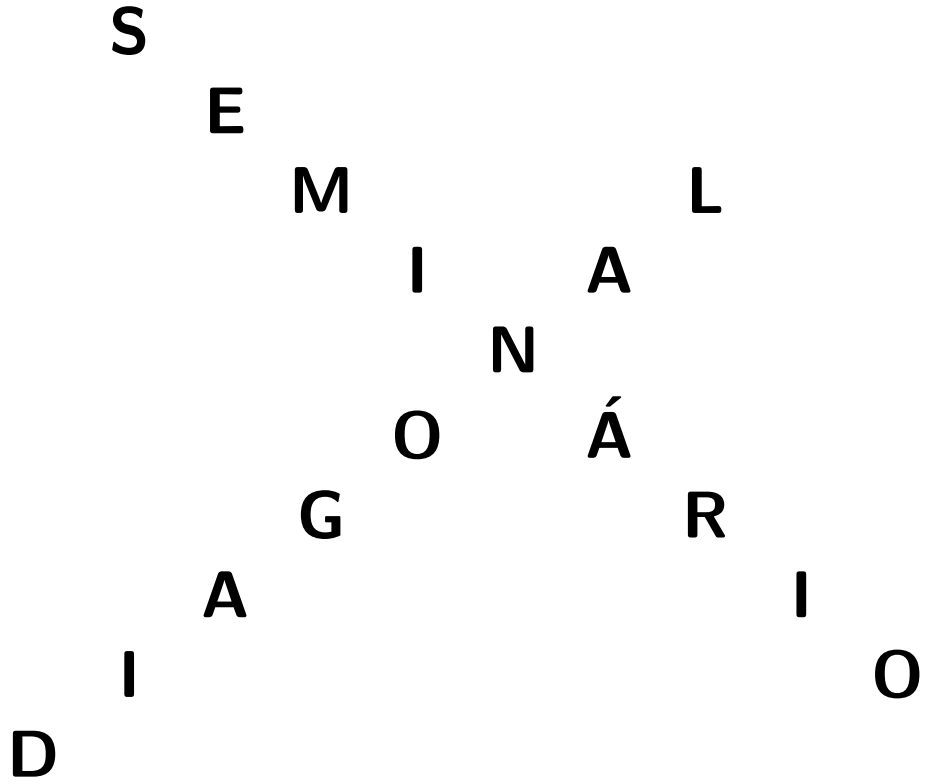


Proceedings IST 2000-01

Novembro 2001

Coordenação Editorial:

João Pedro Boavida
Ana Cannas da Silva
Luís Cruz-Filipe
José Luís Fachada
Pedro Manuel Resende

Título ◇ Seminário Diagonal — Proceedings IST 2000–01 ◇ Novembro 2001

Editores ◇ João Pedro Boavida[1,2] ◇ Ana Cannas da Silva[2,3] ◇ Luís Cruz-Filipe[1,4] ◇ José Luís Fachada[2] ◇ Pedro Manuel Resende[2]

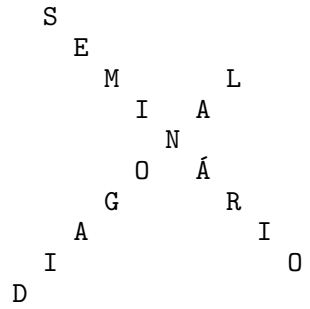
(1) alunos da Licenciatura em Matemática Aplicada do Instituto Superior Técnico em 2000–01 ◇ (2) Departamento de Matemática, IST ◇ (3) presentemente no Institute for Advanced Study, Princeton ◇ (4) Department of Informatics, University of Nijmegen

www ◇ `http://www.math.ist.utl.pt/diagonal/`

Documento LaTeX executado em 18 de Dezembro de 2001, no Departamento de Matemática do Instituto Superior Técnico.

Prefácio

Faz hoje precisamente um ano, por volta das 16 horas, a seguinte mensagem anunciava ao mundo (ou antes, a vários pontos no país) o arranque próximo de um novo seminário no Instituto Superior Técnico da Universidade Técnica de Lisboa:




```
-----------------------------------

O que é?
Novo seminário de estudantes.

Para quem?
Todos os interessados em Matemática.

Sobre quê?
Matemática, no sentido lato.

-----------------------------------

Estreia brevemente em todo o país.
```

Esta mensagem foi seguida, poucas horas mais tarde, pelo anúncio do primeiro seminário, que seria concretizado uma semana depois.[1] Passado um

1 O Seminário Diagonal é uma iniciativa de âmbito nacional, e actualmente realiza-se também na Faculdade de Ciências da Universidade do Porto, na Faculdade de Ciências e Tecnologia da Universidade de Coimbra e na Faculdade de Ciências da Universidade de Lisboa.

ano, parece seguro dizer que as promessas foram cumpridas, e a personalidade do Seminário Diagonal IST está hoje mais clara.

O Seminário tinha sido criado para ser um espaço onde os alunos da Licenciatura em Matemática Aplicada e Computação (LMAC) do IST pudessem trocar ideias matemáticas. O objectivo era mostrar novos assuntos — sobretudo os menos conhecidos — de modo acessível aos alunos dos primeiros anos, ou seja, *diagonalizados*.[2]

O seminário teve a adesão imediata de vários alunos da LMAC e da Licenciatura em Engenharia Física Tecnológica, bem como a presença bem visível de várias pessoas do Departamento de Matemática. Mais tarde contámos também outros alunos na audiência.

De início, os oradores foram escolhidos entre os alunos mais velhos da LMAC (os que estariam menos desconfortáveis no papel de 'pioneiros'). Mais tarde tentámos descer a idade média. Entre os assuntos apresentados contaram-se temas de Trabalhos Finais de Curso, discussões de artigos, ou resultados de investigações no âmbito do Programa Gulbenkian Novos Talentos em Matemática.

Conteúdo

Assim, no final de 2000–01, e em face do sucesso do Seminário, pareceu-nos natural coordenar, editar e promover a publicação de Artigos Diagonais, escritos pelos oradores, a propósito dos seus seminários. O objectivo foi por um lado registar os resultados da experiência, mas por outro trazê-los à consideração de uma audiência mais vasta.

Este é o resultado desse esforço. Estão incluídos oito artigos, aproximadamente correspondentes a oito dos seminários. Infelizmente nem todos os oradores puderam participar, seja por indisponibilidade de tempo, seja simplesmente porque os seus seminários foram baseados em artigos e bastará indicar a referência. Juntamos ainda uma lista dos resumos de todos os seminários realizados. Como todos os oradores/autores são (ou eram) alunos da LMAC, indica-se apenas o ano curricular (em 2000–01) e a especialização.

Agradecimentos

Várias pessoas permitiram que o Seminário crescesse e tivesse o sucesso que teve. O nosso primeiro agradecimento vai assim para os oradores e todos os

2 Tendo em conta que 'semi-simples' é sinónimo de 'diagonalizável', isto significa que os assuntos devem ser apresentados como 'soma directa' de 'partes simples'.

que com eles colaboraram. Vai também para todos os ouvintes, alunos da LMAC, da LEFT, mesmo de outros cursos, e professores do Departamento de Matemática.

Uma pessoa que indirectamente contribuiu para a organização do Seminário foi o João Palhoto Matos. Sem ele provavelmente não existiria a página dos seminários (`http://data.math.ist.utl.pt/sem/diagonal`), nem a produção deste ficheiro no formato PDF teria sido tão rápida.

A Fundação Calouste Gulbenkian, através do Programa Gulbenkian Novos Talentos em Matemática, patrocinou dois dos oradores em 2000–01, cujos artigos estão incluídos nesta colectânea.

Por fim, um agradecimento especial deve ser dirigido às novas co-organizadoras, Ana Luísa Silva e Ana Sofia Vaz, que mantiveram o Seminário a funcionar quase sem apoio, enquanto ultimávamos os pormenores desta edição.

A Organização Diagonal IST 2000–01,

17 de Outubro de 2001,
Lisboa, Nijmegen, Oeiras e Princeton

Conteúdo

Resumos dos Seminários

Habilidades com Somatórios

Luís Cruz-Filipe (5º ano da LMAC — Ciência da Computação)
24 de Outubro

Alguns quebra-cabeças relativamente simples criam por vezes a necessidade de calcular somas pouco atraentes. Nesta apresentação introduzem-se técnicas elegantes que permitem resolver alguns somatórios sem esforço recorrendo, nomeadamente, à introdução de uma notação diferente da habitual. No final, dar-se-ão pistas no sentido de resolver somatórios por métodos análogos aos utilizados na Teoria da Integração.

[1] Luís Cruz-Filipe. Habilidades com somatórios. Neste volume.

Análise Real(mente) Infinitesimal

João Boavida (5º ano da LMAC — Análise, Geometria e Álgebra)
7 de Novembro de 2000

Entre as hipóteses inconscientes na prática matemática habitual, conta-se a possibilidade de provar/refutar o que é verdadeiro/falso num número finito de passos. Assim, o conjunto de fórmulas $\{\varepsilon < 1, \varepsilon < \frac{1}{2}, \varepsilon < \frac{1}{3}, ...\}$ a respeito de um real $\varepsilon > 0$ não pode ser refutado, pelo que deveria existir algum número ε satisfazendo--as, que seria realmente infinitesimal.

Esta observação aparentemente inocente será o nosso ponto de partida para explorar a Matemática Não-Standard.

[1] João Pedro Boavida. Análise real(mente) infinitesimal. Neste volume.

Potências, Países e Padrões Escondidos

Vitor Saraiva (4º ano da LMAC — Análise, Geometria e Álgebra)
28 de Novembro de 2000

Analisando o primeiro dígito da sequência das potências de dois — 1, 2, 4, 8, 1, 3, 6, 1, 2, 5, 1, 2, ... — verifica-se que o algarismo 1 ocorre seis vezes mais frequentemente que o algarismo 9. A análise deste comportamento será o ponto de partida para uma justificação de um padrão escondido nas sequências dos valores numéricos das populações e das áreas dos países do mundo.

[1] Arnold. Repartitioning the world. *Quantum*, Jan/Feb 2000, pp. 34–37.

Códigos Detectores de Erros

Pedro Baptista (5º ano da LMAC — Ciência da Computação)
5 de Dezembro de 2000

Imagine o que acontece se um caixa de banco se engana a escrever o seu número de conta da próxima vez que depositar um cheque. Ou que a telefonista da editora não percebeu correctamente a referência do livro que lhe encomendou. Ou ainda que num supermercado lhe cobram uma lata de caviar por um pacote de manteiga. No mínimo, é desagradável! Resolver estes problemas do quotidiano será o mote para explorar a matemática dos sistemas de identificação.

[1] Jorge Picado. A álgebra dos sistemas de identificação. *Boletim da Sociedade Portuguesa de Matemática*, No. 44, Abril 2001, pp. 39–73.
[2] Jorge Buescu. *O Mistério do Bilhete de Identidade e Outras Histórias*, capítulo 1. Gradiva, 2001.

Modelos, Simulação Numérica e Realidade

Alexandra Moura (5º ano da LMAC — Análise Numérica)
19 de Dezembro de 2000

Quando analisamos fenómenos da vida real, somos muitas vezes conduzidos a modelos matemáticos complexos. Estes envolvem normalmente equações diferenciais nem sempre fáceis de resolver (quando tal é possível!).

Neste seminário vai apresentar-se o método dos Elementos Finitos a fim de obter soluções numéricas para estes problemas, aplicando-o ao estudo da circulação sanguínea e de problemas de aerodinâmica.

Computação Quântica

Alexandre Francisco (4º ano da LMAC — Ciência da Computação)
20 de Março de 2001

Agora vivemos numa nova era, a era quântica! Mas de que forma este novo entendimento do universo influencia o processamento, a aquisição e a transmissão de informação? Agora ouvimos falar em computadores quânticos, mas o que são afinal estas máquinas maravilhosas? Que novo paradigma é este, o da computação quântica? Teremos nós oportunidade de ultrapassar o poder de Turing? Serão agora tratáveis muitos dos problemas cuja resolução apenas sonhámos um dia?

É objectivo deste nosso seminário discutir estas e outras questões, bem como entender melhor o que de novo nos reserva o futuro, talvez mais próximo do que possamos imaginar!

[1] Alexandre P. Lourenço Francisco. Computação quântica. Neste volume.

Nós e Alguns Invariantes

Tiago Requeijo (4º ano da LMAC — Análise, Geometria e Álgebra)
3 de Abril de 2001

Uma questão que provavelmente nunca nos surgiu diz respeito a distinguir dois nós (por exemplo, determinar se um dado nó é cego). Embora possa parecer um assunto de pouca relevância, é crucial saber se um nó se pode desfazer puxando as extremidades. Imagine-se, por exemplo, se os nós da corda de um alpinista se desfizessem ao esticar a corda...

Em matemática um nó não passa de uma corda em que juntamos os extremos. Neste seminário vamos ver alguns invariantes que permitem distingui-los.

Até Onde Podemos Ir?

Luís Cruz-Filipe (5º ano da LMAC — Ciência da Computação)
24 de Abril de 2001

Ao longo do século XX os computadores passaram de inexistentes a indispensáveis, e hoje grande parte das tarefas do dia-a-dia é por eles realizada. Mas haverá um limite para o que podem fazer? Neste seminário veremos alguns resultados clássicos da Teoria da Computação, analisando algumas das suas consequências práticas. No final discutiremos o que os novos paradigmas de computação nos trazem de novo face aos tradicionais, e o que (não) devemos esperar deles.

CRIPTO QUÊ?

Pedro Adão (4º ano da LMAC — Ciência da Computação)
8 de Maio de 2001

Quando queríamos guardar alguma coisa usávamos os cofres; quando queríamos que uma carta chegasse ao destino sem ser aberta, usávamos lacre; quando queríamos garantir que um destinatário recebia uma carta, enviávamo-la com aviso de recepção.

Hoje em dia, no mundo em que vivemos, será possível ter segurança? Podemos ter um cofre na Internet para guardar dinheiro virtual? Podemos assinar documentos virtuais sem que ninguém falsifique a nossa assinatura? Podemos enviar e-mails lacrados? Podemos enviar e-mails com aviso de recepção? Estas e outras questões serão abordadas neste seminário.

[1] Pedro Miguel Adão. Criptologia; contratos e dinheiro virtuais. Neste volume.

CRIPTOGRAFIA!

Tiago Reis (2º ano da LMAC)
15 de Maio de 2001

Será possível que duas pessoas lancem uma moeda ao ar ao telefone? Poderá isto ser feito sem que a pessoa que escolhe cara ou coroa, no caso de perder, não duvide nem um pouco da honestidade do lançamento? Neste seminário veremos qual a solução para este problema e até que ponto é fiável.

Na sequência do seminário anterior, vamos ainda ver o que é um algoritmo de encriptação de chave pública, isto é, um algoritmo em que tanto a chave como o próprio algoritmo são públicos. Por fim, veremos o que é e como funciona o algoritmo RSA, tão amplamente difundido.

[1] Tiago Reis. Criptografia e jogos por telefone. Neste volume.

O TEOREMA DE PITÁGORAS

Luís Russo (3º ano da LMAC — Ciência da Computação)
22 de Maio de 2001

Sabia que Pitágoras não foi o primeiro a descobrir o Teorema de Pitágoras? Sabia que são conhecidas cerca de 380 demonstrações independentes deste resultado que têm fascinado gerações pela sua simplicidade? A abordagem destas questões, bem como algumas curiosidades históricas com elas relacionadas, constitui o tema deste seminário.

[1] Luís Russo. O teorema de Pitágoras. Neste volume.

Espaços Sem Pontos

Pedro Baptista (5º ano da LMAC — Ciência da Computação)
29 de Maio de 2001

A propósito de uma das definições clássicas de número real, devida a Dedekind, veremos uma forma pouco habitual de definir a recta real com base em determinadas propriedades algébricas dos subconjuntos abertos de $\mathbb{R}$. Este exemplo servirá de mote para falar de topologia e ilustrar as ideias básicas da chamada topologia sem pontos, bem como aplicações a áreas da matemática onde por vezes é útil raciocinar construtivamente.

Um Passeio Pouco Aleatório

João Boavida (5º ano da LMAC — Análise, Geometria e Álgebra)
5 de Junho de 2001

Normalmente não nos apercebemos como é frequente que fenómenos que em pequena escala são totalmente deterministas se revelem verdadeiramente aleatórios na escala 'de todos os dias'. Basta pensar na trajectória de um grão de poeira, ou na imagem de um raio nos céus.

Neste seminário vamos descrever o movimento browniano e usá-lo como modelo de ruído em equações diferenciais, o que, como veremos, nos trará algumas surpresas. No final, um passeio curto por Monte-Carlo para explorar algumas propriedades das funções harmónicas.

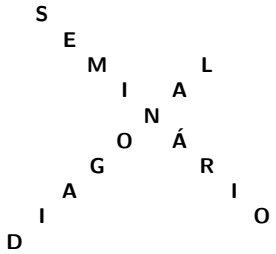
[1] João Pedro Boavida. Um passeio pouco aleatório. Neste volume.

Grupos, Variedades e Relatividade

Patrícia Engrácia (3º ano da LMAC — Análise, Geometria e Álgebra)
12 de Junho de 2001

Os grupos estão muito relacionados com a geometria: há grupos que são espaços geométricos muito ricos e há estruturas geométricas a que podemos associar grupos. Também na física as perspectivas de observadores distintos se relacionam por acção de elementos de grupos. Neste seminário vamos olhar para alguns exemplos e brincar um pouco com a relatividade de Einstein.

[1] Patrícia Engrácia. Grupos, variedades e relatividade. Neste volume.



Habilidades com Somatórios

Luís Cruz-Filipe
5º ano da LMAC — Ciência da Computação
lcf@math.ist.utl.pt

Palavras Chave

somatório, característica, notação de Iverson.

Resumo

Alguns quebra-cabeças relativamente simples criam por vezes a necessidade de calcular somas pouco atraentes. Neste apontamento introduzem-se técnicas elegantes que permitem resolver alguns somatórios sem esforço recorrendo, nomeadamente, à introdução de uma notação diferente da habitual.

1 Introdução

É bem conhecida a lenda daquele soberano da Índia que, enfadado e sem nada com que se entreter, ordenou aos seus conselheiros que inventassem algo para ele se divertir. Dias depois, um desses sábios apareceu com um novo jogo: o xadrez. O soberano, maravilhado com a grandiosidade dessa invenção, disse ao inventor que pedisse a recompensa que quisesse, ao que o sábio respondeu que se contentaria com

> um grão de trigo pela primeira casa do tabuleiro, dois grãos pela segunda, quatro pela terceira, e assim por diante até à sexagésima-quarta casa.

Reza a tradição que o soberano manifestou o seu desagrado por o sábio pedir uma tão singela recompensa, pensando que um saco de trigo chegaria para a satisfazer, e ordenou que esta fosse providenciada; mas quando se fizeram as contas à quantidade de trigo necessária concluiu-se que esta seria suficiente para cobrir toda a Terra com uma camada de trigo com um metro de altura.

A quantidade de trigo pedida pelo sábio (em grãos) pode ser escrita simplesmente como

$$\sum_{k=1}^{64} 2^{k-1}, \tag{1}$$

mas qual o valor exacto desta soma? Todo o aluno dos últimos anos do liceu sabe (ou devia saber!) que se trata de somar os primeiros 64 termos de uma progressão geométrica de razão 2, pelo que o valor daquela expressão é

$$2^{64} - 1.$$

Outra lenda da comunidade matemática passa-se no século XIX, quando o pequeno Gauss estudava na escola primária. Um dia, o professor resolveu entreter os seus alunos durante algum tempo mandando-os somar os cem primeiros números naturais; ele esperava decerto ter algum tempo de sossego enquanto eles se entretinham, mas Gauss chegou rapidamente à resposta. Como? É fácil: ele agrupou os termos pedidos aos pares e concluiu que obtinha 50 pares que somavam, cada um deles, 101; multiplicando estes dois números obteve o resultado pretendido.

$$\begin{array}{ccccc} 1 & + & 100 & \to & 101 \\ 2 & + & 99 & \to & 101 \\ \vdots & \vdots & \vdots & & \vdots \\ 50 & + & 51 & \to & 101 \end{array}$$

Este método é, aliás, o método normalmente utilizado no Secundário para deduzir o valor da soma de n termos consecutivos de uma progressão aritmética. No caso de Gauss, o que ele tinha de calcular era

$$\sum_{k=1}^{100} k. \tag{2}$$

Podemos generalizar ambas as equações (1) e (2), somando um número de termos arbitrário, e não será muito difícil convencermo-nos de que se tem

$$\sum_{k=0}^{n} k = \frac{n(n+1)}{2}; \tag{3}$$

$$\sum_{k=0}^{n} 2^k = 2^{k+1} - 1. \tag{4}$$

Cada uma destas expressões corresponde à soma dos primeiros $(n+1)$ termos duma sucessão; no primeiro caso, esta sucessão é definida simplesmente por

$$u_n = n;$$

no segundo caso, trata-se de

$$v_n = 2^n.$$

Olhando para a complexidade aparente da segunda sucessão o leitor desprevenido será talvez tentado a pensar que calcular somas de n termos de uma sucessão conhecendo o seu termo geral não é afinal assim tão difícil. Mas desengane-se: calcular uma coisa tão simples como

$$\square_n = \sum_{k=1}^{n} k^2 \tag{5}$$

já não é trivial. Na sequência vamos descrever métodos para calcular somatórios bastante mais complicados que nos permitirão, em particular, determinar o valor de $\square_n$ e perceber a importância duma boa notação. Estes métodos permitirão alargar muito o nosso espectro de somas calculáveis, mas, como seria de esperar, não são universais.

2 Notação de Iverson

Antes de prosseguirmos no cálculo de somatórios, vamos introduzir notação. Iverson, em [2], observou que quando se trabalha com funções que não estão sempre definidas é muitas vezes penoso indicar explicitamente as condições em que o estão, e utilizou no seu livro uma convenção para evitar fazê-lo. Esta notação passou quase despercebida, até Knuth a ter apresentado em [1] frisando as suas vantagens.

A ideia é muito simples. Seja P um predicado;[1] então $[P(x)]$ denota a função característica de P no ponto x, ou seja,

$$[P(x)] = \begin{cases} 1 & \text{se } P(x) \\ 0 & \text{c.c.} \end{cases}$$

Alguns casos particulares desta notação são muito utilizados; por exemplo, se A for um conjunto então $[x \in A]$ é simplesmente a função característica de A; se i e j forem números naturais então $[i = j]$ é simplesmente δ_{ij}, o tão utilizado delta de Kronecker. A vantagem da notação de Iverson (na versão melhorada de Knuth, aqui apresentada) é permitir representar estes conceitos de forma uniforme.

Note-se que o valor 0 em $[P(x)]$ é um zero computacionalmente forte, no sentido em que se $[P(x)] = 0$ então $[P(x)]f(x) = 0$, independentemente do valor de $f(x)$ (que pode ser infinito ou não estar sequer definido). A aplicação imediata é podermos escrever, por exemplo,

$$\sum_{i=1}^{n} f(i) = \sum_{i} [1 \leq i \leq n] f(i),$$

1 Ou seja, uma função total que devolve verdadeiro ou falso.

sendo o somatório à direita uma série em que só um número finito de parcelas não são nulas.

Qual o interesse desta notação? Como toda a notação, não permite fazer nada que não conseguíssemos fazer antes; porém, muitas coisas passam a poder ser feitas de uma forma muito mais simples. Ilustraremos isto com um exemplo: seja $f = f(i,j)$ uma função definida em $\mathbb{N}^2$; vamos provar que

$$\sum_{i=1}^{n}\sum_{j=i}^{n} f(i,j) = \sum_{j=1}^{n}\sum_{i=1}^{j} f(i,j).$$

Este resultado é conhecido, mas a sua prova exige normalmente alguma atenção às possíveis maneiras de escolher pares de naturais nos limites dos somatórios considerados. Com a notação de Iverson temos simplesmente:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=i}^{n} f(i,j) &= \sum_{i,j}[1 \le i \le n][i \le j \le n] f(i,j) \\
&= \sum_{i,j}[1 \le i \le j \le n] f(i,j) \\
&= \sum_{i,j}[1 \le i \le j][1 \le j \le n] f(i,j) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{j} f(i,j)
\end{aligned}$$

e tudo o que se utilizou foi manipulação de desigualdades e expressões equivalentes.

Recorrendo somente a esta notação é já espantosa a quantidade de somas que se conseguem calcular. Apresentamos em seguida dois exemplos.

Exemplo 1. Calcular $\sum_{k=1}^{n} k \times 2^k$.

Comecemos por observar que $k \times 2^k = \sum_{j=1}^{k} 2^k$; com isso em mente, podemos calcular:

$$\begin{aligned}
\sum_{k=1}^{n} k \times 2^k &= \sum_{j,k}[1 \le k \le n][1 \le j \le k] 2^k \\
&= \sum_{j,k}[1 \le j \le k \le n] 2^k \\
&= \sum_{j,k}[1 \le j \le n][j \le k \le n] 2^k
\end{aligned}$$

$$
\begin{aligned}
&= \sum_j [1 \leq j \leq n] \left(2^{n+1} - 2^j\right) \\
&= n \times 2^{n+1} - \left(2^{n+1} - 2\right) \\
&= (n-1)2^{n+1} + 2,
\end{aligned}
$$

onde tudo o que utilizámos foi o resultado (4) e a linearidade dos somatórios.

3 Métodos Formais

O exemplo anterior ilustra uma aplicação *naïve* dum método de cálculo de somatórios semelhante ao uso do Teorema de Fubini para cálculo de integrais múltiplos: dado um somatório em duas variáveis, trocar a ordem pela qual as somas são feitas, tentando obter uma expressão mais simples de somar. O recurso à notação de Iverson, no entanto, torna o método tão transparente que nos dispensamos de o apresentar mais formalmente, incluindo apenas um exemplo de uma situação concreta em que nos permite calcular uma soma com bastante mau aspecto.

Exemplo 2. Suponhamos que dispomos de um algoritmo que calcula o valor de $f(k)$ num tempo que é o valor aproximado por defeito de $\sqrt{k}$. Quanto tempo demorará esse algoritmo a tabelar os valores de f entre 1 e n?

Este problema é o exemplo típico dos problemas encontrados em Teoria da Complexidade. Neste contexto, tal como em Matemática Discreta em geral, é muito utilizada a notação $\lfloor x \rfloor$ para denotar "o valor aproximado por defeito de x". Um pouco de reflexão permite concluir que se tem a relação $\lfloor k \rfloor = \sum_j [1 \leq j \leq k]$. Assim, temos:

$$
\begin{aligned}
\sum_{1 \leq k \leq n} \lfloor \sqrt{k} \rfloor &= \sum_k [1 \leq k \leq n] \lfloor \sqrt{k} \rfloor \\
&= \sum_{j,k} [1 \leq k \leq n][1 \leq j \leq \sqrt{k}] \\
&= \sum_{j,k} [1 \leq j \leq \sqrt{k} \leq \sqrt{n}] \\
&= \sum_{j,k} [1 \leq j \leq \sqrt{n}][j^2 \leq k \leq n] \\
&= \sum_j [1 \leq j \leq \sqrt{n}](n - j^2 + 1) \\
&= n\lfloor \sqrt{n} \rfloor - \square_{\sqrt{n}} + \lfloor \sqrt{n} \rfloor \\
&= (n+1)\lfloor \sqrt{n} \rfloor - \square_{\sqrt{n}}.
\end{aligned}
$$

(relembre-se a definição de $\square_n$ em (5)).

Coloca-se naturalmente a questão seguinte: determinámos o valor de $\sum_{1\leq k\leq n}\lfloor\sqrt{k}\rfloor$ em função do de $\square_{\sqrt{n}}$; será que só recorrendo a esta notação conseguimos determinar o valor de $\square_n$?

A resposta é sim, mas de uma forma muito trabalhosa. Deixa-se aqui o desafio ao leitor mais interessado;[2] calcularemos mais à frente o valor de $\square_n$ por outro método.

Repare-se que nos dois exemplos apresentados foi necessário reescrever um somatório em k por forma a obter um somatório em duas variáveis. Esta ideia de reescrita está também subjacente a um outro método, conhecido como o **método da perturbação**, que descrevemos de seguida.

A ideia por trás deste método é a seguinte: tentar obter duas expressões diferentes que tenham o mesmo valor, "perturbando" levemente a soma a calcular. Um exemplo ilustra o funcionamento deste método:

Exemplo 3. Suponhamos que pretendemos calcular o valor de $\square_n$. Vamos perturbar levemente a sua definição e tentar escrever $\sum_{k=1}^{n+1} k$ de duas formas diferentes.

Obviamente, tem-se a relação

$$\sum_{k=1}^{n+1} k^2 = (n+1)^2 + \sum_{k=1}^{n} k^2$$

ou, equivalentemente,

$$\sum_{k=1}^{n+1} k^2 = (n+1)^2 + \square_n. \tag{6}$$

Por outro lado, uma mudança de variável permite escrever

$$\begin{aligned}\sum_{k=1}^{n+1} k^2 &= \sum_{k=0}^{n}(k+1)^2\\ &= \sum_{k=0}^{n}(k^2+2k+1)\\ &= \sum_{k=1}^{n} k^2 + 2\sum_{k=1}^{n} k + \sum_{k=0}^{n} 1,\end{aligned}$$

2 A técnica é análoga à utilizada nos dois exemplos apresentados; a dada altura é possível determinar uma equação de primeiro grau em $\square_n$, donde se retira facilmente o valor pretendido.

o que nos permite concluir directamente

$$\sum_{k=1}^{n+1} k^2 = \square_n + 2\sum_{k=1}^{n} k + (n+1). \tag{7}$$

Chegados a este ponto, parece que não fizemos muitos progressos. Se juntarmos as equações (6) e (7), os termos em $\square_n$ cancelam-se... *mas conseguimos calcular o valor de* $\sum_{k=1}^{n} k$ *a partir dessa mesma equação!!!*

Sugere-se naturalmente um caminho a tentar. Se aplicando o método da perturbação a $\square_n$ conseguimos uma fórmula para $\sum_{k=1}^{n} k$, que tal aplicar o método da perturbação a $\sum_{k=1}^{n} k^3$?

Temos por um lado a relação óbvia

$$\sum_{k=1}^{n+1} k^3 = \sum_{k=1}^{n} k^3 + (n+1)^3$$

e por outro

$$\begin{aligned}
\sum_{k=1}^{n+1} k^3 &= \sum_{k=0}^{n} (k+1)^3 \\
&= \sum_{k=0}^{n} \left(k^3 + 3k^2 + 3k + 1\right) \\
&= \sum_{k=1}^{n} k^3 + 3\sum_{k=1}^{n} k^2 + 3\sum_{k=1}^{n} k + \sum_{k=0}^{n} 1 \\
&= \sum_{k=1}^{n} k^3 + 3\square_n + \tfrac{3}{2}n(n+1) + (n+1).
\end{aligned}$$

Igualando ambas as expressões, obtemos

$$\sum_{k=1}^{n+1} k^3 + (n+1)^3 = \sum_{k=1}^{n+1} k^3 + 3\square_n + \tfrac{3}{2}n(n+1) + (n+1).$$

Esta última equação pode ser resolvida em relação a $\square_n$:

$$\begin{aligned}
&\sum_{k=1}^{n+1} k^3 + (n+1)^3 = \sum_{k=1}^{n+1} k^3 + 3\square_n + \tfrac{3}{2}n(n+1) + (n+1) \\
&\qquad \Leftrightarrow (n+1)^3 = 3\square_n + \tfrac{3}{2}n(n+1) + (n+1)
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow 3\square_n = (n+1)\left[(n+1)^2 - \tfrac{3}{2}n - 1\right] \\
&\Leftrightarrow \square_n = \tfrac{1}{3}(n+1)(n^2+2n+1-\tfrac{3}{2}n-1) \\
&\Leftrightarrow \square_n = \tfrac{1}{6}(n+1)(2n^2+4n-3n) \\
&\Leftrightarrow \square_n = \tfrac{1}{6}n(n+1)(2n+1).
\end{aligned}$$

Estes exemplos parecem sugerir que se pode obter uma fórmula genérica para $\sum_{k=1}^{n} k^m$ aplicando o método de perturbação a $\sum_{k=1}^{n} k^{m+1}$; de facto assim é, desde que sejam conhecidos os coeficientes binomiais de ordem $m+1$ bem como todas as expressões para $\sum_{k=1}^{n} k^j$, para $j < m$. Ainda assim, é um método recursivo facilmente implementável para determinação destes valores.

4 Outros Exemplos

Para terminar, apresentamos aqui dois exemplos de como utilizar estas técnicas. O primeiro é um pequeno quebra-cabeças, que confesso me pareceria impossível sem conhecer o conteúdo deste artigo; o segundo é um exemplo que vale por si.

Exemplo 4. Para quantos números naturais n entre 1 e 1000 se tem que n é múltiplo de $\lfloor \sqrt[3]{n} \rfloor$?

Designemos por N o valor que pretendemos calcular; então temos

$$N = \sum_{n=1}^{1000} [\lfloor \sqrt[3]{n} \rfloor | n],$$

onde $x|y$ denota, como habitualmente, a proposição 'x divide y'.

Podemos reescrever esta equação como

$$N = \sum_{n,k} [k = \lfloor \sqrt[3]{n} \rfloor][k|n][1 \le n \le 1000]$$

ou, equivalentemente, como

$$N = \sum_{n,k} [k \le \sqrt[3]{n} < k+1][k|n][1 \le n \le 1000].$$

Vamos agora introduzir uma nova variável com base no seguinte raciocínio: k divide n se e somente se existir m tal que $n = km$; então $[k|n] = \sum_m [n = km]$. Assim, obtemos

$$N = \sum_{n,k,m} [k^3 \le n < (k+1)^3][n = km][1 \le n \le 1000],$$

ou ainda, uma vez que só não são nulos os termos em que $n = km$,

$$N = \sum_{n,k,m} [k^3 \leq km < (k+1)^3][n = km][1 \leq km \leq 1000].$$

Ora nesta soma só há um termo não nulo — quando $n = km$ —, pelo que podemos escrever ainda

$$N = \sum_{k,m} [k^3 \leq km < (k+1)^3][1 \leq km \leq 1000].$$

Dividindo por k na primeira parcela obtemos ainda

$$N = \sum_{k,m} \left[k^2 \leq m < \frac{(k+1)^3}{k}\right] [1 \leq km \leq 1000].$$

No caso $km = 1000$ é fácil verificar que só $k = 10$ e $m = 100$ satisfazem ambas as condições; nos restantes casos, necessariamente $k \leq 9$, donde podemos simplificar a última equação e obter

$$N = \sum_{k,m} \left[k^2 \leq m < \frac{(k+1)^3}{k}\right] [1 \leq k \leq 9] + 1.$$

Desenvolvendo a fracção obtemos

$$N = \sum_{k,m} \left[k^2 \leq m < k^2 + 3k + 3 + \frac{1}{k}\right] [1 \leq k \leq 9] + 1.$$

É fácil verificar que existem exactamente $3k+4$ valores em cada intervalo $\left[k^2, k^2 + 3k + 3 + \frac{1}{k}\right)$, pelo que obtemos simplesmente

$$\begin{aligned} N &= \sum_{k=1}^{9} (3k+4) + 1 \\ &= \frac{3 \times 9 \times 10}{2} + 36 + 1 \\ &= 172. \end{aligned}$$

Ao leitor mais céptico recomenda-se que faça uma tabela e verifique a validade do resultado.

O exemplo final é retirado de [1], e não resisti a apresentá-lo por ser totalmente inesperado. As passagens são imediatas, pelo que o apresento

sem comentários. Note-se que $\lg k$ denota o logaritmo de base 2 de k ($\log_2 k$).

$$\begin{aligned}
\sum_{k\geq 1} \binom{n}{\lfloor \lg k \rfloor} &= \sum_{k\geq 1} \binom{n}{m} [m = \lfloor \lg k \rfloor] \\
&= \sum_{k,m} \binom{n}{m} [m = \lfloor \lg k \rfloor][k \geq 1] \\
&= \sum_{k,m} \binom{n}{m} [m \leq \lg k < m+1][k \geq 1] \\
&= \sum_{k,m} \binom{n}{m} [2^m \leq k < 2^{m+1}][k \geq 1] \\
&= \sum_{m} \binom{n}{m} (2^{m+1} - 2^m)[m \geq 0] \\
&= \sum_{m} \binom{n}{m} 2^m 1^{n-m} \\
&= 3^m.
\end{aligned}$$

Se alguém conhecer uma forma alternativa de calcular esta última expressão teria curiosidade em conhecê-la.

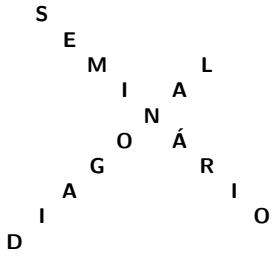
5 Agradecimentos

Gostaria de agradecer especialmente ao professor José Luís Fachada por me ter introduzido à arte da Combinatória e por me ter facilitado o material que serviu de base ao seminário que este texto sintetiza, bem como pelas sugestões quanto à elaboração do mesmo. Também à professora Ana Cannas da Silva gostaria de deixar o meu agradecimento por todas as sugestões que permitiram simplificar em muito algumas das passagens.

Referências

[1] Knuth, Donald E., Two notes on notation, *American Mathematical Monthly*, Vol. 99, No. 5, May 1992, pp. 403–422.

[2] Iverson, Kenneth E., *A Programming Language*, Wiley, 1962.



Análise Real(mente) Infinitesimal

João Pedro Boavida
5º ano da LMAC — Análise, Geometria e Álgebra
jboavida@math.ist.utl.pt

Palavras Chave

número infinitesimal, número infinitamente grande, função contínua, matemática não-standard, modelos, compacidade.

Resumo

Entre as hipóteses inconscientes na prática matemática habitual, conta-se a possibilidade de provar/refutar o que é verdadeiro/falso num número *finito* de passos. Assim, o conjunto de fórmulas

$$\{\varepsilon < 1, \varepsilon < \tfrac{1}{2}, \varepsilon < \tfrac{1}{3}, \dots\}$$

a respeito de um real $\varepsilon > 0$ não pode ser refutado, pelo que deveria existir algum número ε satisfazendo-as, que seria realmente infinitesimal.

Esta observação aparentemente inocente será o nosso ponto de partida para explorar a Matemática Não-Standard.

Introdução

Ninguém desconhece[1] que a ideia de número infinitesimal esteve sempre presente no desenvolvimento da análise matemática como a conhecemos hoje. Nem é preciso referir nomes como Leibniz, recordar os fantasmas de quantidades desaparecidas do arcebispo Berkeley [1], ou ainda manuais famosos como os escritos por Euler [5], de l'Hospital [2], Cauchy. Por outro lado, a tradição associa ao abandono de uma tal noção nomes como Bolzano, Weierstrass ou Cantor.[2]

Evidentemente o conceito de número infinitesimal não foi posto de lado sem boas razões. Afinal, espera-se de um número infinitesimal que

(1) tenha exactamente as mesmas propriedades que um real habitual (princípio de Leibniz),

1 Pelo menos entre os potenciais leitores deste texto...

2 Para uma visão histórica diferente da habitual, vale a pena consultar o último capítulo de [7].

mas ao mesmo tempo

(2) seja (em módulo) menor que qualquer real positivo habitual,

o que decerto não é uma das propriedades dos reais habituais. Dos números infinitamente grandes espera-se algo análogo em estranheza. É exigir demais.

Porém, no início dos anos 60, o lógico Abraham Robinson conseguiu explorar algumas das subtilezas da relação entre descrições e modelos (afinal ele era um dos criadores da teoria de modelos) para explicitar quais as propriedades que estão em causa em (1), e de modo a que (2) não seja uma dessas propriedades. Neste artigo tentamos perceber o que se passa, e explorar algumas das consequências.[3]

Antes de começar, vale a pena fazer um aviso: pode não ser seguro usar métodos não-standard nos exames de matemática (quanto mais não seja, seria preciso justificar todas as construções).

1 Lógica: Modelos e Compacidade

O nosso objectivo é descrever como foram reabilitados os conceitos de infinitesimal e infinitamente grande. Para tanto, é preciso explicar como é possível conciliar as exigências (1) e (2). E é por isso que faremos primeiro um passeio por território aparentemente distante: nesta primeira secção vamos fazer uma digressão pela lógica proposicional, a lógica com quantificadores e os respectivos modelos.

1.1 Modelos para a Lógica Proposicional

Na lógica proposicional temos uma família de *símbolos*, ditos *proposicionais*, que representam os factos ou a informação que pretendemos estudar, podem assumir os valores lógicos $\top$ ('verdadeiro') ou $\bot$ ('falso'), e constituem a *linguagem*.

Exemplo. Se tivermos um mapa, podemos usar o símbolo verde_i para representar o facto de o país i estar pintado de verde.

Os símbolos podem ser combinados com $\wedge$ ('e'), $\vee$ ('ou'), $\Rightarrow$ ('implica'), $\Leftrightarrow$ ('sse'), $\neg$ (negação) ou outros operadores lógicos para obter *fórmulas proposicionais* que descrevem informação mais complexa.

3 Tendo em conta que alguns pormenores são de facto especialmente subtis, vamos quase sempre não nos preocupar demasiado com eles, remetendo-os para notas de rodapé ou para 'Exercícios'. Nem é preciso dizer, a ideia de 'exercício' não deve ser levada demasiado a sério...

Exemplo. Naturalmente nenhum país está pintado de verde e de amarelo simultaneamente. Generalizando os símbolos já discutidos, esse facto pode ser representado pelas fórmulas $\mathsf{verde}_i \Rightarrow \neg\mathsf{amarelo}_i$.

Por fim, um *modelo* $\mathfrak{M}$ é a descrição completa de uma situação,[4] ou seja, a atribuição de um valor lógico a cada símbolo proposicional. Nem é preciso dizê-lo, os valores lógicos das fórmulas ficam automaticamente determinados.

Dizemos que $\mathfrak{M}$ é um modelo para a fórmula φ se lhe atribui o valor $\top$, e um modelo para o conjunto de fórmulas K se atribui $\top$ a cada fórmula $\varphi \in K$. Estas relações escrevem-se, respectivamente, $\mathfrak{M} \models \varphi$ e $\mathfrak{M} \models K$.

Exemplo. Se só estiverem disponíveis as cores 'amarelo', 'azul', 'encarnado' e 'verde', uma coloração de um mapa é um modelo de $K = \big\{\mathsf{amarelo}_i \vee \mathsf{azul}_i \vee \mathsf{encarnado}_i \vee \mathsf{verde}_i,\ \mathsf{amarelo}_i \Rightarrow \neg(\mathsf{azul}_i \vee \mathsf{encarnado}_i \vee \mathsf{verde}_i),\ \cdots \mid$ para cada país $i\big\}$.

Exercício (Teorema das Quatro Cores). Para cada mapa finito (i.e., com um número finito de países) no plano ou numa superfície esférica[5] existe sempre um modelo do conjunto $K \cup L$, onde $L = \big\{\mathsf{amarelo}_i \Rightarrow \neg\mathsf{amarelo}_j,\ \mathsf{azul}_i \Rightarrow \neg\mathsf{azul}_j,\ \mathsf{encarnado}_i \Rightarrow \neg\mathsf{encarnado}_j,\ \mathsf{verde}_i \Rightarrow \neg\mathsf{verde}_j \mid i \text{ e } j \text{ são adjacentes}\big\}$. O que são os modelos de $K \cup L$?

Outra relação útil é $K \models \varphi$: todos os modelos de K são modelos de φ. Sucede que se $K \models \varphi$, é possível demonstrá-lo num número finito de passos — se olharmos para K como uma colecção de hipóteses suficientes para provar φ, basta usar um número finito dessas hipóteses.[6]

Por outro lado, se K não é *contraditório* (ou seja, se não é possível a partir dele provar disparates como $\varphi \wedge \neg\varphi$) então existe algum modelo $\mathfrak{M} \models K$.[7] Assim, temos (porquê?) o seguinte

Metateorema (Compacidade da Lógica Proposicional). *Um conjunto de fórmulas tem um modelo sse cada subconjunto finito tiver.*

Alguns exemplos tradicionais de aplicação:

Exercício (Grupos Infinitos). Como existem grupos de qualquer ordem finita, existem necessariamente grupos de ordem infinita.

4 Ou melhor, *tão completa quanto possível* nos limites da linguagem em uso.

5 Isto é, usando a linguagem mais comum em geometria, uma 2-esfera (o 2 refere-se à dimensão).

6 Embora possa parecer, este facto está longe de ser trivial, e codifica algumas suposições fundamentais sobre o que pode ser demonstrado e em que contextos. Em qualquer caso, alguns poderão reconhecê-lo como 'completude de ⟨*algum sistema dedutivo*⟩ para a lógica proposicional'.

7 Outro facto não trivial...

Exercício (Teorema das Quatro Cores para Mapas Infinitos). Um mapa (finito ou infinito) no plano ou na esfera pode sempre ser colorido usando quatro cores, e sem que países adjacentes fiquem com a mesma cor.

Exercício (Princípio do Ultrafiltro). Um *filtro próprio* na álgebra de Boole $\mathfrak{B}$ é um subconjunto $\mathfrak{F} \subsetneq \mathfrak{B}$ fechado para $\wedge$ e tal que se $a \in \mathfrak{F}$ e $a \leq b$ então $b \in \mathfrak{F}$. Um *ultrafiltro* é um filtro próprio maximal. Qualquer filtro próprio $\mathfrak{F}$ está contido num ultrafiltro $\mathfrak{U}$.
Sugestão: Os elementos de $\mathfrak{B}$ podem ser usados como símbolos proposicionais, e as propriedades de um filtro próprio podem ser descritas por fórmulas proposicionais. Um filtro é maximal sse cada $a \in \mathfrak{B}$ ou o seu complemento $\overline{a} \in \mathfrak{B}$ está em $\mathfrak{F}$.

Vale a pena referir que este exemplo é profundamente relevante para a teoria de modelos da lógica proposicional. Esse facto pode ser ilustrado acrescentando que é equivalente ao Metateorema da Compacidade.[8] Tem também relações exóticas com o exemplo anterior (Cf. [6]).

1.2 Modelos para a Lógica com Quantificadores

A lógica proposicional não tem poder expressivo suficiente para descrever sem ambiguidades todas as situações interessantes.

Exemplo. Se alargarmos a linguagem dos exemplos que vimos com coloração de mapas de modo a descrever mais países, *não* é verdade que o conjunto K continue a ter como modelos (só) as colorações, pois K só dá informação sobre os símbolos (sobre os países) a que se refere. Mais concretamente, se a for um país adicional que não é referido pelas fórmulas de K, é claro que não temos garantias que um modelo de K não satisfaça simultaneamente $\mathsf{amarelo}_a$ e verde_a.

O que queremos é poder dizer, por exemplo, que 'qualquer país' está pintado de uma só cor. O que queremos é ter um universo (ou universos) de objectos que podem satisfazer propriedades (chamadas *predicados*), e podermos referir-nos a eles usando quantificadores $\forall$ ('para todo') e $\exists$ ('existe'). Queremos poder dar nomes (*constantes*) a objectos, ou até mesmo estudar *funções* sobre eles.[9]

8 E já agora, para os que tenham familiaridade com a linguagem, também é equivalente ao Teorema de Tychonoff para espaços Hausdorff (afinal o nome 'compacidade' não foi dado por acidente...).

9 Em boa verdade, estas duas últimas exigências são perfeitamente dispensáveis (Cf. [7]). É só por comodidade que não fazemos tudo com predicados.

Os *modelos* serão agora especificações dos universos, complementadas por identificações dos objectos nomeados pelas constantes e funções, e por descrições completas dos valores lógicos dos predicados.

Exemplo. No caso dos mapas podemos descrever a mesma situação de modo muito mais conciso referindo-nos a modelos de $K' = \big\{(\forall i)(\text{`}i \text{ é amarelo'} \vee \text{`}i \text{ é azul'} \vee \text{`}i \text{ é encarnado'} \vee \text{`}i \text{ é verde'}),\ (\forall i)(\text{`}i \text{ é verde'} \Rightarrow \neg\text{`}i \text{ é amarelo'}),\ \ldots\big\}$, onde usámos símbolos como '$\cdot$ é verde' para representar os predicados.

Exemplo (O Corpo dos Reais). O corpo $\mathbb{R}$ dos números reais é um modelo de uma colecção bem conhecida de axiomas $\{\text{propriedades dos reais}\} = \big\{(\forall x)(\forall y)(\exists z)(x+y=z),\ (\forall x)(\forall y)(\forall z)(\forall w)\big((x+y=z \wedge x+y=w) \Rightarrow z=w\big),\ (\forall x)(x+0=x),\ (\forall x)(\forall y)(x+y=y+x),\ \ldots\big\}$. Ou seja, se x e y são reais, existe um real z (chamado soma), e é único; a soma de um real x com a constante 0 (um real específico) é x; a soma é comutativa; etc.[10]

Em alguns casos a situação é mais sofisticada, já que temos diferentes tipos de objectos:

Exemplo (Espaço Vectorial Real). Temos dois tipos de objectos: os escalares (cujo universo vamos designar $\mathbb{R}$) e os vectores (cujo universo vamos designar V), e um espaço vectorial real é um modelo do conjunto de axiomas $\big\{0 \in \mathbb{R},\ \mathbf{0} \in V,\ (\forall \lambda \in \mathbb{R})(\forall \mathbf{v} \in V)(\forall \mathbf{w} \in V)(\lambda\mathbf{v} + \lambda\mathbf{w} = \lambda(\mathbf{v}+\mathbf{w})),\ (\forall \mathbf{v} \in V)(0\mathbf{v} = \mathbf{0}),\ \ldots\big\}$.

É comum usar convenções tipográficas diferentes (como fizemos ainda agora) para representar objectos de universos diferentes. E é o que faremos no seguimento.

O facto importante que nos interessa é que, mais uma vez, se tem

Metateorema (Compacidade da Lógica com Quantificadores). *Um conjunto de fórmulas tem um modelo sse cada subconjunto finito tiver.*

Um exemplo clássico:

Exercício (Princípio do Ideal Máximo). Se A é um anel com identidade e $J \subset A$ é um subanel próprio, então existe sempre um ideal maximal $I \supset J$. *Sugestão:* O ideal I é maximal sse a identidade $1 \in A$ é combinação 'linear' de *qualquer* elemento fora de I com algum elemento em I.

Mas este é um exemplo muito inocente. Vamos tentar perceber até que ponto as consequências podem ser exóticas, considerando o modelo *standard* (isto é, 'aquele' com que estamos habituados a trabalhar) $\mathfrak{N}$ dos naturais,

$\mathfrak{N} \models \{\text{propriedades dos naturais}\},$ |

10 Mas... quais são as propriedades da igualdade? E como escrever o axioma do supremo?

que tem como universo o conjunto $\mathbb{N}$ dos naturais *standard*, os quais satisfazem diversas propriedades — como a existência de uma operação $+$ comutativa, associativa e com um elemento neutro designado 0, ou a validade do princípio de indução.[11] Na nossa descrição do modelo standard podemos naturalmente incluir os nomes de todos os naturais standard, como 33 (que pode ser identificado como *o único* natural que é obtido somando $1+1+1+\cdots+1$ trinta e três vezes) ou $10^{23} - 4^{16}$.

O que é surpreendente é que se alargarmos a linguagem para incluir uma nova constante N, vai haver um modelo[12] para um outro conjunto:

$$^*\mathfrak{N} \models \{\text{propriedades dos naturais},\ N > 1,\ N > 2,\ N > 3,\ \ldots\}.$$

Com efeito, cada subconjunto finito de fórmulas F tem claramente um modelo: basta considerar $\mathfrak{N}$ e escolher qualquer N que satisfaça todas as condições adicionais presentes em F, *que são em número finito*.

Este novo modelo $^*\mathfrak{N}$ vai conter *todos* os naturais standard, como 33 e $10^{23} - 4^{16}$, pois *eles têm nomes* que têm que ser (re)interpretados como objectos do novo universo $^*\mathbb{N}$. E *todos* esses naturais têm as mesmas propriedades (descritíveis!) que tinham antes, sendo portanto indistinguíveis das suas encarnações em $\mathbb{N}$.

Notação. *Por vezes*, vamos preceder os objectos de $\mathfrak{N}$ reinterpretados em $^*\mathfrak{N}$ pelo sinal *. Faremos o mesmo para os predicados.

Exemplo. O novo modelo $^*\mathfrak{N}$ tem objectos antigos, como ^{*33}, mas tem pelo menos um novo objecto $N \in {}^*\mathbb{N}$.

Como qualquer natural é par ou ímpar,

N, que é *natural, é *par ou *ímpar.

Mas esta não é a propriedade mais interessante de N:

N é um *natural *maior* que todos os naturais standard.

É crucial observar que isto *não* significa que N é maior que todos os *naturais: *não é!!* Só significa que é maior que todos os que estão numa subclasse particular: a classe dos que foram importados de $\mathfrak{N}$.[13]

Conseguimos portanto duas coisas aparentemente contraditórias: (1) N satisfaz exactamente as propriedades de todos os naturais *identificáveis na linguagem*,[14] (2) mas ainda assim é maior que todos eles. Ou seja, N é um

11 Como escrevê-lo?

12 Na verdade, vai haver tantos modelos que nem sequer constituem um conjunto!!

13 Isto ilustra uma das muitas subtilezas de que é feita a análise não-standard: $^*\mathbb{N} \neq \{^*n \mid n \in \mathbb{N}\}$.

14 Note-se bem que * *não* está na linguagem!

*natural *infinitamente grande*.

DEFINIÇÃO. Um *natural N diz-se *infinitamente grande* se for maior que todos os *naturais standard. De contrário diz-se *finito*.

O truque muito subtil está na forma como a caracterização que fazemos de N não está ao alcance da linguagem com que o tentamos descrever:

Exemplo. Não é possível dizer algo como $(\forall M)(N > M)$, pois nesse caso N teria que ser maior que todos os naturais, *incluindo todos os que sejam acrescentados ao modelo*, o que seria claramente impossível — afinal $(\forall M)(M + 1 > M)$ é uma propriedade dos (*)naturais, que se aplica, em particular, a N, permitindo-nos concluir que $N + 1 > N$.

Mais geralmente,

OBSERVAÇÃO FUNDAMENTAL. *Não é possível descrever um objecto não-standard usando um número* finito *de fórmulas, a não ser que essas façam referência explícita a objectos não-standard. Não é possível discriminar entre objectos standard e não-standard sem fazer referência explícita a objectos não-standard.*

As fórmulas ou propriedades que não usem operações como * ou predicados como '$\cdot$ é standard' são designadas *internas*. Entre essas, as que não façam referência explícita a objectos não-standard são designadas *standard*. Partindo de um modelo $\mathfrak{M}$, construímos um modelo não-standard $^*\mathfrak{M}$ juntando objectos descritos por uma colecção infinita de fórmulas.

METATEOREMA (PRINCÍPIO DE TRANSFERÊNCIA). *Se $\mathfrak{M}$ é um modelo, $^*\mathfrak{M}$ um modelo não-standard correspondente, e φ uma fórmula standard, então $\mathfrak{M} \models \varphi$ sse $^*\mathfrak{M} \models \varphi$. Dito de outro modo, as propriedades standard dos objectos standard são* precisamente *as de todos os objectos (standard e não-standard).*

Observação. Para perceber porquê a restrição a fórmulas standard, é útil observar que essas são *precisamente* as fórmulas que fazem sentido nos dois mundos $\mathfrak{M}$ e $^*\mathfrak{M}$.

Veremos no seguimento as consequências profundas deste facto.

Notação. Usamos $^{\text{st}}x$ para indicar que o objecto x do modelo não-standard é standard. Pelo contrário, usamos $^*x$ para referir o correspondente no modelo não-standard do objecto x do modelo standard.

2 A Recta $^*\mathbb{R}$eal

Ainda não verificámos se existem números infinitesimais. Outro exemplo pode ajudar a elucidar essa questão; consideremos o modelo standard:

$\mathfrak{R} \models \{\text{propriedades dos reais}\}$

Juntando uma constante ε, teremos um modelo para um novo conjunto:

$$^*\mathfrak{R} \models \{\text{propriedades dos reais}, \varepsilon > 0, \varepsilon < 1, \varepsilon < \tfrac{1}{2}, \varepsilon < \tfrac{1}{3}, \ \ldots\}$$

ε é um *real positivo *menor* que todos os reais positivos standard:[15] é um *número infinitesimal*.

DEFINIÇÃO. Dizemos que x é *infinitamente grande*, e escrevemos $x \sim \infty$, se $|x|$ é maior que algum *natural infinitamente grande. Dizemos que ε é *infinitesimal*, e escrevemos $\varepsilon \simeq 0$, se $|\varepsilon|$ é menor que qualquer *real positivo standard.

Aplicando o princípio de transferência:[16]

	Seja $\varepsilon > 0$ um *real infinitesimal,
Qualquer real $\varepsilon > 0$ é invertível.	Portanto, ε é invertível.
$0 < \varepsilon < 1 \Leftrightarrow \frac{1}{\varepsilon} > 1$	Logo, $\frac{1}{\varepsilon} > 1$
$0 < \varepsilon < \frac{1}{2} \Leftrightarrow \frac{1}{\varepsilon} > 2$	$\frac{1}{\varepsilon} > 2$
$0 < \varepsilon < \frac{1}{3} \Leftrightarrow \frac{1}{\varepsilon} > 3$	$\frac{1}{\varepsilon} > 3$
$\vdots$	$\vdots$
	Resumindo, $\frac{1}{\varepsilon}$ é infinitamente grande, pois é maior que todos os naturais standard.

Acabamos assim de mostrar

PROPOSIÇÃO. *Se $\varepsilon \simeq 0$ é infinitesimal, então $\frac{1}{\varepsilon} \sim \infty$.*

Como quaisquer *reais têm uma soma, se $x \in {}^*\mathbb{R}$ então para cada $\varepsilon \simeq 0$ existe $x + \varepsilon$. Escrevemos $x \simeq y$, e dizemos que x e y estão *infinitamente próximos* se $|x - y| \simeq 0$.

15 É possível identificar inequivocamente os reais standard? Como?

16 Note-se como até as variáveis podem ter significados diferentes nos dois mundos $\mathfrak{R}$ e $^*\mathfrak{R}$.

DEFINIÇÃO. A *vizinhança infinitesimal* de $x \in {}^*\mathbb{R}$ é o conjunto $\{y \in {}^*\mathbb{R} \mid y \simeq x\}$.

DEFINIÇÃO. Um *real $x \in {}^*\mathbb{R}$ diz-se *quase-standard* (também *próximo--standard* ou *finito*) sse $(\exists y \in \mathbb{R})(x \simeq {}^*y)$. Nesse caso escreveremos ainda $y = {}^\circ x$. Se x for infinitamente grande positivo escreveremos ${}^\circ x = +\infty$, e ${}^\circ x = -\infty$ se for infinitamente grande negativo.

Assim, um número infinitesimal é um *real x tal que ${}^\circ x = 0$.

Todos os *reais têm uma expansão decimal.

Exemplo. Todos os dígitos (de ordem finita ou infinitamente grande) da expansão de ${}^*\frac{1}{3}$ são iguais a 3. Por outro lado, um *real x é infinitesimal sse todos os dígitos de ordem finita são 0. Com efeito, o mesmo é dizer que $|x| < 10^{-n}$ para cada natural finito ${}^{\mathrm{st}}n \in {}^*\mathbb{N}$.

Vale a pena verificar mais alguns resultados nesta nova linguagem.

PROPOSIÇÃO. *Se $x \not\simeq \infty$ e $\varepsilon \simeq 0$, então ${}^\circ(x+\varepsilon) = {}^\circ x \in \mathbb{R}$ e $\varepsilon x \simeq 0$.*

Demonstração. Com efeito, dizer que $x \not\simeq \infty$ é dizer que $(\exists {}^{\mathrm{st}}n)(|x| < n)$. Logo $|x+\varepsilon| \leq |x| + |\varepsilon| < n+1$ e $x+\varepsilon$ é quase-standard.

Além disso, dado qualquer $y \in \mathbb{R}^+$, temos também $\frac{y}{n} \in \mathbb{R}^+$, pelo que podemos dizer $|\varepsilon x| \leq |\varepsilon||x| < \frac{y}{n}n = y$ e concluir $\varepsilon x \simeq 0$ (pela definição de infinitesimal). □

Pode ser útil verificar também as seguintes propriedades algébricas:

Exercício. Se $x, y \not\simeq \infty$ e $\varepsilon, \delta \simeq 0$, então $x+y, x+\varepsilon, xy \not\simeq \infty$ e $\varepsilon\delta \simeq 0$. Se além disso $x \not\simeq 0$, então $\frac{\varepsilon}{x} \simeq 0$.

Exercício. Será possível dizer alguma coisa sobre $\frac{\varepsilon}{\delta}$? O quê?

Já ganhámos alguma familiaridade com a mentalidade não-standard, e em especial já temos uma justificação para usar sem problemas expressões como 'infinitesimal' e 'infinitamente grande'. Comecemos então a explorar o que isso traz de diferente à análise em $\mathbb{R}$.

*Exercício.** Ir pensando o que mudaria em $\mathbb{R}^n$, em espaços de Banach, espaços topológicos, ou outros contextos com que se tenha familiaridade.

2.1 Abertos e Fechados

Será possível descrever os abertos de $\mathbb{R}$ na linguagem não-standard?

$$U \text{ é } \textit{aberto} \text{ sse } (\forall x \in U)(\exists \varepsilon > 0)(\forall y)(|y-x| < \varepsilon \Rightarrow y \in U).$$

> Seja ${}^{\mathrm{st}}x \in {}^{\mathrm{st}}U$ e ${}^{\mathrm{st}}\varepsilon > 0$ correspondente. Se $y \simeq x$, então $|y-x| < {}^{\mathrm{st}}\varepsilon$, donde $y \in U$.

Ou seja, se ${}^{\mathrm{st}}U$ é um aberto standard, então $(\forall {}^{\mathrm{st}}x \in U)(\forall y \simeq x)(y \in U)$. O recíproco também é verdade; dado ${}^{\mathrm{st}}x \in U$,

> escolha-se $\varepsilon \simeq 0$ qualquer. Então $|y - x| < \varepsilon \Rightarrow y \simeq x \Rightarrow y \in U$. É portanto verdade que

$$(\exists \varepsilon > 0)(\forall y)(|y - x| < \varepsilon \Rightarrow y \in U),$$

> ou seja, já que esta última fórmula é satisfeita para qualquer $x \in U$, é verdade que o conjunto U é aberto.

Em resumo,

PROPOSIÇÃO. *O conjunto* ${}^{\mathrm{st}}U$ *é aberto sse* $(\forall {}^{\mathrm{st}}x \in U)(\forall y \simeq x)(y \in U)$.

Exemplo. O *intervalo *aberto ${}^*(1, 2)$ contém as vizinhanças infinitesimais dos seus pontos standard. Mas ^{*1}, apesar de estar infinitamente próximo de alguns pontos de ${}^*(1, 2)$, não pertence ao conjunto.

Exemplo. As coisas são diferentes com conjuntos não-standard. Se $\varepsilon \simeq 0$ é infinitesimal, o conjunto $[1 + \varepsilon, 2 - \varepsilon]$ é *fechado. Porém, contém as vizinhanças infinitesimais de todos os seus pontos standard (verificar).

Exercício. Qual será a caracterização para conjuntos *fechados*?
Sugestão: F é fechado sse F^{c} é aberto.

2.2 Sucessões e Limites de Sucessões

De modo análogo podemos descrever os limites de sucessões standard.

$$u \to \lambda \text{ sse } (\forall \varepsilon > 0)(\exists p \in \mathbb{N})(\forall n > p)(|u_n - \lambda| < \varepsilon).$$

> Escolha-se então um ${}^{\mathrm{st}}\varepsilon > 0$ e um ${}^{\mathrm{st}}p$ correspondente. Se $N \sim \infty$, então $N > {}^{\mathrm{st}}p$, donde $|u_N - {}^{\mathrm{st}}\lambda| < \varepsilon$. Como ${}^{\mathrm{st}}\varepsilon > 0$ é arbitrário, vemos que $u_N \simeq \lambda$.

E, uma vez mais, também o recíproco se verifica: se ${}^{\mathrm{st}}u$ é uma sucessão standard, ${}^{\mathrm{st}}\lambda \in {}^*\mathbb{R}$ e $(\forall N \sim \infty)(u_N \simeq \lambda)$, então $u \to \lambda$. Com efeito, dado ${}^{\mathrm{st}}\varepsilon > 0$,

> escolha-se qualquer $p \sim \infty$. Então $n > p \Rightarrow n \sim \infty \Rightarrow u_n \simeq \lambda \Rightarrow |u_n - \lambda| < {}^{\mathrm{st}}\varepsilon$. É portanto verdade que

$$(\exists p \in \mathbb{N})(\forall n > p)(|u_n - \lambda| < \varepsilon),$$

ou seja, já que a última fórmula é satisfeita para qualquer $\varepsilon > 0$, é verdade que $u \to \lambda$.

Em resumo,

PROPOSIÇÃO. *A sucessão standard* ${}^{\mathrm{st}}u$ *converge para o real standard* ${}^{\mathrm{st}}\lambda$ *sse* $(\forall N \sim \infty)(u_N \simeq \lambda)$.

Exemplo. Um cálculo típico, na abordagem não-standard ($N \sim \infty$):

$\lim_n \dfrac{n^3 + 4n^2 - 7}{3n^3 - 2n + 3}$

$\dfrac{N^3 + 4N^2 - 7}{3N^3 - 2N + 3} = \dfrac{1 + \frac{4}{N} - \frac{7}{N^3}}{3 - \frac{2}{N^2} + \frac{3}{N^3}}$

numerador $\simeq 1$, denominador $\simeq 3$

fracção $\simeq \frac{1}{3}$

logo, $\lim = \frac{1}{3}$.

Exercício. A sucessão ${}^{\mathrm{st}}u$ é limitada sse $(\forall N \sim \infty)(u_N$ é finito).

Exercício. ${}^{\mathrm{st}}\lambda$ é sublimite da mesma sucessão u sse $(\exists N \sim \infty)(u_N \simeq \lambda)$.
Sugestão: Escrever 'λ é sublimite de u' usando uma fórmula com quantificadores.

Estes dois factos permitem demonstrar facilmente o

TEOREMA (BOLZANO–WEIERSTRASS). *Se u é uma sucessão limitada, então u tem uma subsucessão convergente.*

Exercício. Verificar o Teorema de Bolzano–Weierstrass.
Sugestão: Pelo princípio de transferência basta provar isto para sucessões (e como tal sublimites) standard.

3 Continuidade e Diferenciabilidade de Funções

No seguimento limitar-nos-emos a usar as caracterizações não-standard, e a mostrar as suas aplicações. Naturalmente o leitor interessado poderá verificá-las sem dificuldade.

PROPOSIÇÃO. *Se* ${}^{\mathrm{st}}f$ *é uma função real de variável real e* ${}^{\mathrm{st}}a, {}^{\mathrm{st}}x_0 \in {}^*\mathbb{R}$ *são reais standard, então* $\lim_{x \to x_0} f(x) = a$ *sse* $(\forall x \simeq x_0)(f(x) \simeq a)$.

PROPOSIÇÃO. *A função* ${}^{\mathrm{st}}f$ *é contínua no ponto* ${}^{\mathrm{st}}x_0 \in {}^*\mathbb{R}$ *sse* $(\forall x \simeq x_0)$ $(f(x) \simeq f(x_0))$.

Exemplo. A função f definida por $0 \neq x \mapsto \frac{\sin x}{x}$ e $0 \mapsto 1$ é contínua no ponto 0. Com efeito, se $x \simeq 0$, ou bem que $x \neq 0$ e $f(x) \simeq 1$, ou bem que $x = 0$ e $f(x) = 1 \simeq 1$. Em qualquer caso $f(x) \simeq f(0)$.[17]

Exemplo. A função $x \mapsto \sin \frac{1}{x}$ não pode ser estendida por continuidade ao ponto $x = 0$. Se escolhermos $N \sim \infty$ *inteiro, em $x = \frac{1}{2N\pi} \simeq 0$ temos $f(x) = 0$, e em $x = \frac{2}{(4N+1)\pi} \simeq 0$ temos $f(x) = 1$. Um eventual prolongamento contínuo $f(0)$ teria que satisfazer simultaneamente $f(0) \simeq 0$ e $f(0) \simeq 1$.

Exercício. Pensar numa *função *contínua que não satisfaça a proposição anterior (vai necessariamente não ser standard).

3.1 Continuidade em Intervalos

Já explorámos um pouco do que passa com a continuidade num ponto. Será que podemos fazer o mesmo para a continuidade num intervalo? Será que a função ${}^{\mathrm{st}}f$ é contínua no intervalo ${}^{\mathrm{st}}I$ sse $(\forall x, y \in I)(x \simeq y \Rightarrow f(x) \simeq f(y))$?

Exemplo. A função $f : \mathbb{R}^+ \to \mathbb{R} : x \mapsto \frac{1}{x}$ é contínua no intervalo $(0, +\infty)$. Se $\varepsilon \simeq 0$, então $\varepsilon \simeq \frac{\varepsilon}{2}$. Mas $f(\varepsilon) = \frac{1}{\varepsilon} \not\simeq \frac{2}{\varepsilon} = f(\frac{\varepsilon}{2})$, pois $\left|\frac{1}{\varepsilon} - \frac{2}{\varepsilon}\right| = \frac{1}{\varepsilon} \sim \infty$.

Este problema não existiria se f estivesse definida em 0 e fosse contínua, pois nesse caso teríamos $f(\varepsilon) \simeq f(0) \simeq f(\frac{\varepsilon}{2})$ e portanto $f(\varepsilon) \simeq f(\frac{\varepsilon}{2})$. Mais geralmente, vamos dizer que

DEFINIÇÃO. O conjunto ${}^{\mathrm{st}}K$ é *compacto* sse $(\forall x \in K)(\exists {}^{\mathrm{st}}x_0 \in K)(x \simeq x_0)$.

DEFINIÇÃO. A função ${}^{\mathrm{st}}f$ é *uniformemente contínua* sse $(\forall x, y)(x \simeq y \Rightarrow f(x) \simeq f(y))$.

Nesse caso, para ${}^{\mathrm{st}}f$ contínua num compacto ${}^{\mathrm{st}}K$,

> sejam $x, y \in K$ com $x \simeq y$. Seja ainda ${}^{\mathrm{st}}x_0 \in K$ com $x \simeq x_0$. Temos $f(x) \simeq f(x_0)$ e $f(y) \simeq f(x_0)$, pelo que $f(x) \simeq f(y)$.

Isto prova[18]

TEOREMA (HEINE–CANTOR). *Uma função contínua num compacto K é uniformemente contínua em K.*

Outro resultado célebre é

17 Observe-se que, pelo princípio de transferência, *todas* as funções standard (em particular as funções trigonométricas) mantêm *todas* as suas propriedades standard.

18 O leitor atento notará que há um pequeno pormenor que *não* referimos. Qual?

Teorema (Heine–Borel). *Um conjunto K é compacto sse é limitado e fechado.*

Exercício. Verificar o Teorema de Heine–Borel.
Sugestão: ${}^{\mathrm{st}}K$ é limitado sse $(\forall x \in K)(x$ é finito). Além disso, se $x \simeq {}^{\mathrm{st}}x_0$ e $x \simeq {}^{\mathrm{st}}x_1$, então $x_0 = x_1$.[19]

Exercício. Já agora, a definição não-standard sugere uma caracterização sequencial *standard* para a continuidade uniforme. Qual?

Demonstramos também o

Teorema (Bolzano). *Se f é contínua em $[a,b]$, $f(a) > 0$ e $f(b) < 0$, então existe $\xi \in [a,b]$ tal que $f(\xi) = 0$.*

Demonstração. Supomos, por transferência, que f, a e b são standard. Sendo $N \sim \infty$, dividimos $[a,b]$ em N partes iguais, com extremos $x_i = a + i\frac{b-a}{N}$, onde $0 \leq i \leq N$. Como $f(x_0) > 0$ e $f(x_N) < 0$, deve haver um primeiro i tal que $f(x_i) < 0$ — afinal $\{i \in {}^*\text{inteiro} \mid 0 \leq i \leq N\}$ é um conjunto *finito!

Temos $f(x_{i-1}) > 0$. Seja $x_{i-1} \simeq x_i \simeq {}^{\mathrm{st}}\xi$ ($[a,b]$ é compacto). Virá $f(\xi) \simeq f(x_i) < 0$ e $f(\xi) \simeq f(x_{i-1}) > 0$. O único real standard (e $f(\xi)$ é standard!) satisfazendo essas duas condições é 0 (verificar!), portanto $f(\xi) = 0$. □

Outro teorema célebre é

Teorema (Weierstrass). *Um função contínua num compacto K tem máximo e mínimo em K.*

Exercício. Como seria uma demonstração não-standard deste teorema?

Outra pergunta natural é

Exercício. Qual é a caracterização não-standard da continuidade num intervalo?

3.2 Diferenciabilidade

Definição. A função ${}^{\mathrm{st}}f$ é *diferenciável* no ponto ${}^{\mathrm{st}}x_0$ com derivada ${}^{\mathrm{st}}a$ sse $(\forall x \neq x_0)\big(x \simeq x_0 \Rightarrow \frac{f(x)-f(x_0)}{x-x_0} \simeq a\big)$.

Será que dizer que ${}^{\mathrm{st}}f$ é diferenciável num intervalo ${}^{\mathrm{st}}I$ com derivada f' é o mesmo que dizer $(\forall x)(\forall y)\big((x \simeq y \simeq {}^{\mathrm{st}}x_0 \wedge x \neq y) \Rightarrow \frac{f(x)-f(y)}{x-y} \simeq f'(x_0)\big)$?

19 Esta propriedade é conhecida como Hausdorff-separação.

Exemplo. Definimos a função f por $x \neq 0 \mapsto x^2 \cos \frac{1}{x}$ e $0 \mapsto 0$. Usando as regras de derivação concluímos que é diferenciável em $x_0 \neq 0$, com derivada $f'(x_0) = 2x_0 \cos \frac{1}{x_0} + \sin \frac{1}{x_0}$. Por outro lado, se $x \simeq 0$ temos $\dfrac{x^2 \cos \frac{1}{x} - 0}{x - 0} = x \cos \frac{1}{x} \simeq 0$, pelo que $f'(0) = 0$.

Porém, substituindo para $\frac{1}{2N\pi} \simeq \frac{1}{(2N+1)\pi}$ obtemos

$$\frac{f\left(\frac{1}{2N\pi}\right) - f\left(\frac{1}{(2N+1)\pi}\right)}{\frac{1}{2N\pi} - \frac{1}{(2N+1)\pi}} = \frac{\frac{1}{(2N\pi)^2} + \frac{1}{((2N+1)\pi)^2}}{\frac{1}{2N(2N+1)\pi}} = \frac{1}{\pi}\left(\frac{2N+1}{2N} + \frac{2N}{2N+1}\right) \simeq \frac{2}{\pi}.$$

Claramente $\frac{2}{\pi} \not\simeq f'(0) = 0$, apesar de f ser diferenciável.

Exercício. A propriedade acima referida equivale a dizer que f é diferenciável e a sua derivada é contínua.
Sugestão: Num dos sentidos pode ser útil utilizar o Teorema de Lagrange.

4 Integral de Riemann

Se quisermos medir a área debaixo do gráfico de uma função ${}^{\mathrm{st}}f$ definida no intervalo $[{}^{\mathrm{st}}a, {}^{\mathrm{st}}b]$, podemos cortar o intervalo em fatias de espessura infinitesimal: $a \equiv x_0 \leq x_1 \leq \cdots \leq x_N \equiv b$, tal que $\Delta x_i = x_{i+1} - x_i \simeq 0$ (por exemplo, $x_i = a + i\frac{b-a}{N}$), sendo $N \sim \infty$. A área pode ser estimada[20] pelo seguinte *somatório (*finito!):

$$\sum_{0 \leq i < N} f(x_i)\Delta x_i.$$

DEFINIÇÃO. Se todas essas estimativas forem infinitamente próximas do mesmo real standard, esse real será chamado o *integral* (de Riemann) da função f entre a e b, e será representado (u é uma variável muda)

$$\int_a^b f(u)du.$$

Exemplo. A função $f = \chi_{\mathbb{Q}\cap[0,1]}$ não é integrável. Com efeito, se escolhermos $x_i = \frac{i}{N}$ será $\sum f(x_i)\Delta x_i = \sum_{0 \leq i < N} 1 \cdot \frac{1}{N} = 1$.

Já com ε *irracional satisfazendo $0 < \varepsilon < \frac{1}{N}$, se escolhermos $x_0 = 0$, $x_i = \frac{i}{N} + \varepsilon$ para $0 < i < N$ e $x_N = 1$ será $\sum f(x_i)\Delta x_i = f(x_0)\Delta x_0 + \sum_{0<i<N} f(x_i)\Delta x_i = 1 \cdot (\frac{1}{N} + \varepsilon) + 0 \simeq 0$.

20 O que acontece nos pontos em que a função assume valores negativos?

Note-se que $\sum_{0\leq i<N} f(x_i)\Delta x_i$ é a área por baixo do gráfico de uma função seccionalmente constante (a função que em $[x_i, x_{i+1})$ vale $f(x_i)$). Se f for contínua, essa função estará infinitamente próxima, em cada ponto, de f. Isto sugere a seguinte observação

Exercício. Se f é contínua, os valores de duas estimativas são infinitamente próximos, isto é, f é integrável.

Por outro lado, se f é integrável, ponhamos $F(x) = \int_a^x f(u)du$.

Exercício. Mostrar que $F'(x) = f(x)$.

Vejamos outra classe de funções integráveis:

Exemplo. Seja ${}^{\mathrm{st}}F$ uma função com derivada contínua $f = F'$. Sendo x_i os pontos que determinam uma divisão de $[a, b]$, temos $\sum_{0\leq i<N} f(x_i)\Delta x_i \simeq \sum \frac{F(x_{i+1})-F(x_i)}{x_{i+1}-x_i}\Delta x_i = \sum(F(x_{i+1}) - F(x_i)) = F(x_N) - F(x_0)$. Resumindo, f é integrável e $\int_a^b f(u)du = F(b) - F(a)$.

Exercício. Verificar a aproximação do somatório.
Sugestão: Uma diferença é infinitesimal se for menor em módulo que todo o real positivo standard. O produto é distributivo em relação à soma.

Concluímos portanto

Teorema (Fundamental do Cálculo). *Uma função contínua f é sempre integrável, e um seu integral $F(x) = \int_a^x f(u)du$ tem derivada $F'(x) = f(x)$. Pelo contrário, se uma função F é diferenciável com derivada contínua $F'(x) = f(x)$ então o seu integral satisfaz $\int_a^x f(u)du = F(x)$.*

Porém, se não temos cuidado a aplicar o teorema, podemos ter surpresas:

Exemplo. Seja H a função não-standard definida por $(-\infty, -\varepsilon] \ni x \mapsto 0$, $[-\varepsilon, \varepsilon] \ni x \mapsto \frac{x+\varepsilon}{2\varepsilon}$ e $[\varepsilon, +\infty) \ni x \mapsto 1$; onde $0 < \varepsilon \simeq 0$. Uma função como esta descreve uma variação rápida — por exemplo, a variação da temperatura através de uma janela de espessura 2ε. Na escala normal, uma tal variação parece instantânea. Isso corresponde a uma aproximação standard valendo 0 em $(-\infty, 0)$ e 1 em $(0, +\infty)$ — que é uma boa aproximação macroscópica para a mesma variação de temperatura.

Podemos derivar H para obter $H' = 0$ em $(-\infty, -\varepsilon)\cup(+\varepsilon, +\infty)$ e $H' = \frac{2}{\varepsilon}$ em $(-\varepsilon, +\varepsilon)$. Chamemos δ a esta função. O seu significado físico é de novo intuitivo, bem como o da sua aproximação standard, 0 em $\mathbb{R} \setminus \{0\}$. Até aqui tudo é normal. Bem... nem tudo:

Exercício. Se f é contínua, então $\int \delta(u)f(u)du = \int_{-\varepsilon}^{+\varepsilon} \delta(u)f(u)du \simeq f(0)$.

*Exercício.** Porque é que isto é surpreendente? O que está a acontecer?

Agradecimentos

Tendo em conta a dificuldade em explicar matemática feita simultaneamente em mundos tão diferentes como $\mathfrak{M}$ e ${}^*\mathfrak{M}$, a ajuda de algumas pessoas foi muito útil para tornar a explicação mais clara (embora eu desconfie que não soube aproveitar essa ajuda da melhor forma...).

Na preparação do seminário o apoio do Luís Cruz-Filipe e da Ana Cannas da Silva foram determinantes. Na revisão do texto, onde contei de novo com o apoio da Ana Cannas, os comentários críticos do Rui Loja Fernandes obrigaram-me a reformular várias passagens menos claras (na opinião dele, e não só...), espero que para melhor. Nos dois casos, a opinião especializada do João Paulo Teixeira foi igualmente útil. Ficam aqui expressos os agradecimentos. Como nem todas as sugestões foram seguidas, é possível que tenham restado alguns erros, e — claro! — essas pessoas não têm qualquer responsabilidade neles.

Bibliografia

O objectivo deste artigo/seminário era provocar a curiosidade. Assim, se tiver sido bem sucedido, por esta altura o leitor gostaria de procurar mais informação sobre a matemática não-standard. Antes de dar algumas pistas a esse respeito, não é demais alertar que um tal leitor se debaterá — se não se debateu já durante a leitura deste texto! — muitas vezes com algumas subtilezas. Isso não deve ser razão para desânimo. Afinal, repetindo palavras do Resumo, essas mesmas subtilezas se escondem por trás de hipóteses inconscientes na prática matemática habitual.[21]

Um bom ponto de partida, extremamente acessível, é [4], uma colectânea de artigos sobre matemática não-standard. Não se exige qualquer conhecimento prévio, tanto mais que o primeiro desses artigos é uma introdução à teoria IST (Internal Set Theory) de Nelson, uma abordagem axiomática à *matemática.

Outro ponto de partida muito acessível é [3]. Aí o leitor reencontrará os mesmos temas deste texto, muito mais desenvolvidos.

Há excelentes referências para quem queira aprofundar os seus conhecimentos de lógica. Entre elas conta-se [8], de leitura extraordinariamente agradável, sem no entanto deixar de assegurar o desenvolvimento desejável. Van Dalen refere vários exemplos de aplicação noutras áreas da matemática.

21 Por exemplo, o que é igualdade de objectos? Ou o que é um conjunto? Pior (porquê?!): o que é um subconjunto?

Porém, não pode duvidar-se que *o* livro de Análise Não-Standard é [7], onde Robinson discute quer a construção dos modelos não-standard, quer as aplicações em diversas áreas da matemática. Esta referência exige mais sofisticação que as anteriores. Em particular o segundo capítulo, em que a teoria de modelos não-standard é discutida, pode parecer muito estranho. Desde a utilização de ultrafiltros, à notação lógica usada, muitas coisas parecem conspirar para o tornar incompreensível. Porém, olhando com mais cuidado, *todas* as construções têm uma razão e um significado muito naturais. Em qualquer caso, o leitor que já chegou a este ponto pode provavelmente saltar esse capítulo numa primeira leitura.

Referências

[1] George Berkeley. "The Analyst, A Discourse Addressed to an Infidel Mathematician". 1734. James R. Newman (editor). *The World of Mathematics*, volume 1, pages 281–288. 1956.

[2] G. de l'Hospital. *Analyse des Infiniment Petits, Pour l'Intelligence des Lignes Courbes*. 1696.

[3] Augusto Franco de Oliveira. *Infinitesimais: Passado, Presente e Futuro*. Departamento de Matemática da Universidade de Évora, 1987.

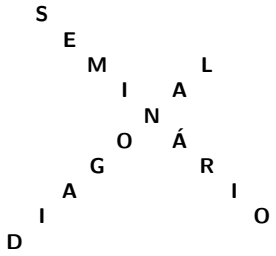
[4] Francine Diener and Marc Diener, editors. *Nonstandard Analysis in Practice*. Universitext. Springer, 1995.

[5] Leonhard Euler. *Introduction in Analysin Infinitorum*. 1748. John D. Blanton (trans.). *Introduction to Analysis of the Infinite*, book 1. Springer, 1988.

[6] Paul Howard and Jean E. Rubin. *Consequences of the Axiom of Choice*. AMS, 1998.

[7] Abraham Robinson. *Non-Standard Analysis*. Studies in Logic and the Foundations of Mathematics. North-Holland, 1966.

[8] D. van Dalen. *Logic and Structure*. Springer, 2nd ed., c.1985.



Computação Quântica

Alexandre P. Lourenço Francisco
4º ano da LMAC — Ciência da Computação
aplf@math.ist.utl.pt

Palavras Chave

computador quântico, computação reversível,
computação paralela, qubit, entanglement, factorização.

Resumo

Quando falamos em Computação Quântica e em Computadores Quânticos, todos pensamos em máquinas ultra-rápidas. No entanto não é apenas o tempo de processamento que está em causa. Ao nível da computação teórica têm ocorrido algumas surpresas, estudam-se novos algoritmos estruturalmente diferentes dos usuais e a complexidade parece não obedecer aos padrões clássicos. Contudo, para o utilizador comum, a surpresa maior virá a ocorrer aquando do primeiro computador quântico utilizável: os códigos criptográficos até então seguros serão facilmente quebrados. Teremos como objecto de discussão estes e outros pontos associados à Computação Quântica, tais como o processamento de informação, o hardware, os novos algoritmos e a sua complexidade.

1 Introdução

Os computadores quânticos foram discutidos primeiro por Benioff em 1980, no âmbito da simulação de uma máquina de Turing clássica por um sistema quântico. Benioff concluiu que os processos computacionais quânticos são pelo menos tão poderosos como os processos computacionais clássicos.

Em 1982 Feynman considerou o problema inverso, ou seja, até que ponto podem os computadores clássicos simular os sistemas quânticos. Feynman observou que os computadores clássicos sofrem invariavelmente uma desaceleração exponencial quando simulam sistemas quânticos, mas que os sistemas quânticos podem simular-se entre si sem que tal ocorra.

Coube no entanto a Deutsch, em 1985, notar que a sobreposição quântica poderia permitir a um sistema quântico em evolução realizar computações clássicas em paralelo, tendo Deutsch desenvolvido a máquina de Turing

quântica universal. Analisaremos adiante de que forma ocorre esse processamento paralelo.

Bernstein e Vazirani verificaram em 1993 que uma máquina de Turing quântica universal pode simular qualquer outra máquina de Turing quântica em tempo polinomial.

Entre 1994 e 1997, Simon mostrou que o computador quântico é exponencialmente mais rápido do que o computador clássico, mas apenas para alguns problemas. Neste mesmo período, Shor desenvolveu um algoritmo, destinado a correr num computador quântico, que realiza a factorização inteira em tempo polinomial.

Em 1995, Cirac e Zoller afirmaram que, com a tecnologia disponível no momento, era possível iniciar a construção de portas e circuitos básicos para o processamento quântico de informação.

Lloyd, em 1996, mostrou que a observação de Feynman estava correcta: os computadores quânticos podem efectivamente simular outros sistemas quânticos.

Em 1999 Lloyd e Braunstein verificaram que o processamento quântico pode ser realizado bastante bem quando trabalhamos com sistemas quânticos que correspondem a espaços de Hilbert de dimensão infinita.

É claro que muitos outros nomes poderiam ainda ser indicados, assim como os numerosos resultados conseguidos (ver [1]). Mas mesmo o breve resumo que fizemos permite constatar como é recente a Computação Quântica.

2 Computação Reversível

Com os computadores quânticos as computações serão efectuadas à escala do átomo e, quando trabalhamos a esta escala, surgem alguns problemas associados à construção e operação de tais sistemas. Actualmente o maior problema conhecido na construção dos computadores, que recorrem a uma maior miniaturização dos componentes, é a dissipação de calor.

Em 1961 Landauer estudou as limitações físicas colocadas pela dissipação de calor e, surpreendentemente, mostrou que todas as operações necessárias na computação são *reversíveis*,[1] excepto uma. A primeira condição para que qualquer sistema determinista seja reversível é que o input e o output se determinem univocamente entre si, ou seja, que haja *reversibilidade lógica*. Se, para além de ser logicamente reversível, também for possível fazer o sistema evoluir em ambas as direcções temos *reversibilidade física*. Portanto, a segunda lei da termodinâmica garante que não ocorre dissipação de calor.

1 Ou seja, sem consumo de energia e portanto sem dissipação de calor.

Qualquer computação razoável levada a cabo por um computador comum pode ser escrita em termos de expressões booleanas, e qualquer expressão booleana pode ser construída tendo por base um conjunto de portas ou conectivos lógicos. A um conjunto de conectivos que permita construir todas as expressões booleanas, por exemplo AND, OR e NOT, chamamos *conjunto universal*. De facto podemos considerar apenas dois conectivos, nomeadamente o AND e o NOT, ou o OR e o NOT. Também podemos considerar outros conectivos, por exemplo o XOR, e verificar que o XOR e o AND constituem um conjunto universal.

A	B	A AND B	A OR B	A XOR B	NOT B
0	0	0	0	0	1
0	1	0	1	1	0
1	0	0	1	1	1
1	1	1	1	0	0

Tabela 1: Tabela de Verdade dos Conectivos

No que diz respeito à reversibilidade, é fácil verificar que o AND, o XOR e o OR não são reversíveis, pois têm dois inputs mas apenas um output, não se podendo assim saber quais os inputs conhecendo apenas o output.

Por outro lado, as portas lógicas atrás indicadas, embora suficientes para expressar as expressões lógicas inerentes a uma computação, não o são para construir uma máquina. Um computador útil necessitará também das portas FANOUT e ERASE. A porta FANOUT permitirá duplicar o input, ou seja, dado um input devolve dois outputs iguais ao input, e a porta ERASE eliminará o conteúdo do registo de input, não devolvendo output.

Consideremos a porta FANOUT. Será reversível? Certamente não é destruída informação, portanto há pelo menos reversibilidade lógica, e Landauer mostrou que também se pode conseguir reversibilidade física. Podemos encontrar um modelo simples desta porta em [5].

A porta ERASE é necessária para limpar a memória do computador periodicamente. Um tipo de apagamento pode ser efectuado reversivelmente: se possuirmos uma cópia de alguma informação, podemos apagar as restantes cópias invertendo o FANOUT. Contudo temos um problema quando pretendemos eliminar a última cópia.

Landauer concluiu, tendo por base argumentos gerais inerentes à compressão do espaço de fases, que apagar um bit de informação a uma temperatura T requer a dissipação de uma energia maior ou igual a $k_B T \ln 2$.[2]

2 Aqui k_B é a constante de Boltzmann, cujo valor aproximado é $1{,}38\times10^{-23}\,\mathrm{J\,K^{-1}}$. Por exemplo, à temperatura ambiente $T = 20^{\circ}\mathrm{C} = 293{,}15\,\mathrm{K}$, a energia será pelo menos $2{,}81 \times 10^{-21}\,\mathrm{J}$.

Porém, a primitiva ERASE não é imprescindível. Landauer observou que este obstáculo pode ser ultrapassado recorrendo ao facto de qualquer função f poder ser tornada bijectiva se conservarmos uma cópia do input:

$$a \longmapsto (a, f(a))$$

Chegamos assim à solução desenvolvida por Toffoli e Fredkin para realizar computações reversíveis, conhecida como *porta de Toffoli* (Figura 1). Tendo

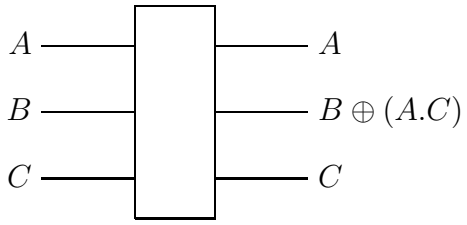


Figura 1: Porta de Toffoli

em conta que $\oplus$ e . correspondem respectivamente ao XOR e AND, observamos que com a porta de Toffoli podemos obter as portas lógicas usuais:

$$B \oplus (A.C) = \begin{cases} A.C & \text{se } B = 0 \quad \text{(AND)} \\ A \oplus B & \text{se } C = 1 \quad \text{(XOR)} \\ \overline{A} & \text{se } B = C = 1 \quad \text{(NOT)} \\ A & \text{se } B \neq C = 1 \quad \text{(FANOUT)} \end{cases}$$

Como notou Landauer, existe um pequeno problema neste procedimento. À medida que utilizamos cada vez mais portas maior é número de bits em "lixo" acumulado, pois a cada utilização temos de guardar o input para manter a reversibilidade. Bennett resolveu este problema mostrando que esse lixo pode ser eliminado através do uso de FANOUT. Para mais detalhes podemos ver [5].

Concluímos esta breve discussão acerca da reversibilidade indicando que em 1973 Bennett provou a existência de uma máquina de Turing reversível.

3 Alguns Fundamentos

Um exemplo de sistema quântico simples é aquilo a que em física se chama uma partícula de spin $\frac{1}{2}$.

O *espaço de estados* de um tal sistema é $\mathbb{C}^2$, o espaço de Hilbert de dimensão 2. $\mathbb{C}^2$ tem uma base ortonormada $\{(1,0),(0,1)\}$, cujos vectores costumam ser representados em mecânica quântica por $|0\rangle$ e $|1\rangle$. Qualquer

estado é uma combinação linear de $|0\rangle$ e $|1\rangle$, aquilo a que se chama habitualmente uma *sobreposição* de $|0\rangle$ e $|1\rangle$.

Um tal sistema é o correspondente quântico da noção clássica de sistema com dois estados. Há dois estados "especiais", nomeadamente $|0\rangle$ e $|1\rangle$, mas além desses há uma infinidade de outros possíveis (todas as combinações lineares). Um exemplo clássico de um sistema com dois estados é uma célula de memória cujos valores podem ser apenas 0 ou 1, vulgarmente designada *bit*. Por isso o correspondente sistema quântico é designado *quantum bit*, ou *qubit*.

Na verdade apenas interessa considerar combinações lineares $\alpha|0\rangle + \beta|1\rangle$ com $|\alpha|^2 + |\beta|^2 = 1$, o que significa que o verdadeiro espaço de estados pode ser identificado com a esfera $\mathbb{S}^3$. O significado físico de um tal estado é o seguinte: depois de fazer uma medição sobre o sistema, em que por *medição* se entende um processo destinado a averiguar se o estado do sistema é $|0\rangle$ ou $|1\rangle$, o estado é necessariamente um destes dois, sendo $|\alpha|^2$ a probabilidade de obter $|0\rangle$ e $|\beta|^2$ a probabilidade de obter $|1\rangle$.[3] Esta interpretação mantém-se para qualquer outra medição relativa a outra base ortonormada; isto é, se $\{|0'\rangle, |1'\rangle\}$ for outra base e o mesmo vector for da forma $\alpha'|0'\rangle + \beta'|1'\rangle$ nesta nova base, então uma medição em relação a esta base tem probabilidade $|\alpha'|^2$ de terminar no estado $|0'\rangle$ e $|\beta'|^2$ de terminar no estado $|1'\rangle$. Mas *há sempre apenas dois estados finais possíveis*, o que justifica que este seja considerado um sistema "com dois estados".

O processamento ideal da informação quântica é então levado a cabo em espaços de Hilbert, que correspondem ao conjunto de estados de sistemas quânticos isolados. A evolução do sistema quântico é traduzida pela sucessiva aplicação de operadores *unitários* ao estado inicial do sistema, onde os operadores unitários não são mais que transformações lineares representadas por matrizes unitárias.

Generalizando, o espaço de estados de um sistema de N qubits é $\mathbb{C}^{2^N}$, onde agora os vectores da base canónica se representam por sequências em $\{0,1\}^N$:

$$|00\ldots0\rangle,\ |10\ldots0\rangle,\ |01\ldots0\rangle,\ |11\ldots0\rangle,\ \ldots,\ |11\ldots1\rangle.$$

Também é comum usar notações como $|x\rangle|y\rangle$ ou $|x\rangle \otimes |y\rangle$ para representar $|xy\rangle$.

Na realidade, o processamento da informação quântica é efectuado num ambiente com muito ruído onde a interacção e a descoerência[4] dominam, não

3 Desta forma, do ponto de vista da medição, tanto faz usar $|0\rangle$ ou $-|0\rangle$, já que as probabilidades serão iguais.

4 Iremos analisar e exemplificar o que se entende por interacção e descoerência numa das próximas secções.

sendo possível trabalhar com um sistema quântico isolado. Deste modo os elementos básicos para trabalhar ao nível da realidade são mais complexos. Para mais detalhes consultar, por exemplo, [2].

4 Portas Lógicas para Qubits

Tendo em conta o que foi dito, iremos agora ver quais os operadores a considerar para construir a porta de Toffoli. Porém veremos em primeiro lugar portas mais simples e esclareceremos alguns pontos importantes.

Começando com um único qubit, representaremos os estados $|0\rangle$ e $|1\rangle$ como $\begin{pmatrix}1\\0\end{pmatrix}$ e $\begin{pmatrix}0\\1\end{pmatrix}$, respectivamente. Temos também que o operador

$$R_\theta = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}, \quad \theta \in \mathbb{R},$$

é unitário e que

$$R_{\frac{\pi}{2}}|0\rangle = |1\rangle \quad \text{e} \quad R_{\frac{\pi}{2}}|1\rangle = -|0\rangle,$$

o que corresponde a uma porta que recebe um qubit e devolve o qubit obtido do primeiro por aplicação do operador.[5] Outra importante porta que recebe apenas um qubit é $R_{\frac{\pi}{4}}$, a qual dado um qubit no estado $|0\rangle$ devolve um qubit na sobreposição intermédia

$$R_{\frac{\pi}{4}}|0\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}.$$

Consideremos agora um sistema com duas partículas, isto é, com dois qubits. Escrevendo os estados base como vectores temos:

$$|00\rangle = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}; \quad |10\rangle = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}; \quad |01\rangle = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}; \quad |11\rangle = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}.$$

Podemos agora representar o operador XOR, o qual recebe dois qubits e devolve também dois qubits, pela matriz

$$U_{\text{XOR}} = \begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&0&1\\0&0&1&0\end{pmatrix}.$$

5 Recorde-se que, do ponto de vista da medição, $|0\rangle$ e $-|0\rangle$ são estados equivalentes.

É importante notar que a primeira partícula ou qubit não sofre alteração, ficando a segunda com o resultado da operação XOR. Talvez seja interessante comparar os resultados da aplicação do operador aos estados clássicos com os indicados na Tabela 1. Indicamos na Figura 2 o circuito quântico correspondente a este operador, onde $\bigoplus$ representa $\oplus$. O circuito lê-se da

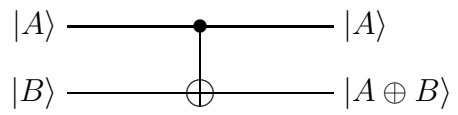


Figura 2: Circuito Quântico para o Operador XOR

seguinte forma: dado como input um qubit com valor A e um qubit com valor B, temos como output o primeiro qubit inalterado e o segundo, antes com valor B, agora com valor $A \oplus B$.

Como exemplo de aplicação sucessiva do operador XOR, consideremos o circuito da Figura 3, que troca os estados de dois qubits, e mostra como mover informação de um qubit para outro.

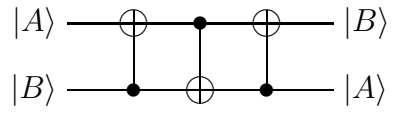


Figura 3: Circuito Quântico que Troca Informação entre Dois Qubits

Estamos agora em condições de indicar um circuito correspondente à porta de Toffoli. Aplicando os operadores U_{XOR}, $R_{-\frac{\pi}{8}}$ e $R_{\frac{\pi}{8}}$ vistos anteriormente, chegamos ao circuito da Figura 4. Notamos que, embora a porta

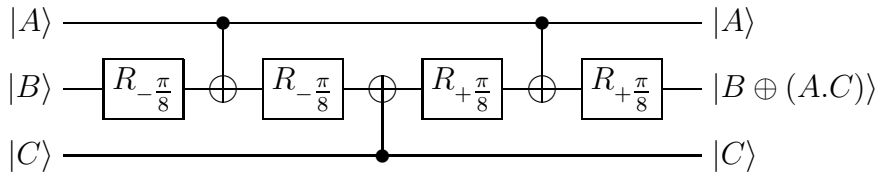


Figura 4: Porta de Toffoli

de Toffoli receba três qubits, apenas precisamos de um operador que recebe dois qubits e de um outro operador que recebe apenas um.

Devemos também indicar que não só é possível construir todas as portas lógicas de um computador quântico, como foi feito em relação à porta de Toffoli, mas que o operador XOR permite ainda construir operadores unitários para qualquer número finito de qubits. Para mais detalhes ver por exemplo [5] ou [2].

5 Paralelismo Quântico

Como observámos na Introdução, Deutsch notou que a sobreposição quântica poderia permitir a um computador quântico efectuar computações clássicas em paralelo. Iremos então agora ver como tal acontece.

A sobreposição confere aos qubits uma propriedade extraordinária. Como já referimos, um bit clássico ou é 0 ou é 1. Contudo, um qubit tem os dois estados clássicos e um número infinito de sobreposições. Numa primeira observação isto pode parecer irrelevante, mas na realidade é como se estivéssemos a codificar um bit clássico com o valor 0 e 1 em simultâneo. Logo, quando aplicamos um operador a um destes bits realizamos duas operações de uma vez só e estamos portanto perante uma forma de *paralelismo quântico*.

Consideremos um exemplo para melhor compreender esta propriedade. Suponhamos uma caixa negra que calcula uma função $f : \{0,1\} \to \{0,1\}$; da qual não sabemos nada a não ser que demora 24 horas a calcular o output. Sabemos no entanto que existem quatro possibilidades, pois $f(0)$ e $f(1)$ podem ser um de dois valores.

Para sabermos o resultado da função temos que determinar $f(0)$ e $f(1)$, o que nos fará esperar 48 horas. Mesmo se não temos tanto tempo e nos basta saber se f é constante ou não, temos ainda assim que esperar 48 horas, pois precisamos dos dois valores para os poder comparar.

Suponhamos agora que a nossa caixa negra é um dispositivo quântico. Evidentemente f poderá não ser invertível, mas a acção de um computador quântico é sempre invertível, pois corresponde a uma transformação unitária. Para realizarmos a computação de f reversivelmente precisamos então de uma transformação U_f que recebe dois qubits e devolve também dois qubits:

$$U_f : |x\rangle|y\rangle \mapsto |x\rangle|y \oplus f(x)\rangle,$$

onde a transformação muda o segundo qubit se $f(x) = 1$ e não muda se $f(x) = 0$.

Podemos determinar se f é constante ou não utilizando o dispositivo duas vezes, mas isso continuará a demorar 48 horas. Será possível saber se f é constante com apenas uma utilização do dispositivo? Esta questão corresponde ao problema de Deutsch.

Como a caixa negra é um computador quântico podemos escolher o input como sendo uma sobreposição de $|0\rangle$ e de $|1\rangle$. Se o segundo qubit for

preparado inicialmente no estado $|-\rangle = \frac{1}{\sqrt{2}}|0\rangle - \frac{1}{\sqrt{2}}|1\rangle$, temos que

$$\begin{aligned} U_f\big(|x\rangle \otimes |-\rangle\big) &= U_f\left(|x\rangle \otimes \frac{|0\rangle - |1\rangle}{\sqrt{2}}\right) \\ &= |x\rangle \otimes \frac{|f(x)\rangle - |1 \oplus f(x)\rangle}{\sqrt{2}} \\ &= |x\rangle \otimes (-1)^{f(x)} \frac{|0\rangle - |1\rangle}{\sqrt{2}} \\ &= (-1)^{f(x)}\ |x\rangle \otimes |-\rangle, \end{aligned}$$

isolando desta forma a função $f(x)$ no valor do sinal. Suponhamos agora que preparamos o primeiro qubit no estado $|+\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$. Então a caixa negra actua da seguinte forma:

$$U_f\big(|+\rangle \otimes |-\rangle\big) = \frac{(-1)^{f(0)}|0\rangle + (-1)^{f(1)}|1\rangle}{\sqrt{2}} \otimes |-\rangle.$$

Efectuamos agora uma medição que projecta o primeiro qubit na base $\{|+\rangle, |-\rangle\}$, obtendo, a menos de sinal, $|+\rangle$ se a função f é constante e $|-\rangle$ caso contrário.

É possível trabalhar de modo semelhante com funções de N bits.

Importa contudo referir que, no caso de pretendermos determinar todos os valores da função, o computador quântico não trará vantagens sobre um computador clássico. Como vimos, no estado final temos codificados todos os valores da função. Porém, quando o observamos ocorre a projecção num dos subespaços de forma aleatória, por exemplo no subespaço de $|x\rangle|f(x)\rangle$. Uma vez ocorrida a projecção é-nos impossível observar qualquer outro resultado: no exemplo anterior não teríamos possibilidade de observar $|y\rangle|f(y)\rangle$ para qualquer outro $y \neq x$.

6 Entrelaçamento

Os qubits têm uma outra propriedade ainda mais estranha, mas que poderá ser muito útil. Imaginemos um processo físico que emite dois fotões, um para a direita e outro para a esquerda, tal que os dois fotões têm polarizações opostas. Até ser detectada, a polarização de cada um deles é indeterminada. Porém, no instante em que uma pessoa observe a polarização de um dos fotões, o estado de polarização do outro fica univocamente determinado, não interessando a que distância estão um do outro.

Este fenómeno permite aos sistemas quânticos desenvolver uma espécie de ligação designada por *entanglement* ou *entrelaçamento*. Poderíamos mesmo suspeitar de enormes vantagens no que diz respeito a comunicações, pois uma partícula pode determinar univocamente o estado de uma outra partícula instantaneamente e a qualquer distância.

Na experiência anterior não foi transmitida qualquer energia e no entanto parece ter sido transmitida alguma informação entre os dois fotões, ou seja, parece que entramos em contradição com a teoria da relatividade restrita, segundo a qual nada pode viajar com uma velocidade superior à da luz [8].

Todavia, tendo por base esta propriedade dos sistemas quânticos, alguns grupos de investigação afirmam a possibilidade do *teletransporte quântico*, tendo já apresentado alguns resultados práticos. Podemos consultar o trabalho de um desses grupos em [6].

Até recentemente o teletransporte não tinha sido levado a sério, porque pensava-se que o Princípio de Incerteza de Heisenberg limitaria a sua aplicabilidade. Segundo este princípio verificamos que ao recolher informação de um sistema quântico, quanto maior for a precisão procurada maior será a interacção com o próprio sistema, e logo passamos a ter um sistema diferente do inicial. Desta forma, podemos chegar a um ponto em que o sistema está completamente alterado, sem que tenhamos informação suficiente para realizar uma cópia perfeita.

Porém, alguns cientistas encontraram uma maneira de contornar este obstáculo, recorrendo a uma célebre propriedade da mecânica quântica, o efeito Einstein–Podolsky–Rosen (EPR). Na prática descobriram uma maneira de obter alguma informação do objecto A a teletransportar, transmitindo com o efeito EPR a informação desconhecida para um outro objecto C que nunca esteve em contacto com A. Posteriormente, aplicando a C um tratamento dependente da informação obtida, é possível colocar C no mesmo estado que A, conseguindo-se assim uma cópia exacta de A. Importa referir ainda que A não permanece no seu estado original devido à leitura de informação realizada. Podemos assim falar em teletransporte como desmaterialização num local e materialização em outro local, não ocorrendo replicação.

Para se conseguir utilizar o efeito EPR como referimos, temos de em primeiro lugar possuir um objecto intermediário B que interage primeiro com C e depois com A. Quando C interage com B, ficam entrelaçados. Depois fazemos o objecto B interagir com A e recolhemos parte da informação do sistema resultante interacção. Como referimos acima, esta leitura de A e de B determina univocamente o estado de C, devido ao seu entrelaçamento com B. Desta forma conseguimos transmitir a informação que faltava obter de A, pois este estava em interacção com B.

Porque não utilizar apenas o efeito EPR para transmitir toda a informação? Um facto bem conhecido é que não é possível explorar o efeito EPR para esse fim, pois não consegue por si só transmitir uma mensagem de forma controlável; embora consiga transmitir exactamente a informação que não conseguimos obter ao fazer a leitura do objecto.

Podemos encontrar alguns desenvolvimentos interessantes em [6] e [7].

7 Descoerência Quântica e Correcção de Erros

Um computador quântico explorará os estados entrelaçados. Porém estas correlações são extremamente frágeis e tendem a desaparecer muito rapidamente. O problema é que o nosso sistema quântico está em contacto com um sistema muito maior, o ambiente envolvente. Como é virtualmente impossível isolar um sistema completamente, este tende a estabelecer correlações não locais com o exterior.

Eventualmente, a informação quântica que codificamos num computador quântico tende a estar codificada nas correlações do computador com o ambiente. Logo se quisermos observar o sistema temos também de observar o ambiente, ou seja, na prática a informação passa a estar irremediavelmente perdida. A este fenómeno de introdução de erros na informação presente num computador quântico por parte do ambiente chamamos *descoerência*.

Contudo a descoerência não é o nosso único problema. Mesmo que consigamos um isolamento perfeito em relação ao ambiente, não podemos esperar que um computador quântico funcione na perfeição. As portas lógicas quânticas efectuam transformações unitárias e o protocolo de execução não virá a ser infalível da mesma forma que num computador clássico ocorrem erros no processamento, pelo que ao fim de algum tempo terão ocorrido vários erros no processamento. Por outro lado a acumulação de erros é ainda mais catastrófica na computação reversível, pois eliminar erros significa deitar fora informação e temos um processo dissipativo, logo não reversível.

Tem vindo a ser desenvolvida toda uma maquinaria para lidar com os erros: a teoria de correcção de erros. O que se pretende é aplicar essa mesma teoria aos computadores quânticos. Existem no entanto algumas dificuldades, estamos a lidar com vários tipos de erros possíveis que não surgem nos computadores clássicos. Por outro lado a medição ou observação do sistema causa distúrbios no mesmo. Verificamos também o designado Teorema da Não-Clonagem, ou seja, um estado quântico não pode ser copiado na perfeição.

8 Algoritmo de Factorização de Shor

Hoje em dia utilizamos métodos criptográficos em qualquer transacção na Internet que pretendamos segura. O esquema mais conhecido e utilizado na maior parte das aplicações é o RSA. Uma análise detalhada deste algoritmo pode ser encontrada no artigo do Tiago Reis [9]; apenas nos interessa reter que os melhores métodos para o quebrar recorrem à factorização de inteiros. Logo a segurança do RSA depende da dificuldade de factorizar números muito grandes.

Porém, Peter Shor da AT&T desenvolveu um algoritmo de factorização para correr num computador quântico em $O\big((\log N)^3\big)$ passos, um número muito inferior ao conseguido pelo melhor algoritmo desenvolvido para correr num computador clássico, $O\big(\exp\big(4 \cdot 9^{-1/3}(\ln N)^{1/3}(\ln \ln N)^{2/3}\big)\big)$. Por exemplo, a factorização de um número com 400 algarismos levará cerca de 10^{10} anos no computador clássico mais rápido da actualidade, porém num computador quântico levará menos de 3 anos.

Se quisermos factorizar um número N basta-nos encontrar um dos factores e reduzir o problema a um mais simples. Em primeiro lugar seleccionamos um número x, e utilizando o algoritmo de Euclides podemos determinar os factores comuns a x e N, reduzindo assim o nosso problema. Assumamos que escolhemos x tal que x e N são coprimos e, consideremos as potências x^n para n natural:

$$1,\ x,\ \ldots,\ x^{r-1},\ x^r,\ x^{r+1},\ \ldots;$$

e calculando $x^n \pmod N$:

$$1,\ x,\ \ldots,\ x^{r-1},\ 1,\ x,\ \ldots,\ 1,\ x,\ \ldots.$$

O número r é o expoente da primeira potência não trivial de x tal que $x^r \equiv 1 \pmod N$ e, observando mais atentamente, verificamos que a sequência acima é periódica com período r. Utilizando os algoritmos standard é difícil encontrar o período de uma sequência longa; porém, com um computador quântico o período é calculado eficientemente. Para detalhes podemos consultar [5].

Suponhamos que obtemos o período r usando um computador quântico. Se r for par continuamos o algoritmo, se for ímpar escolhemos outro x e recomeçamos.

Prosseguindo com x tal que o período r é par, reescrevemos a expressão $x^r \equiv 1 \pmod N$ como a diferença de dois quadrados:

$$(x^{r/2})^2 - 1 \equiv 0 \pmod N,$$

a qual podemos ainda reescrever como

$$(x^{r/2}+1)(x^{r/2}-1) \equiv 0 \pmod{N}.$$

Esta última expressão diz-nos que o produto dos dois termos da esquerda é múltiplo de N, o qual queremos factorizar. Como um destes termos tem pelo menos um factor comum com N, o último passo do algoritmo consiste em calcular o maior divisor comum entre os dois termos e N individualmente. É importante notar que existem algoritmos eficientes para determinar o máximo divisor comum entre dois números; neste caso qualquer divisor comum não trivial é um dos factores pretendidos.

O código respeitante ao algoritmo a correr num computador quântico pode ser encontrado por exemplo em [5], onde se utilizam apenas ferramentas e propriedades referidas nas secções anteriores.

9 Criptografia Quântica

Como vimos, a criptografia actual parece não vir a garantir a segurança necessária quando trabalharmos com computadores quânticos. Porém em simultâneo com os computadores quânticos surge a *criptografia quântica*.

Suponhamos que duas pessoas pretendem comunicar entre si de forma segura e têm acesso a um canal de fibra óptica. O emissor pode enviar qubits cujos estados são escolhidos aleatoriamente de um conjunto de estados não ortogonais, por exemplo $\{|0\rangle, |1\rangle, |+\rangle, |-\rangle\}$. O receptor por sua vez pode observar os qubits na base $\{|0\rangle, |1\rangle\}$ ou na base $\{|+\rangle, |-\rangle\}$, escolhendo aleatoriamente uma delas.

Utilizando um canal clássico e um protocolo especial, o emissor e o receptor podem verificar os qubits em que concordam; estes fornecem uma chave pública e coincidem em 50% das vezes. Por outro lado se a linha for espiada, o espião não pode adquirir informação pois não sabe em que base são transmitidos os estados, apercebendo-se o receptor e o emissor, em cada observação efectuada, dos distúrbios produzidos nos estados. Estes podem mesmo enviar alguns qubits extra para verificarem a segurança da linha e garantirem que o espião não adquire qualquer informação.

Existem inúmeros protocolos desenvolvidos neste sentido, onde muitas vezes se utilizam também as propriedades oferecidas pelo entrelaçamento, por exemplo o teletransporte.

Importa também referir que num computador quântico temos um verdadeiro gerador de números aleatórios, pois dado um estado quântico o resultado da sua observação é perfeitamente aleatório. Contrariamente, num

computador clássico nunca conseguimos na realidade obter a aleatoriedade e utilizamos em vez disso geradores de números pseudo-aleatórios.

Recorrendo então às propriedades físicas e à mecânica quântica conseguimos novas formas de comunicar com tanta ou mais segurança que actualmente; como vimos é possível gerar de forma incondicional chaves seguras.

10 Complexidade

Os resultados de Feynman referidos na Introdução implicam que a noção de *computável* é a mesma que na computabilidade clássica. No entanto a complexidade dos algoritmos parece não ser a mesma e temos como exemplo o algoritmo de Shor que é polinomial, ao contrário dos algoritmos clássicos exponenciais.

Todavia não podemos concluir que todos os algoritmos sofrem uma desaceleração exponencial. Consideremos por exemplo o algoritmo de procura de Groover, que consiste em procurar numa lista não ordenada um dado elemento, o que classicamente ocorre num número de passos da ordem de $O(N)$, onde N é o número de elementos da lista.

O algoritmo de Groover por sua vez consegue em média encontrar o elemento em $\Theta(\sqrt{N})$ passos, em que para tal é associado a cada elemento um estado clássico de um dado sistema quântico e, preparado o estado inicial numa sobreposição com distribuição uniforme das probabilidades para cada elemento. O restante algoritmo consiste em conseguir que a probabilidade associada ao elemento procurado seja superior e logo, quando efectuamos a observação, este estado tem maior probabilidade de ser observado.

A computação quântica leva assim a uma nova classificação dos algoritmos quanto à sua complexidade, embora não ocorra sempre um desaceleramento exponencial. Quanto à classe de funções computáveis verificamos que é a mesma, pois não existe uma função não computável classicamente que seja computável num computador quântico.

11 Hardware

Para construir um computador quântico há que satisfazer alguns requisitos básicos. Temos de ser capazes de armazenar qubits por longos períodos de tempo para que seja completada uma computação razoável. Os qubits têm de ser isolados do ambiente para minimizar os erros devidos à descoerência. Precisamos de medir os qubits eficientemente. Tem de ser possível manipular os estados quânticos dos qubits individualmente, para que seja possível

construir as portas lógicas. Estas têm de ser implementadas com precisão para que o computador funcione correctamente.

Têm sido realizadas tentativas com redes de iões, cavidade QED e ressonância magnética nuclear (NMR). Não iremos entrar em detalhes, referindo apenas que estão a ser utilizadas inúmeras novidades tecnológicas, como por exemplo lasers para excitar os átomos. Para mais detalhes podemos consultar [7].

Até hoje as experiências realizadas têm fornecido bons resultados e os sucessos são animadores. Porém o número de qubits que se consegue manipular é muito pequeno não sendo suficiente para levar a cabo qualquer computação interessante. Espera-se vir a possuir um computador quântico funcional nos próximos dez anos.

12 Conclusão

A Computação Quântica irá permitir uma nova classificação da complexidade, fundada nas leis físicas — contrariamente à classificação clássica. O nosso maior problema, os erros quânticos, parece ter um fim à vista com bons resultados a surgirem. A construção física de um computador quântico é a cada dia que passa uma realidade mais próxima de ser alcançada, estamos a assistir a avanços significativos.

Também já existem inúmeros algoritmos desenvolvidos que nos permitem inferir o sucesso de tais máquinas; podemos encontrar código para máquinas quânticas em muitas das referências indicadas.

Terminamos aqui este artigo tendo consciência de que não explorámos os assuntos apresentados na totalidade e, que nos falta explorar muitos outros. Esperamos contudo que as referências dadas possam esclarecer questões que advenham da leitura deste artigo.

13 Agradecimentos

À organização do Seminário Diagonal agradeço o convite bem como a oportunidade de apresentar o seminário. Aos professores Francisco Coelho, José Félix Costa e Pedro Resende agradeço a ajuda na preparação do seminário, a qual foi indispensável. Ao Luís Cruz-Filipe e ao João Boavida o meu obrigado pelas indicações e sugestões ao longo da preparação. Agradeço ainda à minha namorada, Cátia Vaz, pela paciência e ajuda na preparação do seminário e deste artigo.

Obrigado a todos.

Referências

[1] B. Engquist and W. Schmid (eds.), *Mathematics Unlimited: 2001 and Beyond*, Springer-Verlag, 2000.

[2] Jozef Gruska, *Quantum Computing*, McGraw-Hill, June 1999.

[3] P. W. Shor, Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer, *SIAM J. Computing*, 26, pp. 1484–1509 (1997).

[4] Neil Gershenfeld and Isaac L. Chuang, Quantum computing with molecules, *Scientific American*, June 1998.
url: `http://www.sciam.com/1998/0698issue/0698gershenfeld.html`

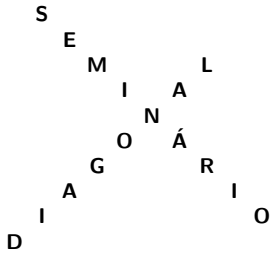
[5] Samuel L. Braunstein, Quantum computation, *Encyclopedia of Applied Physics*, Wiley-VCH, Berlin, 1999.

[6] *Quantum Information and Information Physics at IBM Research Yorktown.*
url: `http://www.research.ibm.com/quantuminfo/teleportation/`

[7] John Preskill, *Course Information for Physics 219/Computer Science 219, Quantum Computation (Formerly Physics 229), 2000–01*, Caltech Particle Theory Group, California Institute of Technology.
url: `http://theory.caltech.edu/~preskill/ph229/`

[8] Patrícia Engrácia, Grupos, variedades e relatividade, neste volume.

[9] Tiago Reis, Criptografia e jogos por telefone, neste volume.



Criptologia; Contratos e Dinheiro Virtuais

Pedro Miguel Adão
4º ano da LMAC — Ciência da Computação
pad@math.ist.utl.pt

Palavras Chave

sistema criptográfico, chave pública, chave privada, assinatura digital, dinheiro virtual.

Resumo

Quando queríamos guardar alguma coisa usávamos os cofres; quando queríamos que uma carta chegasse ao destino sem ser aberta, usávamos lacre; quando queríamos garantir que um destinatário recebia uma carta, enviávamo-la com aviso de recepção.

Hoje em dia, no mundo em que vivemos, será possível ter segurança? Podemos ter um cofre na Internet para guardar dinheiro virtual? Podemos assinar documentos virtuais sem que ninguém falsifique a nossa assinatura? Podemos enviar *e-mails* lacrados? Podemos enviar *e-mails* com aviso de recepção?

Estas e outras questões são o tema deste artigo.

1 Introdução

Todos nós já ouvimos falar de criptografia e codificação de mensagens. Muitos de nós até já usámos aquilo a que chamaremos sistemas criptográficos nos nossos jogos de crianças. Por isso este artigo não será uma explicação do que é a criptografia mas sim uma formalização de alguns conceitos relacionados com este tema.

Começaremos então por definir o que é um *sistema criptográfico* e em seguida falaremos de criptografia de chave privada dando alguns exemplos. Na secção 3 trataremos da cripto-análise de algumas dessas cifras, i.e., dada uma mensagem codificada, tentar descobrir qual é a chave que está a ser utilizada.

A secção 4 é dedicada à criptografia de chave pública e exemplos da sua utilização. Por fim, a última secção é uma tentativa de criação de um

protocolo para a implementação de dinheiro virtual para fazer compras *on-line*. Vamos então começar por formalizar alguns conceitos e notação que vamos utilizar ao longo deste artigo.

1.1 Conceitos e Notação

Como o objectivo principal da criptografia é permitir que duas pessoas comuniquem por um canal inseguro, por exemplo uma linha telefónica ou uma rede de computadores, de tal forma que um intruso não consiga perceber o que está a ser dito, vamos precisar de dois comunicadores e de um intruso. Vamos chamar aos comunicadores *A*na e *B*runo e ao intruso, que vai tentar interceptar a comunicação, *C*arlos.

À mensagem original chamaremos *mensagem* e representaremos por letras minúsculas e à mensagem que a Ana envia de facto chamaremos *mensagem codificada* e será representada por letras maiúsculas.

Vamos então começar por ver como é que os comunicadores podem codificar e descodificar as mensagens.

DEFINIÇÃO 1. Um *sistema criptográfico* é um quíntuplo $\langle X, Y, K, E, D\rangle$ em que

- X é o conjunto das *mensagens*, finito;
- Y é o conjunto das *mensagens codificadas*, finito;
- K é o conjunto de todas as chaves possíveis, finito;
- $E = \{e_k : X \to Y\}_{k\in K}$ é o conjunto das funções de codificação;
- $D = \{d_k : Y \to X\}_{k\in K}$ é o conjunto das funções de descodificação;

tal que qualquer que seja $k \in K$

$$\forall x \in X \quad d_k(e_k(x)) = x. \tag{1}$$

A Ana, para codificar uma mensagem x com a chave k_1, faz $e_{k_1}(x)$. O Bruno, para descodificar uma mensagem y com a chave k_2, faz $d_{k_2}(y)$.

Vemos facilmente que e_k tem de ser injectiva, o que sai directamente de (1). Se assim não fosse a descodificação poderia ser ambígua. Suponhamos que existiam duas mensagens x e x' tais que $e_k(x) = e_k(x') = y$. Como deveria o Bruno descodificar y? Como x ou como x'?

Também temos que decidir o que é o conjunto X. Podemos pensar em X, conjunto das mensagens, como sendo todas as palavras do dicionário, ou podemos pensar em X como sendo o conjunto das letras do alfabeto. Neste

último caso teríamos de definir o que seria $e_k(x_1x_2\ldots x_n)$, ou seja, como é que codificaríamos as palavras. Podemos ainda considerar X um conjunto de números e a cada letra associar um número, $\mathtt{a} \mapsto 0$, $\mathtt{b} \mapsto 1$, $\ldots$, $\mathtt{z} \mapsto 25$. Mais à frente veremos exemplos em que usamos esta técnica.

Um dos objectivos é que os nossos sistemas criptográficos sejam práticos e fáceis de utilizar. Para isso é necessário que as funções e_k e d_k sejam facilmente computáveis.

Outro objectivo é que o sistema criptográfico seja seguro: ao ver uma mensagem codificada y, o Carlos não deve conseguir descobrir $k \in K$ e consequentemente a mensagem original x. Vamos então ver alguns exemplos de sistemas criptográficos.

2 Criptografia de Chave Privada

A criptografia de chave privada é uma forma de codificação em que os dois comunicadores, a Ana e o Bruno, escolhem uma chave $k \in K$ quando estão juntos, ou quando têm um canal seguro à disposição para as suas comunicações. O protocolo de comunicação é muito simples: se a Ana quiser enviar uma mensagem x ao Bruno usa a função e_k para codificar a mensagem, $e_k(x) = y$, e o Bruno de seguida usa a função d_k para descodificar a mensagem codificada, $d_k(y) = d_k(e_k(x)) = x$.

$$\boxed{A \xrightarrow{y=e_k(x)} B}$$

Como o Carlos não conhece a chave, não consegue descodificar as mensagens. Na próxima secção veremos como é que o Carlos pode tentar descobrir as chaves utilizadas pela Ana e pelo Bruno, técnica que é denominada cripto-análise.

Vamos então ver algum exemplos de cifras. Começamos pela mais antiga: a *cifra de translação*.[1]

Definição 2 (Cifra de Translação). Seja $X = Y = K = \mathbb{Z}_m$. Para cada $k \in K$, $0 \leq k \leq m$ define-se[2]

$$e_k(x) = x + k \bmod m;$$
$$d_k(y) = y - k \bmod m.$$

Podemos usar o número m que quisermos dependendo do número de caracteres que queremos codificar. Normalmente esta cifra é usada com $m = 26$

1 Shift cipher.

2 Quando usamos $k = 3$ esta cifra tem o nome de *cifra de César* precisamente por ter sido usada pelo imperador romano na codificação das suas mensagens.

pois são os caracteres do alfabeto. Não é costume codificar os espaços e outros caracteres, pois se os usarmos facilitamos a cripto-análise.

Vamos usar a convenção de a cada letra associar um número de acordo com a Tabela 1.

A	B	C	D	E	F	G	H	I	J	K	L	M
0	1	2	3	4	5	6	7	8	9	10	11	12

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Tabela 1: Conversão entre Números e Letras

Esta cifra é muito simples de utilizar. Se quisermos codificar uma mensagem $x_1x_2\ldots x_n$ fazemos $e_k(x_1x_2\ldots x_n) = e_k(x_1)e_k(x_2)\ldots e_k(x_n)$.

Exemplo 1. Usando $k = 4$ a mensagem `ola` será codificada como

$$\begin{array}{lll} \texttt{o} \mapsto 14 & e_4(\texttt{o}) \mapsto e_4(14) = 14+4 \equiv 18 \bmod 26 & 18 \mapsto \texttt{S} \\ \texttt{l} \mapsto 11 & e_4(\texttt{l}) \mapsto e_4(11) = 11+4 \equiv 15 \bmod 26 & 15 \mapsto \texttt{P} \\ \texttt{a} \mapsto \ 0 & e_4(\texttt{a}) \mapsto e_4(0) \ = \ 0+4 \equiv \ 4 \bmod 26 & 4 \mapsto \texttt{E} \end{array}$$

dando origem à mensagem codificada `SPE`.

A *cifra afim*[3] é uma generalização da cifra de translação:

DEFINIÇÃO 3 (CIFRA AFIM). Seja $X = Y = \mathbb{Z}_m$ e $K = \{(a,b) \in \mathbb{Z}_m \times \mathbb{Z}_m : \gcd(a,m) = 1\}$.[4] Para cada $(a,b) \in K$ define-se

$$\begin{aligned} e_k(x) &= ax + b \bmod m; \\ d_k(y) &= a^{-1}(y-b) \bmod m. \end{aligned}$$

Exemplo 2. Vamos codificar de novo a mensagem do exemplo anterior usando $k = (5,3)$ como chave,

$$\begin{array}{lll} \texttt{o} \mapsto 14 & e_{(5,3)}(\texttt{o}) \mapsto e_{(5,3)}(14) = 5\times 14+3 \equiv 21 \bmod 26 & 21 \mapsto \texttt{V} \\ \texttt{l} \mapsto 11 & e_{(5,3)}(\texttt{l}) \mapsto e_{(5,3)}(11) = 5\times 11+3 \equiv \ 6 \bmod 26 & 6 \mapsto \texttt{G} \\ \texttt{a} \mapsto \ 0 & e_{(5,3)}(\texttt{a}) \mapsto e_{(5,3)}(0) \ = 5\times \ 0+3 \equiv \ 3 \bmod 26 & 3 \mapsto \texttt{D} \end{array}$$

obtendo-se assim a mensagem codificada `VGD`.

3 Affine cipher.

4 K não é definido como sendo $\mathbb{Z}_m \times \mathbb{Z}_m$ pois é necessário existir a^{-1} para que a equação esteja bem definida e isto é verificado precisamente quando $\gcd(a,m) = 1$.

A cifra de translação é um caso particular desta fazendo $a = 1$ e $b = k$. Uma das cifras mais usadas em puzzles cripto-numéricos de jornais e revistas é a cifra de que vamos falar agora, a *cifra de substituição*.[5]

Definição 4 (Cifra de Substituição). Seja $X = Y = \mathbb{Z}_m$ e $K = \{\pi : \pi \text{ é uma permutação sobre } \mathbb{Z}_m\}$. Para cada permutação $\pi \in K$ define-se

$$e_k(x) = \pi(x);$$
$$d_k(y) = \pi^{-1}(y).$$

Exemplo 3. Vamos codificar de novo a mensagem `ola` usando a seguinte permutação π:

a	b	c	d	e	f	g	h	i	j	k	l	m
Y	G	N	S	U	H	Q	A	O	I	W	C	J

n	o	p	q	r	s	t	u	v	w	x	y	z
T	M	R	X	D	L	Z	E	V	F	K	B	P

$$e_\pi(\texttt{o}) = \texttt{M}, \quad e_\pi(\texttt{l}) = \texttt{C}, \quad e_\pi(\texttt{a}) = \texttt{Y}$$

i.e., a codificação da mensagem `ola` dá origem à mensagem codificada `MCY`.

A cripto-análise desta cifra parece ser bastante difícil devido ao número de chaves possíveis, no entanto veremos na próxima secção que, recorrendo a métodos estatísticos, é bastante fácil.

Os sistemas que usámos até agora codificam sempre um dado símbolo da mesma maneira (quando usamos a mesma chave). Estes sistemas são chamados *sistemas criptográficos mono-alfabéticos*.

Vamos referir ainda a título de curiosidade duas outras cifras bastante conhecidas que não são mono-alfabéticas, a *cifra Vigenère* e a *cifra matricial* ou *cifra de Hill*.[6]

Definição 5 (Cifra Vigenère). Seja $X = Y = K = (\mathbb{Z}_m)^n$, para um n fixo. Para cada $k = (k_1, k_2, \ldots, k_n) \in K$ define-se

$$e_k(x_1, x_2, \ldots, x_n) = (x_1 + k_1 \bmod m, \ldots, x_n + k_n \bmod m);$$
$$d_k(y_1, y_2, \ldots, y_n) = (y_1 - k_1 \bmod m, \ldots, y_n - k_n \bmod m).$$

Definição 6 (Cifra Matricial). Seja $X = Y = (\mathbb{Z}_m)^n$ para um n fixo, $K = \{\text{matrizes invertíveis } n \times n \text{ com componentes em } \mathbb{Z}_m\}$. Para cada $k \in K$ define-se

$$e_k(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n) \cdot k;$$
$$d_k(y_1, y_2, \ldots, y_n) = (y_1, y_2, \ldots, y_n) \cdot k^{-1}.$$

5 Substitution cipher.
6 Lester S. Hill.

Estas duas cifras codificam blocos de comprimento n. Por codificarem blocos em vez de caracteres isolados, é possível que um dado carácter tenha duas codificações diferentes. A estes sistemas que podem codificar um símbolo de mais do que uma maneira usando a mesma chave chamam-se *sistemas criptográficos poli-alfabéticos*.

Exemplo 4. Se codificarmos a palavra `ele` $\mapsto (4, 11, 4)$ usando a cifra Vigenère com a chave $k = (1, 2, 3)$ obtemos como resultado $(4, 11, 4)+(1, 2, 3) \equiv (5, 13, 7) \bmod 26$, que corresponde à mensagem codificada `FNH`, ou seja o primeiro `e` foi transformado em `F` e o segundo em `H`.

3 Cripto-Análise

Quando se estuda um sistema criptográfico em termos da sua segurança é importante definirmos à partida aquilo que um intruso conhece do sistema. Normalmente estes estudos têm subjacente o chamado *princípio de Kerckoff*, que diz que qualquer indivíduo que queira interceptar a comunicação entre outros dois pode ter toda a informação possível acerca do sistema criptográfico utilizado por eles.

Poderíamos não assumir isto e então grande parte da segurança da comunicação estaria associada ao desconhecimento sobre o sistema envolvido. Não queremos que a tarefa do intruso seja facilitada pelo simples facto de descobrir o sistema criptográfico em utilização. Sendo assim, a tarefa do cripto-analista é a seguinte: dada uma mensagem codificada tentar descobrir qual a chave de codificação usada pelos dois comunicadores sabendo o sistema criptográfico que estão a usar. Vamos então referir quatro tipos possíveis de ataque a um sistema criptográfico.

1. **Conhecimento de uma mensagem codificada:**[7] o intruso tem conhecimento de um exemplo de uma mensagem codificada, i.e., sabe $d_k(x)$ para algum $x \in X$.

2. **Conhecimento de uma mensagem:**[8] o intruso tem conhecimento de uma mensagem $x \in X$ e da sua codificação $e_k(x)$.

3. **Possível escolha de mensagens:**[9] o intruso tem acesso temporário à máquina de codificação e consegue codificar as mensagens $x \in X$ que quiser obtendo a sua codificação $e_k(x)$.

7 Ciphertext-only attack.
8 Known plaintext attack.
9 Chosen plaintext attack.

4. **Possível escolha de mensagens codificadas:**[10] igual ao anterior mas usando a máquina de descodificação, i.e., descobrindo os pares $(y, d_k(y))$.

Independentemente do ataque que faremos, temos de decidir qual a estratégia a seguir para cripto-analisar uma mensagem codificada. Um método possível para a cripto-análise é o chamado *método da força bruta* que consiste em testar todas as chaves possíveis para ver qual é que está a ser utilizada. No entanto, quando o número de chaves possíveis é muito elevado seguimos por vezes outra estratégia que é o *método da análise de frequências* das letras na língua. Apesar deste método ter grandes vantagens face ao anterior, a sua aplicação não é sempre possível.

3.1 Método da Força Bruta

Vamos então fazer a cripto-análise de um dos exemplos da secção anterior usando o método da força bruta. Escolhamos o Exemplo 1.

Ao olhar para a definição de cifra de translação verificamos que o seu número de chaves é reduzido, tem apenas m chaves. No nosso caso como queremos codificar as letras do alfabeto temos apenas 26 chaves. Este é um caso em que a técnica da *força bruta* é uma boa maneira de tentar descobrir a chave utilizada. Esta cifra é muito frágil mesmo com o ataque mais simples como se pode ver no exemplo que se segue.

Relembremos que a mensagem `ola` foi codificada na mensagem `SPE`. Vamos então fazer $d_k(\texttt{SPE})$ com $k = 0,\ 1, \ldots,\ 25$.

$k = 0$	`spe`	$k = 7$	`lix`	$k = 14$	`ebq`	$k = 20$	`yvk`
$k = 1$	`rod`	$k = 8$	`khw`	$k = 15$	`dap`	$k = 21$	`xuj`
$k = 2$	`qnc`	$k = 9$	`jgv`	$k = 16$	`czo`	$k = 22$	`wti`
$k = 3$	`pmb`	$k = 10$	`ifu`	$k = 17$	`byn`	$k = 23$	`vsh`
$k = 4$	`ola`	$k = 11$	`het`	$k = 18$	`axm`	$k = 24$	`urg`
$k = 5$	`nkz`	$k = 12$	`gds`	$k = 19$	`zwl`	$k = 25$	`tqf`
$k = 6$	`mjy`	$k = 13$	`fcr`				

Verificamos que o único valor que faz sentido é $k = 4$, logo descodificaremos `SPE` para `ola` pois é a única solução aceitável. Esta técnica é eficiente, mas quando o número de chaves aumenta torna-se impraticável.

10 Chosen ciphertext attack.

3.2 Método da Análise de Frequências

Este método consiste em estabelecer uma tabela da frequência das letras na língua portuguesa, através da leitura de jornais, revistas, etc., contando-se quantas vezes cada letra aparece nas respectivas notícias. Esta contagem não é feita recorrendo apenas ao dicionário pois este apresenta todas as palavras da língua portuguesa, mas não indica a frequência com que estas são usadas. Além da análise da frequência de cada uma das letras, devemos também analisar a frequência dos digramas e dos trigramas na língua, sequências de duas e três letras respectivamente.

Nota 5. Para aplicar esta técnica deve ser criada uma tabela recorrendo a textos da matéria em questão. Refere-se como exemplo a língua portuguesa mas se a mensagem fosse sobre aviões deveríamos recorrer a uma tabela de frequências criada a partir da análise de textos sobre aviões e se por acaso fosse sobre um tema científico específico deveríamos usar conhecimento sobre textos desse domínio. De referir ainda que a primeira análise das letras de um texto foi feita pelos árabes que contaram a frequência das letras no Corão.

Letra	**Freq. Relativa**	**Letra**	**Freq. Relativa**
a	0.1356	l	0.02760
e	0.1241	v	0.01460
o	0.1092	g	0.01215
s	0.0779	q	0.01009
i	0.0686	f	0.00980
r	0.0678	b	0.00934
n	0.0557	h	0.00724
d	0.0528	z	0.00427
t	0.0522	j	0.00365
c	0.0436	x	0.00225
m	0.0418	k	0.00052
u	0.0404	y	0.00036
p	0.0280	w	0.00029

Tabela 2: Frequências das Letras na Língua Portuguesa

A Tabela 2 foi obtida através da contagem das letras das palavras saídas em todos os jornais com publicações *on-line* desde 1991.[11] Apresenta-se ainda a mesma contagem para os digramas e trigramas mais comuns na língua portuguesa, na Tabela 3.[12]

11 Dados fornecidos pelo Laboratório de Sistemas de Linguagem Falada do INESC.

12 As tabelas apresentadas foram construídas recorrendo à análise de aproximadamente 100 000

Digrama	**Freq. Relativa**	**Trigrama**	**Freq. Relativa**
de	0.0249	ent	0.0142
es	0.0224	que	0.0118
os	0.0199	nte	0.0093
nt	0.0196	res	0.0067
ra	0.0191	est	0.0065
en	0.0185	nto	0.0065
do	0.0181	com	0.0063
co	0.0173	con	0.0063
te	0.0166	ado	0.0062
ar	0.0165	sta	0.0059
re	0.0165	ara	0.0058
as	0.0164	par	0.0057

Tabela 3: Digramas e Trigramas Mais Comuns na Língua Portuguesa

O objectivo desta técnica é analisar a frequência de uma dada letra na mensagem codificada e tentar inferir através das tabelas acima qual será a sua descodificação. Assim começamos por procurar a letra mais frequente na mensagem codificada e é natural supor que essa letra seja a codificação do a pois esta é a letra mais frequente na língua portuguesa. Devemos ter sempre em conta que esta é a hipótese mais provável, no entanto podem existir mensagens em que a frequência das letras não corresponde à frequência apresentada na tabela.

Repetimos o processo até determinar a descodificação de todas as letras da mensagem codificada. Podemos ainda usar a tabela dos digramas e dos trigramas para inferir sobre a mensagem original.

Esta técnica é particularmente adequada para fazer a cripto-análise das mensagens codificadas com a cifra de substituição pois esta cifra tem $m!$ chaves o que torna a técnica anterior impraticável.

Também é possível fazer a cripto-análise da cifra afim usando esta técnica. Para isso escolhemos as duas letras mais comuns na mensagem codificada, l_1 e l_2, (chamemos y_1 e y_2 aos seus códigos de acordo com a Tabela 1) e as duas letras mais comuns na língua portuguesa, a e e, como se vê na Tabela 2 (cujos códigos são respectivamente $x_1 = 0$ e $x_2 = 4$). Seguidamente tentamos resolver o sistema

$$\begin{cases} y_1 = ax_1 + b \bmod 26 \\ y_2 = ax_2 + b \bmod 26 \end{cases}.$$

palavras diferentes. Podemos referir a título de curiosidade que a palavra de foi a palavra que apareceu mais vezes, 18 859 628 vezes.

Tentamos assim ver se l_1 é a codificação da letra `a` e l_2 é a codificação de `e`.

1. Se $\gcd(a, 26) = 1$, a e b são possíveis candidatos a chave e então descodificamos toda a mensagem com esta chave.
   (a) Se o resultado da descodificação for satisfatório aceitamos esta chave.
   (b) Se o resultado não for satisfatório, descodificámos para uma mensagem sem sentido, vamos procurar outra chave escolhendo por exemplo a primeira e terceira letra mais comuns no alfabeto `a` e `o` e tentamos resolver de novo o sistema usando $x_1 = 0$ e $x_2 = 14$, respectivamente os códigos de `a` e `o`. Repetimos o processo até encontrarmos a chave certa, escolhendo letras diferentes mas de modo a que as frequências destas na mensagem e na Tabela 2 sejam semelhantes.
2. Se $\gcd(a, 26) \neq 1$ escolhemos um novo par de letras como indicado acima.[13]

A cripto-análise da cifra matricial parece bastante complicada. No entanto, se em vez de usarmos o ataque **conhecimento de uma mensagem codificada**, como até agora, usarmos um ataque do tipo **conhecimento de uma mensagem**, verificamos que a única dificuldade para calcular a chave é fazer a inversão de uma matriz.

Já vimos como fazer a cripto-análise de algumas cifras. Será possível fazer a cripto-análise para qualquer mensagem? Nos anos 50 Shannon provou que para conseguir segurança incondicional é necessário que o comprimento da chave seja igual ao comprimento da mensagem. Este facto não seria um grande problema, pois poderíamos combinar uma chave muito grande e usá-la em qualquer comunicação. No entanto outro factor necessário à segurança é que cada chave seja usada uma só vez. Logo a segurança incondicional tornou-se impossível de obter. Ou seja, qualquer mensagem que seja transmitida usando um sistema criptográfico de chave privada, usando uma chave que não satisfaça os requisitos apresentados, pode ser cripto-analisada.

4 Criptografia de Chave Pública

O paradigma da *criptografia de chave pública* apareceu em 1976, criado por Diffie e Hellman. Esta nova forma de codificação das mensagens consistia

13 Se $\gcd(a, 26) \neq 1$ é óbvio que não temos uma chave pois, por definição de K, as chaves são os pares da forma $(a, b) \in \mathbb{Z}_{26} \times \mathbb{Z}_{26}$ tais que $\gcd(a, 26) = 1$.

em codificar as mensagens usando funções de um tipo especial, as *funções de sentido único*.[14] Este conceito é definido informalmente como se segue:

DEFINIÇÃO 7. Uma função $f : X \to Y$ diz-se uma *função de sentido único*, ou simplesmente *fsu*, se é fácil computar $f(x)$ para qualquer $x \in X$ e é difícil calcular $f^{-1}(y)$ para a maioria dos elementos $y \in Y$ escolhidos aleatoriamente.

Observação. Nada se sabe sobre a existência de *fsu* pois este facto está intrinsecamente relacionado com um dos problemas fundamentais da ciência da computação, o problema de saber se $P = NP$.[15] Sabe-se que se $P = NP$ não existem *fsu*; mas se $P \neq NP$ nada se sabe. Quem quiser aprofundar esta questão pode consultar [Lub96].

Apesar de esta ideia apenas ter sido explorada a partir de 1976, a primeira referência ao uso de *fsu* aparece num artigo de 1968. Nesse artigo a criptografia era aplicada ao armazenamento de passwords em computadores. A ideia era muito simples: ao introduzir um novo utilizador, a password deste em vez de ser gravada como foi introduzida, era codificada usando uma certa função de codificação e esse valor ficava associado ao *login* do utilizador na lista de passwords que se encontra no disco do computador. O processo de autorização tornava-se assim bastante simples. Para aceder ao computador, um utilizador introduzia o seu *login* e a sua password, aplicava-se de novo a mesma função de codificação e verificava-se se o resultado coincidia com o que estava nos registos.

O processo de violação das passwords parece bastante fácil pois qualquer intruso teria acesso à lista das passwords codificadas e ainda ao algoritmo de codificação. No entanto as funções usadas para codificar eram do tipo referido acima, logo muito difíceis de inverter. Pela primeira vez aparecia o conceito de *fsu* aplicado a uma área particular da criptografia.

O primeiro algoritmo deste tipo, perfeitamente detalhado, é da autoria de Prudy, em 1974. Neste caso as passwords eram inteiros $\mathrm{mod} 2^{64} - 59$ e a função usada era

$$f(x) = x^{2^{24}+17} + a_1 x^{2^{24}+3} + a_2 x^3 + a_3 x^2 + a_4 x + a_5,$$

onde $a_1, a_2, \ldots, a_5$ são inteiros arbitrários com 19 dígitos.

As funções de codificação que iremos usar são uma subclasse das *fsu* pois tem de ser possível ao "legítimo" receptor descodificá-las, i.e., a inversão

14 One-way function.

15 P é a classe dos problemas que são resolúveis em tempo polinomial por uma máquina de Turing. NP é a classe dos problemas que podem ser resolvidos em tempo polinomial por uma máquina de Turing não determinista.

destas tem de ser impossível para todos à excepção do destinatário da mensagem. Assim as funções de codificação são *fsu* desde que nos falte alguma informação, a chave de descodificação. Sabendo esta, a inversão torna-se bastante simples.

Como é óbvio estas funções de codificação não são escolhidas ao acaso. Normalmente a sua inversão tem subjacente um problema que se sabe ser de difícil resolução, usualmente NP-completo. O primeiro e mais famoso algoritmo de chave pública é o RSA, baseado na factorização prima de um número grande, com aproximadamente 200 dígitos (cf. [Rei01]). Adiante veremos um outro algoritmo, o El Gamal, que é baseado no problema do logaritmo discreto, também este de difícil resolução.

Apesar de ter sido inventado apenas nos anos 70, o RSA é um algoritmo que usa matemática anterior ao século XX. Porque é que a ideia da chave pública e o algoritmo RSA não apareceram mais cedo?

Duas respostas podem ser dadas para esta pergunta. Uma primeira é que até à década de 70 a criptografia apenas era usada pelos serviços diplomáticos e militares e talvez por isso o sistema de chave privada fosse eficaz, pois não havia a necessidade de todos comunicarem com todos. Outra resposta possível é que a eficácia do algoritmo RSA depende da utilização de números primos grandes. Sem os computadores o cálculo da potenciação de grandes números é difícil.

Uma das grandes vantagens da *criptografia de chave pública* face à *criptografia de chave privada* é que quaisquer duas pessoas podem enviar mensagens confidenciais sem terem combinado uma chave de codificação. Mais ainda, uma pessoa pode enviar mensagens codificadas a outra sem nunca a ter contactado.

Como se processa então a codificação usando um protocolo de chave pública? Basta que o receptor da mensagem publique o seu algoritmo de codificação (parte pública) e todas as pessoas que queiram comunicar com ele utilizem esse algoritmo para codificar as mensagens. A única pessoa que conhece o algoritmo de descodificação (parte privada) é o receptor, e logo apenas ele pode descodificar as mensagens enviadas usando aquele algoritmo de codificação.

Depois desta breve introdução podemos pensar como é que a chave pública nos pode ajudar a resolver alguns problemas que nos são colocados hoje em dia, nomeadamente:

- Transmissão confidencial de mensagens.
- Autenticação: poder verificar que uma mensagem foi de facto enviada por quem se identifica como emissor.

- Não-repúdio: uma pessoa não poder negar que enviou de facto uma mensagem.
- Troca de chaves: duas pessoas combinarem publicamente uma chave para usar num sistema de chave privada.
- Partilha de segredos: um sistema só poder funcionar com k autorizações e se apenas tivermos $k-1$ não conseguimos fazer nada.

Abordaremos seguidamente os quatro primeiros pontos, dando exemplos para cada um.

4.1 Transmissão Confidencial de Mensagens

Como exemplo vamos ver o algoritmo de codificação El Gamal. Este algoritmo é baseado na dificuldade de resolução do problema do logaritmo discreto, PLD. Sem entrar em grandes detalhes sobre as condições do problema, assumindo que são de fácil verificação, o PLD pode ser posto como se segue.

Definição 8 (Problema do Logaritmo Discreto). Seja p primo, $\alpha \in \mathbb{Z}_p$ um elemento primitivo[16] e $\beta \in \mathbb{Z}_p^*$.[17] O objectivo é encontrar o único inteiro a com $0 \leq a \leq p-2$ e tal que

$$\alpha^a \equiv \beta \bmod p.$$

Dizemos neste caso que $a = \log_\alpha \beta$.

Este problema é considerado de difícil resolução porque não existe nenhum algoritmo polinomial que o resolva para valores de p elevados (números com aproximadamente 150 dígitos) e tais que $p-1$ tenha pelo menos um factor primo grande.

Como é que nós podemos dizer que este problema nos dará uma boa função de codificação? Tal como é posto, o problema de calcular o logaritmo discreto de um número é complicado, mas calcular a potência, operação inversa, é muito simples. Logo podemos usar como função de codificação a potência, fácil de calcular, e como função de descodificação o logaritmo. Dito de outra forma, o logaritmo é uma *fsu*.

Definição 9 (Cifra El Gamal). Seja p primo tal que o PLD é intratável em $\mathbb{Z}_p$ e $\alpha \in \mathbb{Z}_p^*$ um elemento primitivo. Seja ainda $X = \mathbb{Z}_p^*$, $Y = \mathbb{Z}_p^* \times \mathbb{Z}_p^*$ e $K = \{(p, \alpha, a, \beta) : \beta \equiv \alpha^a \bmod p\}$. Os parâmetros públicos

16 α diz-se um elemento primitivo de $\mathbb{Z}_p$ se $\{\alpha^i : 0 \leq i \leq p-2\} = \mathbb{Z}_p^*$.
17 $\mathbb{Z}_p^* = \{x \in \mathbb{Z}_p : x \neq 0\}$.

são p, α e β. O número a é privado. Seja ainda $z \in \mathbb{Z}_p$ um número aleatório escolhido pelo emissor. Definimos então

$$e_k(x, z) = (y_1, y_2) \text{ em que } \begin{cases} y_1 = \alpha^z \bmod p \\ y_2 = x\beta^z \bmod p \end{cases} ; \tag{2}$$

$$d_k(y_1, y_2) = y_2(y_1^a)^{-1} \bmod p. \tag{3}$$

Vamos ver que quem souber a consegue descodificar a mensagem.

$$\begin{aligned} y_2(y_1^a)^{-1} &\equiv x\beta^z(\alpha^{az})^{-1} && \text{por (2)} \quad (4a) \\ &\equiv x\beta^z(\beta^z)^{-1} && \text{por definição de } K \quad (4b) \\ &\equiv x \bmod p \quad (4c) \end{aligned}$$

O que o Bruno faz ao usar este algoritmo é escolher um $z \in \mathbb{Z}_p$ aleatório e "esconder" a mensagem x usando β^z. Depois envia este valor, $x\beta^z$, juntamente com α^z o que permite à Ana calcular β^z, (4b), parte fundamental do algoritmo. A Ana consegue assim descodificar a mensagem pois conhece a, enquanto o Carlos não. Assim o Carlos tem de calcular z tentando resolver o PLD pois sabe o valor de α e $y_1 = \alpha^z$.

Como vemos, se alguém souber o valor de z consegue descodificar a mensagem pois pode começar o processo a partir de (4b); por isso o valor de z também tem de ser guardado pelo emissor. Um factor que pode dificultar ainda mais a tentativa de descodificação "ilegítima" desta cifra é o facto de ser não determinística. A mensagem `ola` não é codificada sempre da mesma maneira, pois depende do z que o emissor escolher.

4.2 Assinaturas Digitais e Problemas de Autenticação

Um problema que existe na comunicação digital é que devido ao facto de os comunicadores não se verem, as mensagens que estes trocam podem estar a ser enviadas ou alteradas por alguém. Consideremos então estes dois problemas:

A. Problema da Assinatura: a Ana pretende ter a certeza que a mensagem enviada pelo Bruno não foi alterada pelo Carlos.

B. Problema da Autenticação: a Ana quer ter a certeza que foi de facto o Bruno que lhe enviou uma dada mensagem e não o Carlos fingindo ser o Bruno.

Temos assim que discutir uma nova tarefa da criptografia: garantir que o nosso emissor ou receptor é de facto quem pensamos. Para isso temos de

definir, mais uma vez informalmente, um novo conceito que é o conceito de *função de dispersão*.[18] Estas funções são normalmente de domínio público.

DEFINIÇÃO 10. Diz-se que $H : X^k \to X^l$ com $k > l$ é uma *função de dispersão* se tiver as seguintes propriedades:

1. É fácil calcular $H(x)$ para qualquer $x \in X^k$;
2. É difícil encontrar dois valores $x, x' \in X^k$ tais que $H(x) = H(x')$;
3. Dado $y \in \mathrm{Im}(H)$ não é fácil determinar $x \in X^k$ tal que $H(x) = y$.

Estas funções são utilizadas para produzir assinaturas digitais pois *comprimem* as mensagens, ou seja, recebem mensagens de k símbolos e transformam-nas em mensagens de l símbolos, $k > l$. A mensagem assim obtida pode ser usada como assinatura da mensagem original, pois é mais pequena que esta.

A segunda condição da definição acima faz com que, apesar de a função não ser obrigatoriamente injectiva, seja complicado encontrar duas mensagens que tenham a mesma assinatura. A terceira condição faz com que seja difícil arranjar uma mensagem sabendo qual é a sua assinatura.

Suponhamos então que o Bruno pretende enviar uma mensagem para a Ana. Como é que ela pode ter a certeza que a mensagem recebida foi mesmo enviada por ele e não foi alterada pelo Carlos (problemas A e B identificados acima)?

Problema da Assinatura

O que o Bruno tem a fazer para resolver este problema é em vez de enviar apenas a mensagem, enviar a mensagem e a sua assinatura, i.e., enviar o par $(x, H(x))$. Para a Ana verificar que a mensagem não foi alterada, basta-lhe aplicar a mesma função H à primeira componente do par (a mensagem) e comparar o resultado com a segunda componente (a assinatura). Se for igual, é porque a mensagem não foi alterada. Mas como é que perante esta igualdade podemos ter a certeza que a mensagem não foi alterada?

Podemos, porque pela segunda condição da Definição 10 o Carlos não consegue encontrar outra mensagem x' tal que $H(x) = H(x')$. Ele pode apenas tentar alterar ambas as componentes do par $(x, H(x))$ gerando uma mensagem x' e calculando a respectiva assinatura. Temos por isso de resolver o problema que resulta do facto de o Carlos poder enviar uma mensagem totalmente nova, $(x', H(x'))$.

18 Hash-function.

Problema da Autenticação

Para resolver este problema necessitamos apenas que o Bruno use um sistema criptográfico endomórfico.

DEFINIÇÃO 11. Um sistema criptográfico $\langle X, Y, K, E, D\rangle$ diz-se *endomórfico* se $X = Y$.

LEMA 12. *Seja $f : I \to I$ uma função injectiva e I um conjunto finito. Então f é um isomorfismo.*

Observação. Devido ao sistema criptográfico usado pelo Bruno ser endomórfico e a respectiva função de codificação e_B ser injectiva, o Lema 12 garante que e_B é um isomorfismo e que a sua inversa é d_B. Temos então $e_B(d_B(x)) = x$ e $d_B(e_B(x)) = x$.

Depois desta observação é muito simples entender como é que a Ana pode estar certa que foi o Bruno que enviou a mensagem. Basta que o Bruno, em vez de enviar o par $(x, H(x))$, envie $(x, d_B(H(x)))$. O que a Ana tem a fazer para verificar que foi o Bruno é aplicar a função e_B (pública) ao segundo elemento da mensagem que recebeu. Assim obtém o valor de $H(x)$ pois pela observação anterior $e_B(d_B(H(x))) = H(x)$. Seguidamente faz o mesmo que fazia para verificar se a assinatura era válida, ou seja aplica a função H ao primeiro elemento do par e verifica se dá o valor que tinha acabado de calcular.

O único problema que ainda não tínhamos resolvido era o "risco" do Carlos alterar as duas componentes da mensagem. Verificamos agora que lhe é impossível fazer tal coisa, pois só o Bruno conhece d_B. Assim, mesmo que o Carlos encontre um novo par $(x', H(x'))$ não conseguirá calcular $d_B(H(x')$. Deste modo a Ana estará certa que quem enviou a mensagem foi de facto o Bruno. Por outro lado, o Bruno não pode negar que enviou a mensagem pois só ele conhece a função d_B, e por isso só ele poderá usá-la para "codificar" a assinatura. Estabelecemos então uma relação de compromisso e não-repúdio.

Observação. Por simplicidade referi que o Bruno tinha de usar um sistema criptográfico endomórfico, pois era necessário existir uma função que apenas o Bruno conhecesse, neste caso d_B. No entanto poderíamos apenas exigir que existissem duas funções $f, g : X \to X$ tais que f fosse fácil de calcular e injectiva, $f \circ g = id_X$ e g fosse apenas do conhecimento do Bruno.

4.3 Troca de Chaves Privadas Através de Canais Públicos

Suponhamos que a Ana e o Bruno querem combinar uma chave para um sistema criptográfico de chave privada e que não se podem encontrar pessoalmente nem têm nenhuma maneira de contactar seguramente. Terão por

esse motivo de combinar a chave através de um canal público. Como poderão fazê-lo?

1. Ana e Bruno escolhem um primo p e um elemento primitivo $g \in \mathbb{Z}_p$ primitivo. O Carlos pode conhecer p e g, pois foram combinados através de um canal público.

2. Em seguida a Ana escolhe um valor $k_A < p$ e envia $g^{k_A} \bmod p$ para o Bruno.

3. O Bruno faz o mesmo e envia $g^{k_B} \bmod p$ para a Ana.

   Neste momento o Carlos pode saber os valores p, g, g^{k_A} e g^{k_B}, pois todos eles foram combinados através de um canal público.

4. Como em $\mathbb{Z}_p$ temos $(g^{k_A})^{k_B} \equiv (g^{k_B})^{k_A} \equiv g^{k_A k_B} \bmod p$, a Ana e o Bruno escolhem $g^{k_A k_B} \bmod p$ para chave.

   O Bruno consegue calcular esta chave pois conhece k_B e $g^{k_A} \bmod p$. A Ana também consegue calcular a chave pois conhece k_A e $g^{k_B} \bmod p$.

   No entanto será que o Carlos consegue calcular a chave? Será que, dados p, g, g^{k_A} e g^{k_B}, consegue calcular $g^{k_A k_B} \bmod p$? É fácil ver que não consegue, pois teria que saber k_A ou k_B — e descobrir estes números seria o mesmo que resolver o Problema do Logaritmo Discreto. Logo o Carlos nunca conseguirá descobrir a chave se a Ana e o Bruno escolherem um número primo p tal que o PLD seja intratável em $\mathbb{Z}_p$.

5 O Dinheiro Virtual e a Criptografia

O comércio virtual é um mundo em forte expansão nos nossos dias. No entanto, esta expansão tem um entrave: o meio de pagamento. Normalmente existem duas maneiras possíveis de fazer esse pagamento. Uma é o envio à cobrança, a outra é através de cartão de crédito. Enquanto a primeira é mais segura (só pagamos a encomenda quando a recebemos e o nosso número de cartão de crédito não anda a circular pela Internet, mesmo que codificado) a segunda é muito mais eficiente, barata e prática. Temos por isso de contrabalançar a segurança com a eficiência. De momento o comércio virtual ainda está a dar os primeiros passos, mas quando se desenvolver terá de haver um meio muito mais prático e seguro de fazer as transacções.

É com este objectivo que nasceu a ideia de dinheiro virtual. Será possível ter "dinheiro virtual"? Existirão "bancos virtuais"? Claro que não poderão existir bancos meramente virtuais, pois as pessoas necessitam de dinheiro

físico para a sua vida quotidiana; mas será possível termos uma conta virtual, i.e., uma conta onde cada um de nós tem dinheiro para gastar em compras *on-line*? Imaginemos que queremos comprar um livro numa livraria *on-line*. Será possível em vez de enviarmos o número do nosso cartão de crédito, enviarmos por exemplo, um ficheiro informático que valha dinheiro? Onde guardamos esse dinheiro virtual? Num "cofre virtual"?

Este será o problema que vamos abordar nesta secção. Vamos usar os termos "dinheiro físico" e "dinheiro virtual" caso se trate do dinheiro que estamos habituados ou deste novo tipo de dinheiro que queremos inventar. Vamos usar sempre algoritmos de codificação seguros. Para simplificar a abordagem ao problema vamos primeiro fazer algumas considerações sobre o dinheiro em geral e pensar em algumas propriedades que o dinheiro físico tem e que seja importante o dinheiro virtual também ter. Baseados nestas ideias criaremos um protocolo, que apesar de apresentar algumas limitações é um primeiro passo na criação de dinheiro virtual.

Antes de pensarmos como guardar o nosso dinheiro virtual, pensemos como é que este vai ser. A ideia imediata, e se calhar a única, é que as notas virtuais sejam ficheiros. Assim, teremos ficheiros nos nossos computadores que representam dinheiro e de preferência queremos que toda a gente os possa ver, ou seja, estes têm de estar numa parte pública dos nossos computadores.

Contudo, se estiverem numa zona pública dos nossos computadores poderão ser roubados. Então como é que os guardamos? O ideal é que o dinheiro virtual não necessite de ser guardado em cofres virtuais; este dinheiro devia ser ele próprio à prova de roubo. Como conseguir isso? Fazendo com que o nosso dinheiro tenha algumas propriedades anti-roubo.

5.1 Propriedades do Dinheiro Virtual

Pensemos então em algumas propriedades que o dinheiro virtual deverá ter.

- Todas as notas são emitidas por um banco, logo também no campo virtual temos de ter uma entidade emissora/reguladora.
- Todas as notas têm um número de série que as torna únicas, aqui também teremos essa numeração.
- As notas usuais são, ou pretende-se que sejam, difíceis de copiar, mas para já o dinheiro virtual corresponderá a ficheiros informáticos e consequentemente fácil de copiar. Assim a nossa entidade emissora terá também uma função reguladora: verificar que cada nota só pode ser usada para pagar uma operação, i.e., tem que manter registo do proprietário do ficheiro em cada instante.

Para escapar a este último ponto, de complicada resolução, e para simplificar o nosso problema, vamos apenas pensar nos *traveller cheques*.

- À partida temos uma grande vantagem: enquanto os cheques são pessoais e só podem ser usados uma vez, o mesmo não se passa com as notas.
- Os cheques, além de pessoais, têm o nome do proprietário. Assim, cada ficheiro terá também a informação do seu proprietário.
- No caso dos *traveller cheques* temos ainda que os assinar quando os recebemos, e assim garantimos que só pode usar aquele cheque quem reproduzir a nossa assinatura. Neste caso, será uma assinatura virtual que tem subjacente um algoritmo de assinatura seguro.

5.2 Protocolo para Implementação de Dinheiro Virtual

Na nossa transacção temos um comprador, a *A*na, um *B*anco e um *V*endedor. Temos ainda um outro elemento, o *L*adrão, que pretende "desviar" o dinheiro em alguma parte da comunicação.

Admitimos como certo que o Banco é uma entidade honesta, que sabe quem são os seus clientes e guarda as chaves públicas dos mesmos. Sempre que seja necessário o uso de chaves públicas, é ao Banco que elas vão ser pedidas. Vamos usar ainda e_X e d_X para representar as chaves de codificação e descodificação do elemento X. Vejamos então o protocolo.

1. Ana pede o dinheiro ao Banco, identificando-se.
2. O Banco gera um número aleatório n_1 que será o código da nota produzida. Mas, tal como foi referido atrás, esse código deve ser apenas conhecido pela Ana, como se fosse uma assinatura. Sendo assim, em vez de assinar a nota com o código n_1, o Banco assina-a com o código $e_A(n_1)$. Para manter a privacidade o Banco apaga n_1 dos seus registos e guarda o par $(nota, e_A(n_1))$.

Notas emitidas	**Notas em circulação**
$(nota, e_A(n_1))$	

   Neste momento o Banco envia para a Ana a mensagem $(nota, e_A(n_1))$.

$B \xrightarrow{(nota, e_A(n_1))} A \qquad V$

Isto termina a fase da emissão da nota, a qual só entrará em circulação (i.e., só será válida) quando o Banco receber a confirmação de que a Ana recebeu a nota.

3. Entramos então na fase de confirmação da recepção da nota por parte da Ana. O Banco só passa a dar a nota como válida se alguém lhe enviar o valor n_1, associado à informação da nota. Neste momento só a Ana sabe qual é o valor n_1, pois por (1) temos que $n_1 = d_A(e_A(n_1))$ e apenas a Ana conhece a função d_A. Mesmo que intercepte a nota, o Ladrão nunca conseguirá reclamá-la, pois $n_1 \neq d_L(e_A(n_1))$.

   Então o que a Ana faz é calcular o valor de n_1 e reenviá-lo codificado para o Banco, i.e., envia para o Banco a mensagem $(nota, e_B(n_1), A)$.

   $B \xleftarrow{(nota, e_B(n_1), A)} A \qquad V$

   É importante observar que só a Ana sabe o valor n_1, e consequentemente só ela consegue calcular $e_B(n_1)$.

4. Ao receber esta mensagem o Banco selecciona a segunda componente da mensagem, x, e calcula $e_A(d_B(x))$. Seguidamente verifica se este é o valor associado à nota, i.e., verifica se $e_A(d_B(x)) = e_A(n_1)$.

   (a) Se for este o valor associado, ou seja, se $x = e_B(n_1)$, é porque a Ana recebeu a nota, e então ela fica válida. É adicionado ao registo das notas em circulação o triplo $(nota, e_A(n_1), A)$ e o par $(nota, e_A(n_1))$ é removido das notas emitidas. A partir deste momento o Banco só paga o valor da nota a quem provar saber o valor n_1.

Notas emitidas	**Notas em circulação**
~~$(nota, e_A(n_1))$~~	$(nota, e_A(n_1), A)$

   (b) Se não for é porque alguém se tentou passar pela Ana. Em consequência disso, a nota fica inválida e é criada uma nova nota com um novo código $(nota', e_A(n_1'))$, que é de novo enviada para a Ana.

Notas emitidas	**Notas em circulação**
~~$(nota, e_A(n_1))$~~	
$(nota', e_A(n_1'))$	

   $B \xrightarrow{(nota', e_A(n_1'))} A \qquad V$

Fica assim confirmada a recepção ou o desvio da nota.

Neste momento a Ana tem o seu dinheiro e mesmo que este seja roubado ninguém o poderá usar pois não conhece o valor n_1. Vamos então passar à fase de pagamento. Esta tem que ser realizada em simultâneo pelo vendedor e pelo comprador.

5. Suponhamos então que a Ana quer comprar um livro ao Vendedor. A Ana gera um número aleatório n_2 e envia para o Vendedor o par $(nota, e_V(n_2))$.

Notas emitidas	**Notas em circulação**
	$(nota, e_A(n_1), A)$

$$B \qquad\qquad A \xrightarrow{(nota, e_V(n_2))} V$$

Segue-se de novo o protocolo de confirmação de recepção (Passos 3 e 4).

6. O Vendedor recebe o par $(nota, e_V(n_2))$ e obtém o valor n_2 calculando $d_V(e_V(n_2)) = n_2$. Em seguida reenvia para a Ana o par $(nota, e_A(n_2))$.

$$B \qquad\qquad A \xleftarrow{(nota, e_A(n_2))} V$$

7. A Ana recebe a mensagem do Vendedor e verifica se o código que vem associado à nota é n_2, calculando $d_A(e_A(n_2))$. Note-se aqui que apenas a Ana e o Vendedor conhecem n_2. Logo só o Vendedor consegue enviar $e_A(n_2)$.

   (a) Se for esse o valor associado o Vendedor recebeu a nota. Então a Ana comunica ao Banco que já não possui a nota. Para isso envia-lhe a mensagem $e_B(nota, n_1, e_V(n_2), V)$, ou seja comunica que a nota está agora em posse do Vendedor e só deverá ser paga a quem mostrar saber o valor n_2. Mostra por outro lado que a nota era dela pois também conhece n_1.

$$B \xleftarrow{e_B(nota, n_1, e_V(n_2), V)} A \qquad\qquad V$$

   O Banco altera então o registo das notas em circulação mudando o anterior registo $(nota, e_A(n_1), A)$ para $(nota, e_V(n_2), V)$.

Notas emitidas	**Notas em circulação**
	~~$(nota, e_A(n_1), A)$~~
	$(nota, e_V(n_2), V)$

(b) Se não for é porque alguém interceptou a mensagem da Ana para o Vendedor. Neste caso a Ana reenvia a mensagem e voltamos ao Ponto 5.

Nesta altura apenas a Ana e o Vendedor sabem o valor de n_2, pelas mesmas razões indicadas no Passo 3. Em todas as comunicações n_2 viajou codificado, uma vez com a chave da Ana e outras duas com a chave do Vendedor. Logo é impossível para qualquer outra pessoa descobrir o valor de n_2.

8. V envia então ao Banco a mensagem $(nota, e_B(n_2))$ e o Banco, à semelhança do que fez atrás calcula $e_V(d_B(e_B(n_2))) = e_V(n_2)$ e verifica se este é o valor associado à nota. A igualdade só se verifica se o segundo elemento do par enviado for $e_B(n_2)$.

$$\boxed{B \xleftarrow{(nota, e_B(n_2))} V}$$

(a) Se for o valor associado à nota, o Banco confirma a transacção e retira a nota de circulação.

Notas emitidas	**Notas em circulação**
	~~$(nota, e_V(n_2), V)$~~

Seguidamente, o Vendedor conclui a transacção enviando a encomenda à Ana pois o Banco já lhe creditou o dinheiro.

(b) Se não for é porque alguém está a querer passar-se por Vendedor.

Notar que é impossível a Ana reutilizar a nota pois esta sai imediatamente de circulação.

Assim damos por terminada a descrição do protocolo. Neste protocolo cada nota é usada apenas uma vez, ou seja, estas notas são semelhantes aos cheques usuais. Se quiséssemos aplicar um protocolo semelhante que permitisse a circulação de uma nota por vários donos, teríamos de alterar o Ponto 8a e em vez de tirarmos a nota de circulação, combinaríamos um novo número n_3 do conhecimento exclusivo do Banco e do Vendedor, para que a Ana não voltasse a reutilizar a nota.

6 Agradecimentos

Começo por agradecer aos Professores Amílcar Sernadas e Paulo Mateus pela troca de ideias sobre a última secção e pela ajuda na preparação do

seminário em geral. Queria ainda agradecer ao Eng. Hugo Meinedo pela disponibilização dos dados para o cálculo das tabelas de frequências apresentadas. Por fim, queria agradecer à Ana pela paciência e pela ajuda na revisão do texto deste artigo, bem como na preparação do seminário.

7 Bibliografia

Quem quiser iniciar o estudo nesta área poderá consultar [Sti95], um bom livro para começar. [Men97] é também um excelente livro sobre esta matéria, mas requer algum conhecimento prévio.

Referências

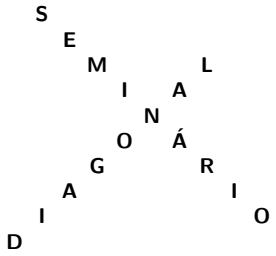
[Sti95] Douglas R. Stinson. *Cryptography — Theory and Practice*. CRC Press, 1995.

[Men97] A. Menezes, P. Van Oorschot e S. Vanstone. *Handbook of Applied Cryptography*. CRC Press, 1997.

[Kob99] Neal Koblitz. *Algebraic Aspects of Cryptography*. Springer, 1999.

[Lub96] Michael Luby. *Pseudorandomness and Cryptographic Applications*. Princeton University Press, 1996.

[Rei01] Tiago Reis. Criptografia e jogos por telefone. Neste volume.



Criptografia e Jogos por Telefone

Tiago Reis
2º ano da LMAC
l48673@lascas.math.ist.utl.pt

Palavras Chave

chave pública, RSA.

Resumo

Será possível que duas pessoas lancem uma moeda ao ar ao telefone? Poderá isto ser feito sem que a pessoa que escolhe cara ou coroa, no caso de perder, não duvide nem um pouco da honestidade do lançamento? Neste artigo veremos uma solução possível para este problema e até que ponto é fiável.

Veremos por fim o que é um algoritmo de encriptação de chave pública, isto é, um algoritmo em que tanto a chave como o próprio algoritmo são públicos. E o que é e como funciona o algoritmo RSA, amplamente difundido.

Introdução

Multiplicar dois números é muito fácil, porém dado um número qualquer descobrir os seus divisores já não é trivial. É certo que existe um método que nos devolve sempre a resposta correcta, basta para isso dividi-lo por todos os inteiros menores que a sua raiz quadrada. Mas este processo de força bruta não é eficiente; para números com muitos algarismos pode demorar um tempo absurdo.

Felizmente temos ao nosso dispor ferramentas bastante poderosas para factorizar números, por exemplo, o Crivo Quadrático e o Método das Curvas Elípticas [1]. Estes métodos são capazes de factorizar números de cerca de 80 a 100 algarismos num tempo razoável.

Nas duas aplicações referidas neste artigo a fiabilidade dos algoritmos utilizados é conseguida explorando a dificuldade de factorizar um número. Os métodos actuais não são eficientes para números arbitrariamente grandes e bem escolhidos. Existem também métodos bastante potentes que recorrem a processamento em paralelo que baixa o tempo de factorização drasticamente, mas ainda assim esta pode não ser conseguida em tempo útil.

1 Lançamento de uma Moeda ao Ar ao Telefone

Imaginemos que duas pessoas querem decidir algo aleatoriamente ao telefone, recorrendo ao lançamento de uma moeda ao ar. Como pode isso ser feito sem que ninguém influencie o resultado?

Vamos chamar às duas pessoas que estão ao telefone Sr. A e Sr. B. Uma estratégia possível é a seguinte:

i) O Sr. A começa por escolher dois números primos distintos p e q da forma $4k+3$, calcula $n = p \times q$ e envia-o ao Sr. B.

ii) Agora o Sr. B escolhe um número x $(0 < x < n)$ relativamente primo a n, calcula $a \equiv x^2 \pmod{n}$ e transmite a ao Sr. A.[1]

iii) Por fim o Sr. A tem de resolver a equação,

$$t^2 \equiv a \pmod{n}, \quad 0 < t < n, \tag{1}$$

que tem quatro soluções agrupadas em $\{x, n-x\}$ e $\{y, n-y\}$. Depois de saber estes dois pares o Sr. A escolhe um e comunica-o ao Sr. B.[2]

Se o Sr. A escolheu o par $\{x, n-x\}$ então é o vencedor, caso contrário o vencedor é o Sr. B que apenas tem de revelar x para o Sr. A se convencer que escolheu o par errado.[3]

Exemplo 1. Para este exemplo foram escolhidos pequenos valores para p e q mas na prática estes números deverão ser muito grandes.

Suponha-se que o Sr. A escolhe $p = 31$ e $q = 23$, de seguida calcula o seu produto e envia $n = 713$ ao Sr. B. Sabendo n o Sr. B escolhe $x = 220$ e calcula $a = 629 \equiv 220^2 \pmod{713}$ e envia de volta a. Resta agora ao Sr. A resolver a equação (1); os pares-solução são $\{220, 493\}$ e $\{654, 59\}$. Se agora o Sr. A escolher o par $\{220, 493\}$ ganha, se escolher o par $\{654, 59\}$ perde.

DEFINIÇÃO (RESÍDUO QUADRÁTICO). Dado um inteiro n e um número primo p, se $\mathrm{mdc}(n, p) = 1$ e se existe um t tal que $n \equiv t^2 \pmod{p}$, então n diz-se um resíduo quadrático módulo p.

PROPOSIÇÃO 1. *Seja $p = 2m + 1$ um primo ímpar. Se n é um resíduo quadrático módulo p então $n^m \equiv 1 \pmod{p}$.*

1 Diz-se que x e y são congruentes módulo n, e escreve-se $x \equiv y \pmod{n}$, se os restos das suas divisões por n são iguais.

2 A escolha aleatória de um destes pares pode ser vista como equivalente a lançar uma moeda ao ar.

3 Veremos adiante porque é que é praticamente impossível o Sr. B fazer batota.

Demonstração. Se n é um resíduo quadrático então existe um t que verifica $n \equiv t^2 \pmod{p}$. Logo,

$$n^m \equiv (t^2)^m \equiv t^{p-1} \equiv 1 \pmod{p},$$

pois temos $\mathrm{mdc}(t, p) = 1$ e podemos aplicar o Teorema 3 [ver à frente]. □

Proposição 2. *Se p é um número primo da forma $p = 4k + 3$ e n é um resíduo quadrático módulo p, então $x \equiv n^{k+1} \pmod{p}$ é solução da equação $x^2 \equiv n \pmod{p}$.*

Demonstração. Da Proposição 1 sabemos que $n^{\frac{p-1}{2}} \equiv 1 \pmod{p}$. Então:

$$(n^{k+1})^2 \equiv n^{2k+2} \equiv n \times n^{\frac{p-1}{2}} \equiv n \pmod{p}.$$ □

Para resolver a equação (1), vamos recorrer ao método de *exponenciação rápida*, o qual permite calcular potências $x^b \pmod{p}$ expeditamente. Este método consiste em escrever b como uma soma de potências de 2 ($b = \alpha_0 2^0 + \alpha_1 2^1 + \cdots + \alpha_m 2^m$) e depois calcular sucessivamente

$$x^b \equiv x^{\alpha_0} \times (x^2)^{\alpha_1} \times \cdots \times (x^{2^m})^{\alpha_m} \pmod{n},$$

reduzindo módulo n após cada multiplicação.

Exemplo 2. Vamos resolver $\gamma^2 \equiv 629 \pmod{31}$. Temos que $31 = 4 \times 7 + 3$; então pela Proposição 2, $\gamma \equiv 629^8 \pmod{31}$. Como na prática a e k são números muito grandes seria muitíssimo difícil calcular a^{k+1} directamente. Recorrendo ao método de exponenciação rápida é muito simples:

$$\begin{aligned} a = 629 &\equiv 9 \pmod{31}, \\ a^2 = 9^2 &\equiv 19 \pmod{31}, \\ a^4 = 19^2 &\equiv 20 \pmod{31}, \\ a^8 = 20^2 &\equiv 28 \pmod{31}. \end{aligned}$$

Obtém-se então $\gamma = 28$.

De facto, a proposição seguinte mostra que, sendo

$$\gamma^2 \equiv a \pmod{p} \quad \text{e} \quad \delta^2 \equiv a \pmod{q},$$

uma vez sabidos γ e δ, as quatro soluções distintas da equação (1) são geradas pela fórmula

$$t \equiv \pm\delta\alpha p \pm \gamma\beta q \pmod{n}, \tag{2}$$

onde os sinais $\pm$ são tomados independentemente e α e β são as soluções da equação diofantina $\alpha p + \beta q = 1$ que podem ser calculadas utilizando o algoritmo de Euclides (ver [1]).

PROPOSIÇÃO. *Os valores* $t = \pm\delta\alpha p \pm \gamma\beta q$ *são soluções da equação* (1).

Demonstração. Sabemos que $\alpha p + \beta q = 1$. Então:

$$t^2 \equiv (\pm\delta\alpha p \pm \gamma\beta q)^2 \equiv (\gamma\beta q)^2 \equiv \gamma^2(1-\alpha p)^2 \equiv \gamma^2 \equiv a \pmod{p}.$$

Analogamente conclui-se que $t^2 \equiv a \pmod{q}$. Logo p e q dividem $t^2 - a$, e portanto conclui-se que $t^2 \equiv a \pmod{n}$. □

Exemplo 3. No exemplo que temos estado a seguir, já sabemos que $\gamma = 28$. Do mesmo modo, obtemos $\delta = 13$. Falta de seguida resolver a equação diofantina $31\alpha + 23\beta = 1$, que vai ter como soluções $\alpha = 3$ e $\beta = -4$. Uma vez sabidos γ, δ, α e β, usando a equação (2) geramos as quatro soluções distintas que formam os pares $\{220, 493\}$ e $\{654, 59\}$.

Vimos como podem ser calculados os pares-solução e como pode ser decidido o vencedor. A impossibilidade do Sr. B fazer "batota" assenta no pressuposto que ele não consegue resolver a equação (1) para poder apresentar sempre ao Sr. A o *outro* par solução. O que impede o Sr. B de resolver a equação?

Ao contrário do Sr. A, o Sr. B não tem conhecimento prévio da factorização de n. Saber a factorização de n é essencial para conseguir determinar os dois pares-solução da equação. Mais, prova-se que saber três soluções distintas equivale a saber a factorização de n. Porém na prática n será o produto de dois números primos muito grandes e, tal como foi discutido no início deste artigo, para números suficientemente grandes os métodos actuais não conseguem devolver uma resposta em tempo útil.

PROPOSIÇÃO. *Saber a factorização de* n *é equivalente a saber pelo menos três soluções distintas da equação* (1).

Demonstração. Já foi apresentado um algoritmo para calcular as soluções sabendo a factorização; falta apenas o outro sentido da equivalência.

Suponhamos que o Sr. B conhece três soluções distintas x, $n - x$ e y da equação.

Claramente $x^2 \equiv y^2 \equiv a \pmod{n}$, logo n divide $y^2 - x^2 = (y - x)(y + x)$ mas não divide $(y - x)$, pois nesse caso teríamos $y \equiv x \pmod{n}$, o que contraria a hipótese inicial; o mesmo se passa para $(y + x)$. Isto significa que p divide $(y - x)$ e q divide $(y + x)$ ou vice-versa, logo:

$$p = \text{mdc}(n, y - x) \quad \text{e} \quad q = \text{mdc}(n, y + x).$$ □

2 Algoritmo RSA

O algoritmo de encriptação RSA foi inventado no ano de 1977 por Rivest, Shamir e Adleman no MIT. É um exemplo de um *algoritmo de chave pública*.

Nos algoritmos de chave pública a chave de encriptação é independente da chave de desencriptação; a chave de encriptação é tornada pública juntamente com o algoritmo ao passo que a chave de desencriptação é apenas conhecida por quem publicou o código. Este conceito fora independentemente proposto por Diffie e Hellman em Stanford e por Merkle na Universidade da Califórnia no ano de 1976.

Os algoritmos de chave pública (ou *de encriptação assimétrica*) têm vantagens claras face aos métodos utilizados antigamente, em que os processos de encriptação e desencriptação eram basicamente os mesmos e portanto todas as entidades capazes de emitir mensagens encriptadas eram um possível alvo para o inimigo tentar obter a chave de desencriptação. Com os algoritmos de chave pública isto já não acontece, pois os dois processos são distintos. Embora teoricamente exista uma relação entre as duas chaves, na prática o código é construído de forma a ser praticamente impossível descobrir em tempo útil esta relação. Novamente vamos explorar a limitação dos métodos actuais de factorização.

Um resultado central que permite estabelecer o algoritmo RSA é um teorema demonstrado por Euler.

TEOREMA 3. *Sejam n e b números inteiros primos entre si. Então:*

$$b^{\phi(n)} \equiv 1 \pmod{n}, \tag{3}$$

onde $\phi(n)$ é a função de Euler definida da seguinte forma:

$$\phi(n) = \#\{k \in \mathbb{N} : \mathrm{mdc}(k, n) = 1 \text{ e } k < n\}.$$

Começamos por escolher dois números primos diferentes p e q e calculamos o seu produto $n = p \times q$. De seguida escolhemos e e d, as chaves de encriptação e desencriptação respectivamente, que devem verificar:

$$e \times d \equiv 1 \pmod{\phi(n)}. \tag{4}$$

Uma forma simples de encontrar um par $\{e, d\}$ que verifique a equação (3) é dada pela proposição seguinte.

PROPOSIÇÃO. *Se e for escolhido de forma a que* $\mathrm{mdc}(e, \phi(n)) = 1$*, então existe um único d que verifica* (4)*, e diz-se que e e d são inversos módulo $\phi(n)$.*

Demonstração. Pelo algoritmo de Euclides é possível encontrar inteiros α e β que verificam

$$e \times \alpha + \phi(n) \times \beta = 1.$$

Logo uma possibilidade é tomar $d = \alpha$. Vejamos que é única: suponhamos que c é uma outra solução distinta de d, então

$$c \times e \equiv 1 \pmod{\phi(n)},$$
$$c \equiv c \times (e \times d) \equiv (c \times e) \times d \equiv d \pmod{\phi(n)}. \qquad \square$$

Finalmente a chave e é publicada juntamente com n, enquanto que d é guardado num local seguro. Denotamos por (e, n) uma chave pública RSA. Temos também de fixar uma forma de codificar texto sob a forma de números, por exemplo:

$$\text{'espaço'} = 99, \quad \texttt{A} = 10, \quad \texttt{B} = 11, \quad \ldots, \quad \texttt{Z} = 35.$$

Para enviar uma mensagem M, as únicas restrições sobre M são:

$$\text{mdc}(M, n) = 1 \quad \text{e} \quad M < n,$$

que podem ser sempre conseguidas se M for menor que p e q. Caso contrário partimos a mensagem M em blocos menores que verifiquem as restrições.

Para encriptar a mensagem basta simplesmente calcular

$$E \equiv M^e \pmod{n},$$

e para desencriptar

$$M \equiv E^d \pmod{n};$$

como se segue do seguinte resultado.

PROPOSIÇÃO. *Se* $e \times d \equiv 1 \pmod{\phi(n)}$ *e* $\text{mdc}(M, n) = 1$*, então:*

$$M \equiv (M^e)^d \pmod{n}.$$

Demonstração.

$$(M^e)^d \equiv M^{e \times d} \equiv M^{k\phi(n)+1} \equiv M \times M^{k\phi(n)} \equiv M \times 1 \equiv M \pmod{n}$$

por aplicação do Teorema 3. $\square$

Exemplo 4. Vamos encriptar a palavra `TEOREMA`. O primeiro passo é fixar uma chave.

Escolhemos $p = 31$ e $q = 23$, logo $n = 713$. Calculamos $\phi(713) = 22 \times 30 = 660$ e escolhemos um e tal que $\text{mdc}(e, 660) = 1$ para garantirmos que existe um d que verifica (4). Por exemplo, $e = 7$. Aplicando o algoritmo de Euclides obtemos $d = 283$, publicamos a chave pública $(7, 713)$ e guardamos d num local seguro.

O segundo passo é traduzir a mensagem para uma forma numérica, utilizando a relação anteriormente sugerida, temos:

$$\texttt{T} = 29, \quad \texttt{E} = 14, \quad \texttt{O} = 24, \quad \texttt{R} = 27, \quad \texttt{E} = 14, \quad \texttt{M} = 22, \quad \texttt{A} = 10.$$

A mensagem é $M = 29142427142210$ e não verifica $M < n$, pelo que tem de ser partida em k blocos mais pequenos $M_1, \ldots, M_k$ tais que $\text{mdc}(M_k, n) = 1$.

$$M_1 = 029, \quad M_2 = 142, \quad M_3 = 427, \quad M_4 = 142, \quad M_5 = 210.$$

Facilmente se verifica que $\text{mdc}(M_i, 713) = 1$. Para encriptar M calculamos:

$$E_i \equiv M_i^7 \pmod{713};$$
$$E_1 = 647, \quad E_2 = 629, \quad E_3 = 561, \quad E_4 = 629, \quad E_5 = 623.$$

Logo, $M^e = 647629561629623$.

Para ter de volta a mensagem original basta fazer $M_i \equiv E_i^{283} \pmod{713}$.

A relação entre a chave de encriptação e desencriptação é dada pela equação (3). Como e e n são públicos basta calcular $\phi(n)$ e aplicar o algoritmo de Euclides para quebrar o código. O difícil para quem conhece apenas (e, n) é calcular $\phi(n)$; prova-se que se $n = p \times q$ onde p e q são dois números primos distintos, então $\phi(n) = (p-1) \times (q-1)$, logo para calcular $\phi(n)$ de um modo simples é necessário conhecer a factorização de n. Aliás, conhecer $\phi(n)$ é conhecer a factorização.

Proposição. *Se p e q são dois primos distintos e $n = p \times q$ então saber $\phi(n)$ equivale a saber a factorização de n.*

Demonstração. Sem perda de generalidade, assume-se que $p > q$. Se $\phi(n) = (p-1) \times (q-1)$ então podemos calcular $p + q$ e $p - q$,

$$p + q = n - \phi(n) + 1,$$
$$p - q = \sqrt{p^2 + 2pq + q^2 - 4pq} = \sqrt{(p+q)^2 - 4n};$$

por fim,

$$p = \frac{(p+q) + (p-q)}{2} \quad \text{e} \quad q = \frac{(p+q) - (p-q)}{2}. \qquad \square$$

Como p e q são escolhidos suficientemente grandes não vai ser trivial factorizar n.

Uma observação que se pode fazer ao algoritmo RSA é que a desencriptação não tem de ser feita pela ordem inversa da encriptação, isto é, se (e, n) e (e', n') são duas chaves públicas RSA, então $(M^e)^{e'}$ não tem de ser primeiro desencriptado por d'.

Exemplo 5 (Jogar às cartas). Consideremos um jogo de cartas qualquer. Começamos por definir duas chaves públicas RSA, (e, n) e (e', n') e codificamos o baralho da seguinte forma: $\clubsuit = 10$, $\heartsuit = 20$, $\spadesuit = 30$, $\diamondsuit = 40$ seguindo-se um número de 01 a 13 consoante a carta; por exemplo, 1001 denota o ás de paus. O baralho é um conjunto $\mathcal{B} = \{M_1, M_2, \ldots, M_{52}\}$.

Agora temos o Sr. I e o Sr. J numa conversa telefónica e vamos admitir que o Sr. I tem o baralho e encripta-o recorrendo à chave (e, n) passando a $\mathcal{B}^e = \{M_1^e, M_2^e, \ldots, M_{52}^e\}$.

De seguida transmite $\mathcal{B}^e$ ao Sr. J que depois de o baralhar distribui as cartas, uma para cada um. O Sr. J retira M_j^e para si e envia M_i^e ao Sr. I. O Sr. J não sabe qual é o seu jogo, ao contrário do seu adversário que conhece d, mas pode encriptar a carta M_j^e com a sua chave (e', n') e enviar também $(M_j^e)^{e'}$ ao Sr. I.

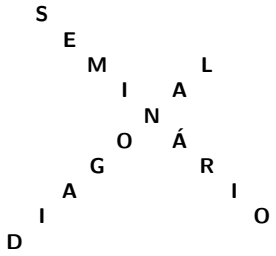
Tendo em conta a observação anterior é imediato reconhecer que o Sr. I pode aplicar d a $(M_j^e)^{e'}$ para ficar com $M_i^{e'}$ e não consegue ver o jogo do Sr. J porque não conhece d'. O Sr. J quando receber $M_i^{e'}$ aplica-lhe d' e passa a saber o seu jogo.

Uma vez que é possível dar cartas é possível jogar qualquer jogo e indo tomando nota de todas as jogadas. No fim ambos os jogadores revelam as suas chaves e verificam se alguém fez batota.

Referências

[1] David M. Bressoud. *Factorization and Primality Testing.* Springer-Verlag, 1989.

[2] Charles Vanden Eynden. Flipping a coin over the telephone. *Mathematics Magazine*, Vol. 62, No. 3, June 1989.



O Teorema de Pitágoras

Luís Russo*
3º ano da LMAC — Ciência da Computação
Luis.Russo@math.ist.utl.pt

Palavras Chave

Pitágoras, triângulo rectângulo, hipotenusa, cateto, Teorema de Pitágoras, demonstrações.

Resumo

Sabia que Pitágoras não foi o primeiro a descobrir o Teorema de Pitágoras? Sabia que são conhecidas cerca de 380 demonstrações independentes deste resultado que tem fascinado gerações pela sua simplicidade? A abordagem destas questões, bem como algumas curiosidades históricas com elas relacionadas, constitui o tema deste artigo.

Introdução

O Teorema de Pitágoras chegou até aos nossos dias sob a forma da Proposição 47 do 1º livro dos "Elementos" de Euclides [1].

Teorema de Pitágoras. *Num triângulo rectângulo, a área do quadrado construído sobre a hipotenusa iguala a soma das áreas dos quadrados construídos sobre os catetos.*

Observando o enunciado deste teorema, podemos concluir que pode ser interpretado segundo duas perspectivas: uma geométrica, que estabelece uma relação entre figuras, e outra algébrica, que traduz uma relação entre números. Os triplos de inteiros que satisfazem esta relação são designados ternos pitagóricos.

Antes de continuar, convém referir que as fontes que constituem a base deste trabalho não são coerentes entre si. Sendo tal incoerência frequente em pesquisas históricas, ainda é mais natural quando se trata do estudo de uma personagem que viveu há mais de dois milénios. Não obstante, segue-se um resumo de uma das mais plausíveis versões da vida de Pitágoras.

* O autor deste artigo foi parcialmente apoiado em 2000–01 por uma bolsa do Programa Gulbenkian Novos Talentos em Matemática.

Pitágoras nasceu cerca de 575 a.C., na ilha de Samos, que se situa no mar Egeu. Com a idade de 18–20 anos, ter-se-á deslocado a Mileto, onde veio a conhecer Tales e Anaximandro, seu aluno. Estes dois filósofos fizeram nascer nele o interesse por ciências como a matemática e a astronomia. Tales, sendo na época já idoso, não chegou a contribuir efectivamente como professor para a formação matemática do jovem Pitágoras, mas tê-lo-á aconselhado a visitar o Egipto, onde poderia aprofundar os seus conhecimentos, como veio a acontecer em relação à Geometria. Assim, cerca de 535 a.C., o Egipto veio a recebê-lo na qualidade de estudante interessado na cultura local. Com efeito, foi visitando templos e conversando com os sacerdotes que adquiriu muitas das suas práticas, tais como o secretismo (que veio a tornar-se uma das principais razões de nenhum dos seus escritos ter chegado aos nossos dias), a recusa em ingerir feijão, a não utilização de vestes confeccionadas com peles de animais e a busca da pureza.

Por volta de 525 a.C., o Egipto é invadido pelo rei da Pérsia e Pitágoras é feito prisioneiro e levado para a Babilónia. É de assinalar que — conforme comprovado por uma placa de barro babilónia actualmente na Universidade de Columbia — este povo já conhecia o famoso teorema cerca de 1000 anos antes de Pitágoras o ter descoberto. Esta placa exibe vários ternos pitagóricos dispostos numa tal ordem que faz suspeitar da utilização de um método operacional. Não se sabe, portanto, se ele o terá descoberto a partir do caso $(3, 4, 5)$ eventualmente conhecido dos egípcios ou se o terá aprendido no cativeiro.

Anos mais tarde veio (não se sabe como) a deixar a Babilónia e tornar-se, talvez cerca de 518 a.C., o fundador de uma Escola filosófica e religiosa em Crotona, Itália. Esta Escola era formada por homens e mulheres que viviam em comunidade, não tinham posses individuais e eram vegetarianos. Pitágoras, o professor e líder, impôs aos seus discípulos a natureza matemática da realidade, a filosofia como purificação espiritual e a obrigatoriedade de todos praticarem a lealdade e o secretismo. Faziam ainda parte da Sociedade alguns membros externos que viviam nas suas próprias casas; não sendo portanto desprovidos de posses, nem obrigados ao vegetarianismo. É importante referir alguns aspectos visionários desta Escola, como o facto de acreditarem que a Terra é redonda e de partilharem a perspectiva heliocêntrica do Sistema Solar.

Como já foi sugerido, a ideologia de Pitágoras concebia os números e as relações entre eles como a base de toda a organização do Universo. Tanto assim que o matemático, também amante da música, veio a dar um valioso contributo a esta arte, estudando a relação entre as razões (proporções) em que se encontram os comprimentos das cordas de alguns instrumentos e os sons que produzem. Vestígios dos seu trabalho chegaram até aos nossos dias,

como a afinação pitagórica ainda usada em alguns instrumentos.

É de assinalar que para Pitágoras todos os números podiam ser expressos como razões de outros, pois os irracionais eram ainda desconhecidos. É interessante referir que, a par do seu estudo de propriedades importantes dos números (como o facto de serem pares, ímpares, primos, ...), Pitágoras atribuía-lhes personalidade própria e outras características, classificando-os como masculinos ou femininos, perfeitos ou incompletos e mesmo bonitos ou feios.

Devido ao mistério que rodeava a Sociedade, é actualmente ténue a distinção entre as descobertas de Pitágoras e as contribuições dos pitagóricos. Sabe-se, porém, que a Matemática da época era bastante diferente da actual. Assim, é difícil nos nossos dias compreender a originalidade do seu trabalho, pois estamos familiarizados com a abstracção matemática e com o acto mental de generalização. Mas, no tempo em que Pitágoras viveu, a passagem da realidade concreta (2 barcos + 2 barcos = 4 barcos) para as relações numéricas abstractas associadas (2 + 2 = 4) foi um grande avanço. Assim, a importância de Pitágoras advém sobretudo do facto de ter sido um dos primeiros (talvez mesmo o primeiro) a formular os conceitos de número, triângulo e ideia abstracta de demonstração. Torna-se, portanto, natural que se tenha tornado no primeiro a provar o famoso teorema.

Inspirados pelo trabalho de Pitágoras, e provavelmente motivados pela beleza e simplicidade do teorema que herdou o seu nome, muitos matemáticos, cientistas, e mesmo simples curiosos têm apresentado inúmeras demonstrações deste resultado. Em 1907, Elisha Loomis [4] fez uma compilação de 386 destas demonstrações (que viria a ser publicada pela primeira vez em 1927), classificando-as em quatro grupos:

- as geométricas, que consistem basicamente em manipular figuras (tipicamente aplicando isometrias) de maneira a que se verifique a relação pretendida entre as suas áreas;

- as algébricas, que partem de figuras geométricas e, através das relações numéricas entre comprimentos e/ou áreas associados, obtêm, após alguns cálculos, o teorema;

- as vectoriais, que consistem em utilizar propriedades de operações entre vectores (como o produto interno);

- as dinâmicas, que se baseiam em princípios físicos.

Seguem-se alguns exemplos, com a respectiva classificação.

1 Demonstrações Geométricas

A demonstração seguinte é a que aparece nos "Elementos", e recorre à proposição auxiliar que se segue, cuja prova não é aqui feita.

PROPOSIÇÃO. *Dadas duas rectas paralelas e dois pontos sobre uma das rectas, quaisquer triângulos cujo terceiro vértice esteja sobre a outra recta têm a mesma área.*

Demonstração (do Teorema). Dado o triângulo rectângulo ABC, com ângulo recto em A, desenhe-se os quadrados $FGAB$ e $EDBC$ sobre os lados AB e BC, respectivamente; como na Figura 1.

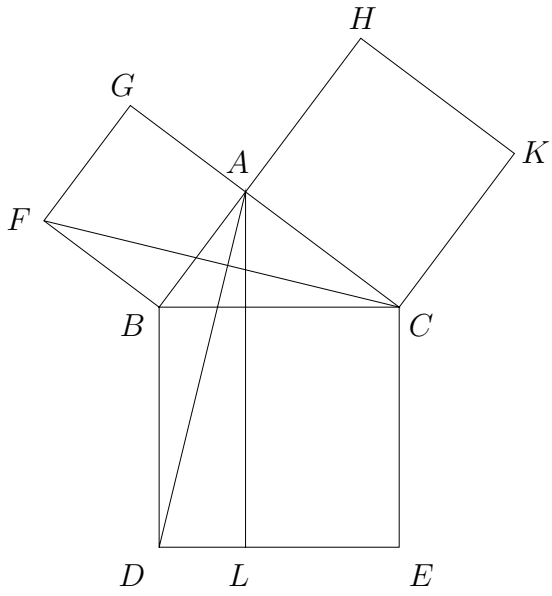


Figura 1: Demonstração nos "Elementos" de Euclides

Baixe-se uma altura de A sobre BC, prolongando-a até DE e chamando L ao ponto de intersecção.

O triângulo FBA mede metade da área de $FGAB$. Logo, pela proposição acima, FBC também mede metade da área de $FGAB$.

Por outro lado, o triângulo BDL tem metade da área do rectângulo que contém os pontos B, D e L como vértices. Logo, novamente pela proposição acima, ABD também tem metade da área do dito rectângulo.

Finalmente, podemos observar que FBC é o triângulo que resulta de aplicar uma rotação de $+90^{\circ}$ a ABD e, portanto, FBC e ABD têm a mesma área.

Assim, concluímos que $FGBA$ tem a mesma área que o rectângulo do qual são vértices D, B e L.

Repetindo o argumento para o outro lado do triângulo, concluímos que a área do quadrado associado ao cateto AC é a mesma que a do rectângulo que tem como vértices C, E e L, obtendo-se assim o resultado desejado. □

A demonstração que se segue é a que se suspeita ter sido descoberta por Pitágoras.

Demonstração. Desenhe-se dois quadrados de lado $a+b$, dividindo-os como na Figura 2.

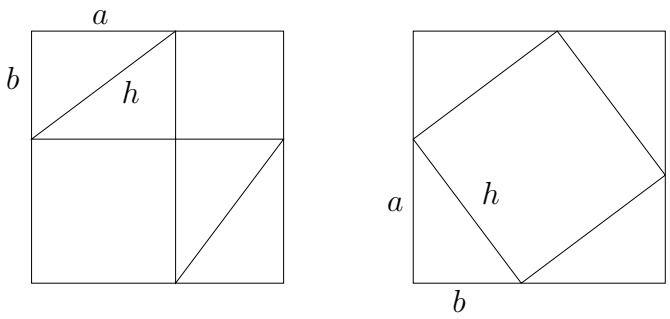


Figura 2: Demonstração atribuída a Pitágoras

É fácil observar que o quadrado da esquerda se decompõe em 4 triângulos e dois quadrados de lados a e b. Por outro lado, o quadrado da direita é formado pelos mesmos 4 triângulos e um quadrado de lado h. Logo, a área do quadrado construído sobre h tem que igualar a soma das áreas dos dois quadrados menores da figura da esquerda, porquanto as áreas dos quadrados exteriores são iguais. □

2 Demonstrações Algébricas

Segue-se uma demonstração que se supõe que Pitágoras veio mais tarde a conhecer.

Demonstração. Tome-se um triângulo rectângulo, de catetos a, b e hipotenusa h.[1] Baixe-se uma altura x sobre a hipotenusa, sendo y a distância do ponto onde a altura intersecta a hipotenusa ao vértice de a sobre a hipotenusa (ver a Figura 3).

Podemos então concluir, por semelhança de triângulos, que

$$a : b : h = y : x : a = x : h - y : b.$$

1 De acordo com a prática comum, confundimos os segmentos com os respectivos comprimentos.

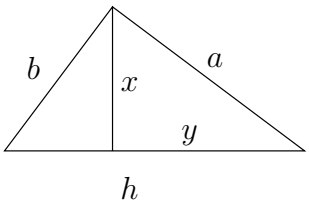


Figura 3: Uma Demonstração Algébrica

A equação acima representa, na realidade, 9 relações de semelhança, das quais é possível extrair 44 provas distintas [4]. Vamos apresentar a menor delas, ou seja, a que exige menos equações e menos passos. Temos

$$\frac{a}{y} = \frac{h}{a};$$

logo,

$$a^2 = hy. \tag{1}$$

Por outro lado,

$$\frac{b}{h-y} = \frac{h}{b};$$

e, tal como em (1),

$$b^2 = h^2 - hy.$$

Somando as equações obtidas, vem

$$a^2 + b^2 = h^2.$$ □

Nota. No livro de Elisha Loomis existe um esquema em que intervêm 7 triângulos semelhantes, e do qual é possível extrair cerca de 65000 demonstrações distintas.

3 Demonstrações Vectoriais

Demonstração. Sejam $\mathbf{a}$ e $\mathbf{b}$ vectores ortogonais e defina-se $\mathbf{h}$ por

$$\mathbf{h} = \mathbf{a} + \mathbf{b}.$$

Assim,

$$\langle \mathbf{h}, \mathbf{h} \rangle = \langle \mathbf{a}, \mathbf{a} \rangle + 2\langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{b}, \mathbf{b} \rangle.$$

Visto que $\mathbf{a}$ e $\mathbf{b}$ são ortogonais, temos que

$$\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{0}.$$

Logo,

$$\langle \mathbf{h}, \mathbf{h} \rangle = \langle \mathbf{a}, \mathbf{a} \rangle + 2 \cdot 0 + \langle \mathbf{b}, \mathbf{b} \rangle = \langle \mathbf{a}, \mathbf{a} \rangle + \langle \mathbf{b}, \mathbf{b} \rangle.$$

E portanto,[2]

$$\|\mathbf{h}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2. \qquad \square$$

As demonstrações que se seguem podem ser classificadas como geométricas e aparecem aqui pela sua beleza e interesse histórico. A primeira delas é devida a Leonardo da Vinci.

Demonstração. Faça-se a construção na Figura 4, que resulta de desenhar os quadrados associados aos lados do triângulo rectângulo ABH, de unir F a E por um segmento de recta e de desenhar um triângulo rectângulo igual ao primeiro sobre o lado oposto do quadrado maior.

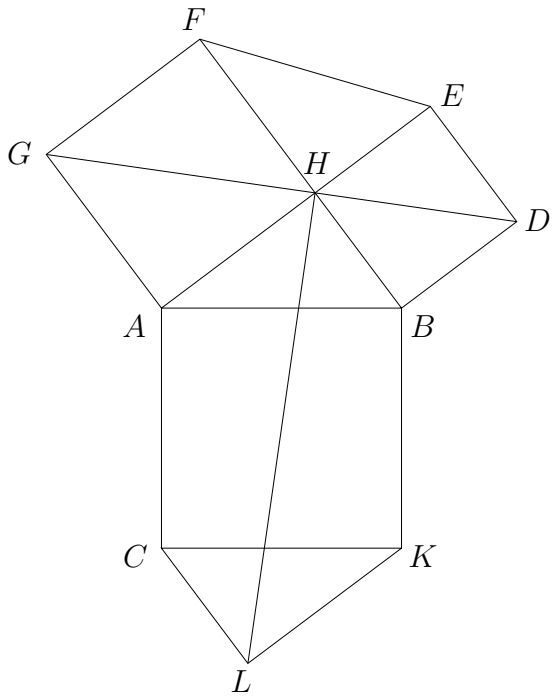


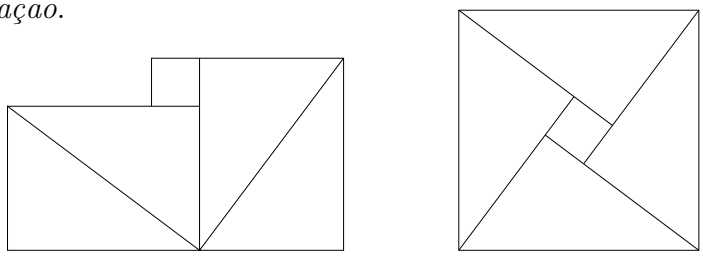
Figura 4: Demonstração de Leonardo da Vinci

2 Analisando atentamente esta demonstração, podemos verificar que apresenta uma certa circularidade, visto que a definição de norma no espaço euclidiano é baseada no Teorema de Pitágoras...

Por um lado, os triângulos AHB, FHE e CLK são iguais (seja x a sua área). Por outro, observando atentamente os trapézios $GFED$, $AGDB$, $AHLC$ e $HBKL$ podemos concluir que são todos iguais e portanto os hexágonos $AGFEDB$ e $AHBKLC$ têm a mesma área. Ora, se esta área se pode decompor quer nos quadrados dos catetos e dois triângulos de área x, quer no quadrado da hipotenusa e dois triângulos com a mesma área x, tal significa que a área do quadrado maior é igual à soma das áreas dos quadrados construídos sobre os catetos. □

Finalmente, apresenta-se a demonstração concebida por Bhaskara (1114–1185), matemático indiano que a fez sem qualquer argumentação explícita. No sentido de preservar esse espírito, característico da matemática hindu da época, aqui fica um esboço da construção que usou:

Demonstração.



□

Agradecimentos

Queria agradecer a algumas pessoas sem as quais, quer a elaboração deste trabalho, quer a apresentação do seminário, teriam sido impossíveis. A saber: Professor Félix Costa, Professor Paulo Almeida, Luís Cruz-Filipe, João Boavida, João Marco Gonçalves e Rita Sanchas.

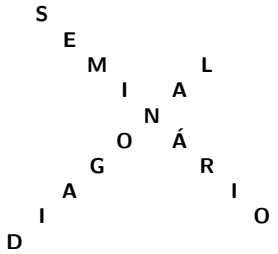
Referências

[1] Euclid. *The Thirteen Books of Euclid's Elements*, Vol. 1. Sir Thomas Heath (ed.). Dover Publications Inc., 1956.

[2] Richard J. Gillings. *Mathematics in the Time of the Pharaohs*. Dover Publications Inc., 1982.

[3] Sir Thomas Heath. *A History of Greek Mathematics*, Vol. 1. Dover Publications Inc., 1981.

[4] Elisha Scott Loomis. *The Pythagorean Proposition*. Washington: National Council of Teachers of Mathematics, 1968.



Um Passeio Pouco Aleatório

João Pedro Boavida
5º ano da LMAC — Análise, Geometria e Álgebra
jboavida@math.ist.utl.pt

Palavras Chave

movimento browniano, equações diferenciais estocásticas, mudança de escala, ruído, funções harmónicas.

Resumo

Normalmente não nos apercebemos como é frequente que fenómenos que em pequena escala são totalmente deterministas se revelem verdadeiramente aleatórios na escala 'de todos os dias'. Basta pensar na trajectória de um grão de poeira, ou na imagem de um raio nos céus.

Neste artigo vamos descrever o movimento browniano e usá-lo como modelo de ruído em equações diferenciais, o que, como veremos, nos trará algumas surpresas. No final, um passeio curto por Monte-Carlo para explorar algumas propriedades das funções harmónicas.

Introdução

'There are Heroisms All Round Us'

(título do primeiro capítulo de [3])

A intenção que preside à organização deste texto é mostrar que os processos estocásticos, em particular os ruídos aleatórios, estão *mesmo* por todo o lado, à nossa volta. Se, como fez o botânico Robert Brown em 1828, seguirmos ao microscópico o movimento de um grão de pólen suspenso numa gota de água, veremos, como ele, um movimento totalmente irregular e imprevisível, hoje chamado *movimento browniano*. É omnipresente.

Se observarmos a trajectória de uma partícula de giz, suspensa no ar, se imaginarmos o que acontece à sua volta, à escala microscópica, perceberemos que o seu mundo é um lugar agitado. De facto, a todo o momento essa partícula é atingida por moléculas do ar, vindas de direcções imprevisíveis, e que com ela chocam com violência também imprevisível. Basta pensar na tempestade causada pelo simples movimento de uma pessoa, ou na torrente de ar que sai da boca de alguém que está, por exemplo, a apresentar um

seminário.[1] Sem falar da eventualidade de ser desviada pelo movimento das asas de alguma mosca de passagem.

A queda de um raio é precedida da ionização de um canal favorável, o qual é determinado por uma sucessão de partículas precursoras, guiadas pelas linhas de força do campo eléctrico (Cf. [7]). Se pensarmos na imagem de um raio nos céus, se tentarmos ver o mundo à volta da precursora, à sua escala, concluímos que não pode ser muito homogéneo. A simples passagem de um avião distante, ou a ignição de um motor de automóvel, podem modificar quase imperceptivelmente o campo eléctrico próximo da partícula e desviá-la noutra direcção, e a passagem de uma partícula ionizada vinda da alta atmosfera pode desorientá-la completamente.

E porque não observar uma mancha de tinta a difundir-se na água, e imaginar o que se passa na escala das moléculas? Afinal, a consideração desses fenómenos permitiu a Einstein [4] explicar o movimento browniano e relacioná-lo com a difusão.

Ou podemos ainda tentar seguir o que acontece a duas populações de seres vivos em competição no meio-ambiente. Podemos considerar mais que duas espécies; ou entrar em linha de conta com os factores meteorológicos, ou (numa escala de tempo maior) climáticos. Ou podemos ainda estudar a influência dos acontecimentos geológicos (a separação de África, Antárctida e Austrália, o choque da Índia com a Ásia, ou um mar interior que seca), já sem falar de acidentes astronómicos (como a queda de um meteoro). Se a tudo isso juntarmos a possibilidade de as próprias espécies se modificarem ao longo do tempo, não há como duvidar que os factores aleatórios possam ter uma palavra importante a dizer no destino das espécies.[2]

Não discutiremos uma abordagem estocástica para um problema tão sofisticado como a teoria da evolução. O nosso objectivo é muito mais modesto:

Vamos começar por descrever o movimento browniano, e estudar algumas das suas propriedades. Aqui o exemplo a ter em mente será a partícula de giz suspensa no ar. Depois desta preparação, veremos alguns exemplos de equações diferenciais em que o movimento browniano é usado como modelo dos ruídos aleatórios que por vezes devem ser tidos em conta: são as chamadas *equações diferenciais estocásticas*.

1 Ou, porque não, da boca de alguém que está a dormitar durante um seminário?

2 Mais exemplos de factores que podem interferir na evolução das espécies, cuidadosamente discutidos, podem ser encontrados num dos mais notáveis livros de ciência e divulgação alguma vez escrito: *On The Origin of Species by Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life* [1].

1 Movimento Browniano como Ruído

Antes de discutir propriedades matemáticas do movimento browniano, convém explicitar aquilo em que estamos a pensar. Trata-se do ruído aleatório acumulado ao longo do tempo. Como isto é ainda demasiado vago, teremos que dizer algo mais antes de podermos avançar. Mas para já os exemplares na Figura 1 podem dar uma ideia daquilo que procuramos descrever.

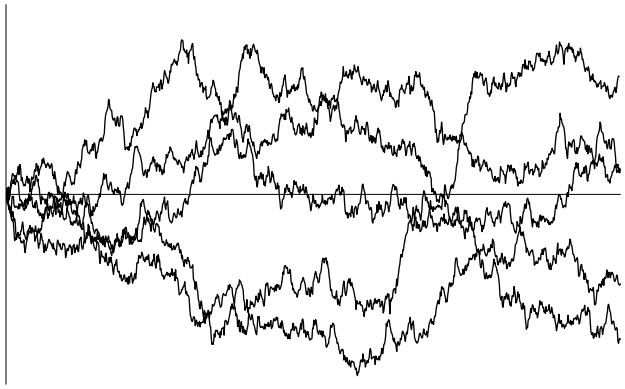


Figura 1: Exemplos de Movimento Browniano como Ruído Acumulado

Não é demais repetir, o exemplo a ter em mente é o movimento aleatório, numa direcção à escolha,[3] de uma partícula de giz ou de um grão de pó suspenso no ar. Vamos chamar B_t à posição da partícula no instante t. Trata-se de uma variável aleatória. Tomamos $\Omega = \{\text{todos os caminhos contínuos } \omega : \mathbb{R}_0^+ \to \mathbb{R} \mid \omega(0) = 0\}$.[4]

1.1 Valor Esperado $E[B_t]$ e Incrementos $B_{t+s} - B_t$

A primeira propriedade relevante do movimento browniano é que o afastamento médio é nulo, isto é, a melhor estimativa possível de $B_t = B_t(\omega)$ sem usar informação adicional é 0. Vamos escrever portanto $E[B_t] = 0$.[5]

3 Poderíamos tentar atacar directamente as 3 dimensões, mas será mais fácil começar por olhar apenas numa direcção específica.

4 Não vamos discutir aqui os fundamentos da teoria das probabilidades, mas a ideia é que temos um espaço Ω de todas as histórias possíveis — neste caso $\Omega = \{$todos os caminhos (contínuos) $\omega : \mathbb{R}_0^+ \to \mathbb{R} \mid \omega(0) = 0\}$, sendo portanto B_t uma abreviatura para $B_t(\omega) = \omega(t)$ — e que conseguimos atribuir probabilidade a alguns subconjuntos específicos de Ω — por exemplo, $\{\omega \mid \omega(t) \geq 0\}$ ou $\{\omega \mid \omega$ sai do intervalo $[2, 5]$ até ao instante $3\}$.

5 Onde usamos $E[\cdot]$ para denotar o *valor esperado*. Também aqui não vamos entrar em pormenores; diremos apenas que, para quem esteja à vontade com a teoria de integração, vale *mesmo* a pena pensar em $E[\cdot]$ como um integral sobre um conjunto Ω com medida total 1.

Outro aspecto importante é que o movimento da partícula de giz parece exactamente igual (excepto no ponto de partida) independentemente do momento em que começamos a observá-lo — dizemos que tem *incrementos estacionários* —, e o passado não parece afectar o futuro (excepto no ponto de partida) — dizemos que tem *incrementos independentes*.[6] Isto mesmo pode ser observado em três dos passeios da figura anterior, que embora pareçam totalmente casuais, são de facto parte de uma mesma trajectória.

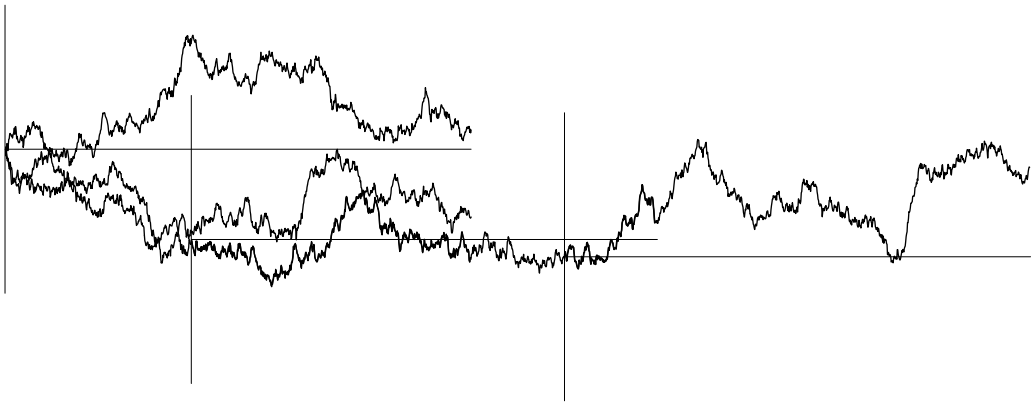


Figura 2: Incrementos Independentes e Estacionários

1.2 Função Variância e Distribuição de B_t

O passo natural depois de compreender a média de B_t, é tentar compreender a sua variância. Uma vez que $E[B_t] = 0$, a variância é simplesmente $E[{B_t}^2]$. Vejamos então (na Figura 3) se as imagens de $B_t(\omega)^2$ (onde ω é um dos cinco caminhos das figuras anteriores) nos ajudam a formar uma ideia.

Não parece muito animador, mas ilustra bem o facto de se tratar de movimentos altamente irregulares, com retornos à origem, e aparentando mudanças bruscas de direcção/sentido. Em qualquer caso não é difícil calcular a função variância $V_t = E[{B_t}^2]$.

$$\begin{aligned} V_{t+s} &= E\left[(B_{t+s} - B_t + B_t)^2\right] \\ &= E\left[(B_{t+s} - B_t)^2\right] + 2E\left[(B_{t+s} - B_t)B_t\right] + E\left[{B_t}^2\right] \\ &= V_s + 0 + V_t = V_t + V_s \end{aligned}$$

6 A *independência* é outro conceito importante na teoria de probabilidades. Dizer que X é independente de Y é dizer que o conhecimento de Y não nos dá qualquer informação sobre X (e reciprocamente). Neste texto a consequência mais importante é que se X e Y são independentes, então $E[XY] = E[X]\,E[Y]$.

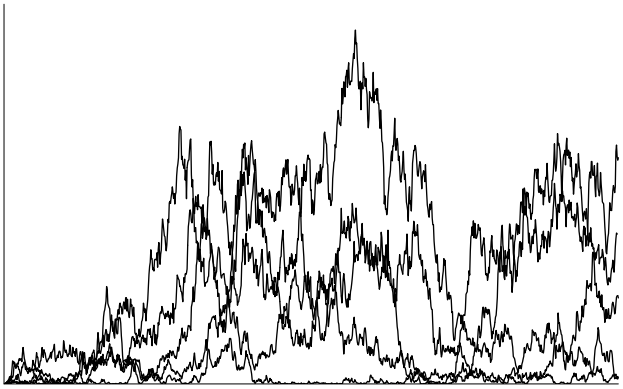


Figura 3: Quadrados dos Passeios Anteriores

Na última linha usámos o facto de os incrementos serem estacionários (donde $E\left[(B_{t+s} - B_t)^2\right] = E\left[{B_s}^2\right]$) e independentes (donde $E\left[(B_{t+s} - B_t)B_t\right] = 0$). Daqui é fácil concluir que $V_q = qV_1$, para $q \in \mathbb{Q}$. Por outro lado, sendo $V_t \geq 0$, concluímos também que V_t é crescente. Estes dois factos em conjunto permitem obter $V_t = \lambda t$, onde $\lambda = V_1$.

Uma primeira constatação é que podemos fixar $V_1 = 1$ e $V_t = t$, pois qualquer caminho com outro valor de λ pode ser obtido ampliando um destes (de $V_1 = 1$) segundo um factor de $\sqrt{\lambda}$. Por outro lado, $B_0 = 0$, $E\left[{B_t}^2\right] = V_t = t$ e a independência dos incrementos são suficientes para determinar a distribuição de $B_t(\omega)$. Uma possibilidade é usar funções geradoras de momentos, ou transformadas de Fourier, e obter de imediato a função de distribuição. Porém, podemos chegar lá de um modo mais intuitivo.

Vamos admitir por simplicidade que B_t satisfaz uma condição de escala $B_t = \sqrt{t}\, B_1$ (o que é compatível com $V_t = t$), e estudar apenas a distribuição de B_1. Seja $\{Z_i \mid i \in \mathbb{N}\}$ uma família de variáveis aleatórias independentes e identicamente distribuídas, com $Z_i \overset{d}{=} B_1$.[7] Usando os argumentos habituais em teoria de probabilidades, temos

$$B_1 = \sum_{0 \leq i < N} \left(B_{\frac{i+1}{N}} - B_{\frac{i}{N}} \right) \overset{d}{=} \sum_{0 \leq i < N} \frac{Z_i}{\sqrt{N}},$$

que, pelo Teorema do Limite Central, converge em distribuição para uma normal com média 0 e variância 1. Explicado por outras palavras, se dividirmos o intervalo $[0, 1]$ em pedaços cada vez mais pequenos, vamos obter

7 A notação $X \overset{d}{=} Y$ significa que as variáveis aleatórias X e Y têm a mesma distribuição de probabilidades (outro conceito importante). Isso implica, em particular, que $E\left[X\right] = E\left[Y\right]$ e que se f é uma função contínua (ou, mais geralmente, mensurável) então $f(X) \overset{d}{=} f(Y)$.

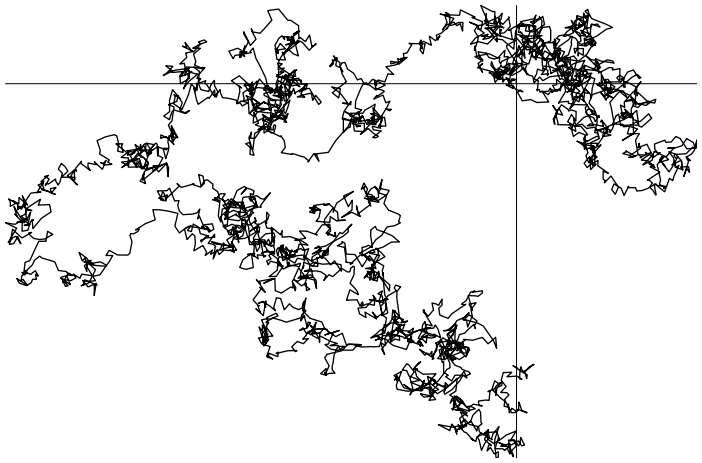


Figura 4: Um Passeio Aleatório em $\mathbb{R}^2$

B_1 como soma de um número cada vez maior de variáveis aleatórias independentes e identicamente distribuídas, e portanto vamos obter, passando ao limite, uma distribuição gaussiana.

Tendo em conta a hipótese de escalamento que fizemos, é também fácil concluir que B_t é uma normal com média 0 e variância t e, mais geralmente, $B_{t+s} - B_t$ é uma normal com média 0 e variância s.

Vale a pena observar que deste modo obtivemos *de facto* um modo de calcular probabilidades de alguns subconjuntos de Ω.[8] Por exemplo, agora é fácil calcular a probabilidade de $B_{3.1} \geq 39.7$, pois é precisamente igual à probabilidade de uma normal de média 0 e variância 3.1 assumir um valor maior ou igual a 39.7.

Isto tem uma consequência inesperada: se fixarmos, como temos feito, $\Omega = \{\text{funções contínuas } \omega : \mathbb{R}_0^+ \to \mathbb{R} \mid \omega(0) = 0\}$, e considerarmos o subconjunto $A = \{\text{funções diferenciáveis em algum ponto}\}$, todos já ouvimos dizer que Ω é muito maior que A. Mas a situação é ainda mais extrema: é possível (mas não imediato) calcular a probabilidade de $\omega \in A$, e essa probabilidade é... 0 !

1.3 Passeios Aleatórios em $\mathbb{R}^2$ e $\mathbb{R}^3$

Já vimos portanto o que queríamos a respeito dos passeios aleatórios em $\mathbb{R}$. Em mais dimensões a situação é inteiramente análoga: um passeio será uma função contínua $\omega : \mathbb{R}_0^+ \to \mathbb{R}^n$ cujas funções componentes são movimentos aleatórios a uma dimensão *independentes*. Vamos chamar $B^1, \ldots, B^n$

8 Trata-se, nem mais nem menos, da medida de Wiener.

às componentes, e portanto usaremos notações como B_t^i para representar $B_t^i(\omega) = \omega^i(t)$.[9] A trajectória de um grão de poeira numa sala é um exemplo de um passeio aleatório em $\mathbb{R}^3$. A Figura 4 mostra uma passeio em $\mathbb{R}^2$, como é o caso da trajectória desse mesmo grão de poeira, vista de cima.

Acontecem coisas estranhas com os movimentos a várias dimensões. Vamos limitar-nos a referir algumas. Por exemplo, com $n = 1$ ou $n = 2$, a probabilidade de um aberto $U \neq \varnothing$ ser visitado um número infinito de vezes é igual a 1. Isto não quer dizer que esteja sempre a ser visitado: o intervalo médio entre visitas é infinito! Já em dimensão 3 não se passa o mesmo, e qualquer caminho aproxima-se de um número comparativamente pequeno de pontos. Por outro lado, seja qual for a dimensão, a probabilidade de sair (independentemente de voltar ou não no futuro) de um aberto limitado é 1.

2 Equações Diferenciais com Ruído

O nosso objectivo, recorde-se, era perceber as equações diferenciais com ruído. Para isso convém começar por perceber as equações diferenciais sem ruído...

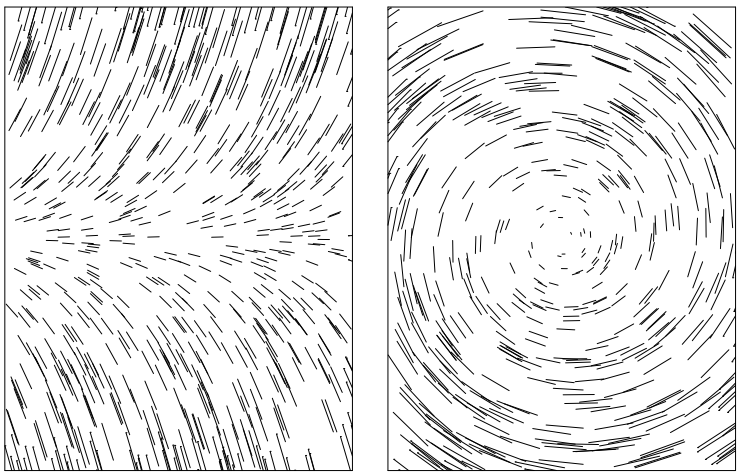


Figura 5: Exemplos de Campos Vectoriais em $\mathbb{R}^2$

A ideia é muito simples: suponhamos que espalhamos uma colecção de vectores em $\mathbb{R}^n$, um vector $v(p)$ em cada ponto p. Vamos supor ainda que esse campo de vectores é definido por uma função suficientemente regular (por

9 Para evitar confusões com potências de expoente i, escreveremos estas num dos formatos ${B_t}^i$, $B_t(\omega)^i$ ou mesmo $(B_t)^i$.

exemplo, com derivadas de todas as ordens). Pode ser um campo de forças, pode ser um campo de direcções. Vamos usar como exemplos os campos da Figura 5, em $\mathbb{R}^2$.

Uma *curva integral* desse campo é um caminho $t \mapsto X_t$ cuja velocidade $\frac{dX_t}{dt}$ em cada instante é igual ao valor $v(X_t)$ do campo vectorial nesse ponto:[10]

$$\frac{dX_t}{dt} = v(X_t).$$

Podemos ver na Figura 6 algumas curvas integrais dos campos da Figura 5.

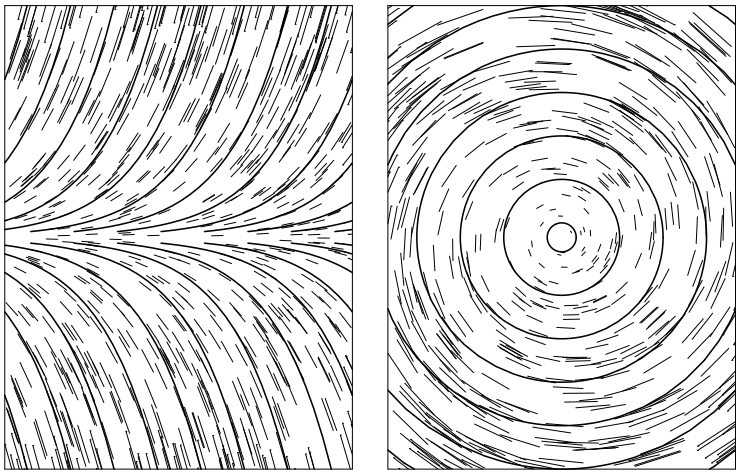


Figura 6: Curvas Integrais dos Campos da Figura 5

Um resultado clássico (Teorema de Picard) garante que, sob certas condições, existe uma e uma só curva integral passando num ponto p no instante 0. Como não é nossa intenção dedicar-nos aos pormenores técnicos, vamos limitar-nos a perceber como essa solução pode ser determinada aproximadamente.

Se escolhermos um intervalo de tempo bastante pequeno $\Delta t \simeq 0$, dizer $\frac{dX_t}{dt} = v(X_t)$ é dizer $\frac{X_{t+\Delta t} - X_t}{\Delta t} \simeq v(X_t)$, ou mais concretamente $X_{t+\Delta t} \simeq X_t + v(X_t)\Delta t$. Assim, não será motivo de espanto que em alguns casos seja possível aproximar[11]

$$X_{N\Delta t} \simeq X_0 + \sum_{0 \leq i < N} v(X_{i\Delta t})\Delta t.$$

10 Não é incomum usar X_p para representar o valor do campo X no ponto p. É igualmente comum escrever X_t para representar um processo estocástico, i.e., uma variável aleatória que varia no tempo, e é com este último significado que tal notação será usada.

11 Note-se que a aproximação depende dos valores anteriores de $X_{i\Delta t}$ para a determinação de v.

A Figura 7 (com valores de Δt propositadamente muito grandes) ilustra em que medida um tal método pode ser ou não fiável.

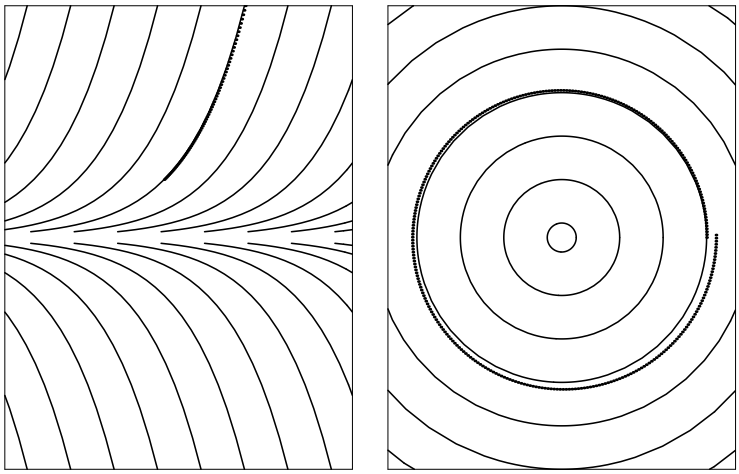


Figura 7: Método de Euler

Pelo menos em certos contextos faz sentido ir mais longe e escrever[12]

$$X_T = X_0 + \int_0^T v(X_t)dt.$$

2.1 Integrais Estocásticos

Mas o que queríamos mesmo era algo como

$$\frac{dX_t}{dt} = v(X_t) + f(t, X_t) \text{ 'ruído acumulado'},$$

e se o 'ruído acumulado' é dado por B_t (o passeio aleatório que definimos há pouco), esperaríamos que o 'ruído instantâneo' fosse dado por $\frac{dB_t}{dt}$ para obtermos algo como

$$\frac{dX_t}{dt} = v(X_t) + f(t, X_t)\frac{dB_t}{dt},$$

e que seria muito semelhante a uma equação sem ruído.

Não obstante, há uma dificuldade grave: não há como definir $\frac{dB_t}{dt}$ sensatamente. Isto porque (consultar de novo as figuras com B_t e ${B_t}^2$) a função B_t varia *mesmo* muito e muito bruscamente. De facto, já tínhamos referido que a probabilidade de B_t ter derivada em algum ponto é 0. Por outro lado $E\left[\frac{B_t}{t}\right] = 0$, parecendo indicar que está tudo bem. Mas basta olhar para a

12 Note-se que aqui se trata de uma igualdade.

variância $E\left[\frac{{B_t}^2}{t^2}\right] = \frac{t}{t^2} = \frac{1}{t}$ para perceber que quanto menor o valor de t, tanto maior a dispersão de $\frac{B_t}{t}$, pelo que o limite não fica bem definido.

Porém, devia fazer sentido escrever

$$\frac{X_{t+\Delta t} - X_t}{\Delta t} \simeq v(X_t) + f(t, X_t)\frac{B_{t+\Delta t} - B_t}{\Delta t}$$

e portanto

$$X_{t+\Delta t} \simeq X_t + v(X_t)\Delta t + f(t, X_t)(B_{t+\Delta t} - B_t),$$

ou até mesmo, em alguns casos,

$$X_{N\Delta t} \simeq X_0 + \sum_{0\leq i<N} v(X_{i\Delta t})\Delta t + \sum_{0\leq i<N} f(i\Delta t, X_{i\Delta t})\Big(B_{(i+1)\Delta t} - B_{i\Delta t}\Big).$$

Na verdade, quase seríamos tentados a escrever

$$X_T = X_0 + \int_0^T v(X_t)dt + \int_0^T f(t, X_t)dB_t.$$

E com efeito, faz sentido definir esta expressão como o limite da anterior quando $\Delta t \to 0$, desde que f satisfaça certas condições (por exemplo, é mais do que suficiente que seja contínua). O integral assim definido é chamado *integral estocástico*.[13] Também é comum escrever

$$dX_t = v(X_t)dt + f(t, X_t)dB_t.$$

É claro que as dificuldades ainda não acabaram. Será que não vamos ter problemas se a variável aleatória $f(t, X_t)$ depender do futuro de B_t? Para evitar tais situações, e já que estamos a descrever sistemas que só podem depender do passado e que não adivinham o futuro distante, vamos *sempre* supor que a variável f é *previsível*, ou seja, que $f(t, \cdot)$ é independente de $B_{t+s} - B_t$ quando $s \geq 0$. Note-se que as funções contínuas são sempre previsíveis.

Isto tem uma consequência altamente desejável: como

$$E\left[f(t, X_t)(B_{t+s} - B_t)\right] = 0,$$

passando ao limite na definição do integral concluímos que

$$E\left[\int_0^T f(t, X_t)dB_t\right] = 0.$$

Ou seja, como devia ser, o efeito médio do ruído aleatório é nulo.

13 Eventualmente alguns leitores familiarizados com a teoria da medida estarão a pensar se não é possível calcular $\frac{dB_t}{dt}$ como uma derivada de Radon-Nikodym deste integral. A resposta segue do facto de o integral estocástico (ou, se se preferir, a medida de Wiener que lhe está subjacente) *não* ser absolutamente contínuo em relação ao integral (à medida) de Lebesgue. Para perceber porquê, observe-se que $\sum_{0\leq i<N}|B_{(i+1)\Delta t} - B_{i\Delta t}|$ cresce sem limite à medida que Δt se aproxima de 0.

2.2 Fórmula de Itô

Tendo as nossas equações estocásticas bem definidas, é altura de começar a experimentar casos. Por exemplo, suponhamos que queremos estudar a variável $X_t = {B_t}^2$. Queremos determinar a sua 'derivada' estocástica. Nada mais simples: $dX_t = 2B_t\, dB_t$, logo $X_T = X_0 + 2\int_0^T B_t\, dB_t$. Podemos aliás (ou olhando para o gráfico da Figura 8 ou usando a previsibilidade) constatar que $E[X_T] = E[X_0] = 0$, certo?

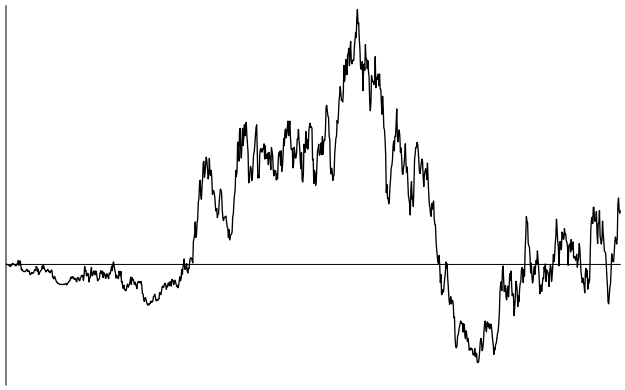


Figura 8: Uma Trajectória de $X_t = {B_t}^2$?

Não, errado! Com efeito, já sabíamos que $X_t \geq 0$ e $E[{B_T}^2] = E[X_T] = T$, e em qualquer caso o gráfico que obtemos é claramente distinto das trajectórias na Figura 3 ou da trajectória real, aproximada na Figura 9. (Mas parece muito semelhante. Em quê?) O que se passou?

O que se passou é que a derivação estocástica não segue as regras habituais, ou de outro modo o integral estocástico seria um integral normal — e não é! Vejamos então:

$$\begin{aligned} X_{t+\Delta t} - X_t = {B_{t+\Delta t}}^2 - {B_t}^2 &= (B_{t+\Delta t} + B_t)(B_{t+\Delta t} - B_t) \\ &= (B_{t+\Delta t} - B_t + 2B_t)(B_{t+\Delta t} - B_t) \\ &= (B_{t+\Delta t} - B_t)^2 + 2B_t(B_{t+\Delta t} - B_t). \end{aligned}$$

Sucede que, quanto mais Δt está próximo de 0, mais $(B_{t+\Delta t} - B_t)^2 \stackrel{d}{=} {B_{\Delta t}}^2$ se aproxima de Δt (já sabíamos que o valor esperado era Δt, mas isto é mais forte).[14] Quer isto dizer que

$$X_{t+\Delta t} - X_t \simeq \Delta t + 2B_t(B_{t+\Delta t} - B_t),$$

14 É fácil concluir isto consultando os momentos da normal.

com a aproximação tanto melhor (comparada com Δt) quanto melhor seja a aproximação $\Delta t \simeq 0$. Na prática isto significa que

$$dX_t = dt + 2B_t\, dB_t,$$

ou, integrando,

$$X_T = X_0 + \int_0^T dt + 2\int_0^T B_t dB_t = T + 2\int_0^T B_t dB_t,$$

o que resolve todas as contradições que tínhamos. Se ainda restam dúvidas, vejamos a aproximação assim obtida na Figura 9.

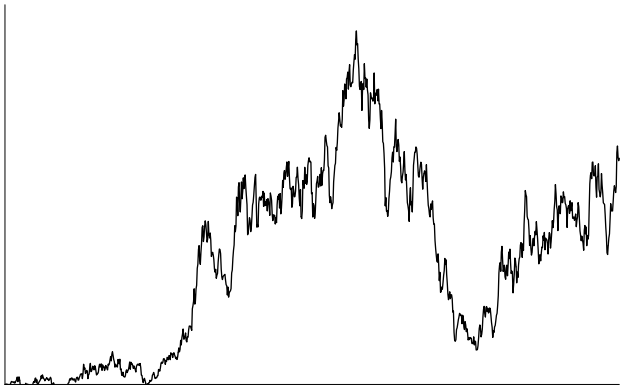


Figura 9: Uma Trajectória de $X_t = {B_t}^2$

Mais geralmente, podemos calcular derivadas estocásticas do mesmo modo que as habituais, com a ressalva de se ter $(dB_t)^2 = dt$. Em particular, $dB_t = O(\sqrt{dt})$. Isto tem implicações, que ficam mais claras com um exemplo:

Se $dX_t = v\, dt + f\, dB_t$ e $Y_t = g(t, X_t)$, então

$$Y_{t+dt} = Y_t + \tfrac{\partial g}{\partial t}dt + \tfrac{\partial g}{\partial x}dX_t + \tfrac{1}{2}\tfrac{\partial^2 g}{\partial x^2}(dX_t)^2 + o(dt).$$

Ora $(dX_t)^2 = f^2 dt + o(dt)$, portanto vem

$$Y_{t+dt} = Y_t + \left(\tfrac{\partial g}{\partial t} + \tfrac{1}{2}\tfrac{\partial^2 g}{\partial x^2} f^2\right) dt + \tfrac{\partial g}{\partial x}dX_t + o(dt),$$

ou, simplificando

$$dY_t = \left(\tfrac{\partial g}{\partial t} + \tfrac{1}{2}\tfrac{\partial^2 g}{\partial x^2} f^2\right) dt + \tfrac{\partial g}{\partial x}dX_t.$$

É certamente esclarecedor experimentar esta fórmula no caso $g(t, x) = x^2$.

Uma fórmula análoga funciona a mais do que uma dimensão. Deve ter-se em conta que as componentes são independentes, o que vai traduzir-se em $dB_t^i\, dB_t^j = \delta^{ij}\, dt$.

Não é de mais insistir que a diferença mais óbvia entre o integral estocástico e o integral usual reside na possibilidade de integrar um termo da ordem de $\sqrt{dt}$. Naturalmente que, havendo duas raízes, não admira que $\int \sqrt{dt}$ varie muito violentamente, e dependa de uma escolha de sinal 'em cada instante' t.[15]

2.3 Exemplos e Aplicações

Agora que 'sabemos' realmente como trabalhar com as equações estocásticas, podemos explorar outro exemplo. Suponhamos que queremos resolver a equação

$$dX_t = X_t dt + \lambda X_t dB_t.$$

Qual será a sua 'primitiva'? Podemos imaginar que tenha a forma $X_t = X_0 e^{at+bB_t}$, mas, admitindo que é um bom palpite, quais os valores de a e b? Derivemos e vejamos o que se obtém:

$$dX_t = aX_t dt + bX_t dB_t + \tfrac{b^2}{2} X_t dt = \left(a + \tfrac{b^2}{2}\right) X_t dt + bX_t dB_t.$$

A conclusão é imediata: $b = \lambda$ e $a = 1 - \frac{\lambda^2}{2}$.

A pergunta natural é: haverá mais soluções? E a resposta é que, tal como nas equações ordinárias, certas condições (análogas à condição de Lipschitz) permitem garantir existência e unicidade de soluções, e essas condições são satisfeitas nesta equação, assim como noutras equações lineares.

Desde o início até aqui fizemos um grande percurso. É talvez a ocasião de referir alguns exemplos de aplicação bem conhecidos. Para mais pormenores recomenda-se vivamente a consulta de [6].

Suponhamos que estamos a observar estados de uma variável X_t. Porém, a observação está sujeita a erros aleatórios (por exemplo, pequenos erros de medida) e queremos filtrar apenas o sinal original. Qual será a melhor estimativa? Se quisermos usar essa informação para tentar fazer correcções a X_t e tivermos um controlo limitado e sujeito a erros aleatórios (por exemplo,

15 É possível tornar estas afirmações vagas mais precisas. Recordamos o argumento intuitivo com que justificámos que B_t tem distribuição normal. Analogamente, podemos construir aproximações a B_t com somas de variáveis de Bernoulli: $Z_i = \pm 1$, ou, se se preferir, $B_{\Delta t} \simeq \pm\sqrt{\Delta t}$. Nesse caso, o integral estocástico é aproximado por $\int_0^T f dB_t \simeq \sum f \cdot (\pm\sqrt{\Delta t})$. É impossível resistir a acrescentar que a abordagem não-standard à integração estocástica é precisamente esta, onde se escolhe $\Delta t \simeq 0$, e se usa $\simeq$ com o significado preciso de 'infinitamente próximo'.

atrito nas transmissões de uma máquina), qual a melhor correcção a fazer em cada instante de modo a assegurar tanto quanto possível que X_t fica dentro de um certo conjunto admissível de valores?

Ou imaginemos que temos vários agentes numa economia, e vários bens com valores variáveis (ou com cotação) que podem ser comprados ou vendidos de acordo com regras pré-determinadas. Qual a sequência de compras ou vendas que garante a maximização do lucro?

Qualquer destes exemplos pode ser traduzido numa equação (ou num sistema de equações) diferencial estocástica, cuja análise pode ser feita por métodos muito similares aos que temos usado. Nem é preciso dizer, os resultados são (mesmo) aplicados a muito do que está à nossa volta.

3 Problema de Dirichlet

Para terminar, vamos considerar uma função harmónica em $\mathbb{R}^2$, ou seja, uma função satisfazendo $\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f = 0$. Um exemplo típico é um potencial electrostático ou gravítico.

Calculemos a derivada estocástica do processo $F_t = f((x,y)+B_t)$ (note-se que aqui B_t é um passeio em $\mathbb{R}^2$ e (x,y) é um ponto, fixo, em $\mathbb{R}^2$). Temos

$$\begin{aligned} dF_t &= \tfrac{\partial f}{\partial t}dt + \tfrac{\partial f}{\partial x}dB_t^1 + \tfrac{\partial f}{\partial y}dB_t^2 + \left(\tfrac{\partial^2 f}{\partial x^2} + \tfrac{\partial^2 f}{\partial y^2}\right) dt \\ &= \tfrac{\partial f}{\partial x}dB_t^1 + \tfrac{\partial f}{\partial y}dB_t^2. \end{aligned}$$

Neste altura estará o leitor a ficar impaciente: qual pode ser o interesse deste cálculo? Observemos que f é previsível, o que significa que

$$E\left[F_t\right] = E\left[F_0\right] = f(x,y)\ !$$

Mais concretamente, f é harmónica sse $E\left[F_t\right] = f(x,y)$. E exactamente o mesmo argumento pode ser usado para qualquer tempo previsível T.

Suponhamos agora que a função f está definida no fecho de um aberto limitado U, e é harmónica em U. Suponhamos ainda que sabemos o seu valor em ∂U e que queremos determiná-lo para outros pontos $(x,y) \in U$.

Já sabemos que $f(x,y) = E\left[F_t\right]$, para qualquer t. Mas só conhecemos o valor de $F_t(\omega)$ se t for tal que $(x,y) + B_t(\omega) \in \partial U$. Além disso o argumento que fizemos acima deixa de se aplicar se $(x,y)+B_t$ sair fora do domínio de f. Deste modo, a nossa única esperança de determinar o valor de $f(x,y)$ (a não ser que tivéssemos uma expressão facilmente calculável para f) é escolher

$$T(\omega) = \text{primeiro instante } t \text{ tal que } (x,y) + B_t(\omega) \in \partial U.$$

Acontece que este T é um tempo previsível, ao qual o argumento acima se aplica! Resumindo

$$f(x,y) = E\left[f\big((x,y) + B_{\text{primeiro instante } t \text{ tal que } (x,y)+B_t \in \partial U}\big)\right].$$

E com isto podemos provar muito elementarmente algumas propriedades célebres das funções harmónicas:

Por exemplo, se U é uma bola e (x,y) o seu centro, constatamos que $f(x,y)$ é a média dos valores de f em ∂U. Mais geralmente, fazendo a média dos vários raios, $f(x,y)$ é a média dos valores de f em U (*princípio da média*). Para conjuntos mais gerais teremos uma média ponderada.

Por fim, se $f(x,y)$ é uma média ponderada dos valores de f em ∂U, certamente será menor ou igual a $\max_{(x',y')\in\partial U} f(x',y')$. Mais, só poderá ser igual se f for constante em ∂U, e por conseguinte em $\overline{U}$ (*princípio do máximo*). Do mesmo modo $f(x,y)$ é sempre maior ou igual a $\min_{(x',y')\in\partial U} f(x',y')$, e só poderá ser igual se f for constante em $\overline{U}$ (*princípio do mínimo*).

Agradecimentos

Na preparação deste texto, bem como do seminário que lhe deu origem, contei com a ajuda de algumas pessoas. O Carlos Alves deu algumas opiniões na preparação do seminário e deu sugestões detalhadas de melhoramentos do texto. Também a Ana Cannas da Silva e o Rui Loja Fernandes deram algumas opiniões. O respeito pela verdade obriga que se diga que embora todas as opiniões tenham sido consideradas, nem todas foram postas em prática. Isso significa que os erros que tenham sobrado se devem à teimosia do autor, e não à desatenção dessas pessoas. Sem a ajuda do João Palhoto Matos talvez o texto não fosse acompanhado por quaisquer figuras.

No prefácio de [2], o zoólogo Richard Dawkins diz que enquanto escrevia tinha em mente três leitores imaginários: o leigo, o especialista e o estudante. Se é certo que não posso ter a ambição de me dirigir ao especialista, ainda assim tive dois leitores imaginários em consideração: o colega estudante que começou recentemente uma licenciatura em matemática (a pensar nele tentei evitar ao máximo toda a sofisticação técnica, mas sem faltar à verdade matemática), mas também o colega dos últimos anos, que sabe algumas coisas mais, e portanto tem outras dúvidas (a pensar nele tentei responder a algumas dessas dúvidas nos comentários de rodapé). A estes dois leitores imaginários cabe também uma parte dos agradecimentos.

Bibliografia

Para o leitor que queira levar mais longe a exploração deste assunto, a melhor referência é, sem qualquer dúvida, [6]. O autor começa por dar alguns exemplos de problemas em que o ruído aleatório é personagem fundamental, e parte daí para a exploração da técnica matemática subjacente.

A forma mais fácil de ter acesso a alguns dos artigos de Einstein sobre o movimento browniano é provavelmente a colectânea [4].

O resumo falava em [métodos de] Monte-Carlo. Se é verdade que fizemos uma passagem breve por lá (e mais não fora prometido), não é menos verdade que não houve tempo para apontar as principais atracções turísticas. Para o leitor que queira saber algo mais sobre esses métodos (incluindo uma discussão da ineficácia numérica da técnica discutida na última secção para encontrar funções harmónicas), um bom ponto de partida é [5].

Referências

[1] Charles Darwin. *On the Origin of Species by Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life*. 1859.

[2] Richard Dawkins. *O Gene Egoísta*. Gradiva, 2ª edição, 1999.

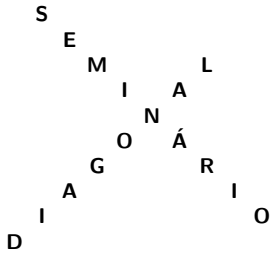
[3] Sir Arthur Conan Doyle. *The Lost World*. 1912.

[4] Albert Einstein. R. Fürth (editor). *Investigations on the Theory of the Brownian Movement*. Dover, 1956.

[5] J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. 1964.

[6] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, 5th edition, 1998.

[7] Jearl Walker. *O Grande Circo da Física*, problema 6.32. Gradiva, 1990.



Grupos, Variedades e Relatividade

Patrícia Engrácia*
3º ano da LMAC — Análise, Geometria e Álgebra
pceng@math.ist.utl.pt

Palavras Chave

grupo, variedade, grupo de Lie, isometria, simetria.

Resumo

Os grupos estão muito relacionados com a geometria: há grupos que são espaços geométricos muito ricos e há estruturas geométricas a que podemos associar grupos. Também na física as perspectivas de observadores distintos se relacionam por acção de elementos de grupos.

Neste artigo vamos olhar para alguns exemplos e brincar um pouco com a relatividade de Einstein.

1 Grupos

Em primeiro lugar vamos começar por relembrar que um *grupo* $(G, *)$ é um conjunto G munido de uma operação binária e associativa $* : G \times G \to G$ com um elemento neutro e tal que cada elemento tem um inverso.

Como exemplos de grupos temos, entre muitos, os inteiros $(\mathbb{Z}, +)$; o grupo das permutações de n elementos S^n com a composição; o espaço n-dimensional $(\mathbb{R}^n, +)$, e a circunferência unitária $\mathbb{S}^1$, que tem a soma de ângulos mod2π como operação de grupo.

Dois grupos $(G, *)$ e $(H, \cdot)$ podem ser "essencialmente" o mesmo, sem no entanto serem iguais, desde que haja uma função bijectiva $i : G \to H$ que preserve a operação de grupo: $i(x*y) = i(x)\cdot i(y)$. Uma tal função designa-se por *isomorfismo* e os grupos dizem-se *isomorfos*. Por exemplo, a esfera $\mathbb{S}^1$ pode ser "dada pelo" (ou seja, é isomorfa ao) grupo das matrizes da forma $R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ ou $R_\theta = e^{i\theta}$, com $R_\theta R_\varphi = R_{\theta+\varphi}$.

* A autora deste artigo foi parcialmente apoiada em 2000–01 por uma bolsa do Programa Gulbenkian Novos Talentos em Matemática.

2 Variedades e Grupos de Lie

Existem grupos de dois tipos: contínuos e discretos. Dos exemplos acima, tanto $\mathbb{R}^n$ como $\mathbb{S}^1$ são conjuntos contínuos e que se podem dizer "regulares", i.e., são espaços geométricos bem comportados que são "suaves", e que não têm vértices nem arestas...

Espaços assim chamam-se *variedades*. De um modo geral e bastante intuitivo, as variedades são conjuntos que localmente são como $\mathbb{R}^n$ (i.e., têm dimensão n). No nosso caso, $\mathbb{S}^1$ é localmente como $\mathbb{R}$ (tem dimensão 1). Para melhor compreender este conceito, podemos ver outro exemplo. Consideremos a esfera $\mathbb{S}^2$ em $\mathbb{R}^3$. Este conjunto é regular e bem comportado, portanto, segundo a nossa maneira intuitiva de ver as coisas, é uma variedade. Pode definir-se variedade de uma maneira rigorosa, mas não vamos tratar disso aqui. É bastante evidente que $\mathbb{S}^2$ é localmente como algum $\mathbb{R}^n$, neste caso como $\mathbb{R}^2$. Se recordarmos um pouco de História reparamos que durante muito tempo não se soube se a Terra é ou não redonda. Para nós é como se estivéssemos a viver num subconjunto do plano. E no entanto, se considerarmos um ponto de partida e começarmos a andar, contornamos a Terra e chegamos ao ponto de onde partimos... Há que notar também que $\mathbb{S}^2$ não é um grupo, apesar de ser uma variedade. Para provar isso usa-se o facto de a esfera não poder ser "penteada".

Aos grupos que são também variedades, e cuja operação é regular chamamos grupos de Lie (1842–1899). Como curiosidade, de entre as esferas unitárias só $\mathbb{S}^1$ e $\mathbb{S}^3$ são grupos, e portanto grupos de Lie.

2.1 Exemplos de Grupos de Lie

Como exemplo, vamos mostrar que o conjunto $SU(2)$ das matrizes unitárias 2×2 é um grupo de Lie (as matrizes *unitárias* são aquelas cuja inversa U^{-1} é a transposta da conjugada, designada *adjunta* e escrita $U^* = \bar{U}^t$).

$$SU(2) = \{U \in M_{2\times 2}(\mathbb{C}) : U^*U = I\}.$$

Se tivermos $U = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, e a sua adjunta $U^* = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}$, então

$$U^*U = \begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix} \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} |a|^2 + |c|^2 & \bar{a}b + \bar{c}d \\ a\bar{b} + c\bar{d} & |b|^2 + |d|^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

e fazendo alguns cálculos, concluímos que $|a|^2 + |b|^2 = 1$ e portanto $SU(2) = \mathbb{S}^3 \subset \mathbb{R}^4 \simeq \mathbb{C}^2$.

Mas já sabemos que $\mathbb{S}^3$ é uma variedade. E $SU(2)$ é um grupo, logo $SU(2)$ é um grupo de Lie (a operação de grupo, i.e., o produto de matrizes, é regular).

3 Aplicações de Grupos em Mecânica Clássica

Como veremos o conceito de grupo pode ter muitas aplicações. Entre elas encontra-se a física. Por agora vamos ver um exemplo relacionado com a mecânica clássica: o grupo das transformações de Galileu. Estas transformações são mudanças de referencial.

Consideremos dois referenciais, um dos quais se move a velocidade constante v em relação ao outro, sendo o tempo igual nos dois.

$$\begin{cases} t' &= t \\ x' &= x - vt \end{cases} \quad \Leftrightarrow \quad \begin{pmatrix} t' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -v & 1 \end{pmatrix} \cdot \begin{pmatrix} t \\ x \end{pmatrix} = O_v \cdot \begin{pmatrix} t \\ x \end{pmatrix}$$

Portanto, o grupo de Galileu é o seguinte:

$$G = \left\{ O_v = \begin{pmatrix} 1 & 0 \\ -v & 1 \end{pmatrix} : v \in \mathbb{R} \right\}$$

(com a multiplicação de matrizes), que é isomorfo a $\mathbb{R}$.

De facto, verificamos que

$$O_v^{-1} = \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} = O_{-v}$$

e que

$$O_v \cdot O_w = \begin{pmatrix} 1 & 0 \\ -v & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -w & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -v-w & 1 \end{pmatrix} = O_{v+w};$$

tal como em $(\mathbb{R}, +)$. Estas transformações preservam os intervalos de tempo e a distância Euclidiana (se tivermos dois acontecimentos (t, x_1) e (t, x_2), então $x_1' = x_1 - vt$ e $x_2' = x_2 - vt$, donde se conclui que $\Delta x' = \Delta x$).

Como se pode ver, a lei da adição de velocidades é dada pela soma das velocidades, como parece ser bastante natural. Por exemplo, se vamos a andar numa passadeira rolante, a nossa velocidade em relação à Terra é a soma da nossa velocidade em relação à passadeira com a velocidade da passadeira. No entanto, isto não se verifica na relatividade restrita, pois se tal acontecesse a velocidade da luz poderia ser ultrapassada.

4 Isometrias

Gostaria agora de introduzir o conceito de isometria. Se considerarmos um espaço vectorial munido de um produto interno, uma transformação T diz-se uma *isometria* se preservar o produto interno, i.e., $\langle T(a), T(b)\rangle = \langle a, b\rangle$. No plano Euclidiano, as transformações lineares que são isometrias são as rotações, as reflexões e suas composições, dadas pelas matrizes *ortogonais*, i.e., pelas matrizes A tal que $A^tA = I$

Estas matrizes formam o grupo ortogonal:

$$O(n) = \{A \in M_{n\times n}(\mathbb{R}) : A^tA = I\}.$$

Se tivermos $A, B \in O(n)$, então $(AB)^tAB = B^tA^tAB = B^tB = I$ e, obviamente, cada elemento de $O(n)$ tem inverso em $O(n)$.

Considerando agora o conjunto $SO(n) = \{A \in O(n) : \det A = 1\}$ obtemos o subgrupo de $O(n)$ das rotações em torno da origem.

Em particular, em $\mathbb{R}^2$, $SO(2)$ é dado pelas matrizes $\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$.

Convém também notar que $O(n) \setminus SO(n)$, que representa as rotações seguidas de reflexões, não é subgrupo de $O(n)$ porque nem sequer contém a identidade.

5 Mecânica Relativística e Grupo de Lorentz

Voltando à física, no final do século XIX fizeram-se umas experiências interessantes (as experiências de Michelson–Morley) em que se mostrou que a velocidade da luz é a mesma em todas as direcções, em qualquer *referencial de inércia* (isto é, um referencial no qual um corpo não sujeito a qualquer força mantém o seu estado de movimento). Contudo, esta conclusão contrariava a regra de adição de velocidades de Galileu. Einstein, em 1905, admitiu a invariância da velocidade da luz. Segundo ele, esta seria a única velocidade absoluta: o seu valor é o mesmo em qualquer referencial de inércia. Einstein foi também levado a reflectir sobre a natureza do espaço-tempo por causa da invariância da velocidade da luz: afinal a velocidade é uma relação entre um espaço e um tempo... Descobriu que, apesar de se pensar o contrário, o tempo é um conceito relativo, não é absoluto ("Por fim, apercebi-me que o tempo era o suspeito!", disse Einstein).

Tinha nascido a Teoria da Relatividade Restrita. Um novo passo tinha sido dado na compreensão do Mundo.

Agora, o papel das transformações de Galileu passou a ser ocupado pelas transformações de Lorentz. Estas dizem-nos como mudar de um referencial

para outro que se mova com velocidade constante em relação ao primeiro. Estas leis de transformação já eram conhecidas como aquelas que preservam as leis do electromagnetismo (i.e., as equações de Maxwell), mas foi Einstein quem lhes deu uma interpretação.

Mais uma vez, vamos considerar dois referenciais, movendo-se um em relação ao outro com velocidade constante.

Os referenciais vão ser (ct, x) e (ct', x') (note-se que vamos considerar referenciais de inércia x e x', mas estamos também a considerar uma coordenada para o tempo que não vai ser independente das coordenadas de espaço) e temos as seguintes relações:

$$\begin{cases} t' = \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t - \frac{v}{c^2}x\right) \\ x' = \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x - vt) \end{cases} \Leftrightarrow \begin{pmatrix} ct' \\ x' \end{pmatrix} = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \begin{pmatrix} 1 & -\frac{v}{c} \\ -\frac{v}{c} & 1 \end{pmatrix} \cdot \begin{pmatrix} ct \\ x \end{pmatrix} = A_v \cdot \begin{pmatrix} ct \\ x \end{pmatrix}.$$

Portanto as transformações de Lorentz são dadas pelas matrizes da forma

$$A_v = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} \begin{pmatrix} 1 & -\frac{v}{c} \\ -\frac{v}{c} & 1 \end{pmatrix},$$

com $v \in \mathbb{R}$ tal que $-c < v < c$, onde c é a velocidade da luz.

Pode-se verificar que $A_v^{-1} = A_{-v}$ e que $A_v A_w = A_{v'}$, onde $v' = \dfrac{v+w}{1+\frac{vw}{c^2}}$.

Estas transformações formam o grupo conhecido como grupo de Lorentz. Como curiosidade, este grupo é comutativo em duas dimensões, mas para dimensões superiores deixa de o ser.

5.1 Algumas Conclusões...

Daqui conseguimos retirar algumas conclusões engraçadas e surpreendentes:

1. Quando os referenciais se movem com velocidade relativa v pequena, a lei de transformação relativística reduz-se à clássica:

   se $v \ll c$, temos que $\frac{v}{c} \to 0$, e

$$t' = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t - \frac{v}{c^2}x\right) \to \frac{1}{\sqrt{1}}(t+0) = t;$$
$$x' = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x - vt) \to \frac{1}{\sqrt{1}}(x - vt) = x - vt.$$

É por esta razão que os nossos movimentos do dia-a-dia podem ser descritos com uma óptima aproximação usando transformações de Galileu.

2. Quando a velocidade é comparável à velocidade da luz, esperam-se algumas diferenças:

   (a) Alteração das distâncias Euclidianas:

   Se considerarmos no primeiro referencial (ct, x) uma distância $\Delta x = x_2 - x_1$, então ficamos com

$$x'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x_1 - vt);$$
$$x'_2 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x_2 - vt);$$
$$\Delta x' = x'_2 - x'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(x_2 - x_1) \Rightarrow \Delta x' > \Delta x.$$

   Logo, a distância Euclidiana não é preservada.

   (b) Alteração dos intervalos de tempo:

   Se tivermos um relógio fixo na origem do referencial (ct, x) que faz duas leituras t_1 e t_2 de dois acontecimentos distintos (ct_1, x) e (ct_2, x), concluímos que

$$t'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t_1 - \tfrac{v}{c^2}x\right);$$
$$t'_2 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\left(t_2 - \tfrac{v}{c^2}x\right);$$
$$\Delta t' = t'_2 - t'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}(t_2 - t_1) \Rightarrow \Delta t' > \Delta t.$$

   Logo, os intervalos de tempo também não são preservados.

   (c) Acontecimentos simultâneos num referencial não o são no outro:

   Sejam (ct, x_1) e (ct, x_2) dois acontecimentos simultâneos no pri-

meiro referencial, com $x_1 \neq x_2$. Então

$$t'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\big(t - \tfrac{v}{c^2}x_1\big);$$
$$t'_2 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\big(t - \tfrac{v}{c^2}x_2\big);$$
$$\Delta t' = t'_2 - t'_1 = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\tfrac{v}{c^2}(x_2 - x_1) \neq 0.$$

Logo estes acontecimentos não são simultâneos no segundo referencial.

(d) Os comprimentos não são preservados:

Para explicar isto vamos considerar um comboio de comprimento L que se desloca à nossa frente a uma velocidade v. Vamos considerar agora os seguintes dois acontecimentos. O primeiro é a passagem da parte da frente do comboio por um observador munido de um cronómetro. Em ambos os referenciais (o do observador e o do comboio) este evento tem coordenadas $(0,0)$. O outro acontecimento é a passagem da parte de trás do comboio pelo observador. Em terra este acontecimento tem coordenadas $(c\frac{L'}{v}, 0)$ onde L' é o comprimento do comboio conforme é percebido em terra (e, portanto, $\frac{L'}{v}$ é o tempo que o comboio leva a passar). No comboio este acontecimento tem coordenadas $(ct', -L)$. Pela transformação de Lorentz,

$$A_v\big(c\tfrac{L'}{v}, 0\big) = (ct, -L),$$

pelo que

$$\begin{cases} t &= \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}}\dfrac{L'}{v} \\ -L &= \dfrac{1}{\sqrt{1-\frac{v^2}{c^2}}}(-L') \end{cases} \quad \Rightarrow \quad L' = L\sqrt{1-\tfrac{v^2}{c^2}}\,.$$

Assim, conclui-se que o comprimento do comboio observado em terra é reduzido pelo factor $\sqrt{1-\frac{v^2}{c^2}}$. É a Contracção de Lorentz.

3. A velocidade da luz não pode ser ultrapassada. Uma consequência muito importante é a de que a velocidade da luz não pode ser ultrapassada e é sempre a mesma relativamente a qualquer referencial que se

mova com velocidade uniforme em relação a um outro referencial (isto vai-nos provar que estas transformações verificam de facto as hipóteses de que partiu a Teoria da Relatividade Restrita):

Pela regra de adição de velocidades, temos que $A_v A_w = A_{v'}$, onde chamamos v' a $\dfrac{v+w}{1+\frac{vw}{c^2}}$. Seja $w \to c$ e $v < c$. Então

$$v' = \frac{v+w}{1+\frac{vc}{c^2}} \to \frac{v+c}{1+\frac{v}{c}} = \frac{c(v+c)}{v+c} = c.$$

5.2 Espaço de Minkowski e o Cone de Luz

Quando se trabalha em relatividade restrita, o espaço utilizado é $\mathbb{R}^4$ com o seguinte pseudo-produto interno ('pseudo' porque não é definido positivo):

$$\mathbf{x} = (x_0, x_1, x_2, x_3) \quad \Rightarrow \quad \|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = x_0^2 - x_1^2 - x_2^2 - x_3^2.$$

A este espaço chama-se Espaço de Minkowski.

Se agora considerarmos o conjunto formado pelos vectores $\mathbf{x}$ tal que $\|\mathbf{x}\| = 0$ obtemos o chamado *cone de luz*. Os vectores situados no interior do cone são aqueles cujo quadrado da norma é positivo, estando os outros no exterior ou sobre o cone.

Se tivermos uma partícula material, a sua trajectória é da forma

$$x_0 = x_0(t) = ct, \qquad x_1 = x_1(t), \qquad x_2 = x_2(t), \qquad x_3 = x_3(t).$$

O vector tangente à trajectória é dado por $\dot{\mathbf{x}} = (c, \dot{x}_1, \dot{x}_2, \dot{x}_3)$, sendo a velocidade dada pelo vector de $\mathbb{R}^3$ $\mathbf{v} = (\dot{x}_1, \dot{x}_2, \dot{x}_3)$. Mas já sabemos que $\|\mathbf{v}\| \leq c \Leftrightarrow c^2 - \dot{x}_1^2 - \dot{x}_2^2 - \dot{x}_3^2 \geq 0 \Leftrightarrow \|\dot{\mathbf{x}}\|^2 \geq 0$. Assim, concluímos que a trajectória da partícula "vive" no interior do cone de luz.

Concluímos também que as transformações de Lorentz são transformações lineares em $\mathbb{R}^n$ que preservam o pseudo-produto interno de Minkowski. Por exemplo, em duas dimensões

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = A_v \begin{pmatrix} ct \\ x \end{pmatrix}$$

e

$$\begin{aligned} \|(ct', x')\|^2 = c^2 t'^2 - x'^2 &= c^2 \frac{1}{1-\frac{v^2}{c^2}} \left(t - \tfrac{v}{c^2} x\right)^2 - \frac{1}{1-\frac{v^2}{c^2}} (x - vt)^2 \\ &= c^2 t^2 - x^2 = \|(ct, x)\|^2. \end{aligned}$$

Portanto, as transformações de Lorentz são as isometrias do espaço de Minkowski.

Agradecimentos

Dirijo os meus agradecimentos ao Professor João Pimentel Nunes, por toda a ajuda e tempo dispendidos.

Referências

[1] B.A. Dubrovin, A.T. Fomenko, S.P. Novikov. *Modern Geometry — Methods and Applications*. Springer-Verlag, 1984.