Problemas com a Balança? Deixe-me ajudar! Versão pesada

Qual estudante que gosta de Matemática nunca quebrou a cabeça com aqueles problemas "encontre a moeda falsa com a balança de dois pratos"? O fato é que esse tipo de problema ainda é bastante estudado em pesquisa, e tem algumas aplicações.

1. Problema inicial

Problema. *Temos dez moedas. Cada uma pode ter dois pesos diferentes. É possível determinar, com três pesagens em uma balança de dois pratos, se todas as moedas têm o mesmo peso?*

Resolução

Numere as moedas de 1 a 10. Suponha que o conjunto $\{1\}$ tem a moedas leves e que o conjunto $\{2,3\}$ tem b moedas leves. Note que $a \in \{0,1\}$ e $b \in \{0,1,2\}$. Todas as pesagens devem resultar equilibradas; caso contrário é imediato que nem todas as moedas têm o mesmo peso. A primeira pesagem é $\{1,2,3\} \times \{4,5,6\}$. Assim, $\{4,5,6\}$ tem $a+b$ moedas leves. A segunda pesagem é $\{1,4,5,6\} \times \{7,8,9,10\}$. Então $\{7,8,9,10\}$ tem $2a+b$ moedas leves. A terceira e última pesagem é $\{2,3,4,5,6\} \times \{1,7,8,9,10\}$. O lado esquerdo tem $b+(a+b) = a+2b$ moedas leves e o lado direito tem $a+(2a+b) = 3a+b$ moedas leves. Logo $a+2b = 3a+b \iff b = 2a$. Se $a = 0$, $b = 0$ e só há moedas pesadas; se $a = 1$, $b = 2$ e só há moedas leves.

A ideia por trás dessa resolução acaba sendo utilizada na sua generalização:

Problema de pesquisa. *Qual é a maior quantidade $m(n)$ de moedas, cada uma podendo ter dois pesos, de modo que existe um algoritmo que determina se todas as $m(n)$ moedas têm o mesmo peso?*

Só um parêntesis: vamos supor que as moedas estão em "posição geral", ou seja,

Condição $(*)$**.** *Sendo $w_1, w_2, \ldots, w_k$ os pesos distintos que podem aparecer, então se $a_1, a_2, \ldots, a_k$ são inteiros tais que $\sum_i a_i = 0$ e $\sum_i a_i w_i = 0$ então $a_1 = a_2 = \cdots = a_k = 0$.*

Isso equivale a dizer que se a balança se equilibrou para dois conjuntos de moedas então as quantidades de moedas de cada peso em cada conjunto são iguais.

Uma resolução parcial do problema surgiu em 1997, no artigo [2], embora haja outros artigos que atacaram o problema anteriormente. O problema foi resolvido (assintoticamente) em 2002, no artigo [3], incluindo uma generalização para k pesos.

O resultado principal é

Teorema 1.1. *$m(n) = n^{\left(\frac{1}{2}+o(1)\right)n}$. Aqui, como usual, $o(1)$ é um termo que tende a zero quando n tende a infinito.*

Vamos, primeiro, modelar o problema com Álgebra Linear. Do exemplo, você já deve suspeitar como vamos fazer isso.

2. Álgebra Linear e Balanças?

Seja S o conjunto das moedas, $|S| = m$ e (A_i, B_i) os conjuntos de moedas colocadas na balança na pesagem i, $1 \leq i \leq n$. Temos $|A_i| = |B_i|$ (os pesos podem ser muito próximos uns dos outros, então não vale a pena pesar quantidades diferentes de moedas). Seja $C \subset S$ o conjunto das moedas leves. Então a condição $(*)$ é o mesmo que dizer que $|C \cap A_i| = |C \cap B_i|$, já que todos os pares de conjuntos se equilibraram.

Vamos modelar com vetores agora. Embora possa parecer mais natural modelar as pesagens em função das moedas que estão na pesagem, faremos o contrário (você vai entender por que logo): para cada moeda

$x \in S$, seja v_x o vetor cuja i-ésima coordenada é 1 se $x \in A_i$ (ou seja, a moeda x está no lado esquerdo da balança na i-ésima pesagem), -1 se $x \in B_i$ (a moeda x está no lado direito da balança na i-ésima pesagem) e 0 caso contrário (x não participou da i-ésima pesagem).

Note que há $|A_i|$ 1s e $|B_i|$ -1s nos v's. Logo $|A_i| = |B_i|$ é o mesmo que dizer que a soma dos vetores é zero:

$$\sum_{x \in S} v_x = 0$$

A solução do problema das dez moedas sugere que vale a pena dividir as moedas em grupos (naquele caso, $\{1\}$, $\{2,3\}$, $\{4,5,6\}$ e $\{7,8,9,10\}$). Faremos algo parecido aqui: considere os 3^n vetores de tamanho n de $W = \{-1,0,1\}^n$. Note que $v_x \in W$, e que eles podem aparecer repetidos (de fato, aparecem repetidos dentro dos grupos, já que cada grupo sempre fica no mesmo lado da balança). Então podemos reescrever a igualdade acima como

$$\sum_{w \in W} \lambda_w w = 0,$$

sendo λ_w o número de vezes que w aparece. Obviamente, $\lambda_w \geq 0$. Por fim, note que a soma dos λ_w's é o total de moedas (cada λ_w acaba sendo o tamanho dos grupos). Ou seja, queremos maximizar a soma das coordenadas de λ.

Como fica a condição $|C \cap A_i| = |C \cap B_i|$? Isso simplesmente quer dizer que a balança se equilibra se só tivermos moedas de C, ou seja,

$$\sum_{x \in C} v_x = 0$$

Agora, se quisermos trabalhar com o vetor λ cujas componentes são os λ_w (note que λ tem 3^n entradas), isso quer dizer que não é possível determinar que as moedas são do mesmo peso se, e somente se, existem vetores α e β tais que

$$\lambda = \alpha + \beta$$

com $\sum_{w \in W} \alpha_w w = \sum_{w \in W} \beta_w w = 0$ e $\alpha, \beta \neq 0$. O vetor α representa as moedas de C e o vetor β, as moedas de $S \setminus C$.

Ou seja, as pesagens determinam que todas as moedas têm o mesmo peso quando não é possível expressar λ como soma de dois vetores α e β com as mesmas características.

Agora vamos trabalhar em termos de matrizes. Seja M a matriz $n \times 3^n$ cujas colunas são os vetores de W, em qualquer ordem. Seja T o kernel de M (ou seja, o espaço dos vetores x que satisfazem $Mx = 0$). Note que $\dim T = 3^n - n$. Enfim, seja K a interseção de T com o cone definido pelas equações $x_i \geq 0$, $i = 1, 2, \ldots, 3^n$. Então K é um cone poliédrico (o conjunto solução de $Ax \geq 0$ para alguma matriz A; note que a única exigência é que as inequações que compõem $Ax \geq 0$ sejam lineares e homogêneas; inequações do tipo $f \leq 0$ podem ser transformadas em $-f \geq 0$ e iguaidades $f = 0$ podem ser substituídas pelas inequações $f \geq 0$ e $-f \leq 0$ – em particular, todo kernel é um cone poliédrico). Enfim, seja I os vetores de K com coeficientes todos inteiros. Note que λ, α e β são vetores de I.

3. Um pouco de geometria

Para analisar a existência (ou inexistência) dos vetores α e β, precisamos de uma noção geométrica de cones poliédricos e alguns resultados sobre eles, bastante encontrados em programação linear inteira.

Definição 3.1. *Seja K um cone poliédrico e I o conjunto de vetores de K com todas as coordenadas inteiras. Um subconjunto finito $H \subset I$ é uma base integral de Hilbert de K se todo vetor de I pode ser escrito como combinação linear* **inteira não negativa** *de elementos de H.*

Note que, considerando a definição de cone poliédrico, de certo modo uma base de Hilbert se comporta como uma base de um espaço vetorial: se $\alpha, \beta \geq 0$ e $v, w \in K$, sendo A a matriz que gera K temos $A(\alpha v + \beta w) = \alpha Av + \beta Aw \geq 0$.

Nem todo cone poliédrico admite base integral de Hilbert (ou seja, às vezes precisamos de infinitos vetores), ou seja, a soma das coordenadas poderia não ser limitada. Todavia, o seguinte teorema mostra a sua existência para cones poliédricos definidos por equações com coeficientes racionais, que é o nosso caso.

Teorema 3.1. *Todo cone poliédrico K admite uma base integral de Hilbert. Se K tem ponta (ou seja, existe um vetor a, não necessariamente em K, tal que $a^t x > 0$ para todo $x \in K$, $x \neq 0$ – geometricamente falando, K não contém uma reta) então existe uma única base integral de Hilbert minimal (ou seja, nenhum subconjunto próprio é uma base integral de Hilbert).*

Demonstração

Seja G o conjunto dos vetores que geram o cone, ou seja, tal que todo elemento de K é da forma $\sum_{g\in G} a_g g$, $a_g \geq 0$. Como o cone é racional, podemos supor, sem perda de generalidade, que as coordenadas dos vetores de G são inteiras (elas são racionais, então basta multiplicar pelo mmc dos denominadores). Note que G é finito. Sendo n o número de coordenadas,

$$H = \left\{ \sum_{g\in G} a_g g \in Z^n \mid 0 \leq a_g \leq 1 \right\}$$

O conjunto H é uma base integral de Hilbert de K. Note que H é finito, porque a soma dos módulos das coordenadas é limitada pela soma dos módulos das componentes de todos os vetores de G. Basta provar que todo vetor de coordenadas inteiras de K é combinação linear inteira de vetores de H. De fato, como G gera K então todo vetor $b \in K \cap Z^n$ pode ser escrito na forma

$$b = \sum_{g\in G} b_g g = \sum_{g\in G} \lfloor b_g \rfloor g + \sum_{g\in G} \{b_g\} g$$

Como $G \subset H$ (tome $a_g = 1$ para somente um vetor g e $a_{g'} = 0$ para os demais) e, pela definição de g, $\sum_{g\in G}\{b_g\}g \in H$, o resultado segue.

Agora, vamos à unicidade. Seja H^* o conjunto de vetores de coordenadas inteiras de K que não podem escritos como soma de outros dois vetores de coordenadas inteiras de K. Então esses vetores devem fazer parte de qualquer base integral de Hilbert de K. Basta então provar que H^* é também uma base integral de Hilbert. Seja a um vetor tal que $a^t x > 0$ para $x \in K \setminus \{0\}$. Suponha por absurdo que exista um vetor u que não é gerado por H^*, e tome u tal que $a^t u$ é mínimo (existe pelo princípio da boa ordem). Então, pela definição de H^* existem vetores v, w de K com coordenadas inteiras tais que $u = v + w$. Mas $a^t v > 0$, $a^t w > 0$ e $a^t u = a^t v + a^t w$; logo, pela escolha de u, os vetores v e w são gerados por H^* e u também, absurdo. ∎

Enfim, mais alguns fatos simples cujas demonstrações serão omitidas (elas são bem intuitivas). Entenderemos "semirreta" como conjuntos da forma αv, $\alpha \geq 0$, v vetor. Além disso, uma semirreta gera um cone se é a interseção do cone com um hiperplano.

Definição 3.2. *O fecho convexo de um conjunto de semirretas gerado pelos vetores $v_1, v_2, \ldots, v_k$ é o conjunto de vetores*

$$\left\{ \sum_{i=1}^{k} \alpha_i v_k \mid \alpha_i \geq 0, 1 \leq i \leq k \text{ e } \sum_{i=1}^{k} \alpha_i = 1 \right\}$$

Lema 3.1. *Se K é um cone gerado por um número finito de semiplanos, então um conjunto finito de semirretas gera K. Além disso, K é o fecho convexo dessas semirretas.*

Note que, desse lema, segue que um vetor x está contido em uma semirreta geradora se, e somente se, não pode ser escrito na forma $x = u + v$, com u e v vetores do cone não múltiplos escalares de x.

Lema 3.2. *Se K é o fecho convexo de p semirretas em um espaço de dimensão k, $p > k$, então todo $x \in K$ está contido no fecho convexo de um subconjunto de k elementos das p semirretas (ou seja, o cone pode ser "triangulado" em cones simpliciais).*

A demonstração desse lema, conhecido como *teorema de Carathéodory*, lembra vagamente o problema 2 da Olimpíada Iberoamericana de Matemática Universitária de 2009 (veja o problema 1) e pode ser encontrada em [5].

Por fim, note que todo vetor gerador mínimo de um cone poliédrico racional (isto é, um vetor v que está contido em uma semirreta geradora, com entradas inteiras cujo mdc é 1) é um elemento de toda base integral de Hilbert.

4. Voltando às moedas

Vamos retomar rapidamente as ideias do problema: a ideia é procurar um vetor λ com entradas inteiras não negativas de tamanho 3^n tal que $M\lambda = 0$, sendo M a matriz cujas colunas são os elementos de $W = \{-1, 0, 1\}^n$. É possível determinar se todas as moedas têm o mesmo peso se, e somente se, λ nao pode ser escrito como a soma de dois vetores α e β com entradas inteiras não negativas tais que $M\alpha = M\beta = 0$. O total de moedas é igual à soma das entradas de λ. Vamos denotar a soma das entradas de λ como $w(\lambda)$.

Para isso, chamamos de K o conjunto dos vetores do kernel de M cujas entradas são todas não negativas.

Como nenhum vetor de uma base integral de Hilbert minimal pode ser escrito como soma de outros dois vetores da base, o nosso problema se torna o seguinte:

Novo problema. *Encontre o maior valor de $w(\lambda)$, entre todos os vetores λ pertencentes à base integral de Hilbert minimal de K.*

Note que o problema se baseia somente na matriz M, de modo que podemos fazer uma busca exaustiva. Mas é possível simplificar bastante o estudo.

Primeiro, vamos caracterizar os vetores geradores de K. Para cada vetor x de K, seja x_+ o vetor obtido removendo as componentes nulas de x e M_x a submatriz de M obtida retirando-se as colunas correspondentes (só para poder multiplicar o pedaço pertinente de M com x_+).

Lema 4.1. *O vetor x é gerador de K se, e somente se, as duas condições a seguir são satisfeitas simultaneamente:*

(a) $M_x x_+ = 0$.

(b) O posto de M_x é ℓ, sendo $\ell + 1$ o comprimento do vetor x_+.

Demonstração

Seja x um gerador de K. Então (a) é imediato, pois $Mx = 0$ e efetuar $M_x x_+$ é o mesmo que calcular Mx, só que desprezando as entradas nulas de x. Agora, suponha que o posto de M_x é menor do que ℓ. Então a dimensão do kernel de M_x é pelo menos 2, e existe um vetor $x'_+ \neq x_+$, independente de x_+, tal que $M_x x'_+ = 0$. Construa x' de tamanho 3^n a partir de x'_+ preenchendo as coordenadas correspondentes aos zeros de x com zeros, mantendo as demais coordenadas de x'_+. Sendo todas as entradas não negativas, existem números positivos α e β tais que $u = \alpha x - x'$ e $v = \beta x - \alpha x + x'$ são vetores com entradas não negativas. Note que $Mu = Mv = 0$, u e v são independentes e $\beta x = u + v$, o que é uma contradição, pois βx também é um gerador.

Agora suponha que x satisfaz (a) e (b). Então vamos provar que x não pode ser escrito como soma de dois outros vetores u e v de K independentes de x. Suponha o contrário. Como x satisfaz (a) e $x = u + v$, então $x_i = 0$ implica $u_i = v_i = 0$. Então $x_+ = u_* + v_*$, sendo que as coordenadas de u, v fora de u_*, v_* são zero. Além disso, $Mu = Mv = 0$ e $M_x u_* = M_x v_* = 0$. Mas, por (b), M_x é uma matriz $n \times (\ell + 1)$ com posto ℓ, logo $M_x y = 0$ só tem soluções da forma αt, sendo t um vetor. Logo u e v são múltiplos escalares de x, e x deve ser gerador. ∎

Por que estamos trabalhando com os geradores? A resposta é simples: os geradores são os elementos mais fáceis de encontrar nas bases integrais de Hilbert. Agora é só tomar matrizes $\ell \times (\ell+1)$ com entradas iguais a -1, 0 e 1 e procurar limitantes superiores e inferiores para a soma das entradas dos elementos de coordenadas não negativas do kernel dessas matrizes. De fato, seja x um vetor gerador de K. Note que x tem entradas racionais, então podemos supor sem perda de generalidade que suas entradas são inteiros não negativos com mdc igual a 1. A matriz M_x corresponendente é $n \times (\ell+1)$ e posto ℓ, então tem uma submatriz L de tamanho $\ell \times (\ell+1)$ com posto ℓ. Não é difícil ver que as coordenadas de x_+ são

$$x_{+_i} = \frac{|\det L_i|}{\operatorname{mdc}(|\det L_j|)_{1 \le j \le \ell+1}},$$

sendo L_i a matriz obtida retirando-se a coluna i de L. Para isso, basta notar que x_+ satisfaz $Lx_+ = 0$ (use o bom e velho desenvolvimento de Laplace).

Note que

$$w(x) = w(x_+) = \frac{\sum_{i=1}^{\ell+1} |\det L_i|}{\operatorname{mdc}(|\det L_j|)_{1 \le j \le \ell+1}} = g(L)$$

Seja então

$$\gamma(n) = \operatorname{máx} g(L),$$

sobre o conjunto das submatrizes L de M com tamanho $\ell \times (\ell+1)$ de posto $\ell \le n$ que admite uma solução y de $Ly = 0$ com entradas positivas. Na verdade, como estamos tirando módulo de $\det L_i$ podemos tomar todas as submatrizes $\ell \times (\ell+1)$ de M com posto ℓ. Caso a solução de $Ly = 0$ tenha alguma entrada igual a zero é só retirar a coluna correspondente e uma linha qualquer e caso tenha alguma entrada negativa, é só trocar o sinal da coluna correspondente.

Agora, podemos chegar a um resultado bastante decisivo no nosso problema.

5. Finalmente, limitantes! E o limitante superior

Teorema 5.1. *Seja $f(n)$ a maior soma das entradas de um vetor da base integral de Hilbert minimal H de K. Então*

$$\gamma(n) \le f(n) \le \frac{3^n - 1}{2}$$

Note que $m(n) = f(n)$.

Demonstração

A primeira desigualdade é trivial porque todo vetor gerador minimal pertence a H. Agora, vamos à outra desigualdade. Note que se z é uma coluna de M então $-z$ também é; então todo vetor y da base integral de Hilbert minimal tem no máximo metade das coordenadas não nulas (ou seja, no máximo $\frac{3^n-1}{2}$). Para ver isso, seja y um vetor de H e considere a matrix M_y correspondente, com a quantidade mínima de colunas; M_y não pode ter colunas simétricas, pois se tivesse, digamos, as duas primeiras colunas simétricas, y_+ poderia ser escrito como $y_+ = v + (y_+ - v)$ com $v = (1, 1, 0, 0, \ldots, 0)$, absurdo (lembre que todas as entradas de y_+ são inteiros positivos). Então M_y tem no máximo $\frac{3^n-1}{2}$ colunas (você não vai contar a coluna nula, certo?).

Seja K' a interseção de K com os hiperplanos $y_i = 0$, sendo y_i as coordenadas nulas de y. Então y está na base integral de Hilbert de K' também: se y é combinação linear positiva (inteira ou não) de alguns elementos de K também, esses elementos também aparecem em K'; além disso, $\dim K' \le \frac{3^n-1}{2}$.

Agora, vamos aplicar os fatos sobre geradores: y está em um cone simplicial, ou seja, no fecho convexo de $\frac{3^n-1}{2}$ vetores geradores mínimos. Então y pode ser escrito como combinação linear deles, ou seja,

$$y = \sum_{i=1}^{(3^n-1)/2} \alpha_i x_i, 0 \le \alpha_i \le 1, x_i \text{ vetores geradores mínimos}$$

Como $x_i \in H$ para todo i então $y - x_i \notin K$, de modo que $\alpha_i \neq 1$ para todo i. Logo

$$w(y) \leq \sum_{i=1}^{(3^n-1)/2} \alpha_i w(x_i) < \sum_{i=1}^{(3^n-1)/2} w(x_i) \leq \frac{3^n-1}{2}\gamma(n)$$

∎

Com isso, podemos provar que $m(n) \leq n^{n\left(\frac{1}{2}+o(1)\right)}$. Antes, um pequeno fato, que é bastante útil:

Desigualdade de Hadamard. *Seja $A = (a_{ij})_{n\times n}$ uma matriz quadrada. Então*

$$|\det A| \leq \prod_{1\leq j\leq n} \left(\sum_{1\leq i\leq n} a_{ij}^2\right)^{1/2}$$

Demonstração

Para quem gosta de geometria: $|\det A|$ é o volume do paralelepípedo determinado pelas colunas de A, e é limitado pelo produto de seus módulos, que é o segundo membro da desigualdade. ∎

Para quem gosta de álgebra e análise: coloque, em cada coluna, $\left(\sum_{1\leq i\leq n} a_{ij}^2\right)^{1/2}$ em evidência. Agora basta provar o resultado para matrizes cujas colunas têm módulo 1, ou seja, $|\det A| \leq 1$. Para isso, considere A^tA, que é simétrica e positiva semidefinida. Então seus autovalores $\lambda_1, \lambda_2, \ldots, \lambda_n$ são não negativos. Pela desigualdade das médias, e considerando que o traço de A^tA é a soma dos quadrados dos módulos das colunas de A, ou seja, é n,

$$\det A^tA = \lambda_1\lambda_2\ldots\lambda_n \leq \left(\frac{\lambda_1+\lambda_2+\cdots+\lambda_n}{n}\right)^n = \left(\frac{\operatorname{tr} A}{n}\right)^n = 1 \implies (\det A)^2 \leq 1 \iff |\det A| \leq 1$$

∎

Lema 5.1. *$\gamma(n) \leq (n+1)^{\frac{n+1}{2}}$. Logo $m(n) \leq \frac{3^n-1}{2}(n+1)^{\frac{n+1}{2}} = n^{n\left(\frac{1}{2}+o(1)\right)}$.*

Seja L uma matriz $\ell \times (\ell+1)$. Então $\sum_{i=1}^{\ell+1} |\det L_i| = \det L'$, em que L' é obtida de L adicionando-se uma linha de -1s e 1s, sendo a escolha feita de modo a acertar o sinal (utilizando de novo o bom e velho desenvolvimento de Laplace). Sendo L' quadrada de ordem $\ell+1$, com entradas -1, 0 ou 1, pela desigualdade de Hadamard e sendo $\ell \leq n$,

$$\det L' \leq (\ell+1)^{\frac{\ell+1}{2}} \leq (n+1)^{\frac{n+1}{2}},$$

e logo $g(L) \leq \det L' \leq (n+1)^{\frac{n+1}{2}}$. ∎

6. O limitante inferior

Para provar o limitante inferior, basta encontrar uma matriz L com $g(L) = n^{n\left(\frac{1}{2}+o(1)\right)}$.

Começamos trocando $g(L)$ por um limitante mais simples: lembre que

$$g(L) = \frac{\sum_{i=1}^{\ell+1} |\det L_i|}{\operatorname{mdc}(|\det L_j|)_{1\leq j\leq \ell+1}}$$

como $\operatorname{mdc}(|\det L_j|)_{1\leq j\leq \ell+1} \leq \operatorname{mín}_{1\leq j\leq \ell+1}(|\det L_j|)$ e $\sum_{i=1}^{\ell+1} |\det L_i| \geq \operatorname{máx}_{1\leq j\leq \ell+1}(|\det L_j|)$,

$$m(n) \geq g(L) \geq \max_{1\leq i,j\leq \ell+1, \det L_j \neq 0} \left|\frac{\det L_i}{\det L_j}\right| = \epsilon(L)$$

Vamos construir uma matriz L de tamanho $n \times (n+1)$ tal que $\epsilon(L) = n^{n\left(\frac{1}{2}+o(1)\right)}$. A ideia é tomar uma matriz inversível C com entradas iguais a 1 e -1 (sim, vamos desprezar o zero!) e acrescentar uma coluna $(a_1, a_2, \ldots, a_n)^t$ à sua direita. Então, sendo C_{ij} a matriz obtida retirando a i-ésima linha e a j-ésima coluna,

$$\epsilon(L) \geq \left|\frac{\det L_1}{L_{n+1}}\right| = \left|\frac{\sum_{i=1}^{n}(-1)^{n+1}a_i C_{i1}}{\det C}\right|$$

escolhendo de novo os a_is de modo a tornar todos os termos da última soma não negativos, temos

$$\epsilon(L) \geq \left|\frac{\sum_{i=1}^{n} C_{11}}{\det C}\right|$$

Mas $\left|\frac{\sum_{i=1}^{n} C_{11}}{\det C}\right|$ é o módulo da entrada $1,1$ da matriz inversa de C. Então agora vamos estimar qual é o maior valor de uma entrada de matrizes inversas de matrizes A com entradas iguais a -1 e 1. Matrizes que têm esse valor alto são chamadas de *matrizes mal-condicionadas*, porque elas tornam a resolução do sistema $Ax = b$ bastante sensíveis a alterações em b.

7. Estimando as entradas da inversa

Seja A uma matriz inversível quadrada de ordem n. Então $\chi(A)$ é a maior entrada, em módulo, da inversa A^{-1} e $\chi(n)$ é o máximo de $\chi(A)$ entre todas as matrizes de ordem n. Provaremos o seguinte resultado:

Teorema 7.1. *No universo das matrizes A de entradas iguais a ± 1,*

$$2^{\frac{1}{2}n\log n - n(2+o(1))} \leq \chi(A) \leq 2^{\frac{1}{2}n\log n - n(1+o(1))}$$

Aqui e até o fim desse artigo, os logaritmos são na base 2. Além disso, construiremos uma matriz A tal que

$$\chi(A) \geq 2^{\frac{1}{2}n\log n - n(2+o(1))}$$

Note que isso quer dizer que $\chi(A) = n^{n\left(\frac{1}{2}+o(1)\right)}$, e isso termina o problema das moedas: basta tomar como a matriz C (lembra dela?) uma matriz com $\chi(C) = n^{n\left(\frac{1}{2}+o(1)\right)}$ e cuja maior entrada da inversa seja a da posição $1,1$ (é só permutar umas linhas e colunas em C, se for necessário).

O limitante superior não é difícil: todo co-fator é um determinante de ordem $n-1$ com entradas iguais a ± 1, que é menor ou igual a $(n-1)^{(n-1)/2} = 2^{\frac{1}{2}n\log n - o(n)}$ pela desigualdade de Hadamard. Além disso, se somarmos a primeira linha às demais linhas de A, obtemos $n-1$ linhas com entradas iguais a -2, 0 ou 2. Logo $\det A$ é divisível por 2^{n-1} e, portanto, $|\det A| \geq 2^{n-1}$ e cada entrada da inversa de A é menor ou igual a

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

O limitante inferior vai ser demonstrado nos seguintes passos:

(1) Construímos uma matriz A de ordem 2^m com $\chi(A)$ na ordem certa;

(2) Mostramos que a função $\chi(n_1 + n_2) \geq \chi(n_1)\chi(n_2)$;

(3) Construímos uma matriz A de ordem n qualquer com $\chi(A)$ na ordem certa, usando os resultados anteriores.

7.1. Passo 1: Matriz de ordem 2^m

A construção é explícita e usa um resultado da combinatória. Seja $n = 2^m$ e ordene os subconjuntos $\alpha_1, \alpha_2, \ldots, \alpha_n$ de um conjunto de m elementos da seguinte forma: $|\alpha_i| \leq |\alpha_{i+1}|$ e $|\alpha_i \,\Delta\, \alpha_{i+1}| \leq 2$, ou seja, em ordem não decrescente de cardinalidade e de modo que, de α_i para α_{i+1}, ou colocamos um elemento (nesse caso, $|\alpha_i| < |\alpha_{i+1}|$ – em particular, $\alpha_i \subset \alpha_{i+1}$ – e $|\alpha_i \,\Delta\, \alpha_{i+1}| = 1$) ou trocamos um elemento por outro (nesse

caso, $|\alpha_i| = |\alpha_{i+1}|$ e $|\alpha_i \,\Delta\, \alpha_{i+1}| = 2$). Você pode provar que tal ordenação existe com uma indução (tente, não é difícil!).

Defina as seguintes matrizes:

- Seja $A_i = \alpha_{i-1} \cup \alpha_i$ para $i > 1$. Defina $F_i = \{\alpha_s \mid \alpha_s \subset A_i \text{ e } |\alpha_s \cap (\alpha_{i-1} \,\Delta\, \alpha_i)| = 1\}$ se $|\alpha_{i-1} \,\Delta\, \alpha_i| = 2$ e $F_i = \{\alpha_s \mid \alpha_s \subset A_i\}$ se $|\alpha_{i-1} \,\Delta\, \alpha_i| = 1$. Ou seja, no primeiro caso, tomamos os subconjuntos de $\alpha_{i-1} \cup \alpha_i$ que contêm exatamente um dos elementos que saiu de α_{i-1} ou entrou em α_i; no segundo caso, tomamos simplesmente os subconjuntos de α_i. Em ambos os casos, temos $|F_i| = 2^{|\alpha_i|}$ (no primeiro caso, são duas vezes a quantidade de subconjuntos de $\alpha_{i-1} \cap \alpha_i$, que tem $|\alpha_i| - 1$ elementos).

A matriz L é definida por $\ell_{ij} = 0$ quando $\alpha_j \notin F_i$ e, para as demais $2^{|\alpha_i|}$ entradas da linha i, $\ell_{i,i-1} = \frac{1}{2^{|\alpha_i|-1}} - 1$ e $\frac{1}{2^{|\alpha_i|-1}}$ para as demais. Além disso, $\ell_{11} = 1$ é o única entrada não nula da primeira linha. Pela ordenação, $\ell_{ij} = 0$ para $j > i$. Isso é imediato se $|\alpha_i| > |\alpha_{i-1}|$; no outro caso, também é imediato se $|\alpha_j| > |\alpha_i|+1$. Se $|\alpha_j| = |\alpha_i|+1$, o único conjunto que se deve verificar é $\alpha_{i-1} \cup \alpha_i$, mas $|(\alpha_{i-1} \cup \alpha_i) \cap (\alpha_{i-1} \Delta \alpha_i)| = 2$. Além disso, os únicos conjuntos de $|\alpha_i|$ elementos que estão em F_i são α_{i-1} e α_i, o que conclui a verificação para $|\alpha_j| = |\alpha_i|$. Isso quer dizer que L é triangular inferior.

- A matriz Q é definida por $q_{ij} = (-1)^{|\alpha_i \cap \alpha_j|}$. A matriz Q é uma matriz de Hadamard simétrica (ou seja, uma matriz com entradas iguais a ± 1 cujas linhas são duas a duas ortogonais – exceto se elas forem iguais, é claro), ou seja, $Q^2 = nI$, sendo I a identidade. Vamos provar isso: seja $Q^2 = (r_{ij})_{n \times n}$. Então

$$r_{ij} = \sum_{k=1}^{n} (-1)^{|\alpha_i \cap \alpha_k| + |\alpha_k \cap \alpha_j|} = \sum_{k=1}^{n} (-1)^{|\alpha_i \cap \alpha_k| + |\alpha_k \cap \alpha_j| - 2|\alpha_i \cap \alpha_k \cap \alpha_k \cap \alpha_j|} = \sum_{k=1}^{n} (-1)^{|(\alpha_i \Delta \alpha_j) \cap \alpha_k|}$$

  e essa soma é zero porque estamo repetindo várias vezes a soma sobre todos os subconjuntos de $\alpha_i \,\Delta\, \alpha_j$, e todo conjunto não vazio tem a mesma quantidade de subconjuntos de cardinalidades par e ímpar, e $r_{ij} = 0$. Se $\alpha_i \,\Delta\, \alpha_j = \emptyset$ então $i = j$ e todos os termos são iguais a 1, e $r_{ij} = n$.

- A nossa matriz A desejada é $A = LQ$. Vamos provar que A só tem entradas 1 e -1:

$$\begin{aligned} a_{ij} = \sum_{s=1}^{n} \ell_{is} q_{sj} &= \sum_{s, \ell_{i,s} \neq 0} \frac{1}{2^{|\alpha_i|-1}} (-1)^{|\alpha_s \cap \alpha_j|} + (-1)(-1)^{|\alpha_{i-1} \cap \alpha_j|} \\ &= \underbrace{\frac{1}{2^{|\alpha_i|-1}} \sum_{s, \ell_{i,s} \neq 0} (-1)^{|\alpha_s \cap \alpha_j|}}_{\Sigma_{ij}} + (-1)^{1+|\alpha_{i-1} \cap \alpha_j|} \end{aligned}$$

Agora, lembre que α_s é subconjunto de $A_i = \alpha_{i-1} \cup \alpha_i$. Então a_{ij} na verdade depende de como α_j e A_i estão relacionados. Vejamos os casos:

Se $\alpha_j \cap A_i = \emptyset$, $\alpha_s \cap \alpha_j = \alpha_{i-1} \cap \alpha_j = \emptyset$ sempre e toda parcela de Σ_{ij} é $\frac{1}{2^{|\alpha_i|-1}}$, de modo que $\Sigma_{ij} = 2$. Mas $(-1)^{1+|\alpha_{i-1} \cap \alpha_j|} = -1$ e $a_{ij} = 1$.

A partir de agora, vamos supor que $\alpha_j \cap A_i \neq \emptyset$. Se $|\alpha_{i-1} \Delta \alpha_i| = 1$, então $A_i = \alpha_i$ e estamos considerando (possivelmente com repetições, mas todos com a mesma quantidade de repetições) todos os subconjuntos de $\alpha_i \cap \alpha_j$, que não é vazio. Metade tem cardinalidade ímpar e metade tem cardinalidade par, e $\Sigma_{ij} = 0$, e $a_{ij} = (-1)^{1+|\alpha_{i-1} \cap \alpha_j|}$. Se $|\alpha_{i-1} \,\Delta\, \alpha_i| = 2$ e α_j tem interseção não vazia com $\alpha_{i-1} \cap \alpha_i$ o mesmo acontece (faça a soma em três partes: as que contém cada elemento de $\alpha_{i-1} \,\Delta\, \alpha_i$ e a que não contém elementos de $\alpha_{i-1} \,\Delta\, \alpha_i$). Se α_j contém somente um elemento de $\alpha_{i-1} \,\Delta\, \alpha_i$ então o mesmo novamente ocorre, pois metade dos elementos de F_i contém esse elemento e metade não contém. Enfim, se $\alpha_j \cap (\alpha_{i-1} \cup \alpha_i) = \alpha_{i-1} \,\Delta\, \alpha_i$, então $|\alpha_s \cap \alpha_j| = 1$ (o único elemento de $\alpha_{i-1} \,\Delta\, \alpha_i$ contido em α_s) e toda parcela de $\sum_{ij}$ é $-\frac{1}{2^{|\alpha_i|-1}}$, de modo que $\Sigma_{ij} = -2$. Mas $|\alpha_{i-1} \cap \alpha_j| = 1$ também, e $(-1)^{1+|\alpha_{i-1} \cap \alpha_j|} = 1$, de modo que $a_{ij} = -1$.

Agora, vamos encontrar um elemento de A^{-1} com módulo grande. Para isso, seja i_0 a primeira posição tal que α_{i_0} tem três elementos e considere o sistema $Ay = \delta$, em que δ é o vetor cuja única entrada não nula é a i_0-ésima, que é 1 (ou seja, y é a i_0-ésima coluna de A^{-1}).

Vamos usar a fatoração $A = LQ$ e começar com o sistema, já conveniente escalonado, $Lx = \delta$. Como L é triangular inferior temos $x_i = 0$ para $i < i_0$. A i-ésima equação é

$$\sum_{\alpha_j \in F_i} \frac{1}{2^{|\alpha_i|-1}} x_j - x_{i-1} = \delta_i \iff x_i = (2^{|\alpha_i|-1} - 1)x_{i-1} - \sum_{\alpha_j \in F_i \setminus \{\alpha_{i-1}, \alpha_i\}} x_j + 2^{|\alpha_i|-1}\delta_i$$

Temos $x_{i_0} = 2^{|\alpha_{i_0}|-1} \cdot 1 = 2^{3-1} = 4$, $x_{i_0+1} = (2^{|\alpha_{i_0+1}|-1} - 1)x_{i_0} = 3x_{i_0}$ (vamos supor n grande, já que estamos trabalhando com ordem de grandeza). Vamos provar por indução que $|x_i| > (2^{|\alpha_i|-1} - 2)|x_{i-1}|$ para $i > i_0$. A base está feita no começo do parágrafo. Se o resultado é verdadeiro para $i-1, i-2, \ldots, i_0+1$, então $|x_{i-1}| > 2|x_{i-2}| > 4|x_{i-3}| \ldots$, de modo que

$$\sum_{\alpha_j \in F_i \setminus \{\alpha_{i-1}, \alpha_i\}} |x_j| < \sum_{t=1}^{\infty} \frac{1}{2^t} |x_i| = |x_i|$$

Logo

$$|x_i| \geq (2^{|\alpha_i|-1} - 1)|x_{i-1}| - \sum_{\alpha_j \in F_i \setminus \{\alpha_{i-1}, \alpha_i\}} |x_j| > (2^{|\alpha_i|-1} - 1)|x_{i-1}| - |x_{i-1}| = (2^{|\alpha_i|-1} - 2)|x_{i-1}|$$

Não é difícil ver também que $x_i \geq 0$ para todo i. Então, telescopando (e lembrando que há $\binom{m}{k}$ conjuntos α_i com k elementos) encontramos que

$$x_n > \prod_{k=3}^{m} (2^{k-1} - 2)^{\binom{m}{k}} = \prod_{k=3}^{m} 2^{(k-1)\binom{m}{k}} \prod_{k=3}^{m} \left(1 - \frac{2}{2^{k-1}}\right)$$

O primeiro produto é fácil de calcular: lembrando que $\sum_{k=0} \binom{m}{k} = 2^m$ e $\sum_{k=0} k\binom{m}{k} = m2^{m-1}$,

$$\prod_{k=3}^{m} 2^{(k-1)\binom{m}{k}} = 2^{\sum_{k=1}^{m} (k-1)\binom{m}{k} - \binom{m}{2}} = 2^{m2^{m-1} - 2^m + 1 - O(m^2)} = 2^{\frac{1}{2} n \log n - n - O(\log^2 n)} = 2^{\frac{1}{2} n \log n - n(1+o(1))}$$

Isso já é bom, x_n ser grande. Mas queremos y. O que fazer? Temos $Ay = \delta \iff LQy = \delta$. Logo $Qy = x \iff y = Q^{-1}x$. Mas $Q^2 = nI \iff Q^{-1} = Q/n$, de modo que $y = Qx/n$. Fazendo a conta temos $y_i = \frac{1}{n} \sum_j qijx_j$. Mas $|q_{ij}| = 1$ e $x_n > 4x_{n-1} > 8x_{n-2} > \ldots$, e

$$|y_i| > \frac{1}{n} \left(x_n - \sum_{j=2}^{\infty} \frac{x_n}{2^j} \right) = \frac{1}{n} \frac{x_n}{2}$$

Dividir por $2n$ não vai mudar a ordem de grandeza, então

$$|y_i| = 2^{\frac{1}{2} n \log n - n(1+o(1))}$$

Note que todos os elementos da i_0-ésima coluna têm módulo grande. De fato, não é difícil conseguir mais colunas com elementos grandes: basta escolher $j_0 > i_0$ de modo que $\prod_{k=|\alpha_{j_0}|}^{m} 2^{(k-1)\binom{m}{k}}$ é grande. Com isso, chegamos a

$$\chi(n) \geq 2^{\frac{1}{2} n \log n - n(1+o(1))} \quad \text{para } n = 2^m$$

7.2. Passo 2: $\chi(n_1 + n_2) \geq \chi(n_1)\chi(n_2)$

A ideia agora é "grudar" matrizes bacanas (a partir de agora, esse vai ser o nome das matrizes com inversas com entradas grandes, porque elas são bacanas, certo?) para obter matrizes bacanas maiores. Na verdade provaremos algo um pouco mais forte: $\chi(n_1 + n_2 - 1) \geq 2\chi(n_1)\chi(n_2)$.

Dadas duas matrizes S, quadrada de ordem n_1, e T, quadrada de ordem n_2, ambas com entradas iguais a ± 1, suponha sem perda de generalidade que a última linha e a última coluna de S são compostas de 1s (se for o caso, troque os sinais de algumas linhas e colunas) e que a primeira linha e a primeira coluna de T são compostas de 1s, com exceção de $t_{21} = -1$ (idem). Além disso, suponha que $\chi(S) = \left|\frac{\det S_{1n_1}}{\det S}\right|$ e $\chi(T) = \left|\frac{\det T_{2n_2}}{\det T}\right|$. Finalmente, defina R como a matriz de ordem $n = n_1 + n_2 - 1$ que tem S no seu canto superior esquerdo, T no seu canto inferior direito e demais entradas iguais a 1. Note que as áreas de S e T têm uma entrada de interseção, mas como $s_{n_1,n_1} = t_{11} = 1$ não há problema.

Subtraindo a linha n_1 das linhas $1, 2, \ldots, n_1 - 1$, pode-se verificar que $\det R = \det S \det T$: de fato, a matriz fica diagonal em dois blocos: o primeiro bloco é a matriz S' obtida de S subtraindo a sua última linha das demais e o segundo bloco é a matriz T. Temos $\det S' = \det S$, então o resultado segue. Além disso, subtraindo a linha n_1 das linhas $n_1+1, n_1+2, \ldots, n_1+n_2-1$, não é difícil obter $\det R_{1n} = 2 \det S_{1n_1} \det T_{2n_2}$ (faça as operações e veja você mesmo!). Então

$$\chi(R) \geq \left|\frac{\det R_{1n}}{\det R}\right| = \left|\frac{2 \det S_{1n_1} \det T_{2n_2}}{\det S \det T}\right| = 2\chi(S)\chi(T)$$

7.3. Passo 3: Matriz de ordem n

Podemos adaptar a operação do item anterior: agora, ajeite S para que sua última linha só tenha -1s e que as demais entradas da última linha sejam 1, como anteriormente; o resto se mantém. Seja $S \diamondsuit T$ a matriz $n_1 + n_2$ com S no seu canto superior esquerdo, T no seu canto inferior direito e todas as demais entradas iguais a 1, exceto a da linha $n_1 + 1$ e coluna n_1, que é igual a -1. Então $\chi(S \diamondsuit T) \geq \chi(S)\chi(T)$.

Agora, para conseguir uma matriz bacana de ordem n, escreva n como soma de potências de 2 (base binária, ao resgate!): $n = \sum_{i=1}^{r} 2^{q_i}$, com $q_1 > q_2 > \ldots > q_r \geq 0$. Seja $n_i = 2^{q_i}$ e construa matrizes bacanas $A_1, A_2, \ldots, A_r$ de tamanhos $n_1, n_2, \ldots, n_r$, respectivamente. Seja $A = A_1 \diamondsuit (A_2 \diamondsuit (\ldots (A_{r-1} \diamondsuit A_r) \ldots))$. Pela definição de $\diamondsuit$, A tem ordem n. Pelo passo anterior,

$$\chi(A) \geq \prod_{i=1}^{r} \chi(A_i) = 2^{\frac{1}{2}\sum_{i=1}^{r} n_i \log n_i - \sum_{i=1}^{r} n_i(1+o(1))}$$

Vamos estimar a primeira soma:

Lema 7.1. *Sejam $q_1 > q_2 > \ldots > q_r \geq 0$ inteiros, $n_i = 2^{q_i}$ e $N = \sum_{i=1}^{r} n_i$. Então*

$$\zeta(N) = \frac{1}{N}\left(N \log N - \sum_{i=1}^{r} n_i \log n_i\right) \leq 2$$

Demonstração

Primeiro, vamos deixar $\zeta(N)$ mais agradável:

$$\zeta(N) = \frac{1}{N}\left(\sum_{i=1}^{r} n_i \log N - \sum_{i=1}^{r} n_i \log n_i\right) = \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{N}{n_i}$$

Diremos que um conjunto $X = \{q_1, q_2, \ldots, q_r\}$ é *cheio* quando é igual a $\{0, 1, 2, \ldots, q_1\}$. Provaremos os seguintes dois fatos que demonstram, em conjunto, o lema:

Fato 1. $\zeta(N) \leq 2$ *para conjuntos cheios.*

Fato 2. $\zeta(N) \leq \zeta(N + n_*)$, *sendo* $n* = 2^q$, $q \notin X$, $q < q_1$.

O lema é demonstrado porque se X não é cheio então nós o "enchemos" usando o fato 2.

O fato 1 é provado com uma conta mesmo: temos $N = 2^{q_1+1} - 1 < 2^{q_1+1}$ e $n_i = 2^{q_1+1-i}$. Então

$$\zeta(N) = \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{N}{n_i} < \sum_{i=1}^{r} \frac{2^{q_1+1-i}}{N} \log 2^i = \frac{\sum_{i=1}^{r} i 2^{q_1+1-i}}{N}$$

A soma no numerador é

$$S = 2^{q_1} + 2 \cdot 2^{q_1-1} + 3 \cdot 2^{q_1-2} + \cdots + (q_1 + 1)$$

Mas

$$\frac{1}{2}S = 2^{q_1-1} + 2 \cdot 2^{q_1-2} + 3 \cdot 2^{q_1-3} + \cdots + q_1 + \frac{q_1+1}{2}$$

de modo que

$$S - \frac{1}{2}S = 2^{q_1} + 2^{q_1-1} + 2^{q_1-2} + 1 - \frac{q_1+1}{2} \iff S = 2N - (q_1 + 1) < 2N$$

e portanto

$$\zeta(N) < \frac{S}{N} < \frac{2N}{N} = 2$$

e o fato 1 está provado.

Agora, o fato 2. Vamos mexer mais um pouquinho em $\zeta(N)$, usando n_1 (a maior potência de 2) como referência agora:

$$\begin{aligned}\zeta(N) &= \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{N}{n_i} = \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{N}{n_i} = \sum_{i=1}^{r} \frac{n_i}{N} \left(\log \frac{n_1}{n_i} + \log \frac{N}{n_1} \right) \\ &= \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{n_1}{n_i} + \log \frac{N}{n_1} \sum_{i=1}^{r} \frac{n_i}{N} = \sum_{i=1}^{r} \frac{n_i}{N} \log \frac{n_1}{n_i} + \log \frac{N}{n_1}\end{aligned}$$

Assim,

$$\zeta(N + n_*) = \sum_{i=1}^{r} \frac{n_i}{N + n_*} \log \frac{n_1}{n_i} + \log \frac{N + n_*}{n_1} + \frac{n_*}{N + n_*} \log \frac{n_1}{n_*}$$

e

$$\begin{aligned}\zeta(N + n_*) - \zeta(N) &= \log \frac{N + n_*}{n_1} + \frac{n_*}{N + n_*} \log \frac{n_1}{n_*} - \log \frac{N}{n_1} + \sum_{i=1}^{r} \left(\frac{n_i}{N + n_*} \log \frac{n_1}{n_i} - \frac{n_i}{N} \log \frac{n_1}{n_i} \right) \\ &= \log \frac{N + n_*}{N} + \frac{n_*}{N + n_*} \log \frac{n_1}{n_*} - \sum_{i=1}^{r} \frac{n_* n_1}{N(N + n_*)} \log \frac{n_1}{n_i}\end{aligned}$$

Provaremos que $\zeta(N + n_*) - \zeta(N) > 0$ mostrando que

$$\frac{n_*}{N + n_*} \log \frac{n_1}{n_*} > \sum_{i=1}^{r} \frac{n_* n_1}{N(N + n_*)} \log \frac{n_1}{n_i}$$

que é equivalente a

$$N\log\frac{n_1}{n_*} > \sum_{i=1}^{r} n_i \log\frac{n_1}{n_i} \iff \sum_{i=1}^{r} n_i \log\frac{n_1}{n_*} > \sum_{i=1}^{r} n_i \log\frac{n_1}{n_i} \iff \sum_{i=1}^{r} n_i \log\frac{n_i}{n_*} > 0 \iff \sum_{i=1}^{r} 2^{q_i}(q_i - q) > 0$$

Note que basta provarmos o fato para $q = q_1 - 1$ (lembre que $q < q_1$); nesse caso o único termo positivo é 2^{q_1} e os termos negativos somam no máximo, em módulo, $2^{q_1-2} + 2 \cdot 2^{q_1-3} + 3 \cdot 2^{q_1-4} + \cdots + (q_1 - 1)$. Assim, basta provarmos que

$$2^{q_1} > 2^{q_1-2} + 2 \cdot 2^{q_1-3} + 3 \cdot 2^{q_1-4} + \cdots + (q_1 - 1)$$

Mas isso é simples, pois basta dividir tudo por 2^{q_1-2}:

$$4 > 1 + \frac{2}{2^1} + \frac{3}{2^2} + \cdots + \frac{q_1 - 1}{2^{q_1-2}}$$

Não é difícil ver que isso é verdade, pois

$$1 + \frac{2}{2^1} + \frac{3}{2^2} + \cdots = 4$$

e o lema está demonstrado. ∎

Agora estamos prontos para estimar $\chi(A)$: como do último lema $\sum_{i=1}^{r} n_i \log n_i \geq n \log n - 2n$,

$$\chi(A) \geq 2^{\frac{1}{2}\sum_{i=1}^{r} n_i \log n_i - \sum_{i=1}^{r} n_i(1+o(1))} \geq 2^{\frac{1}{2}(n\log n - 2n) - n(1+o(1))} = 2^{\frac{1}{2}n\log n - n(2+o(1))}$$

e finalmente demonstramos o teorema e o resultado principal. ∎

8. Por que estudar isso?

Na verdade, o teorema da inversa tem aplicações em outras áreas, como geometria (distância mínima em simplexos), redes neurais (*threshold gates*) e teoria dos grafos (hipergrafos regulares).

Os resultados sobre cones poliédricos com coordenadas inteiras são bastante importantes em programação linear inteira (cujos problemas costumam ser NP-hard), para o desenvolvimento de algoritmos. Há muitos problemas em aberto nessa área. Por exemplo, quantos elementos são necessários em uma base de Hilbert? Existe um análogo para o teorema de Carathéodory para inteiros? Por enquanto sabe-se que a quantidade de vetores é menor ou igual a $2n - 1$, mas ninguém sabe qual é o menor valor.

Algoritmos para achar moedas falsas, com a quantidade de matrizes e aplicações de programação linear inteira, aparecem em códigos transmissores de dados e detectores de erros. Veja [7] para um exemplo de algoritmo.

Exercícios

01. Sejam $x_1, x_2, \ldots, x_n$ vetores não nulos de um espaço vetorial V e $\varphi\colon V \to V$ um operador linear nesse espaço tal que $\varphi x_1 = x_1$, $\varphi x_k = x_k - x_{k-1}$ para $k = 2, 3, \ldots, n$. Demonstre que o conjunto de vetores $x_1, x_2, \ldots, x_n$ é linearmente independente.

02. Um vetor gerador com entradas inteiras com mdc maior do que 1 pode ser parte de uma base integral de Hilbert minimal?

03. Prove que é possível ordenar os subconjuntos $\alpha_1, \alpha_2, \ldots, \alpha_n$ de um conjunto de m elementos da seguinte forma: $|\alpha_i| \leq |\alpha_{i+1}|$ e $|\alpha_i \,\Delta\, \alpha_{i+1}| \leq 2$.

04. Prove que $\chi(S \diamondsuit T) \geq \chi(S)\chi(T)$. (Só para você checar: $\det S \diamondsuit T = 2 \det S \det T$ e $|\det(S \diamondsuit T)_{1n}| = |2 \det S_{1n_1} \det T_{2n_2}|$.)

05. Seja K um cone poliédrico em n coordenadas e H uma base integral de Hilbert de K. Seja $c^I(H, x)$ o menor número de coeficientes não nulos utilizados para representar o vetor $x \in K$ de coordenadas inteiras e seja $c^I(H)$ o maior entre os valores de $c^I(H, x)$. Prove que $c^I(H) \leq 2n - 1$.

9. Referências bibliográficas

[1] Tanya Khovanova's blog. A solução do problema inicial foi retirada do blog dela em

`http://blog.tanyakhovanova.com/?p=279`

[2] Dmitry N. Kozlov, Văn H. Vũ. Coins and cones.

[3] Noga Alon, Văn H. Vũ. Anti-Hadamard matrices, coin weighing threshold gates and indecomposable hypergraphs. Como o próprio título sugere, vários outros resultados são demonstrados.

[4] A parte sobre cones poliédricos foi baseada nas notas de aula de um curso em otimização discreta, ministrada por Endre Boros. Disponível em

`http://rutcor.rutgers.edu/~boros/513/711-513.html`

(mais especificamente, `http://rutcor.rutgers.edu/~boros/513/Convexity.pdf`)

[5] A boa e velha Wikipedia:

`http://en.wikipedia.org/wiki/Carathéodory's_theorem_(convex_hull)`

[6] Para quem quer estudar programação linear inteira: Alexander Schrijver, Theory of linear and integer programming. O livro está (parcialmente) disponível no Google Books em

`http://books.google.com.br/books?id=zEzW5mhppB8C`

[7] Algoritmos para problemas de moedas: Optimal Algorithms for the Coin Weighing Problem with a Spring Scale, de Nader H. Bshouty.