there are more than one of them.

Remark 2.2 *We will refer to the 0-cells and 1-cells that do not belong to a boundary component as internal or non-boundary 0-cells and 1-cells.*

To fix ideas we give some examples of cut cellular surfaces (Figure 2), representing the sphere S, the disk D (two versions with the boundary being "in" or "out") and the cylinder C. See [3] for further examples and discussion.

Moves on CCS's. By analogy with the Pachner moves on triangulated manifolds, we introduce moves for passing between different planar representations of the same surface. There are two types of move.

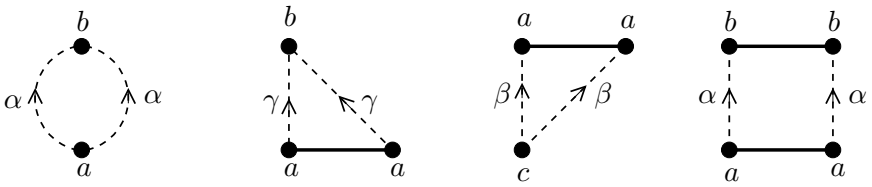


Figure 2: Examples of CCS's

Move I: Introducing a 0-cell into a non-boundary 1-cell, thereby dividing it into two 1-cells, or conversely removing a 0-cell separating two 1-cells, to combine them into a single 1-cell (Figure 3). When this move is applied to a cut 1-cell, the 0-cell is introduced into or removed from both copies of the cut 1-cell in the planar representation.

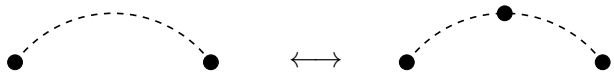


Figure 3: Move I

Move II: Introducing a 1-cell into a 2-cell, thereby dividing it into two 2-cells, or conversely removing a 1-cell separating two 2-cells, to combine them into a single 2-cell (Figure 4). In this figure we have used lines with dots and dashes for the 1-cells bounding the 2-cell to indicate that these are either boundary or internal 1-cells in the planar representation.

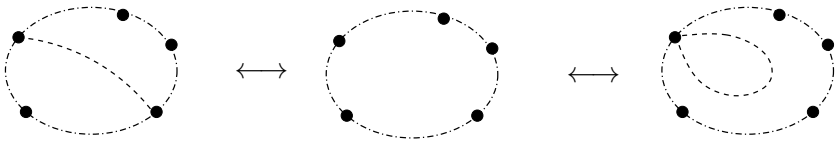


Figure 4: Move II

In [3] it was shown that these moves generate the Pachner moves when M is a triangulated surface without boundary.

3 Invariants for CCS's from finite 2-group colorings

We will be considering colorings of CCS's with finite crossed modules.

Definition 3.1 *A finite crossed module, or finite 2-group,* $\mathcal{G} = (G, H, \partial, \triangleright)$ *is given by:*

- *two finite groups* G *and* H
- *a group homomorphism* $\partial : H \to G$
- *a left action* $\triangleright$ *of* G *on* H *by automorphisms*

such that, for all $h, h_1, h_2 \in H$ *and* $g \in G$*:*

$$\partial(g \triangleright h) = g\,\partial(h)\,g^{-1} \tag{1}$$

$$\partial(h_1) \triangleright h_2 = h_1 h_2 h_1^{-1} \tag{2}$$

Remark 3.2 *An obvious class of examples is given by taking* G *and* H *to be the same, with* ∂ *the identity, and* $\triangleright$ *given by conjugation. For any crossed module* $\ker \partial$ *is contained in the center of* H *and hence is abelian, since for* $h \in \ker \partial$*:* $hfh^{-1} = \partial(h) \triangleright f = 1 \triangleright f = f$*. A further class of examples comes from central extensions. Given a central extension of groups:*

$$1 \to K \to H \xrightarrow{\partial} G \to 1$$

one obtains a crossed module:

$$H \xrightarrow{\partial} G1$$

with lifted action

$$g \triangleright h = fhf^{-1}$$

where $f \in H$ *is any element such that* $\partial(f) = g$*. This action is well-defined because* $K = \ker \partial$ *is central in* H*.*

Fix a finite crossed module $\mathcal{G} = (G, H, \partial, \triangleright)$. Given a CCS, M, we fix orientations on the 1-cells of M, specified as follows with respect to the planar representation:

- the boundary 1-cells are oriented from left to right
- the cut 1-cells are oriented as chosen in Definition 2.1 b)
- the remaining internal 1-cells are oriented arbitrarily.

We also fix a basepoint (0-cell) in the boundary of each 2-cell.

Definition 3.3 *A* $\mathcal{G}$*-coloring of* M *is an assignment of an element* $g_i \in G$ *to each 1-cell labeled* i *and of an element* $h_A \in H$ *to each 2-cell labeled* A*, such that, for each 2-cell in the planar representation, the following condition holds (which we call "fake flatness", in line with terminology from higher gauge theory in physics):*

- *if the 1-cells of the boundary of the 2-cell labeled* A *are labeled* $i_1, \ldots i_k$*, ordered in the anticlockwise direction starting at the basepoint, then*

$$\prod_{j=1}^{k} g_{i_j}^{(-1)} = \partial(h_A) \tag{3}$$

where the factor is g_{i_j} *or* $g_{i_j}^{-1}$*, depending on whether or not the 1-cell* i_j *is oriented compatibly with the positive orientation of the 2-cell.*

See Figure 5 for an example of the fake flatness condition. We have again used dots-and-dashes lines for the 1-cells to indicate that they can be either boundary or non-boundary 1-cells. The basepoint has been shown enlarged in the figure.

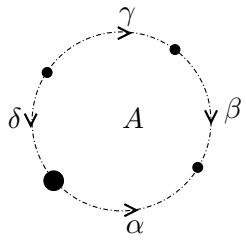


Figure 5: The fake flatness condition here is $g_\alpha g_\beta^{-1} g_\gamma^{-1} g_\delta = \partial(h_A)$.

We can define invariants of CCS's, using $\mathcal{G}$-colorings. Choose elements of G, $g_1, \ldots, g_n$, for the coloring of the "in" boundary components, and $g'_1, \ldots, g'_m$, for the coloring of the "out" boundary components, ordering the boundary components from left to right in the planar representation. Let $|G|$ denote the number of elements of the finite group G, e denote the number of internal edges, i.e. 1-cells, and v denote the number of internal vertices, i.e. 0-cells, of M. Let $\mathrm{Col}(g_1, \ldots, g_n; g'_1, \ldots, g'_m)$ denote the set of all $\mathcal{G}$-colorings of M which have the given assignments on the boundary components.

Definition 3.4 *The following invariants are defined for any choice of boundary colorings:*

$$\langle g'_1, \ldots, g'_m \,|Z_M|\, g_1, \ldots, g_n \rangle = \frac{|H|^{v-e}}{|G|^{\frac{m+n}{2}+v}} \#\mathrm{Col}(g_1, \ldots, g_n; g'_1, \ldots, g'_m). \quad (4)$$

If M has no "in" or "out" components we write the invariants as $\langle \ldots \,|Z_M|\, \emptyset \rangle$ or $\langle \emptyset \,|Z_M|\, \ldots \rangle$.

Remark 3.5 *When M is a triangulated surface without boundary, these are the Yetter invariants [10]. See [5] for an in-depth discussion of Yetter invariants.*

We now discuss in what sense these are invariants. First of all, we have:

Proposition 3.6 *The invariants $\langle g'_1, \ldots, g'_m \,|Z_M|\, g_1, \ldots, g_n \rangle$ are unchanged under changes of orientation of the internal 1-cells.*

Proof. The number of internal vertices and edges is unchanged, and there is a bijection between the respective sets of colorings, given by replacing the element g assigned to any internal 1-cell by g^{-1}, when its orientation is reversed, thus guaranteeing that h can be kept the same to satisfy the fake flatness condition. ■

Likewise the choice of basepoints does not affect the invariant.

Proposition 3.7 *The invariants* $\langle g'_1, \ldots, g'_m \,|Z_M|\, g_1, \ldots, g_n\rangle$ *are unchanged under a change of basepoint in any 2-cell.*

Proof. The number of internal vertices and edges is unchanged, and there is a bijection between the respective sets of colorings, which only differ in the H-coloring of the 2-cell in question. The colorings h and h', corresponding to the first and second choice of basepoint respectively, are related by $h' = g^{-1} \triangleright h$ or equivalently $h = g \triangleright h'$, where g is the ordered multiplication of the group colorings of the edges (taking into account orientation) that link the first basepoint to the second basepoint going round in the anticlockwise direction. This indeed establishes a bijection between the two sets of colorings, since the action of G on H is by automorphisms. The fake flatness condition for the first basepoint may be written as $\partial(h) = gk$, where k represents the ordered multiplication of the group colorings of the edges (taking into acccount orientation) that link the second basepoint to the first basepoint going round in the anticlockwise direction. The fake flatness condition for the second basepoint then follows: $\partial(h') = \partial(g^{-1} \triangleright h) = g^{-1}\partial(h)g = kg$. ■

More importantly we have:

Theorem 3.8 *The invariants* $\langle g'_1, \ldots, g'_m \,|Z_M|\, g_1, \ldots, g_n\rangle$ *are unchanged under moves I and II.*

Proof. Suppose M and M' are related by a move I. Fix a $\mathcal{G}$-coloring for M that assigns g to the 1-cell displayed on the left in Figure 6. Keeping the assignments of all other 1-cells and 2-cells the same, for M' on the right there are $|G|$ compatible $\mathcal{G}$-colorings, since we can choose one assignment, e.g. j, freely in G and the other assignment k is then determined (for the orientations as shown in Figure 6, we have $g = jk$, i.e. $k = j^{-1}g$). Since M' has both an extra internal vertex and an extra internal edge compared to M, the exponent of $|H|$ in (4) is unchanged, and the increase in the exponent of $|G|$ in the denominator is cancelled by the factor $|G|$ relating the respective number of colorings. Thus the invariants (4) are the same for M and M'.

Suppose M and M' are related by a move II, where M' has an extra internal 1-cell compared to M, dividing a 2-cell in M into two 2-cells in M'.

Figure 6: colorings of M and M' for Move I

Using basepoint invariance we may choose the basepoints of the two 2-cells in M' to coincide, and we may choose this same 0-cell as the basepoint of the 2-cell in M. Using invariance under change of orientation, we may choose the extra 1-cell in M' to be oriented so as to have the basepoint as its starting point (see Figure 7, where the basepoint for all 2-cells is the starting point of the 1-cell labeled k_4).

Fix a $\mathcal{G}$-coloring of M that assigns $h \in H$ to the 2-cell we are considering. Keeping the assignments of all other 1- and 2-cells the same, there are $|H|$ corresponding $\mathcal{G}$-colorings of M', since we may choose freely an element $h_2 \in H$ to assign to the 2-cell, say on the left as we follow the subdividing 1-cell in the direction of its orientation, which then determines uniquely the assignment of $h_1 = hh_2^{-1}$ to the other 2-cell, and the assignment of an element $g \in G$ to the subdividing 1-cell, by using the fake flatness condition in either 2-cell. These assignments are compatible with fake flatness, since imposing fake flatness implies $\partial(h_1) = \partial(hh_2^{-1}) = \partial(h)\partial(h_2^{-1})$, which is necessary. Indeed, taking M' on the left in Figure 7 as an example, $\partial(h_1) = k_4k_1g^{-1}$ and $\partial(h)\partial(h_2^{-1}) = k_4k_1k_2k_3\,.\,k_3^{-1}k_2^{-1}g^{-1}$ are the same.

Conversely, given a $\mathcal{G}$-coloring of M' which assigns h_1 and h_2 to the left and right 2-cell respectively, there is a compatible $\mathcal{G}$-coloring of M which assigns $h = h_1h_2$ to the undivided 2-cell and agrees with the $\mathcal{G}$-coloring of M' elsewhere. There are $|H|$ possible $\mathcal{G}$-colorings of M' which give the same $h \in H$, namely $h_1' = h_1h'$ and $h_2' = (h')^{-1}h_2$ for any $h' \in H$.

M' has the same number of internal vertices as M and one extra internal 1-cell. Thus the exponent of $|G|$ in (4) is the same for both M and M', and the increase in the number of colorings for M' by a factor $|H|$ is cancelled by the extra factor $|H|^{-1}$ in (4) coming from the extra internal 1-cell. Thus the invariants are the same for M and M'. ∎

4 Examples

In this section we calculate the invariant for some simple examples. Let K and A denote the kernel and image of ∂, respectively. In Figure 8 below we choose the basepoint to be the bottom 0-cell and on the left, if there is a choice.

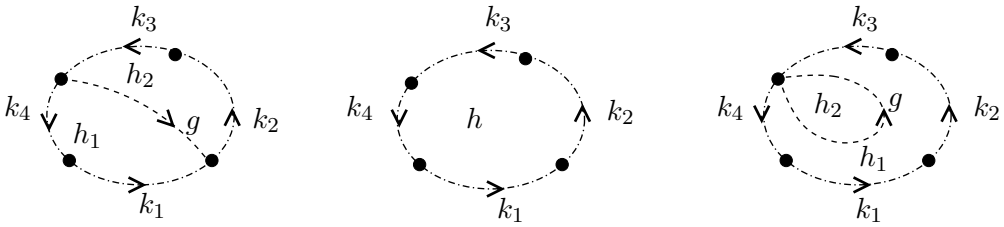


Figure 7: Colorings of M (in the middle) and M' for Move II

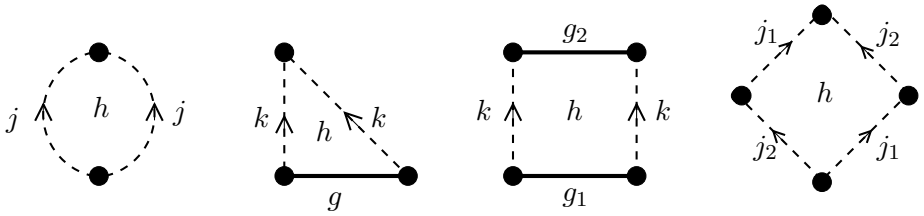


Figure 8: $\mathcal{G}$-colorings for the sphere, disk, cylinder and torus.

Starting with the disk D, the fake flatness condition is $\partial(h) = gkk^{-1} = g$. There are two ways to calculate the overall number of colorings, allowing arbitrary $g \in G$. The first is to fix the 2-cell coloring h. Then g is determined through the condition and the number of colorings is $|H|$. The second is to fix the 1-cell coloring g. If $g \in A$, then one has $|K|$ possible values for h and the number of colorings is $|A|\,|K|$, which is compatible with the previous result, since $|H| = |A|\,|K|$ from group theory. If $g \notin A$, no colorings are possible. The disk has one internal vertex ($v = 1$), one internal edge ($e = 1$), and one boundary component ($m = 0$, $n = 1$). Thus we have:

$$\langle \emptyset | Z_D | g \rangle = \frac{1}{|G|^{1/2}} |K| \mathcal{D}(g) \quad \text{where} \quad \mathcal{D}(g) := \begin{cases} 1, & g \in A \\ 0, & g \notin A \end{cases}$$

For the sphere S, the fake flatness condition is $\partial(h) = jj^{-1} = 1$, i.e. we have $h \in K$, and j is arbitrary in G. Hence the number of colorings is $|K|\,|G|$, which together with $v = 2$, $e = 1$, $m = n = 0$, leads to:

$$\langle \emptyset | Z_S | \emptyset \rangle = \frac{|H|}{|G|^2} |K||G| = \frac{|H||K|}{|G|}$$

For the cylinder C, the fake flatness condition gives $\partial(h) = g_1 k g_2^{-1} k^{-1}$. Note that, if A is the trivial group with one element, this condition expresses that g_1 and g_2 are conjugate to each other, since it is equivalent to: $g_2 = k^{-1} g_1 k$.

We will have more to say about the relation between g_1 and g_2 in section 5. For C we have $v = 0$, $e = 1$, $m = n = 1$, and hence

$$\langle g_2|Z_C|g_1\rangle = \frac{1}{|H||G|}\mathcal{C}(g_1, g_2)$$

where

$$\mathcal{C}(g_1, g_2) = \#\{(h, k) \in H \times G : \partial(h) = g_1 k g_2^{-1} k^{-1}\} \tag{5}$$

Finally, for the torus T, the fake flatness condition is $\partial(h) = j_1 j_2 j_1^{-1} j_2^{-1}$, meaning that h has to lie in the preimage under ∂ of the commutator subgroup of G. Since $v = 1$, $e = 2$, $m = n = 0$, we have:

$$\langle \emptyset|Z_T|\emptyset\rangle = \frac{\#\{(h, g_1, g_2) \in H \times G^2 : \partial(h) = j_1 j_2 j_1^{-1} j_2^{-1}\}}{|G||H|} \tag{6}$$

5 Gluing formula and TQFT

Following our approach in [3], in order to glue two surfaces M_1 and M_2 with matching boundaries, we adopt the following procedure: we identify the shared boundary component furthest to the left (labeled α in Figure 9), and the remaining shared boundary components (just one in the Figure, labeled β) become cut 1-cells in the boundary of the planar representation of a new CCS that we denote by $M_2 \circ M_1$.

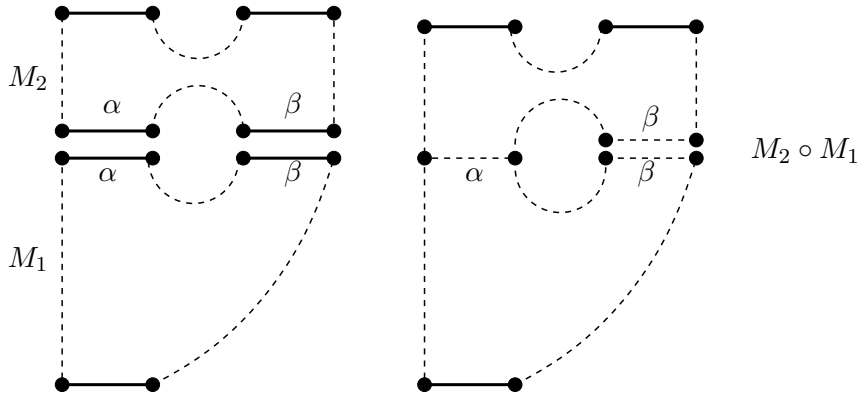


Figure 9: Gluing or composition of two CCS's

Suppose we have M_1 with n incoming boundary components and $m > 0$ outgoing boundary components, and M_2 with m incoming boundary components and p outgoing boundary components. Fixing the colorings of the

"in" boundary components of M_1 and the "out" boundary components of M_2 the colorings of $M_2 \circ M_1$ allow a priori any choice for the colorings of the m intermediate 1-cell components. Thus we arrive at the following property for the invariants.

Proposition 5.1 *(Gluing formula) For any $g_1, \ldots, g_n, i_1, \ldots, i_p \in G$, we have:*

$$\langle i_1, \ldots, i_p \,|Z_{M_2 \circ M_1}|\, g_1, \ldots, g_n \rangle = \sum_{j_1, \ldots, j_m \in G} \langle i_1, \ldots, i_p \,|Z_{M_2}|\, j_1, \ldots, j_m \rangle \, \langle j_1, \ldots, j_m \,|Z_{M_1}|\, g_1, \ldots, g_n \rangle \quad (7)$$

Proof. Since the number of colorings match on both sides of the equation, it remains to check the other factors. Each of the $2m$ boundary components that are glued in M_1 and M_2 gives rise to a factor $\frac{1}{|G|^{1/2}}$ in (4). After gluing $M_2 \circ M_1$ has instead m extra internal vertices, each of which gives a factor $\frac{1}{|G|}$. The factors of $|H|$ are also the same, since for $M_2 \circ M_1$ the m additional internal vertices and the m additional internal edges cancel in the exponent of $|H|$ in (4). ∎

The gluing formula enables us to construct a natural TQFT - see our previous article [3] for an introduction to the notion of TQFT. We assign to each incoming or outgoing boundary of a CCS M, a vector space V_{in} or V_{out} over $\mathbb{R}$, whose basis consists of all G-colorings of the boundary components [3]. The basis elements are written $|\, g_1, \ldots, g_n \,\rangle$ or $\langle\, i_1, \ldots, i_m \,|$, and the dimension of V_{in} and V_{out} is $|G|^n$ and $|G|^m$ respectively. To the CCS itself we assign the linear transformation Z_M from V_{in} to V_{out}, whose matrix elements with respect to these two bases are given by:

$$\langle i_1, \ldots, i_m \,|Z_M|\, g_1, \ldots, g_n \rangle$$

Thus from the gluing formula (7) we have the following fundamental result:

Proposition 5.2 *(TQFT property) For any M_1 and M_2 such that $M_2 \circ M_1$ is defined, we have:*

$$Z_{M_2 \circ M_1} = Z_{M_2} \circ Z_{M_1}. \quad (8)$$

There is an important corollary of (8), which expresses that the cylinder C has assigned to it an idempotent (since $C \circ C$ and C are related by moves I and II, we have $Z_{C \circ C} = Z_C$):

Corollary 5.3 *For the cylinder C, Z_C satisfies*

$$Z_C \circ Z_C = Z_C$$

In terms of the function $\mathcal{C}$ defined in (5), this result is equivalent to

$$\sum_{i\in G} \mathcal{C}\,(g,i)\,\mathcal{C}\,(i,j) = |H||G| \cdot \mathcal{C}\,(g,j)\ , \tag{9}$$

for any $g, j \in G$.

We also give an algebraic proof of (9), which will be useful in what follows. First we define an equivalence relation in G.

Definition 5.4 *We say that two elements g_1 and g_2 of G are 2-conjugate in $\mathcal{G}$, denoted $g_1 \sim g_2$, iff*

$$\mathcal{C}(g_1, g_2) \neq 0\,.$$

Proposition 5.5 *2-conjugacy is an equivalence relation.*

Proof. Let $W(g_1, g_2)$ denote the set $\{(h,k) \in H \times G : \partial(h) = g_1 k g_2^{-1} k^{-1}\}$, which has cardinality $\mathcal{C}(g_1, g_2)$. Then $g_1 \sim g_2$ iff $W(g_1, g_2) \neq \emptyset$.
$\sim$ is reflexive, since $(1_H, 1_G) \in W(g, g)$ for every $g \in G$.
$\sim$ is symmetric: if $(h,k) \in W(g_1, g_2)$, then $(k^{-1} \triangleright h^{-1}, k^{-1}) \in W(g_2, g_1)$, since

$$\begin{aligned} \partial(k^{-1} \triangleright h^{-1}) &= k^{-1}\partial(h^{-1})k \\ &= k^{-1}(k g_2 k^{-1} g_1^{-1})k \\ &= g_2 k^{-1} g_1^{-1} k \end{aligned}$$

$\sim$ is transitive: if $(h,k) \in W(g_1, g_2)$ and $(h', k') \in W(g_2, g_3)$, then we have $(h(k \triangleright h'), kk') \in W(g_1, g_3)$, since

$$\begin{aligned} \partial(h(k \triangleright h')) &= \partial(h)\partial(k \triangleright h') \\ &= \partial(h) k \partial(h') k^{-1} \\ &= g_1 k g_2^{-1} k^{-1} k (g_2 k' g_3^{-1} k'^{-1}) k^{-1} \\ &= g_1 (kk') g_3^{-1} (kk')^{-1} \end{aligned}$$

■

Using the sets $W(g_1, g_2)$ introduced in the previous proof, we can also show the following symmetry.

Proposition 5.6 *For all g_1, $g_2 \in G$, we have $\mathcal{C}(g_1, g_2) = \mathcal{C}(g_2, g_1)$.*

Proof. We establish a bijection $W(g_1, g_2) \underset{\alpha}{\overset{\beta}{\leftrightarrows}} W(g_2, g_1)$ by defining

$$\alpha(h,k) = (k^{-1} \triangleright h^{-1}, k^{-1}) \qquad \beta(h', k') = (k'^{-1} \triangleright h'^{-1}, k'^{-1})$$

From the previous proof, α is well-defined, i.e. $\partial(k^{-1} \triangleright h^{-1})g_2(k^{-1})g_1^{-1}k$, and likewise β is well-defined. α and β constitute a bijection since

$$\begin{aligned}(\beta \circ \alpha)(h,k) &= \beta(k^{-1} \triangleright h^{-1}, k^{-1}) \\ &= (k \triangleright (k^{-1} \triangleright h), k) \\ &= (h,k)\end{aligned}$$

and likewise $(\alpha \circ \beta)(h',k') = (h',k')$. Thus $W(g_1,g_2)$ and $W(g_2,g_1)$ are isomorphic, and hence their cardinality is the same. ■

Using analogous methods we have the following result.

Proposition 5.7 *If $g_1 \sim g_2$, then $\mathcal{C}(g_1,g_2) = \mathcal{C}(g_1,g_1)$.*

Proof. Since $g_1 \sim g_2$, there exists a pair $(h,k) \in H \times G$ such that $\partial(h) = g_1 k g_2^{-1} k^{-1}$. We establish a bijection $W(g_1,g_2) \underset{\alpha}{\overset{\beta}{\leftrightarrows}} W(g_1,g_1)$ by defining

$$\alpha(h',k') = (h'(k'k^{-1}) \triangleright h^{-1}, k'k^{-1}) \qquad \beta(h'',k'') = (h''(k'' \triangleright h), k''k)$$

α is well-defined, since

$$\begin{aligned}\partial(k^{-1} \triangleright h^{-1}) &= k^{-1}\partial(h^{-1})k \\ &= k^{-1}kg_2k^{-1}g_1^{-1}k \\ &= g_2(k^{-1})g_1^{-1}(k^{-1})^{-1}\end{aligned}$$

and likewise β is well-defined. α and β constitute a bijection since

$$\begin{aligned}(\beta \circ \alpha)(h,k) &= \beta(k^{-1} \triangleright h^{-1}, k^{-1}) \\ &= (k \triangleright (k^{-1} \triangleright h), k) \\ &= (h,k)\end{aligned}$$

and likewise $(\alpha \circ \beta)(h',k') = (h',k')$. Thus $W(g_1,g_2)$ and $W(g_1,g_1)$ are isomorphic, and hence their cardinality is the same. ■

Using the previous proposition for the non-zero terms on the l.h.s. of (9), we have $i \sim j$, hence $\mathcal{C}(i,j) = \mathcal{C}(j,j)$. Also $g \sim i \sim j$, and therefore $\mathcal{C}(g,j) = \mathcal{C}(j,j)$. Thus (9) is equivalent to

$$\sum_{i \in G} \mathcal{C}(g,i) = |H||G|$$

which clearly holds since, fixing g, every pair $(h,k) \in H \times G$ belongs to precisely one set of the form $W(g,i)$ with g fixed.

Remark 5.8 *If we denote the 2-conjugacy class of $g \in G$ by $\bar{g}$, we get an equation for the number of elements of $\bar{g}$:*

$$\#\bar{g} = \frac{|G||H|}{\mathcal{C}(g,g)}\,, \tag{10}$$

since $\mathcal{C}(g,g_1) = \mathcal{C}(g,g)$ for all $g_1 \in \bar{g}$, and the number of non-zero terms in the sum on the l.h.s of Eq. (9) is $\#\bar{g}$. Let $2\mathrm{ConjClass}(G)$ denote the set of 2-conjugacy classes of G. Then its cardinality is given by the following equation

$$\#2\mathrm{ConjClass}(G) = \frac{1}{|G|^2|H|^2}\sum_{g,g_1\in G}\mathcal{C}(g,g_1)^2\,. \tag{11}$$

This is clear since the double sum on the r.h.s. decomposes into double sums where g, g_1 both belong to the same 2-conjugacy class $\bar{g}$. Restricting to these terms for a specific 2-conjugacy class $\bar{g}$, the r.h.s. of Eq. (11) becomes:

$$\frac{1}{|G|^2|H|^2}\sum_{g,g_1\in \bar{g}}\mathcal{C}(g,g_1)^2 = \frac{(\#\bar{g})^2\mathcal{C}(g,g)^2}{|G|^2|H|^2} = 1\,, \tag{12}$$

and collecting the contributions from each class, we obtain equation (11).

Our final result is analogous to a proposition obtained in [3]. Consider the invariant for the torus (6), which may be rewritten, using (9), as follows:

$$\begin{aligned}\langle\emptyset|Z_T|\emptyset\rangle &= \frac{\#\{(h,g_1,g_2)\in H\times G^2 : \partial(h) = g_1g_2g_1^{-1}g_2^{-1}\}}{|G||H|}\\ &= \frac{\sum_{g_1\in G}\mathcal{C}(g_1,g_1)}{|G||H|}\\ &= \frac{\sum_{g_1,g_2\in G}\mathcal{C}(g_1,g_2)^2}{|G|^2|H|^2}\,,\end{aligned} \tag{13}$$

This equation reflects the topological fact that the torus is obtained by gluing two cylinders together. See [3] where this point was developed.

For a finite group (not a 2-group) G, its commuting fraction is defined to be the ratio

$$\frac{\#\{(g_1,g_2)\in G^2 : g_1g_2 = g_2g_1\}}{|G|^2}\,,$$

i.e. the ratio of the number of commuting pairs of elements over the number of all pairs of elements. Here we define an analogous fraction for a 2-group $\mathcal{G}$, namely the *generalized commuting fraction* of $\mathcal{G}$

$$\frac{\#\{(h,g_1,g_2)\in H\times G^2 : \partial(h) = g_1g_2g_1^{-1}g_2^{-1}\}}{|H||G|^2}\,. \tag{14}$$

By combining equations (11) and (13) with definition (14), we obtain the following generalization of proposition 6.5 in [3].

Proposition 5.9 *The number of 2-conjugacy classes of G in $\mathcal{G}$ is equal to the generalized commuting fraction of $\mathcal{G}$ times the order of G.*

6 Conclusions and Final Remarks

Viewing our results from the 2-group theory perspective, we have been led to introduce a function $\mathcal{C}$, taking values in the non-negative integers, which depends on two G-elements. The function $\mathcal{C}$ defines an equivalence relation, 2-conjugacy, between G-elements, namely g_1 and g_2 are 2-conjugate iff $\mathcal{C}(g_1, g_2) \neq 0$, but also gives a "measure of the equivalence" between the elements g_1 and g_2 by counting the number of pairs $(h, k) \in H \times G$ such that $\partial(h) = g_1 k g_2^{-1} k^{-1}$. We have derived properties of $\mathcal{C}$ by using topological reasoning, leading us to define the generalized commuting fraction for a 2-group, which we proved to have a property analogous to a property of the standard commuting fraction of a finite group. It should be possible to obtain many further results in the theory of finite 2-groups using a similar topological approach.

Using colorings of cut cellular surfaces with elements of a finite 2-group $\mathcal{G}$, we have found not only invariants for these surfaces, but also a TQFT setting for the invariants. Interestingly these TQFT's do not naturally fit into the standard framework of Atiyah's axioms for TQFT (see [2] and the study by Abrams of 2-dimensional TQFT's [1]), since Z_C for the cylinder is an idempotent, not necessarily the identity. We note that the eigenvalue 1 eigenvectors of Z_C are of the form $g_1 + g_2 + \cdots + g_k$ where the sum runs over all elements of a 2-conjugacy class in G.

As already alluded to, there is an interpretation of these invariants in terms of higher gauge theory based on a finite 2-group $\mathcal{G}$ which we now sketch. First we look at the invariant of our previous article [3] from the point of view of ordinary gauge theory based on a finite group G. In this context we are interested in the moduli space of flat G-connections modulo gauge transformations, or rather the corresponding groupoid whose objects are flat G-connections and morphisms are gauge transformations. Flat G-connections correspond to flat G-colorings of the 1-cells of the surface M, and the gauge transformations are given by assignments of elements of G to the 0-cells of M. For a surface without boundary M, the invariant of [3]

$$\langle \emptyset | Z_M | \emptyset \rangle = \frac{\#\,\text{Flat } G\text{-colorings}}{|G|^v}$$

where v denotes the number of 0-cells, can be understood as the *groupoid cardinality*[1] of the groupoid of flat G-connections on M. In the case of a groupoid coming from the action of a finite group $\widetilde{G}$ on a finite set S, the groupoid cardinality is simply the quotient of the respective cardinalities: $|S|/|\widetilde{G}|$.

In higher gauge theory an analogous picture is emerging, in work by one of us with J. Morton [7, 8]. The higher connections are given by $\mathcal{G}$-colorings of the 1- and 2-cells of M satisfying the fake flatness condition. There are two different types of gauge transformation between these connections, corresponding to assignments of G elements to the 0-cells of M as well as assignments of H elements to the 1-cells of M (taken to be without boundary). In addition there are higher-level transformations between gauge transformations given by assignments of H elements to the 0-cells of M. A satisfying description of all this is in terms of a double groupoid, i.e. a higher algebraic structure having objects, two types of morphisms between objects called horizontal and vertical, and higher morphisms called squares between the morphisms, all morphisms being suitably invertible. The invariant (4), written as follows:

$$\langle \emptyset | Z_M | \emptyset \rangle = \frac{\#\,\text{Fake Flat}\,\mathcal{G}\text{-colorings}\,.\,|H|^v}{|G|^v\,|H|^e}$$

can thus naturally be viewed as the "double groupoid cardinality" of the double groupoid of higher connections. This perspective will be explored in more detail elsewhere.

Acknowledgments

This article is based on a research project carried out by Diogo Bragança under the supervision of Roger Picken. The authors are grateful to the *Fundação Calouste Gulbenkian* for supporting this project through the programme *Novos Talentos em Matemática*, which aims to stimulate undergraduate research in mathematics. Roger Picken is grateful to Dan Christensen and Jeffrey Morton for useful discussions and suggestions. This work was funded in part by the Center for Mathematical Analysis, Geometry and Dynamical Systems (CAMGSD), Instituto Superior Técnico, Universidade de Lisboa, through the projects UID/MAT/04459/2013 and PTDC/MAT-PUR/31089/2017 of the *Fundação para a Ciência e a Tecnologia* (FCT, Portugal).

[1] The groupoid cardinality can be thought of as counting the objects of a groupoid taking into account the number of isomorphisms each object has with other objects. For a nice introduction to the notion of groupoid cardinality, see [9], and for the more general concept of the Euler characteristic of a category, see [6].

References

[1] L. Abrams, "Two-dimensional topological quantum field theories and Frobenius algebras", *J. Knot Theory Ramifications*, Vol. 5, No. 5 (1996), pp. 569-587. https://doi.org/10.1142/S0218216596000333

[2] M. Atiyah, "Topological quantum field theories", *Inst. Hautes Études Sci. Publ. Math.*, Vol. 68 (1988), pp. 175–186. http://www.numdam.org/item?id=PMIHES_1988__68__175_0

[3] D. Bragança and R. Picken, "Invariants and TQFT's for cut cellular surfaces from finite groups", *Bol. Soc. Port. Mat.*, Vol. 74 (2016), pp. 17–44.

[4] R. Dijkgraaf and E. Witten, "Topological Gauge Theories and Group Cohomology", *Comm. Math. Phys.*, Vol. 129, No. 2 (1990), pp. 393–429. https://projecteuclid.org/euclid.cmp/1104180750

[5] J. Faria Martins and T. Porter, "On Yetter's invariant and an extension of the Dijkgraaf-Witten invariant to categorical groups", *Theory Appl. Categ.*, Vol. 18, No. 4 (2007), pp. 118–150. http://www.tac.mta.ca/tac/volumes/18/4/18-04abs.html

[6] T. Leinster, "The Euler characteristic of a category", *Doc. Math.*, Vol. 13 (2008), pp. 21–49. https://www.math.uni-bielefeld.de/documenta/vol-13/02.pdf

[7] J. C. Morton and R. Picken, "Transformation double categories associated to 2-group actions", *Theory Appl. Categ.*, Vol. 30, Paper No. 43 (2015), pp. 1429–1468. http://www.tac.mta.ca/tac/volumes/30/43/30-43abs.html

[8] J. C. Morton and R. Picken, "2-Group Actions on Moduli Spaces of Higher Gauge Theory" (in preparation).

[9] T. Tao, "Counting objects up to isomorphism: groupoid cardinality", What's New (2017) https://terrytao.wordpress.com/2017/04/13/counting-objects-up-to-isomorphism-groupoid-cardinality/

[10] D. N. Yetter, "TQFT's from Homotopy 2-Types", *J. Knot Theory Ramifications*, Vol. 2, No. 1 (1993), pp. 113–123. https://doi.org/10.1142/S0218216593000076

História da Matemática

Editor:
Luís Saraiva

Professor Irineu Bicudo (1940-2018)
In Memoriam

Luís Saraiva
CMAF – Universidade de Lisboa

Faleceu, a 20 de Julho de 2018, o Professor Irineu Bicudo, com 78 anos de idade, antigo docente do Departamento de Matemática do Instituto de Geociências e Ciências Exatas (IGCE), Universidade Estadual Paulista (UNESP), Campus de Rio Claro – SP.

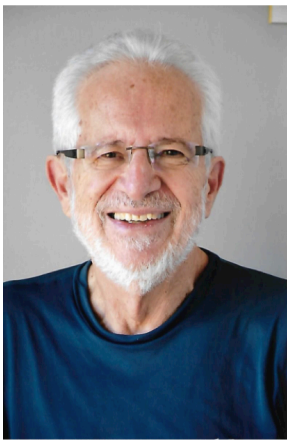


Foto cedida gentilmente por Elisabete Cristina Plombon

O Professor Irineu terminou o seu curso de Matemática na Universidade de São Paulo (USP), em 1963, e doutorou-se em 1973, na Pontifícia Universidade Católica de São Paulo (PUC-SP), com uma tese orientada pelo Professor Mário Tourasse Teixeira. Recebeu o título de Livre-Docente pela Universidade Estadual Paulista, em 1979. De 1989 a 1993 foi Director do IGCE-UNESP.

Começou por fazer pesquisa em Álgebra Universal e Fundamentos da Matemática, e, mais tarde, centrou a sua investigação em História e Filosofia da Matemática e em Filosofia Antiga.

Nos últimos anos tivemos a infelicidade de constatar o falecimento de alguns dos pesquisadores do universo luso-brasileiro em História das Ciências Matemáticas e em Historia do ensino das Ciências Matemáticas. Só para referir os que eu conheci, e de quem era amigo, lembro com saudade o Comandante Estácio dos Reis, entre os portugueses, e Edilson Pacheco e Plínio Tábuas, entre os brasileiros, os três pesquisadores com textos incluídos em *Actas* publicadas dos Encontros Luso Brasileiros de História da Matemática.

O Professor Irineu tinha uma qualidade que, nos dias de hoje, de uma especialização crescente, é cada vez menos frequente: era uma pessoa culta, alguém que via a cultura como um todo, com diferenças, mas sem as barreiras das nacionalidades, via-a na sua complexidade e nos seus múltiplos aspectos, sendo a Matemática e a sua História apenas um deles.

Tive a felicidade de assistir diversas vezes a comunicações suas, e era uma evidência a quem as ouvia que o que tínhamos presenciado era resultado de uma longa prática de reflexão profunda, de toda uma questionação do real, da nossa história humana nas suas infinitas facetas, trabalho de um pensamento crítico que nunca deixou de ser actuante e independente.

Cruzei-me com o Professor Irineu e sua esposa com frequência, quer no Brasil, em Rio Claro, onde residiam, quer nos congressos em que ambos participávamos, quer em Portugal, nas suas habituais peregrinações à Europa. Encontrei sempre a mesma pessoa, simples, directa, amiga, com um espírito jovem e entusiasta a relatar as suas descobertas, que enfrentava com humor e paciência a adversidade que às vezes aparecia.

Conversámos muitas vezes. O Professor Irineu tinha sempre latente uma preocupação de não deixar passar ao lado informação cultural que poderia ser importante para ele. Em Portugal procurava actualizar-se sobre as últimas manifestações e produções culturais do nosso país. Em cada visita acabava invariavelmente por comentar que tinha de parar a compra de livros, não só por já ter as malas cheias deles, mas também por ter a consciência que tinha adquirido muitos, mas mesmo muitos livros. Era sempre obrigatória nas suas vindas a Portugal a visita à sua longa lista de alfarrabistas e livrarias.

Destaco da sua produção a importante tradução directamente do grego dos Elementos de Euclides, um trabalho que lhe levou cerca de uma dezena de anos a realizar, não só por ser intrinsecamente uma tarefa de grande complexidade, mas igualmente por ser um perfeccionista para quem o que era importante era o resultado final e não o tempo gasto na sua realização.

A propósito de Euclides, costumava dizer que, se com Homero a língua grega tinha alcançado a *perfeição*, só com Euclides tinha chegado à *precisão*.

Quando o livro saiu, em 2009, da editora da Unesp, ciente da sua importância, tentei, de acordo com o Professor Irineu, que o livro fosse publicado por uma editora portuguesa. Isso não foi possível. Uma das respostas dadas foi que não se podia publicar em Portugal um livro em português do Brasil. Seria muito importante que alguma das editoras que então nos deu essa resposta revisse a sua decisão e tomasse a iniciativa de propôr a sua publicação em Portugal.

O prefácio deste livro conclui com uma citação de Platão, que, para mim, é não só uma declaração de intenções, que reconheço ter sido uma norma de conduta que sempre o guiou, mas também uma evidência da sua humildade e honestidade intelectual:

Pois tendo aprendido algo, jamais neguei, fazendo o conhecimento ser como que uma descoberta minha, mas louvo como sábio o que me instruiu, tornando públicas as coisas que aprendi com ele.

Perdemos a presença física do Professor Irineu Bicudo, mas ele continuará de algum modo presente em todos os que tiveram a felicidade de o conhecer e privar com ele, e viram a sua vida enriquecida com este contacto.

Obrigado Irineu.
Um abraço do Luís.

Um retrato das mulheres matemáticas em Portugal

Sofia B.S.D. Castro
Faculdade de Economia & Centro de Matemática
Universidade do Porto
e-mail: sdcastro@fep.up.pt

Margarida Mendes Lopes
CAMGSD & Departamento de Matemática
Instituto Superior Técnico, Universidade de Lisboa
e-mail: mmlopes@math.tecnico.ulisboa.pt

Resumo: Neste artigo procuramos descrever a situação relativa a género na comunidade científica matemática em Portugal (dados de 2016).

Abstract: In this article we describe the gender balance in the portuguese mathematical scientific community (with 2016 data).

palavras-chave: dados relativos a matemáticas portuguesas, igualdade de género.

keywords: data concerning Portuguese women mathematicians; gender equality.

1 Introdução

Em anos recentes tem sido prestada especial atenção à distribuição relativa entre homens e mulheres ativos em investigação e docência universitária em várias ciências. Este assunto tem tomado a atenção de várias entidades europeias, havendo mesmo sociedades científicas com departamentos e comissões dedicados a este tema. Especificamente em relação à Matemática, veja-se a título de exemplo as páginas da London Mathematical Society [1, 2], da Unione Matematica Italiana [3] assim como a da European Mathematical Society [4]. Recentemente, foi criada pelo International Council for Science uma página dedicada a assuntos de igualdade de género em ciência em geral e incluindo em particular a Matemática [5].

Olhando para estatísticas internacionais (por exemplo [6]) constata-se que em Portugal a percentagem de mulheres em atividades de investigação e desenvolvimento é superior a muitos outros países desenvolvidos e que no

"global gender report" estão num ranking muito mais alto. Constata-se o mesmo fenómeno no caso particular da Matemática.

Talvez por isso, em Portugal, o assunto de género em Matemática tem sido geralmente ignorado, tendo sido recentemente reavivado numa das sessões temáticas do Encontro Nacional da SPM 2016 (http://enspm16.spm.pt/pt/tematicas), intitulada "Situação das Mulheres Matemáticas (e não só) em Portugal" e organizada pela Catarina Lucas e pela Luísa Castro Guedes.

Neste texto procuramos descrever a situação relativa a género na comunidade científica matemática em Portugal. Restringimos a nossa atenção ao Ensino Superior Universitário público, focamo-nos em atividade científica, e olhamos para indicadores como a presença das mulheres nos vários estádios da carreira docente, em comissões científicas de encontros científicos, em comissões organizadoras e nas sessões plenárias dos mesmos encontros, em orgãos de direção de organizações relevantes e corpos editoriais de revistas.

Este texto é escrito com recurso limitado a dados e a informação que nos foi possível recolher sem dificuldade não é completa ou exaustiva. Não é nosso propósito fechar a discussão da situação das mulheres matemáticas em Portugal. Muito pelo contrário, esperamos com este texto suscitar interesse suficiente para que outros, mais bem posicionados ou mais capazes, recolham mais informação e produzam uma melhor caracterização da realidade. Se a realidade aqui reportada parecer aquém do desejável, esperamos que seja possível modificá-la com medidas que se mostrem eficazes num prazo não muito longo.

2 Contextualização na Europa

De acordo com Hobbs e Koomen [7], a evolução da presença de mulheres matemáticas na Europa entre 1993 e 2005 foi muito positiva. Neste contexto, Portugal destaca-se de todos os restantes apresentando não só a maior percentagem de mulheres matemáticas em 1993 e 2005 mas também a maior percentagem de mulheres matemáticas no topo da carreira em 2005.

A Tabela 1 apresenta os números para Portugal que se encontra em primeiro lugar na Europa para os anos de 1993 e 2005 (informação retirada de [7]) e de 2016 (recolhida por nós e usando a correspondência Professor=Professor Catedrático, Senior lecturer=Professor Associado e Lecturer=Professor Auxiliar). De salientar que em 2005 o segundo lugar na Tabela 2 de [7] cabe à Estónia com $35,2\%$ de mulheres matemáticas, $10,5\%$ de Professoras Catedráticas, $35,3\%$ de Professoras Associadas e $38,1\%$ de

Professoras Auxiliares, números muito abaixo dos registados em Portugal. O segundo país europeu em termos de percentagem de mulheres no topo da carreira em 2005 é a Itália com $15,1\%$ de Professoras Catedráticas. Apenas Espanha e França apresentam percentagens de Professoras Catedráticas acima dos 10%. O distanciamento de Portugal relativamente à Europa diminui nas categorias inferiores da carreira universitária registando-se em Itália $40,3\%$ de Professoras Associadas e $50,4\%$ de Professoras Auxiliares, número igual ao de Portugal em 2005.

Ano	**% Mulheres Matemáticas**	**% Prof. Catedráticas**	**% Prof. Associadas**	**% Prof. Auxiliares**
2016	47,2	32,7	36,1	51,7
2005	47,6	32,1	45,9	50,4
1993	40-45	5		

Tabela 1: Percentagem de mulheres matemáticas em Portugal nos anos de 1993, 2005 e 2016. Os dados relativos a 1993 estão incompletos face aos restantes. Os dados relativos a 1993 e 2005 são os constantes em Hobbs e Koomen [7], os de 2016 foram recolhidos por nós usando a informação disponibilizada na internet pelas seguintes instituições: Universidade do Porto (Faculdades de Ciências, de Economia e de Engenharia), Universidade de Lisboa (Instituto Superior de Economia e Gestão, Instituto Superior Técnico, Faculdade de Ciências) e Faculdades de Ciências e Tecnologia das Universidades de Aveiro, da Beira Interior, de Coimbra, do Minho, Nova de Lisboa e de Trás-os-Montes e Alto Douro.

Sendo notável a posição de Portugal relativamente à Europa tanto em 1993 como em 2005, notamos de 2005 a 2016 uma estagnação nos números. Esta estagnação contraria a tendência anterior particularmente no que diz respeito à presença de mulheres matemáticas no topo da carreira.

Não encontrámos dados europeus relativos a outros indicadores como, por exemplo, presença de mulheres em corpos editoriais de revistas científicas e orgãos de direção de organizações científicas. No entanto, Topaz e Sen [8] estudam a presença de mulheres em corpos editoriais de revistas científicas de matemática em todo o mundo e, num universo de 13067 posições editoriais, observam que $8,9\%$ destas é atribuída a uma mulher. Num texto direcionado aos matemáticos americanos, Martin [9] estuda a presença de mulheres matemáticas em conferências e observa que no International Congress of Mathematicians 2014 houve 20 comunicações plenárias,

tendo uma delas (5%) sido proferida por uma mulher e de um total de 237 oradores plenários e convidados, 35 foram mulheres (14, 8%). Ambos os valores percentuais estão, de acordo com este autor, abaixo da percentagem de doutoramentos atribuídos a mulheres em Matemática nos EUA. Na Europa, recolhemos dados relativos a European Congress of Mathematics (ECM) associados à European Mathematical Society (EMS) que reproduzimos na Tabela 2.

Oradores	**Plenários**			**Convidados**		
ANO	Homens	Mulheres	% M	Homens	Mulheres	% M
2016	7	3	30,0	25	6	19,4
2012	9	1	10,0	30	3	9,1
2008	8	2	20,0	30	5	14,3
2004	6	1	14,3	26	4	13,3
2000	8	1	11,1			

Tabela 2: Homens e mulheres oradores plenários e convidados dos ECM de 2000 a 2016. Não encontrámos dados relativos a oradores convidados em 2000.

3 O caso português

Nesta secção descrevemos a situação das mulheres matemáticas em Portugal de acordo com os dados que nos foi possível facilmente recolher.

Como se pode ver na Tabela 3, o número de doutoramentos produzidos anualmente em Portugal foi, pelo menos até 2009, equilibrado em termos de género. Isto mostra que a posição das mulheres matemáticas em Portugal é muito diferente da que se observa noutros países onde o debate ainda se centra em como atrair mulheres para estudos superiores em Matemática.

O que acontece em Portugal depois do doutoramento pode ser em parte percebido na Tabela 4 onde apresentamos os dados relativos às instituições de ensino superior que de forma agregada constam na Tabela 1. Os dados eram os disponíveis na página de cada instituição em Julho de 2016. Este não é um levantamento exaustivo de todas as instituições portuguesas porque não conseguimos facilmente encontrar este tipo de dados em todas elas. Julgamos no entanto que esta amostragem reflete a realidade.

ano	**Homens**	**Mulheres**	**% Mulheres**
2009	26	31	54.4
2008	32	19	37.3
2007	26	28	51.9
2006	35	36	50.7
2005	15	32	68.1
2004	16	30	65.2
2003	22	20	47.6
2002	15	15	50.0
2001	11	11	50.0
2000	17	18	51.4
1999	4	10	71.4
1998	7	8	53.3
1997	7	6	46.2
1996	7	5	41.7
total	240	269	52.8

Tabela 3: Número de doutoramentos em Matemática realizados em Portugal por género. Fonte: GPEARI — MCTES 14/12/2011 em dados.gov.pt

Observa-se, com 3 exceções[1], um decréscimo na percentagem de mulheres ao passarmos da categoria de Professor Auxiliar para a de Professor Associado. Já na transição entre Professor Associado e Catedrático há um aumento em 6 instituições e uma diminuição muito superior nas outras 6 instituições. Aliás, como se pode ver na Tabela 1, há sempre uma diminuição da percentagem de mulheres matemáticas na transição para categorias superiores.

Entre Julho de 2016 e Junho de 2017, recolhemos alguns indicadores distintos dos da carreira para Portugal. Apresentamo-los nas Tabelas 5–11.

A Tabela 5 mostra uma total ausência das mulheres matemáticas da presidência do Centro Internacional de Matemática (CIM) desde 2000, bem como a completa ausência de mulheres matemáticas nas direções em exercício durante os anos de 1996-2004 e 2011-15, um período total de 6 anos nos 19 anos para os quais temos dados.

Na Tabela 6 apresentamos os dados relativos ao Conselho Científico do CIM de 1996 a 2014. Na contagem distinguimos elementos em instituições

[1] As exceções são a Universidade de Aveiro, a Universidade da Beira Interior e a Faculdade de Economia da Universidade do Porto.

Instituição	Posição na carreira	Homens	Mulheres	% Mulheres
UNIVERSIDADE DE AVEIRO				
	Prof. Auxiliar	19	22	53,7
	Prof. Associado	2	3	60,0
	Prof. Catedrático	6	0	0,0
UNIV. DA BEIRA INTERIOR				
	Prof. Auxiliar	22	12	35,3
	Prof. Associado	1	1	50,0
	Prof. Catedrático	0	1	100,0
UNIVERSIDADE DE COIMBRA				
	Prof. Auxiliar	23	19	45,2
	Prof. Associado	8	5	38,5
	Prof. Catedrático	4	6	60,0
UNIVERSIDADE DE LISBOA				
ISEG	Prof. Auxiliar	9	11	55,0
	Prof. Associado	4	0	0,0
	Prof. Catedrático	2	4	66,7
IST	Prof. Auxiliar	39	23	37,1
	Prof. Associado	17	5	22,7
	Prof. Catedrático	10	3	23,1
Fac.Ciências	Prof. Auxiliar	9	15	62,5
	Prof. Associado	7	4	36,4
	Prof. Catedrático	5	2	28,6
UNIVERSIDADE DO MINHO				
	Prof. Auxiliar	16	29	64,4
	Prof. Associado	7	6	46,2
	Prof. Catedrático	1	1	50,0
UNIV. NOVA DE LISBOA				
	Prof. Auxiliar	33	34	50,7
	Prof. Associado	7	2	22,2
	Prof. Catedrático	2	1	33,3
UNIVERSIDADE DO PORTO				
Fac. Ciências	Prof. Auxiliar	12	14	53,8
	Prof. Associado	6	4	40,0
	Prof. Catedrático	5	0	0,0
Fac. Economia	Prof. Auxiliar	10	7	41,2
	Prof. Associado	1	2	66,7
	Prof. Catedrático	0	1	100,0
Fac. Engenharia	Prof. Auxiliar	6	11	64,7
	Prof. Associado	1	1	50,0
	Prof. Catedrático	2	1	33,3
UNIV. DE TRÁS-OS-MONTES E ALTO DOURO				
	Prof. Auxiliar	9	22	71,0
	Prof. Associado	1	2	66,7
	Prof. Catedrático	0	0	0,0

Tabela 4: Percentagens de mulheres matemáticas em 2016. Os dados foram recolhidos por nós usando a informação disponibilizada na internet de cada instituição.

em Portugal de elementos portugueses mas de instituições estrangeiras que tipicamente fazem parte do Conselho Científico.

Direção	**Presidente**		**Tesoureiro**		**Outros membros**	
Mandato	Homens	Mulheres	Homens	Mulheres	Homens	Mulheres
2016-19	1	0	0	1	3	0
2011-15	1	0	1	0	4	0
2008-11	1	0	0	1	2	1
2004-08	1	0	1	0	1	2
2000-04	1	0	1	0	3	0
1996-2000	1	0	1	0	2	0

Tabela 5: Mulheres e homens na Direção do CIM desde 1996. Observamos a completa ausência de mulheres nos mandatos 2011-15 e nos dois mandatos mais antigos.

Conselho Científico	**Homens (todos)**	**Mulheres (todas)**	**% M (todas)**	**Homens (nac.)**	**Mulheres (nac.)**	**% M (nac.)**
2011-14	9	2	18,2	5	1	16,7
2009-11	10	2	16,7	6	1	14,3
2005-08	13	2	13,3	8	1	11,1
2000-04	16	1	5,9	10	1	9,1
1996-2000	13	3	18,8	7	3	30,0

Tabela 6: Mulheres e homens no Conselho Científico do CIM de 1996 a 2014. Consideramos na contagem separadamente os elementos pertencentes a instituições em Portugal (últimas três colunas). Excetuando o Conselho Científico mais antigo, verifica-se a presença de apenas uma mulher pertencente a uma instituição nacional em cada mandato.

Na Tabela 7 apresentamos as comissões científicas e organizadoras dos Encontros Nacionais da Sociedade Portuguesa de Matemática (SPM) onde se verifica que a presença de mulheres matemáticas é exígua na comissão científica e, por contraste, abundante na comissão organizadora. Sendo muito poucas as sessões plenárias científicas, notamos ainda assim que em nenhum dos encontros houve mais mulheres matemáticas do que homens a fazer uma apresentação plenária, havendo mesmo uma total ausência de mulheres matemáticas neste tipo de sessão em dois dos quatro anos para os quais temos dados.

	Comissão Científica		Comissão Organizadora		Plenárias científicas	
Ano	Homens	Mulheres	Homens	Mulheres	Homens	Mulheres
2016	5	1	0	8	2	1
2014	6	0	2	6	3	0
2012	6	1	2	6	2	1
2010	5	0	2	5	3	0

Tabela 7: Mulheres e homens na organização dos Encontros Nacionais da SPM de 2010 a 2016. Em 2012, a única mulher da Comissão Científica estava explicitamente dedicada a questões de ensino.

No que concerne mulheres matemáticas (baseadas em Portugal) no corpo editorial da Portugaliae Mathematica, a sua presença tem vindo a diminuir ao longo dos anos (ver Tabelas 8 e 9), havendo mesmo uma completa ausência de 2008 a 2017. Esta ausência é muito flagrante nos editores executivos[2] (que têm sido sempre baseados em Portugal).

	Editores executivos		Editores associados	
Ano	Homens	Mulheres	Homens	Mulheres
2017	5	0	2	0
2016	5	0	2	0
2015	5	0	4	0
2014	5	0	4	0
2013	5	0	4	0
2012	5	0	4	0
2011	4	0	3	0
2010	4	0	3	0
2009	4	0	3	0
2008	4	0	3	0

Tabela 8: Mulheres e homens no corpo editorial da Portugaliae Mathematica desde que esta é publicada pela EMS e até 2017. Consideramos apenas os editores baseados em Portugal.

A SPM contou com uma mulher matemática como sua presidente entre 2000 e 2004, tendo havido uma mulher vice-presidente em 2010, conforme consta na Tabela 10. Em anos mais recentes, a tendência parece seguir a apresentada na Tabelas 8 e 9 relativas à Portugaliae Mathematica, havendo uma ausência de mulheres matemáticas na presidência e uma diminuição de mulheres matemáticas vogais. A tesouraria parece ser mais frequentemente atribuída a uma mulher matemática.

[2]Notamos durante a revisão deste artigo (Julho de 2018) o aparecimento de duas mulheres no cargo de Editor executivo.

	Editores executivos		Editores associados	
Ano	Homens	Mulheres	Homens	Mulheres
2007	6	0	6	1
2006	6	0	7	1
2005	6	0	7	1
2004	6	0	8	1
2003	5	0	8	1
2002	6	0	9	2
2001	6	0	13	4
2000	6	0	13	4
1999	5	0	14	4
1998	5	0	13	4
1997	6	0	13	4
1996	6	0	13	4

Tabela 9: Mulheres e homens no corpo editorial da Portugaliae Mathematica antes de ser publicada pela EMS. Consideramos apenas os editores baseados em Portugal.

Direção	**Presidente**		**Vice-Presidente**		**Tesoureiro**		**Vogal**	
Mandato	H	M	H	M	H	M	H	M
2016	1	0	2	0	0	1	3 (1)	4 (2)
2014	1	0	2	0	0	1	3 (2)	3 (1)
2012	1	0	2	0	0	1	2	3 (2)
2010	1	0	1	1	0	1	2	3
2008	1	0	2	0	1	0	1	4 (2)
2006	1	0	2	0	1	0	1	4 (2)
2004	1	0	1	0	0	1 (0)	0	2
2002	0	1	1	0	0	1	1	1
2000	0	1	1	0	0	1	1	0
1998	1	0	1	0	0	1	1	1

Tabela 10: Mulheres e homens na Direção da SPM de 1998 a 2016. Neste período, apenas nos mandatos que se iniciaram em 2000, 2002 e 2010 houve uma mulher na presidência/vice-presidência. Anteriormente só no mandato que se iniciou em 1994 houve uma mulher na presidência. Nas três últimas colunas entre parêntesis indicamos o número quando a contagem se limita ao ensino superior universitário.

A Tabela 11 mostra os números e percentagens relativos a concursos da Fundação para a Ciência e a Tecnologia (FCT). As percentagens observadas neste contexto são as que melhor aproximam o retrato da Tabela 1, havendo

mesmo um enviesamento favorável às mulheres matemáticas na atribuição de financiamento em 2012 e 2013. É relevante referir que, de acordo com dados muito recentes da Direção-Geral de Estatísticas da Educação e Ciência (DGEEC), havia no final de 2016 uma percentagem muito comparável de homens e mulheres integrados em unidades de I&D da FCT: 77% dos docentes homens e 75% dos docentes mulheres pertenciam a uma unidade de I&D da FCT. Em termos de docente equivalente a tempo integral (ETI), havia na mesma data 458 homens e 484 mulheres ETI.

Ano	**Financiados**			**Avaliados**		
	H	M	% M	H	M	% M
2014	8	2	20,0	62	37	37,4
2013	1	8	88,9	27	23	46,0
2012 (E)	1	1	50,0	—	—	—
2012	1	4	80,0	42	32	43,2
2010	11	4	26,7	48	33	40,7
2009	12	4	25,0	38	22	36,7
2008	16	10	38,5	62	41	39,8
2006	15	5	25,0	40	18	31,0

Tabela 11: Mulheres e homens como Investigador Principal (IP) de projetos avaliados e financiados pela FCT. Usamos a designação (E) em 2012 para distinguir os projetos de excelência lançados nesse ano.

4 Conclusões

Notamos que, excetuando os números da Tabela 11, os valores percentuais que encontrámos na realidade portuguesa (e que, reforçamos, podem estar enviesados pela recolha de informação que conseguimos fazer) relativamente a indicadores para além da mera existência estão bastante abaixo dos valores percentuais correspondentes à existência de mulheres matemáticas em Portugal. Não podemos deixar de referir que a participação em comissões científicas de encontros internacionais, a apresentação em sessões plenárias e a presença em corpos editoriais de revistas científicas são frequentemente parâmetros relevantes em concursos da carreira universitária.

A Tabela 3 deixa clara a diferença entre a situação das mulheres matemáticas em Portugal e na Europa: as mulheres em Portugal não evitam formação superior em Matemática, ou pelo menos, não o fazem mais do que

os homens. O que é patente da comparação entre os dados da Tabela 3 com as tabelas subsequentes é que esta formação superior em Matemática não se reflete proporcionalmente na progressão na carreira universitária nem nos outros indicadores, com a exceção da percentagem de projetos financiados pela FCT.

Não é propósito deste texto procurar as razões por detrás deste retrato; deixamos esta tarefa ao futuro e eventualmente a outros. Pensamos que a discussão que possamos suscitar com este artigo, seja qual for o seu resultado, apenas pode beneficiar a Matemática enquanto ciência e os matemáticos portugueses, homens e mulheres.

Agradecimentos

Gostaríamos de agradecer à Catarina Lucas (coordenadora para Portugal na organização European Women in Mathematics) e à Luísa Castro Guedes que organizaram a sessão temática do encontro da SPM no Barreiro, bem como aos participantes (homens e mulheres) dessa sessão. Agradecemos também à Isabel Labouriau e à Maria Manuel Pinho por nos terem feito chegar vários tipos de dados sobre mulheres e matemática.

A primeira autora é membro integrado do CMUP (UID/MAT/00144/2013), financiado pela FCT (Portugal) com fundos estruturais nacionais (MEC) e europeus (FEDER). A segunda autora é membro integrado do CAMGSD financiado pela FCT (Portugal) (UID/MAT/04459/2013).

Referências

[1] https://www.lms.ac.uk/about/committees/women-mathematics-committee

[2] https://www.lms.ac.uk/womeninmaths

[3] http://umi.dm.unibo.it/en/working-group-for-equal-opportunities

[4] https://womenandmath.wordpress.com/

[5] https://icsugendergapinscience.org/

[6] http://reports.weforum.org/global-gender-gap-report-2015/

[7] C. Hobbs and E. Koomen, *Statistics on Women in Mathematics* (2006) https://womenandmath.wordpress.com/past-activities/statistics-on-women-in-mathematics/

[8] C.M. Topaz and S. Sen, *Gender Representation on Journal Editorial Boards in the Mathematical Sciences*, PLOS One (2016) http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165367

[9] G. Martin, *Addressing the underrepresentation of women in mathemathics conferences*, arXiv:1502.06326v1 (2015).

Problemas

Editor:
Jorge Nuno Silva

Notas sobre o Problema anterior e Problemas para meninas

Jorge Nuno Silva

Os leitores são convidados a enviar, para eventual publicação, soluções, comentários, propostas de problemas, etc. Essa correspondência deve ser enviada para a SPM, ao cuidado do editor desta secção. Há livros para sortear entre as soluções recebidas em cada número.

Relembremos o problema do número anterior.

Não julgue um número pelo aspecto!

Os programas doutorais contêm, em algumas universidades, exames de acesso, constituídos por problemas que, ao longo dos anos, vão criando belas colecções, que os estudantes usam para se prepararem. Isto sucede com UC Berkeley, de cujos *Preliminary examination (2016)*, retiramos a nossa proposta de hoje. Este era o primeiro de nove problemas propostos, de que os candidatos deviam resolver seis em três horas.

Mostre que o número

$$\int_4^9 \sqrt{-6+5\sqrt{-6+5\sqrt{-6+5\sqrt{-6+5\sqrt{x}}}}}\,dx$$

é racional.

Seja a função $f:[4,9]\to[2,3]$ definida por

$$f(x)=\sqrt{-6+5\sqrt{-6+5\sqrt{-6+5\sqrt{-6+5\sqrt{x}}}}}$$

É fácil ver que f admite inversa dada por

$$f^{-1}(y)=\left(\left(\left(\left(\left(\left(\left(y^2+6\right)\frac{1}{5}\right)^2+6\right)\frac{1}{5}\right)^2+6\right)\frac{1}{5}\right)^2+6\right)\frac{1}{5}\Bigg)^2$$

que é um polinómio com coeficientes racionais. O integral dado é a área limitada pelo gráfico de f e as rectas verticais $x=4, x=9$ e o eixo das abcissas. A união desta região com a área limitada pelo mesmo gráfico, as rectas horizontais $y=2, y=3$ e o eixo das ordenadas é a diferença de dois rectângulos—um limitado pelas rectas $x=9, y=3$ e os eixos coordenados, o outro limitado pelas rectas $x=4, y=2$ e os eixos. Portanto, temos

$$\int_4^9 f(x)dx+\int_2^3 f^{-1}(y)dy=9\cdot 3-4\cdot 2=19$$

O segundo integral, sendo a função integranda um polinómio com coeficientes racionais, é um número racional, pelo que o primeiro também tem de ser racional.

O Professor Armando Machado, nosso leitor assíduo, foi sorteado, entre os que enviaram soluções para este problema, e receberá um livro de oferta. Parabéns ao feliz contemplado!

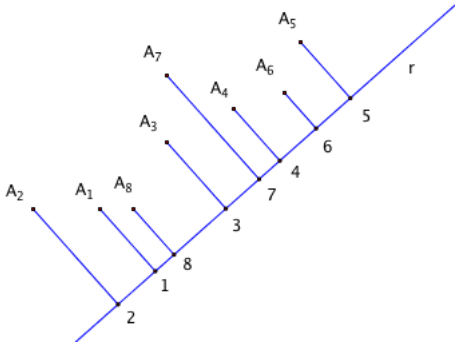
Problemas para meninas

Por esse mundo fora estão implementadas muitas competições matemáticas, algumas com longas tradições. As Olimpíadas—IMO—são as mais conhecidas, sendo que é a SPM a instituição promotora da sua versão portuguesa.

Hoje viajamos até à China, onde esta competição tem versões exclusivamente para *meninas*, se bem que também haja uma versão nacional global. Escolhemos duas questões para meninas (anos 2002 e 2003).

1. Seja n um inteiro positivo e D_n o conjunto dos divisores positivos de n, incluindo 1 e n. Prove que no máximo metade dos elementos de D_n tem 3 como algarismo das unidades.

2. Sejam $A_1, \ldots, A_8$ oito pontos quaisquer do plano. Seja $\overrightarrow{r}$ uma recta (orientada) qualquer. Consideremos os pés das perpendiculares tiradas pelos pontos $A_1, \ldots, A_8$ sobre $\overrightarrow{r}$.



Atendendo à orientação da recta, estes novos pontos definem uma permutação dos índices $1, \ldots, 8$. Por exemplo, na figura a permutação é

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 1 & 8 & 3 & 7 & 4 & 6 & 5 \end{pmatrix}$$

Imaginemos que este processo se repete para todas as rectas do plano que não geram projecções ortogonais coincidentes para pontos diferentes. No máximo, quantas permutações diferentes de $\{1, 2, 3, 4, 5, 6, 7, 8\}$ se podem obter?

Informações aos autores

O Boletim da SPM publica-se, em geral, duas vezes por ano.

O Boletim constitui um espaço diversificado de informação, promove a circulação de ideias e de opiniões, bem como troca de experiências entre os que ensinam, investigam ou aplicam a Matemática.

O Boletim não publica artigos de investigação especializada. Estes trabalhos poderão ser submetidos à Portugaliæ Mathematica, revista de prestígio internacional, também propriedade da SPM e editada pela European Mathematical Society.

Os artigos dedicados a assuntos de natureza pré-universitária deverão, de preferência, ser submetidos à Gazeta de Matemática.

As actividades da SPM, nomeadamente das suas Delegações Regionais e Secções, são noticiadas com regularidade no Boletim.

As opiniões expressas pelos autores dos artigos publicados no Boletim não representam necessariamente posições da SPM.

Os Editores das Secções são os únicos responsáveis pela aceitação de artigos nas Secções que dirigem. A Secção Opinião — Cartas ao Director é da exclusiva responsabilidade da Directora do Boletim. Todos os outros trabalhos serão enviados pelos editores a *Referees* especializados para aconselharem sobre a respectiva publicação (com eventuais alterações).

Os manuscritos devem ser submetidos em `http://revistas.rcaap.pt/BoletimSPM` ou enviados por correio electrónico para um dos editores. Agradece-se o envio da versão PDF do texto. Os autores devem indicar as respectivas instituições, bem como os seus endereços de correio electrónico. Os trabalhos submetidos devem incluir um sumário em português e em inglês, e uma lista de palavras-chave nestas duas línguas. Recomenda-se vivamente que os trabalhos sejam preparados em LaTeX. A bibliografia deve seguir o padrão habitual no LaTeX.

Endereço para correspondência:
Boletim da Sociedade Portuguesa de Matemática
Av. da República, 45–3º E, 1050–187 Lisboa

ISSN 0872–3672

SUMÁRIO

Artigos

Secção de História

Secção de Problemas