Matemática Computacional

Luiz Alberto do Carmo Viana

12 de agosto de 2020

Resumo

Notas de aula para a disciplina de Matemática Computacional.

Sumário

1 Interpolações polinomiais

Algumas funções $f : A \to B$, com $A, B \subseteq \mathbb{R}$, são difíceis de lidar usando um computador. Pode ser que f seja muito cara de computar. Pode também ser o caso de, dadas as limitações de representação numérica do computador, f ser sensível a operações em ponto flutuante, produzindo assim perturbações em algumas operações clássicas do Cálculo.

Nesta seção, descrevemos técnicas de aproximação (ou interpolação) de uma função f por um polinômio. Mas por que interpolar usando um polinômio? Polinômios não são caros de computar, e existe uma noção de "fechamento"na classe dos polinômios para algumas operações do Cálculo.

Exercício 1. *Dado um polinômio $p : \mathbb{R} \to \mathbb{R}$, digamos $p(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$, e $\overline{x} \in \mathbb{R}$, quantas operações aritméticas (somas e multiplicações) são necessárias para computar $p(\overline{x})$?*

Exercício 2. *Considere $\mathcal{P}$ a classe de polinômios. Dado $p(x) \in \mathcal{P}$, $\frac{d}{dx}p \in \mathcal{P}$? $\int p(x)dx \in \mathcal{P}$?*

1.1 Polinômio de Lagrange

Dada $f : A \to B$, com $A, B \subseteq \mathbb{R}$, e um conjunto de "amostragem"$X \subseteq A$ ($X = \{x_1, x_2, \ldots, x_k\}, k \in \mathbb{N}$), denotamos por $p_{L(f,X)} : \mathbb{R} \to \mathbb{R}$ o único polinômio de grau no máximo $k-1$ tal que $p_{L(f,X)}(\overline{x}) = f(\overline{x}), \forall \overline{x} \in X$. Chamamos $p_{L(f,X)}$ o polinômio de Lagrange de f nos pontos X. Como encontramos $p_{L(f,X)}$? E como podemos utilizá-lo para obter boas aproximações de f? Vamos construir um certo p' e deixar a cargo do leitor a prova de que p' é um polinômio e tem as propriedades do polinômio de Lagrange.

Dado $i \in [k]$ (essa notação significa $[k] = \{1, 2, \ldots, k\}$), definimos $l(i)$ como:

$$l(i) = \prod_{j \in [k] : i \neq j} \frac{x - x_j}{x_i - x_j}$$

Exercício 3. *Qual o valor de $l(i)$ quando $x = x_i$? E quando x assume o valor de algum ponto em X diferemte de x_i?*

Com a ajuda dessa notação, descrevemos $p' : \mathbb{R} \to \mathbb{R}$ como:

$$p'(x) = \sum_{i \in [k]} f(x_i) l(i)$$

Exercício 4. *Mostre que p' é um polinômio. Dica: mostre que $l(i)$ é um polinômio e que $\mathcal{P}$ é fechada para a soma e múltiplicação por escalar. Pode ser útil saber que é possível decompor um polinômio por suas raizes.*

Exercício 5. *Mostre que p' tem as propriedades que definem $p_{L(f,X)}$.*

Exercício 6. *Prove a unicidade de $p_{L(f,X)}$, ou seja, $p' = p_{L(f,X)}$. Dica: se $p \in \mathcal{P}$ tem grau n e tem $n+1$ raízes, então $p = 0$ (o polinômio com todos os coeficientes nulos).*

Como exemplo, vamos ilustrar o polinômio de Lagrange da função $f(x) = \cos(x)$ com amostras $X = \{0, \frac{\pi}{2}, \pi\}$. Vamos construir $p_{L(f,X)}(x)$ a comçar pelos termos $l(1), l(2)$ e $l(3)$.

$$l(1) = \frac{(x - \frac{\pi}{2})(x - \pi)}{-\frac{\pi}{2}(-\pi)} \tag{1.1}$$

$$l(2) = \frac{x(x - \pi)}{\frac{\pi}{2}(\frac{\pi}{2} - \pi)} \tag{1.2}$$

$$l(3) = \frac{x(x - \frac{\pi}{2})}{\pi(\pi - \frac{\pi}{2})} \tag{1.3}$$

Utilizando os termos de Lagrange descritos pelas equações (1.1)-(1.3), definimos $p_{L(f,X)}(x)$ como

$$p_{L(f,X)}(x) = f(0)l(1) + f\left(\frac{\pi}{2}\right) l(2) + f(\pi)l(3)$$

Naturalmente, existe mais de um modo de descrever o polinômio de Lagrange. A forma aqui apresentada é dita a *forma de Lagrange*, bastante simples de descrever.

Agora, como obtemos boas aproximações para f a partir de $p_{L(f,X)}$? A verdade é que um polinômio não pode ser uma boa aproximação para uma função arbitrária. Basta pensar em funções exponenciais: suas taxas de crescimento são muito mais altas que as de um polinômio, e portanto nenhum polinômio poderia ser uma boa aproximação para uma função exponencial em todo seu domínio.

No entanto, se pensarmos em aproximar f em parte de seu domínio, talvez tenhamos êxito. Dada $f : A \to B$, com $A, B \subseteq \mathbb{R}$, vamos tomar um intervalo $[a, b] \subseteq A$, com $a, b \in \mathbb{R}$. Digamos que, por um motivo qualquer, estamos interessados em aproximar f em $[a, b]$, e que já temos um conjunto de "amostra" $X = \{x_1, x_2, \ldots, x_k\}, k \in \mathbb{N}$, $X \subset [a, b]$, tal que já conhecemos os valores $f(\overline{x}), \forall\, \overline{x} \in X$. Com isso em mãos, poderíamos considerar $p_{L(f,X)}$ uma aproximação para f no intervalo $[a, b]$.

Como $p_{L(f,X)}$ é um polinômio, e polinômios são contínuos, não se pode esperar que $p_{L(f,X)}$ seja uma boa aproximação para f em $[a, b]$ no caso de f não ser contínua em $[a, b]$. Um outro fator que pode atrapalhar essa aproximação é $\frac{d^2}{dx^2} f$, a *derivada segunda* de f: se a variação da taxa de crescimento de f fica próxima a zero em $[a, b]$, f se torna bastante similar a uma reta em $[a, b]$, e portanto fácil de aproximar; por outro lado, se a taxa de crescimento de f varia bruscamente em $[a, b]$, pode ser o caso de f crescer ou decrescer muito mais rapidamente que $p_{L(f,X)}$, ou ter mais máximos e mínimos locais em $[a, b]$ do que $p_{L(f,X)}$, novamente resultando em uma má aproximação.

Por último, listamos duas escolhas bastante comuns para a construção de X, quando da aproximação de f por $p_{L(f,X)}$ em $[a, b]$. A primeira consiste em tomar os pontos X equidistantes em $[a, b]$, ou seja, $X = \{x_1, x_2, \ldots, x_k\}$

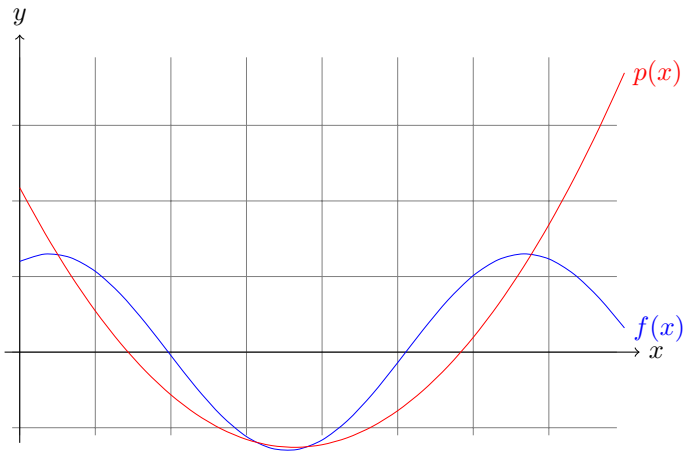

Figura 1: Ilustração do polinômio de Lagrange. Aproximamos $f(x) = 0.5\sin(x) + 1.2\cos(x)$ no intervalo $[0.5, 6.7]$ com o polinômio de Lagrange $p(x)$, construído com amostras $\{0.5, 3.5, 6.7\}$.

seria determinado por $x_i = a + (i-1)h$, onde $h = \frac{b-a}{k-1}$. A segunda seria tomar X como um subconjunto dos pontos de máximo e mínimo locais em $[a, b]$. Enquanto a primeira escolha de X é barata de computar, ela não leva em conta o comportamento de f em $[a, b]$, podendo resultar em uma aproximação não muito boa. Já a segunda escolha de X pode ser bem mais cara de computar (a depender de f), mas ela faz com que a aproximação tenha alguns pontos em comum com f em regiões onde a variação de f tende a ser baixa, o que garante uma melhor aproximação em alguns trechos de $[a, b]$. Quanto a $|X|$, é perceptível que, independente da escolha de construção de X, não se deve tomar mais pontos em $[a, b]$ do que a quantidade de máximos e mínimos locais de f em $[a, b]$.

Exercício 7. *Explique o que é o* fenômeno de Runge

1.2 Polinômio de Taylor

Na subseção anterior, vimos que não se pode esperar que um polinômio seja uma boa aproximação para uma função qualquer em todo seu domínio. Com isso em mente, nos contentamos em tomar aproximações em intervalos do domínio. Agora, vamos apresentar um modo de aproximar uma função na vizinhança de um ponto do domínio, mais uma vez por meio de um polinômio.

Dada $f : A \to B$, com $A, B \subseteq \mathbb{R}$, e um ponto $a \in A$, denotamos por $p_{T(f,a,k)} : \mathbb{R} \to \mathbb{R}$ o único polinômio de grau no máximo $k \in \mathbb{N}$ tal que $p_{T(f,a,k)}(a) = f(a)$ e $\frac{d^i}{dx^i} p_{T(f,a,k)}(a) = \frac{d^i}{dx^i} f(a), \forall\, i \in [k]$. Naturalmente, $p_{T(f,a,k)}$ está bem definido apenas quando f é k-diferenciável em a. Chamamos $p_{T(f,a,k)}$ o polinômio de Taylor de f em a até a k-ésima derivada. Como encontramos $p_{T(f,a,k)}$?

Inicialmente, vamos tratar do caso em que $a = 0$. Considere $p' : \mathbb{R} \to \mathbb{R}$ como $p'(x) = c_0 + c_1 x + c_2 x^2 + \cdots + c_k x^k$ tendo as propriedades de $p_{T(f,0,k)}$. Assim, temos

$$c_0 = p'(0) = f(0)$$

$$c_1 = \frac{d}{dx} p'(0) = \frac{d}{dx} f(0)$$

$$2c_2 = \frac{d^2}{dx^2} p'(0) = \frac{d^2}{dx^2} f(0)$$

$$3!c_3 = \frac{d^3}{dx^3} p'(0) = \frac{d^3}{dx^3} f(0)$$

$$\vdots$$

$$k!c_k = \frac{d^k}{dx^k} p'(0) = \frac{d^k}{dx^k} f(0)$$

Assim, tendo os coeficientes definidos, escrevemos

$$p'(x) = f(0) + \left(\frac{d}{dx} f(0)\right) x + \left(\frac{1}{2}\frac{d^2}{dx^2} f(0)\right) x^2 + \left(\frac{1}{3!}\frac{d^3}{dx^3} f(0)\right) x^3 + \cdots + \left(\frac{1}{k!}\frac{d^k}{dx^k} f(0)\right) x^k$$

Perceba que os coeficientes de p' são determinados de forma única. Como p' tem as mesmas propriedades de $p_{T(f,0,k)}$, temos $p' = p_{T(f,0,k)}$

Note como foi conveniente a suposição de que $a = 0$. Para determinar o coeficiente $c_i, i \in \{0\} \cup [k]$, utilizamos derivação para eliminar os coeficientes com índice menor que i, enquanto $a = 0$ elimina os coeficientes com índice maior que i. No entanto, apenas essa manipulação não é o bastante para resolvermos o caso geral de $p_{T(f,a,k)}$, com um a qualquer.

Exercício 8. *Tome $g(x) = x + a$. Considere $(f \circ g)(x) = f(x + a)$. Note que $(f \circ g)(0) = f(a)$. Quais são os valores das derivadas de $f \circ g$ em 0? Descreva $p_{T(f \circ g,0,k)}$.*

Exercício 9. *Como o valor de $p_{T(f \circ g,0,k)}$ (e de suas derivadas) em 0 se relaciona com o valor de f (e de suas derivadas) em a?*

Exercício 10. *Note que, ao usarmos g, fazemos um "shift"para a direita (para a esquerda, caso $a < 0$) na reta dos reais (o exercício anterior evidencia isso). Assim, seria preciso desfazer esse "shift"para obtermos a propriedade desejada. Como poderíamos modificar $p_{T(f \circ g,0,k)}$ de forma que o "shift"seja desfeito?*

Ao final desses exercícios, esperamos obter

$$p_{T(f,a,k)}(x) = f(a)+\left(\frac{d}{dx}f(a)\right)(x-a)+\left(\frac{1}{2}\frac{d^2}{dx^2}f(a)\right)(x-a)^2+\left(\frac{1}{3!}\frac{d^3}{dx^3}f(a)\right)(x-a)^3+\dots$$

$$\cdots+\left(\frac{1}{k!}\frac{d^k}{dx^k}f(a)\right)(x-a)^k$$

Para evidenciar a noção de vizinhança, podemos descrever x em função de sua distância até a. Fazendo a substituição $x = a + h$, obtemos

$$p_{T(f,a,k)}(a+h) = f(a)+\left(\frac{d}{dx}f(a)\right)(h)+\left(\frac{1}{2}\frac{d^2}{dx^2}f(a)\right)(h)^2+\left(\frac{1}{3!}\frac{d^3}{dx^3}f(a)\right)(h)^3+\cdots+\left(\frac{1}{k!}\frac{d^k}{dx^k}f(a)\right)(h)^k$$

Como exemplo, vamos criar o polinômio de Taylor da função $f(x) = \sin(x)$, no entorno do ponto π, considerando três derivadas de $f(x)$.

$$\frac{d}{dx}f(x) = \cos(x) \tag{1.4}$$

$$\frac{d^2}{dx^2}f(x) = -\sin(x) \tag{1.5}$$

$$\frac{d^3}{dx^3}f(x) = -\cos(x) \tag{1.6}$$

Podemos descrever $p_{T(f,\pi,3)}(x)$ da seguinte forma

$$p_{T(f,\pi,3)}(\pi + h) = \sin(\pi) + \cos(\pi)h - \frac{1}{2}\sin(\pi)h^2 - \frac{1}{6}\cos(\pi)h^3$$

O seguinte resultado, que vamos deixar sem demonstração, é crucial para sabermos medir o erro que $p_{T(f,a,k)}$ comete ao aproximar f em torno do ponto a. A equação apresentada a seguir é chamada de *expansão de Taylor*.

Teorema 1. *Tome $f : A \to B$, com $A, B \subseteq \mathbb{R}$, e f tendo sua k-ésima derivada definida em a, para todo $k \in \mathbb{N}$. Dado $a \in A$, temos que*

$$f(a + h) = \sum_{i=0}^{\infty}\left(\frac{1}{i!}\frac{d^i}{dx^i}f(a)\right)h^i$$

A primeira coisa que notamos é que $p_{T(f,a,k)}$ é a soma das k primeiras parcelas da expansão de Taylor correspondente. Como sempre há outras parcelas desconsideradas por $p_{T(f,a,k)}$, independente do valor de k, percebemos que essa aproximação cometerá erros sempre que as parcelas desconsideradas forem não-nulas.

Exercício 11. *Dê exemplo de uma classe de funções $\mathcal{F}$ tal que $\frac{d^i}{dx^i}f(x) = 0, \forall\, i \geq k, \forall\, f \in \mathcal{F}$, dado um certo $k \in \mathbb{N}$.*

Podemos reescrever o teorema como

$$f(a+h) = p_{T(f,a,k)}(a+h) + \sum_{i=k+1}^{\infty} \left(\frac{1}{i!} \frac{d^i}{dx^i} f(a) \right) h^i$$

Parece que a fórmula do erro da aproximação está tomando forma! Para medirmos assintoticamente o erro, vamos estabelecer que h sempre assume valores no intervalo aberto $(-1, 1)$. Mas por quê? Porque quando consideramos potências de h, ocorre que $O(h) \supseteq O(h^2) \supseteq O(h^3) \ldots$, desde que $h \in (-1, 1)$.

Note, com isso, que podemos medir o erro da aproximação de f por $p_{T(f,a,k)}$ em $(a-1, a+1)$ com base na primeira parcela da expansão de Taylor desconsiderada por $p_{T(f,a,k)}$. Perceba que tal parcela é

$$\left(\frac{1}{(k+1)!} \frac{d^{k+1}}{dx^{k+1}} f(a) \right) h^{k+1} \in O(h^{k+1})$$

E assim, podemos escrever a relação entre f e $p_{T(f,a,k)}$ como

$$f(a+h) = p_{T(f,a,k)}(a+h) + O(h^{k+1})$$

Isso quer dizer que só podemos utilizar $p_{T(f,a,k)}$ como aproximação para f em $(a-1, a+1)$? Claro que não! O que os resultados apresentados dizem é que $p_{T(f,a,k)}$ é uma aproximação de f excelente numa região muito próxima de a, e nessa região temos a garantia de que, quanto mais parcelas da expansão de Taylor utilizamos, menor é o erro cometido. Na prática, $p_{T(f,a,k)}$ é uma aproximação muito boa de f numa região em torno de a tão grande quanto o valor de k que tomarmos, mas que pode se degenerar de forma imprevisível.

Como pontos negativos para o uso do polinômio de Taylor para interpolações, observamos que seu uso é restrito pelo número de derivadas definidas para f. Se f não for diferenciável no ponto de interesse, essa técnica não é aplicável. Mais ainda, usa-se os valores das derivadas de f no ponto de interesse, o que se torna impeditivo quando f não tem derivadas elementares ou essas são computacionalmente caras.

2 Zeros de funções

Nesta seção, vamos apresentar técnicas para encontrar raizes de uma função. Para isso, vamos fazer uso das interpolações descritas na seção anterior. Para fins de revisão, um ponto $\overline{x} \in A$ é dito raiz de $f : A \to B$, com $A, B \subseteq \mathbb{R}$, se $f(\overline{x}) = 0$. Já que vamos lidar com aproximações, definimos $\hat{x} \in A$ uma ϵ-raiz de f, $\epsilon \in \mathbb{R}_+$, se existe uma raiz $\overline{x} \in A$ de f tal que $\hat{x} \in [\overline{x} - \epsilon, \overline{x} + \epsilon]$ ou $|f(\hat{x})| \leq \epsilon$. Naturalmente, esperamos utilizar valores de ϵ suficientemente pequenos nessa definição.

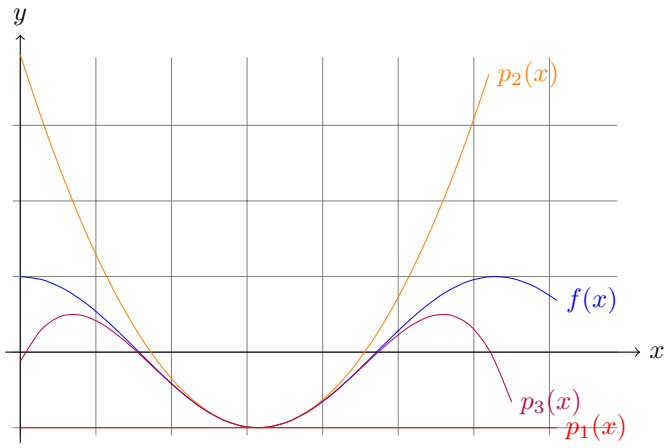

Figura 2: Ilustração de polinômios de Taylor interpolando $f(x) = \cos(x)$ em torno do ponto $a = \pi$. Os polinômios p representam polinômiso de Taylor para certos valores de k: $p_1(x)$ para $k = 0, 1$, $p_2(x)$ para $k = 2, 3$ e $p_3(x)$ para $k = 4$.

2.1 Método da posição falsa

Tome $f : A \to B$, com $A, B \subseteq \mathbb{R}$, uma função contínua. Vamos descrever uma técnica para encontrar uma ϵ-raiz de f em um intervalo $[a, b] \subseteq A$, desde que exista uma raiz de f em $[a, b]$. Antes de tudo, precisamos de condições suficientes para a existência de uma raiz em um intervlao do domínio de f.

Teorema 2. *Seja $f : A \to B$, com $A, B \subseteq \mathbb{R}$, uma função contínua. Admitindo $f(a) \leq f(b)$ ($f(a) \geq f(b)$), dado um $d \in [f(a), f(b)]$ ($[f(b), f(a)]$), existe $c \in [a, b]$ tal que $f(c) = d$.*

O teorema acima é chamado de *Teorema do Valor Intermediário*. Como podemos utilizá-lo para garantir uma raiz em um intervalo $[a, b]$ do domínio de f? O teorema diz que sempre vai haver um ponto c em $[a, b]$ tal que $f(c) = d$, para qualquer $d \in [f(a), f(b)]$. Então basta tomarmos um intervalo $[a, b]$ tal que $0 \in [f(a), f(b)]$. Se $f(a)$ for negativo e $f(b)$ for positivo, ou vice-versa, o Teorema do Valor Intermediário nos garante a existência de uma raiz em $[a, b]$.

Corolário 1. *Seja $f : A \to B$, com $A, B \subseteq \mathbb{R}$, uma função contínua. Se $[a, b] \subseteq A$ é tal que $f(a)f(b) < 0$, então existe uma raiz de f em $[a, b]$.*

Agora que temos condições para determinar a existência de uma raiz de f em um intervalo de seu domínio, vamos desenvolver uma técnica para encontrar uma aproximação para essa raiz, uma ϵ-raiz. Para isso, vamos utilizar um polinômio de Lagrange.

Tome $[a, b] \subseteq A$ tal que $f(a)f(b) < 0$, e portanto com uma raiz de f nesse intervalo. Vamos utilizar $p_{L(f,\{a,b\})}$ como aproximação de f em $[a, b]$.

Exercício 12. *Argumente que $p_{L(f,\{a,b\})}$ tem uma única raiz em $[a, b]$.*

Se $p_{L(f,\{a,b\})}$ for uma boa aproximação para f em $[a,b]$, então sua única raiz pode estar próxima de uma raiz de f em $[a,b]$. Lembramos que as condições desenvolvidas para existência de uma raiz em um intervalo não garantem que tal raiz é a única no dito intervalo. Mas qual é a raiz de $p_{L(f,\{a,b\})}$?

Vamos desenvolver $p_{L(f,\{a,b\})}$

$$p_{L(f,\{a,b\})}(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}$$

Agora precisamos encontrar a raiz de $p_{L(f,\{a,b\})}$ resolvendo a equação

$$f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} = 0$$

$$f(a)\frac{x-b}{a-b} = -f(b)\frac{x-a}{b-a}$$

$$f(a)(x-b)(b-a) = -f(b)(x-a)(a-b)$$

$$(f(a)x - bf(a))(b-a) = (af(b) - f(b)x)(a-b)$$

$$bf(a)x - b^2f(a) - af(a)x + abf(a) = a^2f(b) - af(b)x - abf(b) + bf(b)x$$

$$bf(a)x - af(a)x + af(b)x - bf(b)x = a^2f(b) - abf(b) + b^2f(a) - abf(a)$$

$$(bf(a) - af(a) + af(b) - bf(b))x = (b^2 - ab)f(a) + (a^2 - ab)f(b)$$

$$x = \frac{(b^2 - ab)f(a) + (a^2 - ab)f(b)}{(b-a)f(a) + (a-b)f(b)}$$

Essa fórmula parece muito complexa. Podemos simplificar a obtenção de uma fórmula para a raiz de $p_{L(f,\{a,b\})}$ se o tratarmos de uma outra forma, para além da forma de Lagrange. Basicamente, $p_{L(f,\{a,b\})}$ é a reta que passa pelos pontos $(a, f(a))$ e $(b, f(b))$. Temos que a equação geral da reta é, dado um de seus pontos (x_0, y_0) e um coeficiente de inclinação m

$$r(x) = y_0 + m(x - x_0)$$

Como $p_{L(f,\{a,b\})}$ passa por $(a, f(a))$ e $(b, f(b))$, calculamos

$$m = \frac{f(b) - f(a)}{b-a}$$

Tomando $(x_0, y_0) = (a, f(a))$, obtemos

$$p_{L(f,\{a,b\})}(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

Novamente, buscamos por sua única raiz

$$f(a) + \frac{f(b) - f(a)}{b - a}(x - a) = 0$$

$$\frac{f(b) - f(a)}{b - a}(x - a) = -f(a)$$

$$x - a = f(a)\frac{a - b}{f(b) - f(a)}$$

$$x = a + f(a)\frac{a - b}{f(b) - f(a)}$$

Agora podemos dizer que ficou mais simples. Mas e agora? E se a raiz de $p_{L(f,\{a,b\})}$ não for uma ϵ-raiz de f? Vamos denotar a raiz de $p_{L(f,\{a,b\})}$ como $\hat{x}$. Como $\hat{x} \in [a, b]$ e há uma raiz de f em $[a, b]$, perceba que, se $b - a \leq \epsilon$, então $\hat{x}$ é uma ϵ-raiz de f. Mas e quando $b - a > \epsilon$? $\hat{x}$ ainda pode ser uma ϵ-raiz, desde que $|f(\hat{x})| \leq \epsilon$. Se esse também não for o caso, temos que ou $f(\hat{x}) < 0$ ou $f(\hat{x}) > 0$. Com isso, podemos procurar por uma raiz de f em um intervalo menor: ou $[a, \hat{x}]$ ou $[\hat{x}, b]$, dependendo de qual deles satisfaça a condição de existência de uma raiz. Repetindo o procedimento da posição falsa quantas vezes forem necessárias, há a garantia de que encontraremos uma ϵ-raiz de f.

Exercício 13. *Escreva um pseudocódigo para sintetizar o método da posição falsa.*

Exercício 14. *Existe uma versão mais simples e grosseira do método da posição falsa, chamado método da bisseção. Dado um intervalo $[a, b] \subseteq A$ contendo uma raiz de f, o método da bisseção simplesmente assume $\hat{x} = \frac{a+b}{2}$, desconsiderando o comportamento de f em $[a, b]$. Mostre que, para o método da bisseção, basta que $b - a \leq 2\epsilon$ para que $\hat{x}$ seja uma ϵ-raiz de f.*

Exercício 15. *Usando o método da posição falsa, encontre $\sqrt{7}$ com uma precisão de quatro casas decimais. Quantas iterações foram usadas? Compare com o número de iterações usados pelo método da bisseção. Explique a relação entre os desempenhos dos métodos. Explique também por que não é possível representar $\sqrt{7}$ de forma precisa em um computador.*

2.2 Método de Newton

Tome $f : A \to B$, com $A, B \subseteq \mathbb{R}$, uma função contínua. Vamos descrever um método para encontrar uma ϵ-raiz de f usando um polinômio de Taylor.

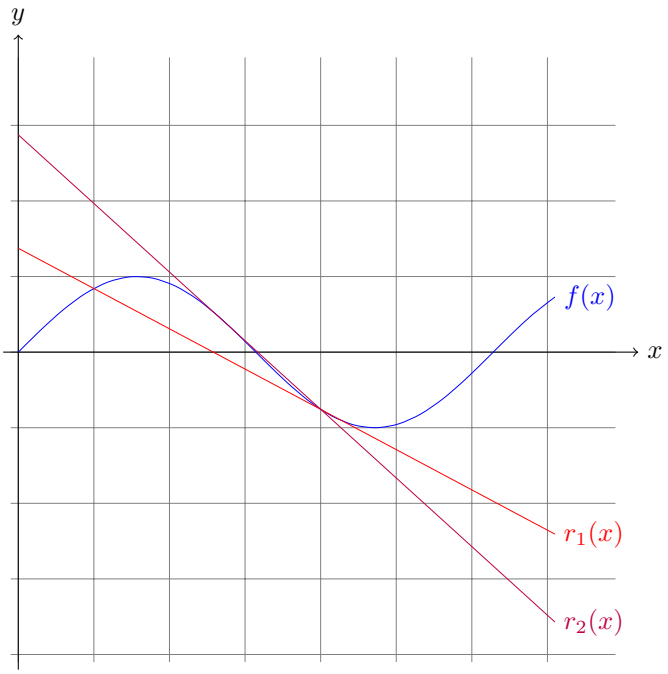

Figura 3: Ilustração de duas iterações do método da posição falsa. Tomamos $f(x) = \sin(x)$ e consideramos a interpolação de sua primeira iteração ($r_1(x)$) e de sua segunda iteração ($r_2(x)$). A primeira iteração considerou o intervalo $[1, 4]$ e a segunda iteração considerou o intervalo $[2.5794, 4]$. Consulte a Tabela 1.

Como sabemos, os polinômios de Taylor são construídos para aproximar f em torno de um ponto $a \in A$. Dado $a \in A$, vamos construir um polinômio de Taylor de grau 1 para aproximar f em torno de a. Desenvolvemos $p_{T(f,a,1)}$ como

$$p_{T(f,a,1)}(x) = f(a) + \left(\frac{d}{dx} f(a)\right)(x - a)$$

Se $p_{T(f,a,1)}$ for uma boa aproximação para f em torno de a, e se houver uma raiz de f próxima de a, então esperamos que a única raiz de $p_{T(f,a,1)}$ seja uma ϵ-raiz de f. Vamos encontrar a raiz de $p_{T(f,a,1)}$.

$$f(a) + \left(\frac{d}{dx} f(a)\right)(x - a) = 0$$

$$\left(\frac{d}{dx} f(a)\right)(x - a) = -f(a)$$

Iteração	a	b	raiz
$r_1(x)$	1	4	2.5794
$r_2(x)$	2.5794	4	3.1664

Tabela 1: Intervalos utilziados na Figura 3. As colunas a e b indicam os pontos em que as retas r interpolam a função f, e a coluna raiz indica a raiz da reta r no intervalo $[a, b]$.

$$x - a = -\frac{f(a)}{\frac{d}{dx}f(a)}$$

$$x = a - \frac{f(a)}{\frac{d}{dx}f(a)}$$

Vamos chamar de $\hat{x}$ a raiz de $p_{T(f,a,1)}$. Como não sabemos da existência de uma raiz de f em uma região de comprimento conhecido, temos de lidar com a segunda parte da definição de ϵ-raiz, ou seja, temos de verificar se $|f(\hat{x})| \leq \epsilon$.

Se $|f(\hat{x})| \leq \epsilon$, então $\hat{x}$ é uma ϵ-raiz de f. Caso não seja, podemos utilizar $p_{T(f,\hat{x},1)}$ para dar continuidade ao método.

Perceba, no entanto, que não foi dada nenhuma garantia de existência de uma raiz de f, muito menos o método de Newton preserva invariantes que garantam a convergência para uma raiz de f. Não é incomum, inclusive, que o método de Newton não convirja para uma raiz de f.

Exercício 16. *Escreva um pseudocódigo sintetizando o método de Newton. Como não há garantia de convergência, use um limite para o número de iterações do método.*

Quanto ao seu comportamento, é verificado que o método de Newton é muito sensível à escolha de $a \in A$. Para diminuir o risco de divergência, geralmente utiliza-se o método da posição falsa para tomar um intervalo pequeno o suficiente contendo uma raiz de f, e então escolhe-se a nesse intervalo para aumentar as chances de convergência do método.

Exercício 17. *Use o método de Newton com $a \in \{-2, -1, 0\}$ para encontrar uma raiz de*

$$f(x) = \frac{(x-1)^2}{x^2+1}$$

Explique os comportamentos apresentados pelo método.

Exercício 18. *Escreva um psudocódigo para sintetizar o método híbrido descrito acima. Use dois parâmetros de precisão, um para a fase da posição falsa e outro para a fase do método de Newton.*

Uma outra observação é que o método de Newton pode terminar de forma imprevisível por conta de divisões por zero. Como a única divisão na fórmula do método tem como divisor $\frac{d}{dx}f(a)$, esse caso ocorre quando o método toma

um ponto a com $\frac{d}{dx}f(a) = 0$. Em outras palavras, se alguma iteração do método considerar um ponto de máximo ou mínimo local (ou algum outro ponto estacionário), o ponto da próxima iteração não poderá ser computado.

Vamos agora analisar o erro do método de Newton, baseado no erro de aproximação de $p_{T(f,a,1)}$. Sabemos que

$$f(a+h) = p_{T(f,a,k)}(a+h) + O(h^{k+1})$$

Então, para $p_{T(f,a,1)}$ temos

$$f(a+h) = p_{T(f,a,1)}(a+h) + O(h^2)$$

Percebemos que $|f(\hat{x})| \in O(h^2)$, sendo h a distância de $\hat{x}$ até a. Se $\hat{x} \in (a-1, a+1)$ e há uma raiz de f nesse intervalo aberto, podemos esperar convergência. Geralmente, o método de Newton apresenta convergência quadrática, ou seja, a distância de $\hat{x}$ até uma raiz de f vai decaindo quadraticamente.

O método de Newton herda as limitações da aplicação do polinômio de Taylor. É esperado que f seja diferenciável em um intervalo de comprimento suficientemente grande em torno de a, uma vez que os pontos tomados pelo método podem estar distantes de a (enfatizamos que uma boa escolha de a costuma evitar grandes saltos por parte do método). Caso f não seja diferenciável em todo seu domínio, e caso o método apresente divergência, este pode se comportar de forma arbitrária.

Iteração	a	$\hat{x}$
1	2	4.185
2	4.185	2.468
3	2.468	3.266
4	3.266	3.140

Tabela 2: Valores numéricos de a e $\hat{x}$ para a execução do método de Newton ilustrada na Figura 4.

3 Integração numérica

Nesta seção, vamos apresentar técnicas de aproximação para o cálculo de integrais definidas de funções $f : A \to B$, com $A, B \subseteq \mathbb{R}$, em um intervalo $[a, b] \subseteq A$, isto é, integrais da forma $\int_a^b f(x)dx$. Para isso, vamos inicialmente fazer uso de polinômios de Lagrange (*quadraturas de Newton-Cottes*) e vamos desenvolver quadraturas para a integração exata de polinômios com grau limitado (*quadrataruas de Gauss-Legendre*).

3.1 Regra do trapézio

Começamos por apresentar a primeira quadratura de Newton-Cottes, uma das regras de aproximação de integrais mais simples. Dada $f : A \to B$, com $A, B \subseteq$

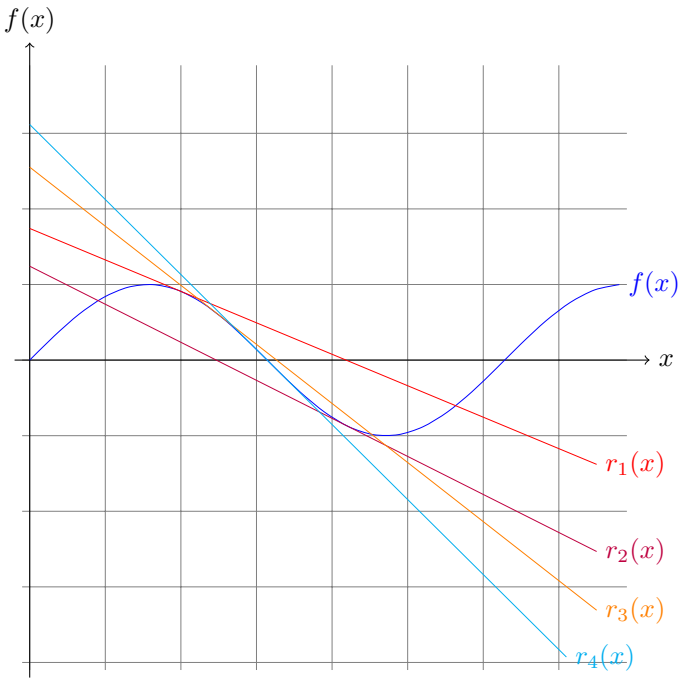

Figura 4: Ilustração do método de Newton. Encontramos uma raiz de $f(x) = \cos(x)$ com ponto inicial $a = 2$. As retas r_1, r_2, r_3 e r_4 são utilizadas como interpolação nas iterações correspondentes. Veja a Tabela 2.

$\mathbb{R}$, e um intervalo $[a, b] \subseteq A$, utilizamos $p_{L(f,\{a,b\})}$ como uma aproximação para f em $[a, b]$. Se $p_{L(f,\{a,b\})}$ for uma boa aproximação para f em $[a, b]$, esperamos que

$$\int_a^b f(x)dx \approx \int_a^b p_{L(f,\{a,b\})}(x)dx$$

Como $p_{L(f,\{a,b\})}$ é uma reta, podemos calcular analiticamente $\int_a^b p_{L(f,\{a,b\})}(x)dx$, esperando obter uma fórmula simples para seu valor. Desenvolvemos $p_{L(f,\{a,b\})}$ como

$$p_{L(f,\{a,b\})}(x) = f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a}$$

Temos então que

$$\int_a^b p_{L(f,\{a,b\})}(x)dx = \int_a^b \left(f(a)\frac{x-b}{a-b} + f(b)\frac{x-a}{b-a} \right) dx$$

Como a integral da soma é a soma das integrais, e como a integral de uma constante multiplicada por uma função é a constante multiplicada pela integral

da função, temos

$$= \frac{f(a)}{a-b}\int_a^b (x-b)dx + \frac{f(b)}{b-a}\int_a^b (x-a)dx$$

Agora encontramos as primitivas das integrais acima, que são bastante elementares.

$$= \frac{f(a)}{a-b}\left[\frac{x^2}{2} - bx\right]_a^b + \frac{f(b)}{b-a}\left[\frac{x^2}{2} - ax\right]_a^b$$

Em seguida, aplicamos o Teorema Fundamental do Cálculo.

$$= \frac{f(a)}{a-b}\left(-\frac{b^2}{2} - \frac{a^2}{2} + ab\right) + \frac{f(b)}{b-a}\left(\frac{b^2}{2} - ab + \frac{a^2}{2}\right)$$

Retiramos o -1 que multiplica a expressão em parênteses do primeiro termo, e colocamos esse -1 multiplicando o denominador $a-b$. Além disso, simplificamos as frações contidas nos parênteses.

$$= \frac{f(a)}{b-a}\left(\frac{a^2+b^2}{2} - ab\right) + \frac{f(b)}{b-a}\left(\frac{a^2+b^2}{2} - ab\right)$$

Colocamos a parte comum dos termos em evidência, e obtemos

$$= (f(a)+f(b))\left(\frac{1}{b-a}\left(\frac{a^2+b^2}{2} - ab\right)\right)$$

Colocamos $-ab$ no numerador da fração mais interna, e em seguida fazemos o produto das duas frações.

$$= (f(a)+f(b))\frac{a^2-2ab+b^2}{2(b-a)} = (f(a)+f(b))\frac{(a-b)^2}{2(b-a)}$$

Como $(a-b)^2 = a^2 - 2ab + b^2 = (b-a)^2$, fazemos a substituição e obtemos

$$(f(a)+f(b))\frac{(a-b)^2}{2(b-a)} = (f(a)+f(b))\frac{(b-a)^2}{2(b-a)} = (b-a)\frac{f(a)+f(b)}{2}$$

Com isso, se $p_{L(f,\{a,b\})}$ for uma boa aproximação para f em $[a,b]$, temos

$$\int_a^b f(x)dx \approx (b-a)\frac{f(a)+f(b)}{2}$$

Exercício 19. *Desenvolva a regra do trapézio, tratando* $\int_a^b p_{L(f,\{a,b\})}(x)dx$ *como a área do trapézio de vértices* $(a,0),(b,0),(a,f(a))$ *e* $(b,f(b))$.

Claro que existe o caso de $p_{L(f,\{a,b\})}$ não ser uma boa aproximação para f em $[a,b]$. Como $p_{L(f,\{a,b\})}$ é uma reta, a regra do trapézio é efetiva para aproximar $\int_a^b f(x)dx$ quando f tem sua derivada segunda assumindo valores próximos a zero em $[a,b]$. Caso esse não seja o caso, podemos nos valer da propriedade

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx, \forall\, c \in [a,b]$$

Assim, podemos aproximar $\int_a^b f(x)dx$ aplicando a regra do trapézio para $\int_a^c f(x)dx$ e $\int_c^b f(x)dx$. Uma escolha bastante comum para c é $c = \frac{a+b}{2}$.

Essa estratégia de particionar o intervalo de integração em subintervalos de igual comprimento pode facilmente ser generalizada. Para particionar $[a,b]$ em k subintervalos de mesmo comprimento, $k \in \mathbb{N}$, utilizamos os pontos $\{a_0, a_1, \ldots, a_k\}$ definidos por $a_i = a + ih, i \in \{0\} \cup [k]$, onde $h = \frac{b-a}{k}$ é o comprimento de cada subintervalo. Usamos a generalização da propriedade apresentada

$$\int_a^b f(x)dx = \sum_{i=0}^{k-1} \int_{a_i}^{a_{i+1}} f(x)dx$$

Com essa generalização, podemos aplicar a regra do trapézio para aproximar as integrais em cada um dos subintervalos, somar seus resultados e obter uma aproximação para a integral no intervalo completo. Podemos iterar, inclusive, sobre valores de k para verificar se há convergência do método ao passo em que consideramos mais subintervalos.

Mas qual seria o critério de convergência? Bom, como não podemos esperar ter o resultado da integral que estamos aproximando (se tivéssemos não faria sentido buscar por aproximações), vamos admitir que o método converge, dada uma margem $\epsilon \in \mathbb{R}_+$, se a diferença entre a aproximação da iteração atual e a da iteração anterior for menor que ϵ, em valores absolutos.

Exercício 20. *Sintetize um método que aproxima uma integral $\int_a^b f(x)dx$, aplicando a regra do trapézio em k subintervalos de mesmo comprimento que particionam $[a,b]$, na sua k-ésima iteração. O método termina quando atingir o critério de convergência apresentado acima. Quanto ao número de operações realizadas em uma iteração, existe alguma vantagem em tomar intervalos de mesmo comprimento?*

Algo interessante de se notar sobre o erro da regra do trapézio simples é sua relação com a concavidade de f em $[a,b]$. Se f é côncava em $[a,b]$, a regra do trapézio subestima sua integral no dito intervalo. No caso de f convexa, a regra do trapézio superestima a mesma integral.

Exercício 21. *Enuncie a definição de função côncava e de função convexa. Como essas definições podem ser utilizadas para argumentar sobre o comportamento do erro da regra do trapézio?*

Exercício 22. *Aproxime $\int_0^\pi cos(x)dx$ usando a regra do trapézio simples. Explique o resultado obtido.*

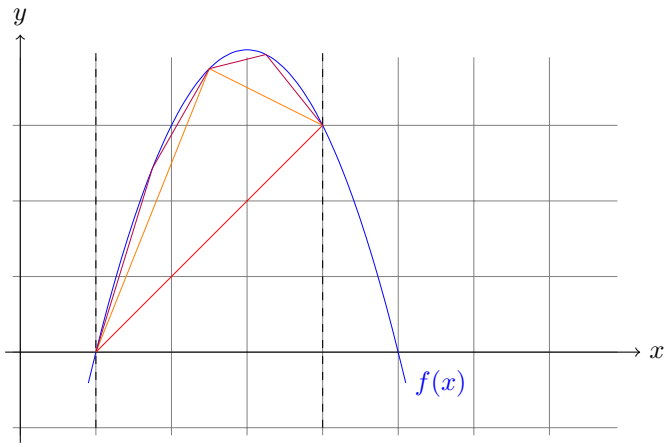

Figura 5: Ilustração da regra do trapézio para $\int_1^4 f(x)dx$, com $f(x) = -(x-1)(x-5)$: regra do trapézio simples em vermelho, considerando dois subintervalos em laranja, e considerando quatro em roxo.

Exercício 23. *Calcule $\int_0^2 \sqrt{x}dx$ usando a regra do trapézio simples. Calcule também usando quatro subintervalos de mesmo comprimento. Compare os resultados das aproximações com o valor exato da integral.*

3.2 Regra de Simpson

Agora, vamos desenvolver a segunda quadratura de Newton-Cottes. Vamos tentar aproximar a integral $\int_a^b f(x)dx$ usando um polinômio de Lagrange de grau 2. Vamos utilizar $p_{L(f,\{a,c,b\})}$, com $c = \frac{a+b}{2}$. Vamos, portanto, desenvolver $p_{L(f,\{a,c,b\})}$

$$p_{L(f,\{a,c,b\})}(x) = f(a)\frac{(x-c)(x-b)}{(a-c)(a-b)} + f(c)\frac{(x-a)(x-b)}{(c-a)(c-b)} + f(b)\frac{(x-a)(x-c)}{(b-a)(b-c)}$$

Agora vamos calcular a integral de $p_{L(f,\{a,c,b\})}$ no intervalo $[a,b]$

$$\int_a^b p_{L(f,\{a,c,b\})}(x)dx = f(a)\int_a^b \frac{(x-c)(x-b)}{(a-c)(a-b)}dx + f(c)\int_a^b \frac{(x-a)(x-b)}{(c-a)(c-b)}dx + f(b)\int_a^b \frac{(x-a)(x-c)}{(b-a)(b-c)}dx$$

Para facilitar as coisas, vamos fazer a substituição $x = a + ht$, onde t vai ser nossa nova variável e $h = \frac{b-a}{2}$. Temos que $dx = hdt$. Perceba que $c = a + h$ e que $b = a + 2h$.

$$\int_a^b p_{L(f,\{a,c,b\})}(x)dx = f(a)\int_0^2 \frac{h(t-1)h(t-2)}{(-h)(-2h)}hdt +$$

$$+f(c)\int_0^2 \frac{h(t)h(t-2)}{(h)(-h)}hdt + f(b)\int_0^2 \frac{h(t)h(t-1)}{(2h)(h)}hdt$$

Agora, agrupamos os termos h para uma posterior simplificação.

$$\int_a^b p_{L(f,\{a,c,b\})}(x)dx = f(a)\int_0^2 \frac{h^3(t-1)(t-2)}{2h^2}dt + f(c)\int_0^2 \frac{h^3t(t-2)}{-h^2}dt + f(b)\int_0^2 \frac{h^3t(t-1)}{2h^2}dt$$

Após a simplificação de h, as constantes (em termos de h e 2) são retiradas das integrais.

$$= \frac{hf(a)}{2}\int_0^2 (t-1)(t-2)dt - hf(c)\int_0^2 t(t-2)dt + \frac{hf(b)}{2}\int_0^2 t(t-1)dt$$

Após escrever os polinômios de t de maneira explícita, calculamos suas primitivas.

$$= \frac{hf(a)}{2}\left[\frac{t^3}{3} - \frac{3t^2}{2} + 2t\right]_0^2 - hf(c)\left[\frac{t^3}{3} - t^2\right]_0^2 + \frac{hf(b)}{2}\left[\frac{t^3}{3} - \frac{t^2}{2}\right]_0^2$$

Aplicamos o Teorema Fundamental do Cálculo. Note que, como o limite inferior de integração é 0, isso consiste apenas em substituir t por 2.

$$= \frac{hf(a)}{2}\left(\frac{8}{3} - 6 + 4\right) - hf(c)\left(\frac{8}{3} - 4\right) + \frac{hf(b)}{2}\left(\frac{8}{3} - 2\right)$$

$$= \frac{hf(a)}{2}\left(\frac{2}{3}\right) - hf(c)\left(-\frac{4}{3}\right) + \frac{hf(b)}{2}\left(\frac{2}{3}\right)$$

$$= \frac{hf(a)}{3} + \frac{4hf(c)}{3} + \frac{hf(b)}{3}$$

$$= \frac{h}{3}(f(a) + 4f(c) + f(b))$$

Substituindo h e c, temos que

$$\int_a^b p_{L(f,\{a,c,b\})}(x)dx = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

E temos desemvolvido a regra de Simpson. Se $p_{L(f,\{a,c,b\})}$ é uma boa aproximação para f em $[a,b]$, então podemos esperar que a regra de Simpson forneça um valor próximo da integral em questão.

Como a regra de Simpson é baseada numa interpolação via um polinômio de Lagrange de grau 2, é óbvio que ela não comete erros quando o integrando é um polinômio de grau 2. O que é muito peculiar é o fato da regra de Simpson não cometer erros quando o integrando é um polinômio de grau 3!

Exercício 24. *Sabendo que a fórmula do erro que a regra de Simpson comete ao aproximar* $\int_a^b f(x)dx$ *é*

$$-\frac{1}{90}\left(\frac{b-a}{2}\right)^5 \frac{d^4}{dx^4}f(c)$$

para algum $c \in [a,b]$*, explique por que a regra de Simpson não comete erros quando o integrando é um polinômio de grau 3.*

Quanto à decomposição da integral em subintervalos, convergência e critérios de parada, podemos perceber que o que foi dito na seção da regra do Trapézio pode ser generalizado para qualquer quadratura. Mais ainda, deve ter ficado evidente que quadraturas de Newton-Cottes são desenvolvidas via polinômios de Lagrange, tomando pontos equidistantes em $[a,b]$.

Exercício 25. *Generalize o pseudocódigo desenvolvido no exercício 20, considerando* $Q(f,a,b)$ *uma quadratura genérica dada como entrada.*

Exercício 26. *Calcule* $\int_0^2 \sqrt{x}dx$ *usando a regra de Simpson simples. Calcule também usando 4 subintervalos de mesmo comprimento. Compare o resultado com o valor exato da integral.*

Dados pontos $X = \{x_1, x_2, \ldots, x_k\}, k \in \mathbb{N}$, equidistantes em $[a,b]$, com $x_1 = a$ e $x_k = b$, dizemos que a quadratura desenvolvida via $p_{L(f,X)}$ é uma quadratura de Newton-Cottes *fechada*. Se tomarmos $X' \subset X$ como $X' = \{x_2, x_3, \ldots, x_{k-1}\}$ e desenvolvermos uma quadratura via $p_{L(f,X')}$, ela é dita uma quadratura de Newton-Cottes *aberta*. Note que as quadraturas fechadas consideram os extremos do intervalo, enquanto as abertas não consideram.

As duas quadraturas de Newton-Cottes aqui apresentadas são fechadas. Como o desenvolvimento das abertas é feito de forma análoga, deixamos como exercício para o leitor.

Exercício 27. *Desenvolva uma quadratura de Newton-Cottes aberta considerando* $X = \{x_1, x_2, x_3\}$*. A quadratura a ser desenvolvida é chamada de regra do quadrado, ou* midpoint rule*.*

Exercício 28. *Desenvolva uma quadratura de Newton-Cottes fechada considerando quatro pontos equidistantes para a interpolação via Lagrange. A quadratura a ser dessenvolvida é conhecida como Simpson "três oitavos".*

Exercício 29. *Dado que você fez o exercício 7, explique por que o fenômeno de Runge desencoraja o uso de quadraturas de Newton-Cottes baseadas em polinômios de Lagrange de grau elevado.*

3.3 Introdução às quadraturas de Gauss-Legendre

Agora, tratamos das quadraturas de *Gauss-Legendre*, que são uma classe de quadraturas "especializadas"em aproximar integrais do tipo $\int_{-1}^1 f(x)dx$. À primeira vista, parece meio bobo desenvolver quadraturas para um intervalo de

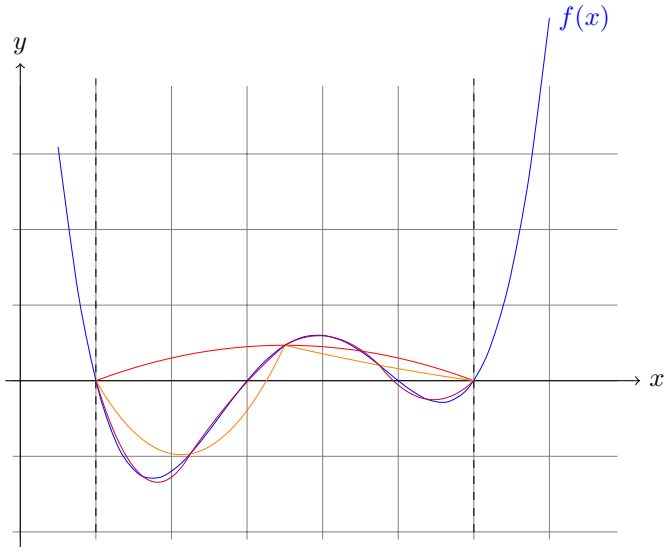

Figura 6: Ilustração da regra de Simpson para $\int_1^6 f(x)dx$, com $f(x) = 0.1(x-1)(x-3)(x-5)(x-6)$. São representadas aplicações com divisão em 1, 2 e 4 subintervalos, ilustradas pelas cores vermelha, laranja e roxa, respectivamente.

integração específico, mas basta um pouco de Cálculo para resolver o exercício a seguir.

Exercício 30. *Qual substituição de variáveis $x(t)$ é tal que $\int_a^b f(x)dx = \int_{-1}^1 g(t)dt$? Será que a reta que passa pelos pontos $(-1, a)$ e $(1, b)$ ajuda de alguma forma?*

Com isso esclarecido, vamos tentar ver quais são as vantagens em trabalhar com aproximações de integrais em um intervalo fixo. Vamos tentar fazer algo bem específico, inicialmente.

Primeiro, vamos tentar construir uma quadratura baseada em pontos distintos x_1 e x_2

$$\int_{-1}^1 f(x)dx \approx w_1 f(x_1) + w_2 f(x_2)$$

de tal forma que, se $f(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ (um polinômio de grau até 3), nossa quadratura não comete erros, ou seja

$$\int_{-1}^1 (c_0 + c_1 x + c_2 x^2 + c_3 x^3)dx = w_1 f(x_1) + w_2 f(x_2)$$

A pergunta é: existem valores para w_i e x_i, com $i \in [2]$, tal que a nossa quadratura tem a propriedade desejada? Vamos utilizar Álgebra Linear para descobrir!

Perceba que a classe de polinômios de grau até 3, denotada $\mathcal{P}^3$, tem umas propriedades que lembram bastante as de um espaço vetorial.

Exercício 31. *Mostre que, se $p_1, p_2 \in \mathcal{P}^3$, então $p_1 + p_2 \in \mathcal{P}^3$. Mostre que, se $p \in \mathcal{P}^3$ e $c \in \mathbb{R}$, então $cp \in \mathcal{P}^3$. Perceba também que o polinômio nulo está em $\mathcal{P}^3$.*

Se $\mathcal{P}^3$ se assemelha a um espaço vetorial, deve haver um conjunto de polinômios em $\mathcal{P}^3$ capaz de "gerar"qualquer polinômio em $\mathcal{P}^3$, uma espécie de "base". Tal "base"é $\{1, x, x^2, x^3\}$. Temos, inclusive, a propriedade

$$\int_{-1}^{1}(c_0+c_1x+c_2x^2+c_3x^3)dx = c_0\int_{-1}^{1}dx+c_1\int_{-1}^{1}xdx+c_2\int_{-1}^{1}x^2dx+c_3\int_{-1}^{1}x^3dx$$

Percebemos que, para nossa quadratura ter a propriedade desejada, basta garantirmos a propriedade para os casos $f(x) = 1, f(x) = x, f(x) = x^2$ e $f(x) = x^3$.

$$\int_{-1}^{1} dx = w_1 + w_2$$

$$\int_{-1}^{1} xdx = w_1x_1 + w_2x_2$$

$$\int_{-1}^{1} x^2dx = w_1x_1^2 + w_2x_2^2$$

$$\int_{-1}^{1} x^3dx = w_1x_1^3 + w_2x_2^3$$

Resolvendo as integrais, obtemos

$$w_1 + w_2 = 2$$

$$w_1x_1 + w_2x_2 = 0$$

$$w_1x_1^2 + w_2x_2^2 = \frac{2}{3}$$

$$w_1x_1^3 + w_2x_2^3 = 0$$

O conjuto-solução do sistema não-linear acima representa todas as quadraturas com a propriedade que buscamos, descrevendo a ponderação w e em que pontos x devemos computar valores de f. Poxa, mas resolver sistemas não-lineares pode ser difícil. Podemos inserir mais restrições sobre nossa quadratura para facilitar a resolução desse sistema (perceba como essa abordagem é experimental: se não encontrarmos uma solução, nada podemos concluir).

Podemos, por exemplo, exigir que x_1 e x_2 sejam simétricos, isto é, $x_1 = -x_2$. Isso nos permite escrever

$$w_1 x_1 - w_2 x_1 = 0$$

$$x_1(w_1 - w_2) = 0$$

Como x_1 e x_2 são distintos e $x_1 = -x_2$, temos $x_1 \neq 0$. Isso implica que $w_1 - w_2 = 0$, e portanto $w_1 = w_2$. Como sabemos que $w_1 + w_2 = 2$, concluímos que $w_1 = w_2 = 1$. Sob a restrição $x_1 = -x_2$, e com w determinado, escrevemos

$$x_1^2 + (-x_1)^2 = \frac{2}{3}$$

$$2x_1^2 = \frac{2}{3}$$

Como estamos forçando x_1 e x_2 a serem simétricos, podemos admitir que $x_1 = \frac{\sqrt{3}}{3}$ e $x_2 = -\frac{\sqrt{3}}{3}$. Com isso, encontramos uma quadratura em dois pontos com a propriedade buscada.

$$\int_{-1}^{1} f(x)dx \approx f\left(\frac{\sqrt{3}}{3}\right) + f\left(-\frac{\sqrt{3}}{3}\right)$$

Acabamos de desenvolver a quadratura de Gauss-Legendre de ordem 2, que não comete erros ao aproximar a integral, no intervalo $[-1, 1]$, de polinômios de grau até 3. No caso geral, a quadratura de Gauss-Legendre de ordem n mantém a precisão para polinômios de grau até $2n - 1$, e pode ser determinada como uma solução de um sistema não-linear envolvendo $2n$ variáveis e $2n$ equações.

Como exemplo, vamos verificar que a nossa quadratura de Gauss-Legendre calcula com exatidão a integral de $p(x) = x^3 + 4x^2 - 2x + 1$ no intervalo $[-1, 1]$.

$$\int_{-1}^{1} p(x)dx = p\left(\frac{\sqrt{3}}{3}\right) + p\left(-\frac{\sqrt{3}}{3}\right)$$

Vamos primeiro calcular o valor da integral do polinômio $p(x)$ em $[-1, 1]$.

$$\left[\frac{x^4}{4} + \frac{4x^3}{3} - x^2 + x\right]_{-1}^{1}$$

$$\frac{1}{4} + \frac{4}{3} - \left(\frac{1}{4} - \frac{4}{3} - 1 - 1\right)$$

$$\frac{8}{3} + 2 = \frac{14}{3}$$

Agora vamos calcular $p\left(\frac{\sqrt{3}}{3}\right) + p\left(-\frac{\sqrt{3}}{3}\right)$.

$$\left(\frac{\sqrt{3}}{3}\right)^3 + 4\left(\frac{\sqrt{3}}{3}\right)^2 - 2\frac{\sqrt{3}}{3} + 1 - \left(\frac{\sqrt{3}}{3}\right)^3 + 4\left(\frac{\sqrt{3}}{3}\right)^2 + 2\frac{\sqrt{3}}{3} + 1$$

$$8\left(\frac{\sqrt{3}}{3}\right)^2 + 2 = 8\frac{3}{9} + 2 = \frac{8}{3} + 2 = \frac{14}{3}$$

A ordem da quadratura essencialmente é o número de pontos em que se toma valores de f. Para n ímpar, 0 é um dos pontos tomados, enquanto os demais são pares simétricos. Para n par, apenas pares de pontos simétricos são tomados.

Exibimos valores de w e x para algumas ordens de quadraturas de Gauss-Legendre:

- Ordem 2: $x_1, x_2 = \pm\frac{\sqrt{3}}{3}$; $w_1, w_2 = 1$
- Ordem 3: $x_1 = 0, x_2, x_3 = \pm\frac{\sqrt{15}}{5}$; $w_1 = \frac{8}{9}, w_2, w_3 = \frac{5}{9}$
- Ordem 4: $x_1, x_2 = \pm\sqrt{\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}}, x_3, x_4 = \pm\sqrt{\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}}$; $w_1, w_2 = \frac{18+\sqrt{30}}{36}, w_3, w_4 = \frac{18-\sqrt{30}}{36}$

Exercício 32. *Escreva a fórmula de uma quadratura, baseada em Gauss-Legendre de ordem 3, para aproximar integrais da forma* $\int_a^b f(x)dx$.

Chamamos a atenção para o fato de quadraturas de Gauss-Legendre serem boas aproximações para funções que muito se assemelham a polinômios de até um certo grau. No entanto, devemos notar também o custo de aplicar a mudança de variável e as operações de divisão que ela acrescenta. Outro ponto de destaque é que os pontos e pesos utilizados por vezes são irracionais ou dízimas, e a falta de uma representação exata para esses já incorre em erros de representação. Na prática, Gauss-Legendre comete pelo menos os erros de representação (e acumulados em suas operações) mesmo nos casos em que deveria cometer nenhum erro.

3.4 Integrais duplas

Até agora, apresentamos métodos de aproximação, ou quadraturas, para integrais de funções em uma variável. Vamos apresentar meios de aproveitar essas quadraturas para aproximar integrais de funções com duas variáveis, isto é, funções com o domínio em $\mathbb{R}^2$ e o contradomínio em $\mathbb{R}$.

Dada $f : A \times B \to C$, com $A, B, C \subseteq \mathbb{R}$, tomamos a integral da forma

$$\int_a^b \left(\int_{c(x)}^{d(x)} f(x, y)dy \right) dx$$

Vamos considerar a função

$$F(x) = \int_{c(x)}^{d(x)} f(x,y)dy$$

Note que, dado um ponto $\overline{x} \in A$, o valor $F(\overline{x})$ é dado pela integral definida $\int_{c(\overline{x})}^{d(\overline{x})} f(\overline{x},y)dy$, e essa nós já sabemos aproximar. Pela definição de F, sabemos aproximar valores de F usando quadraturas. E também podemos escrever

$$\int_a^b \left(\int_{c(x)}^{d(x)} f(x,y)dy \right) dx = \int_a^b F(x)dx$$

Com isso, podemos aproximar a integral dupla usando quadraturas. Dada uma quadratura (vamos chamar essa de "externa") para aproximar $\int_a^b F(x)dx$, determinamos os pontos em que F deve ser computada, e aproximamos o valor de F nesses pontos utilizando também uma quadratura (e essa chamamos de "interna"). Podemos utilizar quadraturas distintas na parte externa e na parte interna da integral.

Exercício 33. *Calcule a seguinte integral dupla. Depois, aproxime-a usando Gauss-Legendre de ordem 2 na integral externa e a regra de Simpson na interna. Compare a aproximação com o valor da integral.*

$$\int_{-1}^{1} \left(\int_0^2 (2x - 3y^2)dy \right) dx$$

Exercício 34. *Descreva um método para aproximar integrais da forma*

$$\int_a^b \left(\int_{c(x)}^{d(x)} \left(\int_{g(x,y)}^{h(x,y)} f(x,y,z)dz \right) dy \right) dx$$

4 Derivação numérica

Nesta seção, vamos tratar de como aproximar o valor da derivada de uma função em um certo ponto de seu domínio. Para tanto, vamos desenvolver fórmulas baseadas em polinômios de Taylor.

4.1 A abordagem mais simples

Tome $f : A \to B$, com $A, B \subseteq \mathbb{R}$. Como bem lembramos do Cálculo, a derivada de f em um ponto $a \in A$ pode ser definida como

$$\frac{d}{dx}f(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

Como não podemos representar grandezas infinitesimais em um computador, o que podemos fazer com essa definição é substituir h por um valor pequeno,

mas devemos estar cientes de que valores muito pequenos podem provocar erros numéricos. Certamente, o valor encontrado dessa forma é uma aproximação de $\frac{d}{dx}f(a)$, mas não sabemos o quão boa ela é. Para sabermos medir o erro cometido por uma aproximação de derivada, vamos fazer uso da Expansão de Taylor.

4.2 Utilizando a Expansão de Taylor

Como visto anteriormente, a Expansão de Taylor de uma função $f : A \to B$, com $A, B \subseteq \mathbb{R}$, em torno de um ponto $a \in A$ é dada por

$$f(a+h) = f(a) + \sum_{i=1}^{\infty} \left(\frac{1}{i!} \frac{d^i}{dx^i} f(a) \right) h^i$$

Retiramos a primeira parcela do somatório, para evidenciar $\frac{d}{dx}f(a)$

$$f(a+h) = f(a) + \frac{d}{dx}f(a)h + \sum_{i=2}^{\infty} \left(\frac{1}{i!} \frac{d^i}{dx^i} f(a) \right) h^i$$

Agora vamos isolar $\frac{d}{dx}f(a)$

$$f(a+h) - f(a) - \sum_{i=2}^{\infty} \left(\frac{1}{i!} \frac{d^i}{dx^i} f(a) \right) h^i = \frac{d}{dx}f(a)h$$

$$\frac{d}{dx}f(a) = \frac{f(a+h) - f(a)}{h} - \sum_{i=2}^{\infty} \left(\frac{1}{i!} \frac{d^i}{dx^i} f(a) \right) h^{i-1}$$

Podemos perceber que utilizar a definição de derivada, substituindo h por um valor pequeno, produz um erro. Já que propomos utilizar valores pequenos para h, podemos admitir $h \in (-1, 1)$. Como argumentado anteriormente, podemos escrever

$$\frac{d}{dx}f(a) = \frac{f(a+h) - f(a)}{h} + O(h)$$

Ou seja, constatamos que a nossa abordagem básica para derivação numérica produz um erro assintoticamente proporcional a h. Será que podemos fazer melhor que isso?

Como estamos lidando com $h \in (-1, 1)$, erros menores que $O(h)$ seriam algo como $O(h^2), O(h^3)$, e assim por diante. Pelo visto, a definição de derivada não vai ser o bastante para atingirmos tais erros. Mas e se considerássemos outros valores de f em torno de a?

Tome os pontos $a + \overline{h}$ e $a - \overline{h}$, para algum $\overline{h} \in (-1, 1)$. Substituimos h por $\overline{h}$ e $-\overline{h}$, respectivamente, na fórmula da Expansão de Taylor (evidenciando as duas primeiras parcelas do somatório)

$$f(a+\overline{h}) = f(a) + \frac{d}{dx}f(a)\overline{h} + \frac{1}{2}\frac{d^2}{dx^2}f(a)\overline{h}^2 + \sum_{i=3}^{\infty}\left(\frac{1}{i!}\frac{d^i}{dx^i}f(a)\right)\overline{h}^i \quad (4.1)$$

$$f(a-\overline{h}) = f(a) - \frac{d}{dx}f(a)\overline{h} + \frac{1}{2}\frac{d^2}{dx^2}f(a)\overline{h}^2 + \sum_{i=3}^{\infty}\left(\frac{1}{i!}\frac{d^i}{dx^i}f(a)\right)(-\overline{h})^i \quad (4.2)$$

Multiplicamos (4.2) por -1

$$-f(a-\overline{h}) = -f(a) + \frac{d}{dx}f(a)\overline{h} - \frac{1}{2}\frac{d^2}{dx^2}f(a)\overline{h}^2 - \sum_{i=3}^{\infty}\left(\frac{1}{i!}\frac{d^i}{dx^i}f(a)\right)(-\overline{h})^i \quad (4.3)$$

Somamos as equações (4.1) e (4.3)

$$f(a+\overline{h}) - f(a-\overline{h}) = 2\frac{d}{dx}f(a)\overline{h} + 2\sum_{i=1}^{\infty}\left(\frac{1}{(2i+1)!}\frac{d^{2i+1}}{dx^{2i+1}}f(a)\right)\overline{h}^{2i+1}$$

Isolamos $\frac{d}{dx}f(a)$

$$f(a+\overline{h}) - f(a-\overline{h}) - 2\sum_{i=1}^{\infty}\left(\frac{1}{(2i+1)!}\frac{d^{2i+1}}{dx^{2i+1}}f(a)\right)\overline{h}^{2i+1} = 2\frac{d}{dx}f(a)\overline{h}$$

$$\frac{d}{dx}f(a) = \frac{f(a+\overline{h}) - f(a-\overline{h})}{2\overline{h}} - \sum_{i=1}^{\infty}\left(\frac{1}{(2i+1)!}\frac{d^{2i+1}}{dx^{2i+1}}f(a)\right)\overline{h}^{2i}$$

Escrevemos de forma simplificada, substituindo $\overline{h}$ por h

$$\frac{d}{dx}f(a) = \frac{f(a+h) - f(a-h)}{2h} + O(h^2)$$

Com isso percebemos que, apesar de continuar utilizando dois valores de f em torno de a, pudemos obter uma melhor aproximação para a derivada de f em a, e nem mesmo utilizamos o valor de f em a!

As fórmulas que estamos desenvolvendo e analisando com o auxílio da Expansão de Taylor têm nome: são chamadas de *diferenças finitas*. Elas tomam pontos equidistantes numa região de a para aproximar a derivada de f em a. Elas podem ser classificadas, inclusive, de acordo com os pontos que utilizam para aproximar a derivada de f em a: quando os pontos utilizados estão à direita de a, dizemos tratar-se de uma *forward difference*; quando estão à esquerda, trata-se de uma *backward difference*; e quando são tomados pares de pontos simétricos em torno de a, estamos falando de uma *central difference*.

Percebemos que nossa abordagem simples, baseada na definição de derivada, consiste em utilizar uma forward difference, e que a fórmula desenvolvida em seguida é uma central difference.

A obtenção de aproximações para derivadas e a análise de seus erros provêm da mera manipulação algébrica da Expansão de Taylor, aplicada em pontos em torno do ponto de interesse. Como a estratégia utilizada para a obtenção das aproximações consiste em isolar $\frac{d}{dx}f(a)$ após algum algebrismo, é natural que possamos desenvolver aproximações não apenas para $\frac{d}{dx}f(a)$, mas para derivadas de quaisquer ordens, já que a Expansão de Taylor relaciona todas elas.

Exercício 35. *Desenvolva uma aproximação para* $\frac{d^2}{dx^2}f(a)$. *Dica: é fácil a partir de* (4.1) *e* (4.2).

Também é natural pensar que, se a derivada de f em a nos diz um pouco sobre como a função se comporta em torno de a, então podemos entender melhor a derivada em a caso tomemos mais pontos de "amostra"em torno de a. Essa intuição nos leva a desenvolver aproximações que utilizam cada vez mais pontos em torno de a, com o objetivo de reduzir o erro cometido.

Vamos considerar agora os pontos $a-2\overline{h}, a-\overline{h}, a+\overline{h}$ e $a+2\overline{h}$, para algum $\overline{h} \in (-1, 1)$. Vamos substituí-los na fórmula da Expansão de Taylor, evidenciando suas quatro primeiras parcelas.

$$f(a-2\overline{h}) = f(a) - 2\frac{d}{dx}f(a)\overline{h} + 2\frac{d^2}{dx^2}\overline{h}^2 - \frac{4}{3}\frac{d^3}{dx^3}\overline{h}^3 + \frac{2}{3}\frac{d^4}{dx^4}\overline{h}^4 + O(h^5) \tag{4.4}$$

$$f(a-\overline{h}) = f(a) - \frac{d}{dx}f(a)\overline{h} + \frac{1}{2}\frac{d^2}{dx^2}f(a)\overline{h}^2 - \frac{1}{6}\frac{d^3}{dx^3}\overline{h}^3 + \frac{1}{24}\frac{d^4}{dx^4}\overline{h}^4 + O(h^5) \tag{4.5}$$

$$f(a+\overline{h}) = f(a) + \frac{d}{dx}f(a)\overline{h} + \frac{1}{2}\frac{d^2}{dx^2}f(a)\overline{h}^2 + \frac{1}{6}\frac{d^3}{dx^3}\overline{h}^3 + \frac{1}{24}\frac{d^4}{dx^4}\overline{h}^4 + O(h^5) \tag{4.6}$$

$$f(a+2\overline{h}) = f(a) + 2\frac{d}{dx}f(a)\overline{h} + 2\frac{d^2}{dx^2}\overline{h}^2 + \frac{4}{3}\frac{d^3}{dx^3}\overline{h}^3 + \frac{2}{3}\frac{d^4}{dx^4}\overline{h}^4 + O(h^5) \tag{4.7}$$

Primeiro, tomamos (4.4) e (4.7). Multiplicamos (4.4) por -1 e somamos a (4.7), obtendo (4.8). Agora, tomamos (4.5) e (4.6). Multiplicamos (4.5) por -1 e somamos a (4.6), obtendo (4.9).

$$f(a+2\overline{h}) - f(a-2\overline{h}) = 4\frac{d}{dx}f(a)\overline{h} + \frac{8}{3}\frac{d^3}{dx^3}f(a)\overline{h}^3 + O(h^5), \tag{4.8}$$

$$f(a+\overline{h}) - f(a-\overline{h}) = 2\frac{d}{dx}f(a)\overline{h} + \frac{1}{3}\frac{d^3}{dx^3}f(a)\overline{h}^3 + O(h^5) \tag{4.9}$$

Em seguida, multiplicamos (4.9) por -8 e somamos a (4.8).

$$f(a+2\overline{h}) - 8f(a+\overline{h}) + 8f(a-\overline{h}) - f(a-2\overline{h}) = -12\frac{d}{dx}f(a)\overline{h} + O(h^5)$$

Isolando $\frac{d}{dx}f(a)$ e substituindo $\overline{h}$ por h, obtemos

$$\frac{d}{dx}f(a) = \frac{-f(a+2h) + 8f(a+h) - 8f(a-h) + f(a-2h)}{12h} + O(h^4)$$

Manipulando suficientemente a Expansão de Taylor, conseguimos obter aproximações para a derivada de f em a tão boas quanto quisermos. Por esse motivo, dispensamos o uso de técnicas iterativas para a minimização do erro cometido por nossas aproximações, uma vez que isso é proporcionado pelas técnicas algébricas.

Por fim, observamos o compromisso que há em utilizar aproximações de derivada de alta precisão. Quanto maior a precisão que utilizamos, mais vezes temos de computar a função cuja derivada nos interessa. Com uma função computacionalmente cara, torna-se proibitivo utilizar boas aproximações de sua derivada. Além disso, se temos o propósito de aproximar a derivada em diversos pontos, é preciso notar que aproximações cada vez melhores exigem sempre mais operações aritméticas, e isso deve ser considerando, mesmo quando a função é computacionalmente barata.

Exercício 36. *A maioria das aproximações aqui apresentadas são central differences. Desenvolva uma forward difference para aproximar* $\frac{d}{dx}f(a)$ *com erro de ordem* $O(h^2)$*. Dica: utilize os pontos* $a, a+\overline{h}$ *e* $a+2\overline{h}$*.*

Exercício 37. *Desenvolva uma aproximação para* $\frac{d^3}{dx^3}f(a)$ *com erro de ordem* $O(h^2)$*.*

5 Álgebra Linear Computacional

Nesta seção, começamos a descrever e analisar técnicas numéricas para lidar com alguns dos tópicos mais elementares de Álgebra Linear. Vamos abordar técnicas para resolução de sistemas de equações lineares, decomposições de matrizes, cálculo de determinantes e suavização de erros. Antes disso, vamos fazer uma pequena revisão dos conceitos de Álgebra Linear.

5.1 Definições

Vamos começar com o básico. Álgebra Linear trata de entidades e operações que ocorrem em certas estruturas algébricas chamadas *espaços vetoriais*. Um espaço vetorial é constituído de um conjunto V de *vetores* e de um corpo F de *escalares* (usamos F porque um corpo é chamado *field* em inglês), e possui duas operações: a *soma* $(+)$ entre elementos de V e a *multiplicação* $(*)$ entre um elemento de V e um elemento de F. Além disso, um espaço vetorial deve satisfazer as seguintes propriedades:

- Deve haver um vetor $0 \in V$ chamado *origem* tal que $0 + v = v$, para todo $v \in V$.
- Para todo vetor $v \in V$, deve existir um vetor $-v \in V$, o *inverso aditivo* de v, tal que $v + (-v) = 0$.
- A operação de soma deve ser comutativa e associativa.
- Deve haver um escalar $1 \in F$ tal que $1 * v = v$, para todo $v \in V$.
- $(ab) * v = a * (b * v)$, para todo vetor $v \in V$, para quaisquer escalares $a, b \in F$.
- A multiplicação deve ser distributiva em relação à adição.
- A multiplicação deve ser distributiva em relação à adição em F.

Como é costumeiro construir espaços vetoriais utilizando números reais, consideramos que V é $\mathbb{R}^n$, para algum $n \in \mathbb{N}$, e que F é $\mathbb{R}$. Esses são chamados de espaços vetoriais reais. Vamos denotar um espaço vetorial real $(\mathbb{R}^n, \mathbb{R})$ apenas como $\mathbb{R}^n$, uma vez que sempre vamos considerar o mesmo corpo de escalares.

Exercício 38. *Dado um espaço vetorial (V, F), prove a unicidade de 0 em V.*

Exercício 39. *Pode haver um espaço vetorial (V, F) com V finito? E com F finito?*

Dizemos que um vetor $u \in \mathbb{R}^n$ é *combinação linear* de um conjunto de vetores $\{v_1, \ldots, v_k\} \subset \mathbb{R}^n, k \in \mathbb{N}$, se podemos escrever $u = \sum_{i=1}^{k} c_i v_i$, para alguns escalares $c_i \in \mathbb{R}, i \in [k]$. Um conjunto de vetores $V' \subset \mathbb{R}^n$ é *linearmente independente* se, para todo $v \in V'$, v não é combinação linear de $V' \setminus \{v\}$ (de outra forma, V' é linearmente dependente). Se um conjunto de vetores $V' \subset \mathbb{R}^n$ é tal que todo vetor em $\mathbb{R}^n$ é combinação linear de V', dizemos que V' é um conjunto *gerador*. Se um conjunto de vetores $B \subset \mathbb{R}^n$ é linearmente independente e gerador, então B é dito uma *base* de $\mathbb{R}^n$. Um espaço vetorial pode ter infinitas bases.

Exercício 40. *Prove que, se $V' \cup \{v\}$ é uma base de $\mathbb{R}^n$, então $V' \cup \{cv\}$ é uma base de $\mathbb{R}^n$, para todo $c \in \mathbb{R}$.*

Exercício 41. *Tome $A \in \mathbb{R}^{m \times n}$. Prove que os vetores-coluna de A são linearmente independentes sse $Ax = 0$ tem apenas $x = 0$ como solução.*

Exercício 42. *Prove que uma base é um conjunto de vetores linearmente independente maximal e gerador minimal.*

Exercício 43. *Prove que, se B_1 e B_2 são bases de $\mathbb{R}^n$, então $|B_1| = |B_2|$.*

Como pode ser provado no exercício acima, bases de um mesmo espaço vetorial têm a mesma cardinalidade. Chamamos de *dimensão* de um espaço vetorial $S = (V, F)$ a cardinalidade de suas bases, denotando-a por $\dim(S)$. Dizemos que a base *canônica* de $\mathbb{R}^n$ é sua base composta pelos vetores-coluna da matriz identidade $I \in \mathbb{R}^{n \times n}$. Inclusive, é fácil perceber que $\dim(\mathbb{R}^n) = n$.

5.2 Sistemas lineares

Vamos abordar o problema de determinar o conjunto-solução de uma equação do tipo $Ax = b$, com $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$. Uma equação matricial dessa forma pode representar um sistema de equações lineares: a i-ésima linha de A seria composta pelos coeficientes da i-ésima equação do sistema linear, cujo termo independente seria a i-ésima posição de b.

Dado $Ax = b$ como descrito acima, perceba que os vetores-coluna de A e o vetor b estão no mesmo espaço, $\mathbb{R}^m$. Vamos chamar o conjunto dos vetores-coluna de A de C_A. Tome uma base B de $\mathbb{R}^m$. Pela definição de base, o vetor b é combinação linear de B. Se temos $B \subseteq C_A$, então b é combinação linear de C_A. Em particular, podemos reordenar as colunas de A e, consequentemente, as variáveis x, de forma a termos

$$Ax = b \iff \begin{bmatrix} A_B & A_N \end{bmatrix} \begin{bmatrix} x_B \\ x_N \end{bmatrix} = b$$

onde $A_B \in \mathbb{R}^{m\times m}$ é formada pelas colunas de A que estão em B e A_N pelas demais colunas. Similarmente, x_B e x_N são as variáveis correspondentes às colunas de A_B e A_N, respectivamente. Perceba que essa permutação de colunas preserva o sistema linear representado justamente porque as variáveis são igualmente permutadas. Podemos escrever, portanto

$$A_B x_B + A_N x_N = b$$

Como as colunas de A_B formam uma base de $\mathbb{R}^m$, sabemos que elas são linearmente independentes. Isso implica que A_B é invertível, e denotamos sua inversa por A_B^{-1}. Prosseguimos com

$$A_B x_B = b - A_N x_N$$

$$A_B^{-1} A_B x_B = A_B^{-1}(b - A_N x_N)$$

$$I x_B = A_B^{-1} b - A_B^{-1} A_N x_N$$

$$x_B = A_B^{-1} b - A_B^{-1} A_N x_N$$

O que podemos concluir com a equação acima? Podemos perceber que, se conhecemos uma base de $\mathbb{R}^m$ formada por algumas colunas de A, podemos usá-la para obter uma solução para $Ax = b$. Perceba também que a equação acima põe x_B em função de x_N. Dada uma valoração $\overline{x}_N$ para x_N, obtemos um vetor $b - A_N\overline{x}_N \in \mathbb{R}^m$, e como as colunas de A_B formam uma base de $\mathbb{R}^m$, $b - A_N\overline{x}_N$ é combinação linear de C_{A_B}. De fato, dada qualquer valoração $\overline{x}_N$ para x_N, a valoração $(\overline{x}_B, \overline{x}_N)$, onde

$$\overline{x}_B = A_B^{-1} b - A_B^{-1} A_N \overline{x}_N$$

é uma soolução para $Ax = b$.

Quando $n > m$, a matriz A_N tem pelo menos uma coluna e há pelo menos uma variável em x_N. Nesse caso, as soluções $(\overline{x}_B, \overline{x}_N)$ do sistema $Ax = b$ são chamadas de *soluções parametrizadas* associadas à base B, e as variáveis x_N são ditas os *parâmetros* (lembra que x_B está em função de x_N?). Como podemos valorar x_N de qualquer forma e obter uma solução para $Ax = b$, o conjunto-solução do sistema é infinito. Dentre as infinitas soluções parametrizadas associadas à base B, destacamos a solução $(\overline{x}_B, 0)$ obtida ao valorar $x_N = 0$, que é chamada de solução *básica*.

Quando $n = m$, a matriz A é quadrada, $A = A_B$ e $x = x_B$. Nesse caso, $Ax = b$ tem apenas a solução básica $\overline{x} = A^{-1}b$, já que não há parâmetros.

Até o presente momento, consideramos que alguma base de $\mathbb{R}^m$ é subconjunto de C_A. Essa condição é suficiente para garantir que $Ax = b$ tem soluções, sejam elas apenas uma (quando A é quadrada) ou infinitas. No caso em que essa condição não é válida, $Ax = b$ ou tem solução e algumas de suas equações são redundantes (implicadas pelas demais), ou é inconsistente, e portanto não tem solução.

Quando temos $Ax = b$ de tal forma que b está em um subespaço S de $\mathbb{R}^m$ e há uma base B de S tal que $B \subseteq C_A$, então o sistema tem solução e algumas de suas equações são redundantes. Quando isso não ocorre, isto é, quando b não está na união dos subespaços gerados pelos subconjuntos de C_A, então o sistema é dito inconsistente.

Exercício 44. *Considere o sistema*

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b$$

Há uma base de $\mathbb{R}^2$ em C_A? Quais subespaços de $\mathbb{R}^2$ são gerados pelos subconjuntos de C_A? Para quais valores de b o sistema tem solução?

Tudo indica que, para encontrarmos soluções de $Ax = b$, temos de saber como encontrar bases de $\mathbb{R}^m$ dentro de C_A ou encontrar sistemas equivalentes a $Ax = b$ que tenham bases evidentes.

Antes de prosseguir, vamos apresentar uma maneira conveniente de representar uma matriz para o desenvolvimento de algoritmos por indução. Enfatizamos que todos os nossos vetores, mesmo fora do contexto matricial, são tratados como vetores-coluna. Dada uma matriz $A \in \mathbb{R}^{m \times n}$, podemos decompor A como

$$A = \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix}$$

onde $a \in \mathbb{R}, r \in \mathbb{R}^{n-1}, c \in \mathbb{R}^{m-1}$ e $A' \in \mathbb{R}^{(m-1)\times(n-1)}$. Note que a representa o primeiro elemento de A, r^T representa o restante da primeira linha de A e c representa o restante da primeira coluna de A.

Primeiro, vamos mostrar que podemos resolver $Ax = b$ quando A é uma matriz triangular superior. Se $A \in \mathbb{R}^{1\times n}$, então $Ax = b$ pode ser escrito como

$a^T x = b$, onde a^T é a única linha de A. Seguindo nossa decomposição, temos

$$a^T x = b \iff \begin{bmatrix} a_B & r^T \end{bmatrix} \begin{bmatrix} x_B \\ x_N \end{bmatrix} = b$$

$$a_B x_B + r^T x_N = b$$

$$a_B x_B = b - r^T x_N$$

$$x_B = \frac{b - r^T x_N}{a_B}$$

Ou seja, temos uma descrição de solução com uma variável básica, tomando as demais variáveis como parâmetros. Essa descrição é válida quando $a \neq 0$, pois podemos permutar as colunas de A e as variáveis x de modo a obter um sistema equivalente com $a_B \neq 0$ (o que caracteriza qualquer base de $\mathbb{R}$, espaço onde b se encontra). Quando $a = 0$, temos que ou o sistema é trivial ($b = 0$) e seu conjunto-solução é $\mathbb{R}^n$ ou o sistema é inconsistente ($b \neq 0$) e não tem solução.

Exercício 45. *Caracterize o conjunto-solução de $Ax = b$, quando A é triangular superior e $A \in \mathbb{R}^{m \times 1}$*

Com um pouco de atenção, percebe-se que descrevemos a base de uma indução (uma parte foi deixada como exercício, na verdade). Agora, precisamos notar que uma matriz A triangular superior pode ser escrita, conforme nossa decomposição, da seguinte forma

$$A = \begin{bmatrix} a & r^T \\ 0 & A' \end{bmatrix}$$

onde A' também é triangular superior. Para resolver $Ax = b$, tratamos

$$\begin{bmatrix} a & r^T \\ 0 & A' \end{bmatrix} \begin{bmatrix} x_1 \\ x' \end{bmatrix} = \begin{bmatrix} b_1 \\ b' \end{bmatrix}$$

Fazendo a multiplicação no lado esquerdo, obtemos

$$\begin{bmatrix} ax_1 + r^T x' \\ 0x_1 + A'x' \end{bmatrix} = \begin{bmatrix} b_1 \\ b' \end{bmatrix}$$

Isso equivale a resolver o sistema

$$ax_1 + r^T x' = b_1$$
$$A'x' = b'$$

Por hipótese (já que A' é triangular superior e tem suas dimensões estritamente menores que as de A), sabemos determinar se $A'x' = b'$ tem solução.

Caso não tenha, o sistema $Ax = b$ também não tem. Caso tenha, seja $\overline{x}'$ tal que $A'\overline{x}' = b'$. Construimos uma solução $(\overline{x}_1, \overline{x}')$ para $Ax = b$, onde

$$\overline{x}_1 = \frac{b_1 - r^T \overline{x}'}{a}$$

Exercício 46. *Quantas operações são necessárias para determinar uma solução do sistema $Ax = b$, quando A é triangular superior?*

Exercício 47. *Descreva como obter uma solução para $Ax = b$ caso A seja triangular inferior.*

Mas e quando A não é triangular? Nesse caso, vamos precisar transformar o sistema $Ax = b$ em um sistema equivalente $PAx = Pb$, de tal forma que PA seja triangular. E quem é P? P é uma matriz construída para representar as operações necessárias para triangularizar a matriz A. Após obtida, basta multiplicar P à esquerda de ambos os lados do sistema $Ax = b$ para obter um sistema equivalente cuja matriz de coeficientes é triangular. Vamos precisar encontrar P.

Antes de continuar o desenvolvimento, vamos introduzir algumas notações. Vamos denotar por I a matriz identidade. Vamos denotar por E_{ij} a matriz quadrada com todas as posições nulas, exceto pela posição (i, j) cujo valor é 1.

Vamos utilziar três tipos de operações para construir P de modo que PA seja triangular: permutação de linhas, multiplicação de linha por escalar e adição de um múltiplo de uma linha a uma outra. A operação de permutação das linhas i e j pode ser representada pela matriz $R_{ij} = I - E_{ii} - E_{jj} + E_{ij} + E_{ji}$. A operação de multiplicação da linha i pelo escalar c pode ser representada pela matriz $S_i^c = I - E_{ii} + cE_{ii}$. Por fim, a operação de adicionar um múltiplo da linha i pelo escalar c à linha j pode ser representada pela matriz $S_{ij}^c = I + cE_{ji}$

Exercício 48. *Dada a matriz*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

calcule $R_{13}A$, $S_2^5 A$ e $S_{31}^2 A$

Com o auxílio das matrizes introduzidas acima, vamos descrever uma sequência de operações para triangularizar A. Vamos codificar tais operações na matriz P. Observamos que se $A \in \mathbb{R}^{1\times n}$, então A já é triangular superior e basta fazermos $P = I$.

Exercício 49. *Para o caso em que $A \in \mathbb{R}^{m\times 1}$, descreva P de modo que PA seja triangular superior.*

Novamente, temos descrito a base de um processo indutivo (com o auxílio do exercício). Tome $A \in \mathbb{R}^{m\times n}$. Decompomos A como

$$A = \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix}$$

Precisamos descrever uma sequência de operações de forma a transformar c em 0. Considere

$$c = \begin{bmatrix} c_2 \\ c_3 \\ \vdots \\ c_m \end{bmatrix}$$

indexado dessa forma porque o primeiro elemento da coluna é a. Precisamos tornar c_i em 0, $2 \leq i \leq m$.

Se $a \neq 0$, adicionamos a primeira linha de A, multiplicada por $s_i = -\frac{c_i}{a}$, à i-ésima linha de A, para $2 \leq i \leq m$. Tal operação é representada pela matriz $S_{1i}^{s_i}$, $2 \leq i \leq m$. Dessa forma, para transformar c em 0, multiplicamos A por

$$P_1 = \prod_{i=2}^{m} S_{1i}^{s_i}$$

Perceba que, se definimos

$$s = \begin{bmatrix} s_2 \\ s_3 \\ \vdots \\ s_m \end{bmatrix}$$

podemos decompor P_1 como

$$P_1 = \begin{bmatrix} 1 & 0^T \\ s & I \end{bmatrix}$$

e temos

$$P_1 A = \begin{bmatrix} 1 & 0^T \\ s & I \end{bmatrix} \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix} = \begin{bmatrix} a & r^T \\ as + Ic & sr^T + IA' \end{bmatrix} = \begin{bmatrix} a & r^T \\ as + c & sr^T + A' \end{bmatrix}$$

Pela definição de s, note que $as = -c$. Assim, substituimos e obtemos

$$P_1 A = \begin{bmatrix} a & r^T \\ 0 & sr^T + A' \end{bmatrix}$$

Tomamos $A'' = sr^T + A'$. Por hipótese, conhecemos $P_2 \in \mathbb{R}^{(m-1)\times(m-1)}$ tal que $P_2 A''$ é triangular superior. Definimos a matriz $P_3 \in \mathbb{R}^{m\times m}$ como

$$P_3 = \begin{bmatrix} 1 & 0^T \\ 0 & P_2 \end{bmatrix}$$

Agora, podemos construir uma matriz $P \in \mathbb{R}^{m\times m}$, baseada em P_1 e P_2, tal que PA é triangular superior. Escrevemos P como $P = P_3 P_1$ e verificamos

$$PA = P_3 P_1 A = \begin{bmatrix} 1 & 0^T \\ 0 & P_2 \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & A'' \end{bmatrix} = \begin{bmatrix} a & r^T \\ 0 & P_2 A'' \end{bmatrix}$$

Como P_2A'' é triangular superior, por hipótese, temos que PA é triangular superior.

Agora, caso tenhamos $a = 0$, não podemos fazer divisões por a e o método acima não se aplica. Se $c \neq 0$, então existe $c_i \neq 0$, para algum $2 \leq i \leq m$, e podemos permutar a primeira e a i-ésima linhas de A, obtendo um sistema equivalente $Cx = d$, com $C = R_{1i}A$ e $d = R_{1i}b$, que pode ser resolvido conforme foi descrito. Caso $c = 0$, então a variável do sistema $Ax = b$ associada à primeira coluna de A não tem influência no sistema (ela é multiplicada por 0 em todas as equações do sistema) e podemos ignorá-la (ela pode ser dita um *parâmetro irrelevante*). Nesse caso, dadas as decomposições

$$A = \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x' \end{bmatrix}$$

onde $x_1 \in \mathbb{R}$ e $x' \in \mathbb{R}^{n-1}$, consideramos o sistema $Cx' = b$ onde

$$C = \begin{bmatrix} r^T \\ A' \end{bmatrix}$$

Assim, se $Cx' = b$ tem uma solução $\overline{x}'$, construímos soluções para $Ax = b$ baseadas em $\overline{x}'$, associando qualquer valor a x_1. Se $Cx' = b$ não tem solução, emtão $Ax = b$ também não tem.

O método aqui descrito é conhecido como *eliminação gaussiana*, que consiste em triangularizar a matriz de coeficientes do sistema linear para facilitar sua resolução. Como envolve divisões e operações de somas de linhas, é esperado que esse método acumule erros numéricos com certa facilidade. Apesar desses erros numéricos acumulados serem inevitáveis, é possível minimzar seu acúmulo, como veremos na próxima subseção.

Exercício 50. *Escreva um pseudocódigo para sintetizar a eliminação gaussiana. Dado um sistema $Ax = b$ com m equações e n variáveis, quantas operações são realizadas pela eliminação gaussiana? Qual sua complexidade assintótica? Não se esqueça de que as operações sobre linhas de A devem ser repetidas também sobre b.*

5.3 Pivoteamentos parcial e total

Quando fazemos uma eliminação gaussiana, estamos realizando diversas operações de divisão (lembra do $s_i = -\frac{c_i}{a}$?). Apesar de, por uma questão de simplicidade, termos feito permutações de linha apenas quando encontramos $a = 0$, podemos nos beneficiar de permutações de linha com propósitos além de evitar divisões por zero.

Sabemos que os computadores não se dão bem com todos os números. Representar π em um computador é impossível, pois π tem infinitas casas decimais significativas, que ocorrem sem padrão aparente, e um computador tem memória finita. Quando utilizamos a representação de números em ponto flutuante, ficamos ainda mais limitados, pois um ponto flutuante tem um limite de casas

decimais que pode representar. Com essa representação, não apenas os irracionais são incompatíveis, mas também as dízimas, que apesar de apresentarem um padrão em suas casas decimais, podem ter apenas algumas delas representadas. Erros numéricos dessa natureza são ditos erros de representação.

Exercício 51. *Abra um interpretador Python e calcule* `1.4 + 0.2`*. Porque esse erro ocorreu? Nenhum número envolvido é irracional ou dízima (pelo menos na base decimal). Mas que base numérica os computadores usam?*

Quando tratamos de uma divisão $\frac{a}{b}, a, b \in \mathbb{R}$, devemos ficar particularmente preocupados com erros numéricos em b. Tomemos como exemplo $\frac{1}{a}$, de tal forma que a é representado, num certo computador, como $\tilde{a} = a + \epsilon, \epsilon \in \mathbb{R}$, onde ϵ é um pequeno erro numérico. Assim, temos

$$\frac{1}{a} \approx \frac{1}{\tilde{a}} \implies \frac{1}{a} \approx \frac{1}{a+\epsilon}$$

Agora vamos medir a em termos de ϵ. Existe $c \in \mathbb{R}$ tal que $a = c\epsilon$. Dessa forma, podemos escrever

$$\frac{1}{c\epsilon} \approx \frac{1}{(c+1)\epsilon} \implies \frac{1}{\epsilon}\frac{1}{c} \approx \frac{1}{\epsilon}\frac{1}{c+1} \implies \frac{1}{c} \approx \frac{1}{c+1}$$

Perceba que c nos dá uma ideia do quão grande é a em relação a ϵ. Perceba também que, quanto maior c, melhor é a aproximação descrita acima, e também que ela se deteriora para valores de c que fazem a muito próximo de ϵ. Quando $c = 1$, a aproximação nos diz que $1 \approx \frac{1}{2}$. Quando $c = 1000$, temos $\frac{1}{1000} \approx \frac{1}{1001}$.

Com isso, podemos perceber que divisões por números muito pequenos podem ser problemáticas, uma vez que o erro numérico do denominador, seja de representação ou acumulado, pode ser comparável ao próprio número. Nesse caso, a operação de divisão pode "ampliar"um erro tão pequeno quanto um erro de representação.

Sempre que pudermos escolher, devemos preferir as operações que minimizam a "ampliação"de erros numéricos. Esta subseção existe porque, durante uma eliminação gaussiana, podemos escolher quais divisões vamos fazer!

Considere o sistema $Ax = b$, com $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$. Tome $\bar{i} \in [m]$ tal que

$$|A(\bar{i}, 1)| = \max_{i\in[m]} |A(i, 1)|$$

Dessa forma, percebemos que a primeira posição da $\bar{i}$-ésima linha de A é maior ou igual à primeira posição de qualquer outra linha, em termos absolutos. Se formos fazer uma divisão com o denominador sendo algum elemento da primeira coluna de A, é preferível que esse elemento seja o $\bar{i}$-ésimo. Sabemos que

$$Ax = b \iff R_{1\bar{i}}Ax = R_{1\bar{i}}b$$

e, com isso, fazemos a eliminação gaussiana no sistema equivlaente. Isso fará com que, nas operações de linha para zerar todas (menos a primeira) posições da primeira coluna, as divisões realizadas sejam as menos danosas possíveis.

Como a eliminação gaussiana se descreve de forma indutiva, essa técnica pode ser aplicada também antes da eliminação gaussiana do sistema linear da hipótese. Tal técnica é conhecida como *pivoteamento parcial*.

Em termos assintóticos, não há aumento da complexidade quando a eliminação gaussiana é precedida pelo pivoteamento parcial. No entanto, o número de operações realizadas aumenta, e as operações realizadas a mais são de comparação, que podem ser tão custosas quanto as aritméticas.

Uma outra técnica, baseada no mesmo princípio, consiste em selecionar o melhor divisor não apenas na primeira coluna, mas na matriz inteira! Dado $Ax = b$, com $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$, seja $(\bar{i}, \bar{j}) \in [m] \times [n]$ uma posição máxima da matriz A em valores absolutos. Já sabemos como permutar linhas, mas para fazer o elemento da posição $(\bar{i}, \bar{j})$ ocupar a posição $(1, 1)$, teremos também de permutar colunas.

Exercício 52. *Tome a matriz*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Calcule $R_{13}A$ *e* AR_{12}

Como o exercício acima ilustra, podemos permutar as colunas i e j de uma matriz se multiplicarmos R_{ij} pela direita. Apesar de termos definido R_{ij} para representar a operação de permutação das linhas i e j, existe uma propriedade útil sobre ela que nos permite utilizá-la para algo mais.

Exercício 53. *Prove que* R_{ij} *é simétrica, isto é,* $R_{ij} = R_{ij}^T$. *Dica: use a definição de* R_{ij} *e propriedades de transposição de matrizes.*

Com isso, prosseguimos com

$$(AR_{ij})^T = R_{ij}^T A^T = R_{ij} A^T$$

para argumentar que $(AR_{ij})^T$ é a matriz A^T com suas linhas i e j permutadas. Mas as linhas de A^T são justamente as colunas de A. Logo, AR_{ij} é a matriz A com suas colunas i e j permutadas.

Dessa forma, temos

$$Ax = b \iff R_{1\bar{i}} A R_{1\bar{j}} x = R_{1\bar{i}} b$$

Note que a permutação de colunas deve ser feita apenas no lado esquerdo, já que b é apenas um vetor-coluna.

A técnica de permutar linhas e colunas de forma que o divisor a ser utilizado na eliminação gaussiana seja o maior possível, em módulo, é chamada de *pivoteamento total*. Assim como no pivoteamento parcial, não há aumento de complexidade assintótica quando a eliminação gaussiana é precedida pelo pivoteamento total. No entanto, o pivoteamento parcial é preferível, na prática, já que envolve menos operações de comparação que o pivoteamento total.

Exercício 54. *Tome o sistema*

$$\begin{bmatrix} \tilde{0} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

onde $\tilde{0}$ *deveria ser* 0*, mas devido a erros numéricos, acabou sendo algo como* $\tilde{0} = 0,01$*. Resolva o sistema linear por eliminação gaussiana, e depois por eliminação gaussiana com pivoteamento parcial.*

Exercício 55. *Resolva o seguinte sistema de equações usando eliminação gaussiana. Em seguida, resolva-o usando pivoteamento parcial*

$$x_1 + 3x_2 + 5x_3 = 14 \tag{5.1}$$
$$2x_1 - x_2 - 3x_3 = 3 \tag{5.2}$$
$$4x_1 + 5x_2 - x_3 = 7 \tag{5.3}$$

5.4 Decomposição LU

Tome $A \in \mathbb{R}^{m \times n}$. Queremos determinar se é possível representar A como $A = LU$, onde $L \in \mathbb{R}^{m \times m}$ e $U \in \mathbb{R}^{m \times n}$ são matrizes triangulares, inferior e superior, respectivamente. A representação de A nessa forma é conhecida como sua *decomposição* LU. Mas qual vantagem essa decomposição nos traz?

Tome m sistemas lineares, por exemplo, $Ax = b_1, Ax = b_2, \ldots, Ax = b_m$, com $b_i \in \mathbb{R}^m, i \in [m]$. Se conhecemos matrizes L e U como descrito acima, podemos utilizar $A = LU$ para resolver os m sistemas sem fazer uma eliminação gaussiana. Dado que

$$Ax = b_1 \iff LUx = b_1$$

consideramos o sistema $Ly = b_1$, e determinamos uma solução $\overline{y}$ com custo $O(m^2)$, dado que L já é triangular. Depois, consideramos o sistema $Ux = \overline{y}$, e determinamos uma solução $\overline{x}$ com custo $O(mn)$, já que U também é triangular. No fim das contas, determinar $\overline{x}$ custa $O(m(m+n))$. Resolver todos os sistemas custa, portanto, $O(m^2(m+n))$. Como uma eliminação gaussiana custa $O(m^2 n)$ para resolver um desses sistemas, resolver todos eles por esse método custa $O(m^3 n)$. Dessa forma, se conhecemos $A = LU$, é computacionalmente mais barato tirar proveito dessa decomposição para resolver os m sistemas do que resolver cada um deles via eliminação gaussiana. Mas quanto custa decompor A em LU? Aliás, como podemos decompor A em LU?

Podemos descobrir um procedimento para decompor A em LU se pensarmos de forma indutiva. Primeiro, consideramos que $A \in \mathbb{R}^{1 \times n}$, ou seja, $A = a^T$ e pode ser decomposta como

$$A = \begin{bmatrix} a_1 & r^T \end{bmatrix}$$

e L deve estar em $\mathbb{R}^{1 \times 1} = \mathbb{R}$, assim como U deve estar em $\mathbb{R}^{1 \times n}$. Podemos decompor L e U como

$$L = \begin{bmatrix} l \end{bmatrix}, U = \begin{bmatrix} u_1 & s^T \end{bmatrix}$$

Para determinar L e U, precisamos resolver

$$LU = A \implies \begin{bmatrix} l \end{bmatrix} \begin{bmatrix} u_1 & s^T \end{bmatrix} = \begin{bmatrix} a_1 & r^T \end{bmatrix}$$

e temos o sistema

$$lu_1 = a_1$$
$$ls^T = r^T$$

Ora, queremos encontrar L triangular inferior e U triangular superior tal que $A = LU$. Podemos muito bem fazer a diagonal de L unitária, sem prejuízo para a sua caracterização. Em outras palavras, não há problema em admitir $l = 1$. Dessa forma, é fácil decompor A em

$$\begin{bmatrix} a_1 & r^T \end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} \begin{bmatrix} a_1 & r^T \end{bmatrix}$$

Exercício 56. *Descreva a decomposição $A = LU$ quando $A \in \mathbb{R}^{m\times 1}$.*

Agora tome $A \in \mathbb{R}^{m\times n}$. Queremos encontrar $L \in \mathbb{R}^{m\times m}$ e $U \in \mathbb{R}^{m\times n}$, triangulares inferior e superior, respectivamente, tal que $A = LU$. Podemos decompor essas matrizes como

$$A = \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix}, L = \begin{bmatrix} l & 0^T \\ d & L' \end{bmatrix}, U = \begin{bmatrix} u & s^T \\ 0 & U' \end{bmatrix}$$

onde $a, l, u \in \mathbb{R}, r, s \in \mathbb{R}^{n-1}, c, d \in \mathbb{R}^{m-1}, L' \in \mathbb{R}^{(m-1)\times(m-1)}$ e $A', U' \in \mathbb{R}^{(m-1)\times(n-1)}$. Além disso, L' e U' também são triangulares inferior e superior, respectivamente.

Dessa forma, para termos $A = LU$, devemos satisfazer

$$A = LU \implies \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix} = \begin{bmatrix} l & 0^T \\ d & L' \end{bmatrix} \begin{bmatrix} u & s^T \\ 0 & U' \end{bmatrix}$$

$$\begin{bmatrix} lu & ls^T \\ ud & ds^T + L'U' \end{bmatrix} = \begin{bmatrix} a & r^T \\ c & A' \end{bmatrix}$$

Organizando como um sistema de equações, temos

$$lu = a$$
$$ls^T = r^T$$
$$ud = c$$
$$ds^T + L'U' = A'$$

Fazendo a diagonal de L unitária, temos que $l = 1$. Isso implica que $u = a, s = r$ e $d = \frac{1}{a}c$. Perceba que $\frac{1}{a}c$ está bem definido apenas quando $a \neq 0$. Vamos fazer essa suposição agora e tratar o caso quando $a = 0$ depois. Nos resta determinar L' e U'. Temos que

$$L'U' = A' - \frac{1}{a}cr^T$$

Perceba que $A' - \frac{1}{a}cr^T \in \mathbb{R}^{(m-1)\times(n-1)}$. Como, por hipótese, sabemos fazer a decomposição LU de $A' - \frac{1}{a}cr^T$ e L' e U' são triangulares inferior e superior, respectivamente, então sabemos determinar L' e U'. Então, temos que a decomposição LU de A é dada por

$$\begin{bmatrix} a & r^T \\ c & A' \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ \frac{1}{a}c & L' \end{bmatrix} \begin{bmatrix} a & r^T \\ 0 & U' \end{bmatrix}$$

Quando $a = 0$, temos de considerar c. Se $c \neq 0$, existe algum $c_i \neq 0$, para algum $2 \leq i \leq m$. Nesse caso, consideramos obter a decomposição LU da matriz $R_{1i}A$. Dessa forma, podemos fazer a decomposição $R_{1i}A = LU$, e sempre que uma matriz de permutação for utilizada, devemos multiplicá-la à esquerda (obteremos assim uma decomposição $PA = LU$, onde P é o produtório das matrizes de permutação utilizadas). Se $c = 0$, então não adianta fazer permutações de linha e utilizar o método descrito. Quando fazemos $l = 1$, podemos evitar definir d com divisões por a se fizermos $u = 0$. Essas valorações implicam que $s^T = r^T$, e podemos fazer $d = 0$, satisfazendo o sistema de equações. Aplicando essas valorações, obtemos

$$\begin{bmatrix} 0 & r^T \\ 0 & A' \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ 0 & L' \end{bmatrix} \begin{bmatrix} 0 & r^T \\ 0 & U' \end{bmatrix}$$

Com isso, vemos que $A' = L'U'$. Por hipótese, conhecemos L' e U' e temos a decomposição LU de A quando $a = 0$ e $c = 0$.

Assim, temos desenvolvido a decomposição LU. A seguir, vamos utilizá-la para calcular determinantes e inversas.

Exercício 57. *A decomposição LDU consiste em decompor uma matriz A como $A = LDU$, onde L e U são triangulares inferior e superior, respectivamente, ambas com diagonal unitária, e a matriz D é diagonal, ou seja, triangular inferior e superior. Descreva a decomposição LDU. Dica: que tal decompor $A = LU$ e em seguida $U = DU'$, onde D e U' tem as propriedades desejadas?*

Exercício 58. *A decomposição LU é muito útil para o cálculo de determinantes. Sabendo que* $\det(AB) = \det(A)\det(B)$ *e que o determinante de uma matriz triangular é dado pelo produtório dos elementos de sua diagonal, descreva como utilizar a decomposição $A = LU$ para calcular* $\det(A)$*. É muito comum que precisemos fazer permutações de linha durante a decomposição LU. Dado que* $\det(R_{ij}) = -1$*, descreva como utilizar a decomposição $PA = LU$ para calcular* $\det(A)$*, onde P é o produtório de matrizes de permutação de linhas.*

Tome uma matriz $A \in \mathbb{R}^{n\times n}$, já que apenas matrizes quadradas têm inversas. Se temos a decomposição $A = LU$, podemos utilizá-la para calcular a inversa de A. Ora, sabemos que se A^{-1} existe, ela é tal que $AA^{-1} = I$. Dessa forma

$$LUA^{-1} = I$$

Se A é invertível e $A = LU$, então L e U também são invertíveis, e portanto L^{-1} e U^{-1} existem. Repare que a contraposição nos diz que se L ou U não é

invertível, então A também não é. Fazemos

$$L^{-1}LUA^{-1} = L^{-1}I$$
$$UA^{-1} = L^{-1}$$
$$U^{-1}UA^{-1} = U^{-1}L^{-1}$$
$$A^{-1} = U^{-1}L^{-1}$$

Com isso, vemos que se A^{-1} existe, então $A^{-1} = U^{-1}L^{-1}$. Se L^{-1} ou U^{-1} não existem, então A^{-1} também não pode existir. Isso nos leva a considerar a decomposição LU para computarmos a inversa de A, e para tanto devemos descrever como determinar L^{-1} e U^{-1}.

Primeiro, consideramos L, que é quadrada. Caso $L \in \mathbb{R}^{1\times 1}$, temos que $L = [l]$ e L é invertível apenas se $l \neq 0$.

Agora tomamos $L \in \mathbb{R}^{n\times n}$. Sabemos que L pode ser decomposta como

$$L = \begin{bmatrix} 1 & 0^T \\ c & L' \end{bmatrix}$$

Queremos encontrar L^{-1} tal que $L^{-1}L = I$. Podemos decompor L^{-1} como

$$L^{-1} = \begin{bmatrix} x & y^T \\ z & M \end{bmatrix}$$

Com isso, temos

$$L^{-1}L = I \implies \begin{bmatrix} x & y^T \\ z & M \end{bmatrix}\begin{bmatrix} 1 & 0^T \\ c & L' \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ 0 & I' \end{bmatrix}$$

Resolvemos o lado esquerdo e obtemos

$$\begin{bmatrix} x + y^T c & y^T L' \\ z + Mc & ML' \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ 0 & I' \end{bmatrix}$$

Escrevemos em forma de sistema de equações

$$x + y^T c = 1$$
$$y^T L' = 0^T$$
$$z + Mc = 0$$
$$ML' = I'$$

Perceba que o sistema tem solução apenas se L' é invertível, com $M = L'^{-1}$. Isso nos permite determinar $z = -L'^{-1}c$. Podemos fazer $y = 0$, o que implica $x = 1$. Fazendo as substituições, obtemos

$$\begin{bmatrix} 1 & 0^T \\ -L'^{-1}c & L'^{-1} \end{bmatrix}\begin{bmatrix} 1 & 0^T \\ c & L' \end{bmatrix} = \begin{bmatrix} 1 & 0^T \\ 0 & I' \end{bmatrix}$$

Temos, portanto, que L é invertível apenas se L' é invertível, e que

$$L^{-1} = \begin{bmatrix} 1 & 0^T \\ -L'^{-1}c & L'^{-1} \end{bmatrix}$$

Exercício 59. *Explique como determinar se U é invertível e, caso seja, descreva sua inversa.*

Exercício 60. *Calcule a fatoração LU da seguinte matriz. Calcule-a também usando pivoteamento parcial.*

$$A = \begin{bmatrix} 5 & -1 & 9 & -2 \\ 0 & 1 & 1 & 0 \\ 8 & 1 & -3 & 1 \\ 3 & 0 & 1 & 0 \end{bmatrix}$$

Exercício 61. *Ainda sobre a matriz do exercício 60, calcule seu determinante a partir de suas decomposições LU, com e sem pivoteamento. Há alguma diferença perceptível?*

Exercício 62. *Calcule a inversa da matriz do exercício 60, da forma que desejar.*

6 Programação Linear

Começamos agora nosso estudo de Programação Linear. Vamos aprender a, dado um sistema de equações lineares, encontrar as melhores soluções de seu conjunto-solução. Mas como determinar se uma solução é melhor que outra?

Quando tomamos o conjunto-solução de um sistema $Ax = b$, todos os seus elementos nos parecem iguais: todos eles satisfazem $Ax = b$. No entanto, podemos utilizar um certo vetor w para ponderar essas soluções. Tome, por exemplo, duas soluções x_1 e x_2 com $Ax_1 = b$ $Ax_2 = b$. Podemos calcular $w^T x_1$ e $w^T x_2$ e termos, por exemplo, $w^T x_1 \leq w^T x_2$. Se estamos procurando a solução "mais barata", preferimos x_1 a x_2. Se estamos em busca da solução "mais rentável", preferimos x_2 a x_1.

Podemos ter um programa linear de minimização (em busca das soluções "mais baratas") representado na forma

$$\begin{aligned} \min \quad & w^T x && (6.1) \\ \text{s.t} \quad & Ax = b && (6.2) \end{aligned}$$

onde $w^T x$ é dita a função-objetivo e $Ax = b$ representa as restrições do programa.

Antes de aprender a resolver programas dessa forma, vamos ver como é possível representar as mais diversas situações com equações lineares. Vamos fazer alguns estudos de caso para constatar que a restrição de trabalhar apenas com equações lineares limita muito pouco a expressividade de um programa linear.

6.1 Modelagem matemática

Começamos por exibir modelos matemáticos de Programação Linear para problemas clássicos. Abordamos desde problemas encontrados por nutricionistas até a determinação das melhores rotas para o escoamento de cargas.

6.1.1 Problema da dieta

Uma dieta viável satisfaz, e pode exceder, requerimentos nutricionais básicos. Uma dieta ótima é uma dieta viável que tem o menor custo possível para o consumidor.

Considere a elaboração de uma dieta, dada a disponibilidade dos seguintes alimentos.

Id	Alimento	Energia	Proteína	Cálcio	Preço da porção	Limite diário
1	Aveia	110	4	2	3	4
2	frango	265	32	12	24	3
3	Ovo	160	13	54	13	2
4	Leite	160	8	285	9	8
5	Torta de cereja	420	4	22	20	2
6	Feijão	260	14	80	19	2

Segundo um nutricionista, uma dieta viável proporciona, diariamente, 2000 Kcal de energia, 55 g de proteína e 800 mg de cálcio. Por motivos de saúde, devemos respeitar os limites diários de cada alimento. Gostaríamos de encontrar uma dieta viável que tenha custo mínimo para o consumidor.

Usamos $x \in \mathbb{R}^6$ como variável de decisão. O valor de x_i indica quantas porções do alimento i vamos consumir em um dia, com $i \in [6]$.

Temos quantidades mínimas de energia, proteína, e cálcio para atender.

$$110x_1 + 265x_2 + 160x_3 + 160x_4 + 420x_5 + 260x_6 \geq 2000$$

$$4x_1 + 32x_2 + 13x_3 + 8x_4 + 4x_5 + 14x_6 \geq 55$$

$$2x_1 + 12x_2 + 54x_3 + 285x_4 + 22x_5 + 80x_6 \geq 800$$

Com isso, construímos o seguinte modelo

$$\begin{aligned}
\min \quad & 3x_1 + 24x_2 + 13x_3 + 9x_4 + 20x_5 + 19x_6 && (6.3)\\
\text{s.t} \quad & 110x_1 + 265x_2 + 160x_3 + 160x_4 + 420x_5 + 260x_6 \geq 2000 && (6.4)\\
& 4x_1 + 32x_2 + 13x_3 + 8x_4 + 4x_5 + 14x_6 \geq 55 && (6.5)\\
& 2x_1 + 12x_2 + 54x_3 + 285x_4 + 22x_5 + 80x_6 \geq 800 && (6.6)\\
& x_1 \leq 4 && (6.7)\\
& x_2 \leq 3 && (6.8)\\
& x_3 \leq 2 && (6.9)\\
& x_4 \leq 8 && (6.10)\\
& x_5 \leq 2 && (6.11)\\
& x_6 \leq 2 && (6.12)\\
& x \geq 0 && (6.13)
\end{aligned}$$

Note que todas as soluções para esse modelo satisfazem os critérios estabelecidos para uma dieta viável. Como estamos buscando, dentre as soluções viáveis para o modelo, aquelas que tem o menor custo, devemos minimizar a função objetivo.

6.1.2 Planejamento da força de trabalho

Em um restaurante, cada trabalhador trabakha cinco dias consecutivos e folga dois dias consecutivos. Sua semana de trabalho pode começar em qualquer dia da semana. O restaurante tem a seguinte demanda de trabalhadores por dia da semana.

	Seg	Ter	Qua	Qui	Sex	Sáb	Dom
Demanda	14	13	15	16	19	18	11

Gostaríamos de suprir a demanda de todos os dias da semana usando o menor número de trabalhadores possível. Tomamos os dias da semana, de segunda a domingo, como os dias de 1 a 7. Vamos utilizar a variável $x_i, i \in [7]$, para indicar quantos trabalhadores começam sua semana de trabalho no dia i.

O que precisamos notar, para a construção do modelo, é que o trabalhador que começa sua semana de trabalho no dia i vai trabalhar até o dia $i + 4$ ou até o dia $(i + 4) \mod 7$, caso $i + 4 > 7$. Com essa observação, construímos o seguinte modelo.

$$
\begin{aligned}
\min \quad & x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 && (6.14)\\
\text{s.t} \quad & x_1 + x_4 + x_5 + x_6 + x_7 \geq 14 && (6.15)\\
& x_1 + x_2 + x_5 + x_6 + x_7 \geq 13 && (6.16)\\
& x_1 + x_2 + x_3 + x_6 + x_7 \geq 15 && (6.17)\\
& x_1 + x_2 + x_3 + x_4 + x_7 \geq 16 && (6.18)\\
& x_1 + x_2 + x_3 + x_4 + x_5 \geq 19 && (6.19)\\
& x_2 + x_3 + x_4 + x_5 + x_6 \geq 18 && (6.20)\\
& x_3 + x_4 + x_5 + x_6 + x_7 \geq 11 && (6.21)\\
& x \in \mathbb{Z}_+^7 && (6.22)
\end{aligned}
$$

Diferente do modelo anterior, esse modelo tem suas variáveis inteiras (não faz sentido lidar com frações de trabalhadores). Quando isso ocorre, temos um modelo matemático de Programação Linear Inteira.

6.1.3 Planejamento de produção

Uma empresa tem 5000 de seus produtos em estoque. Para os próximos quatro meses, espera-se uma demanda de 7000, 15000, 10000 e 8000 de seus produtos, respectivamente. A empresa tem capacidade de produção mensal de 10000 produtos a um custo de 2000 por produto. A empresa pode produzir, mensalmente, 2500 produtos adicionais utilizando hora-extra, a um custo de 2200 cada. Produtos podem ser estocados, a um custo de 100 por mês. Como a empresa fará para atender à demanda dos quatro meses a um custo mínimo?

Vamos utilizar as seguintes variáveis: $t \in [4]$ é um índice que representa os próximos quatro meses; x_t representa a quantidade de produtos produzidos no mês t; y_t é o número de produtos produzidos no mês t, com o uso de hora-extra; por fim, e_t é o número de produtos estocados ao final do mês t.

Para satisfazer a demanda do primeiro mês, temos de ter

$$5000 + x_1 + y_1 - e_1 = 7000$$

Já para o segundo mês, temos de satisfazer

$$e_1 + x_2 + y_2 - e_2 = 15000$$

No terceiro mês, teríamos

$$e_2 + x_3 + y_3 - e_3 = 10000$$

Por fim, o quarto mês nos traz

$$e_3 + x_4 + y_4 = 8000$$

Dessa forma, podemos montar o seguinte modelo matemático para representar essa situação.

$$\begin{aligned}
\min \quad & 2000\sum_{t=1}^{4} x_t + 2200\sum_{t=1}^{4} y_t + 100\sum_{t=1}^{4} e_t && && (6.23)\\
\text{s.t} \quad & x_1 + y_1 - e_1 = 2000 && && (6.24)\\
& e_1 + x_2 + y_2 - e_2 = 15000 && && (6.25)\\
& e_2 + x_3 + y_3 - e_3 = 10000 && && (6.26)\\
& e_3 + x_4 + y_4 = 8000 && && (6.27)\\
& x_t \leq 10000, && \forall t \in [4] && (6.28)\\
& y_t \leq 2500, && \forall t \in [4] && (6.29)\\
& x, y, e \in \mathbb{Z}_+^4 && && (6.30)
\end{aligned}$$

6.1.4 Problema do transporte

Existem I portos, $P_1, P_2, \ldots P_I$, que possuem uma certa mercadoria, e existem J mercados, $M_1, M_2, \ldots, M_J$, para os quais essa mercadoria deve ser enviada. O porto P_i tem uma quantidadde s_i da mercadoria, $i \in [I]$, e o mercado M_j deve receber uma quantidade t_j da mercadoria, $j \in [J]$. Tome c_{ij} como o custo de transportar uma unidade da mercadoria do porto P_i para o mercado M_j. Nosso objetivo é atender à demanda dos mercados a um custo de transporte mínimo.

Definimos a variável x_{ij} como o número de unidades da mercadoria que deve ser transportado do porto P_i para o mercado M_j. Modelamos esse problema como segue.

$$\begin{aligned}
\min \quad & \sum_{i=1}^{I}\sum_{j=1}^{J} c_{ij}x_{ij} && && (6.31)\\
\text{s.t} \quad & \sum_{j=1}^{J} x_{ij} \leq s_i, && \forall i \in [I] && (6.32)\\
& \sum_{i=1}^{I} x_{ij} \geq t_j, && \forall j \in [J] && (6.33)\\
& x \in \mathbb{Z}_+^{IJ} && && (6.34)
\end{aligned}$$

6.1.5 Emparelhamento máximo

Tome um grafo $G = (V, E)$ (V é um conjunto de vértices, E é um conjunto de arestas e cada aresta é um conjunto de dois vértices). Dizemos que $E' \subseteq E$ é um emparelhamento em G quando as arestas de E' não têm extremos em comum. O problema do emparelhamento máximo consiste em determinar um emparelhamento de G com cardinalidade máxima.

Podemos modelar esse problema com as variáveis $x_e \in \mathbb{B}, \forall e \in E$, indicando se a aresta e faz parte de um emparelhamento máximo. Para simplificar a escrita do modelo, definimos $\delta(v) = \{e \in E : e \cap \{v\} \neq \emptyset\}$ como o conjunto de arestas que incidem em $v, \forall v \in V$. Segue o modelo.

$$\max \quad \sum_{e \in E} x_e \tag{6.35}$$

$$\text{s.t} \quad \sum_{e \in \delta(v)} x_e \leq 1, \qquad \forall v \in V \tag{6.36}$$

$$x \in \mathbb{B}^{|E|} \tag{6.37}$$

6.1.6 Cobertura mínima

Tome um grafo $G = (V, E)$. Dizemos que $V' \subseteq V$ é uma cobertura de G se $e \cap V' \neq \emptyset, \forall e \in E$, isto é, se toda aresta de G tem pelo menos um extremo em V'. O problema da cobertura mínima consiste em determinar uma cobertura de G com cardinalidade mínima.

Exercício 63. *Usando as variáveis $x_v \in \mathbb{B}, \forall v \in V$, que indicam se o vértice v faz parte de uma cobertura mínima, modele o problema da cobertura mínima.*

Exercício 64. *Tome um grafo $G = (V, E)$ onde emparelhamentos máximos têm cardinalidade a e coberturas mínimas têm cardinalidade b. Prove que $a \leq b$. Dê exemplo de um grafo para o qual $a < b$.*

Exercício 65. *Tome $a = \min(b, c)$, com $a, b, c \in \mathbb{R}$. Escreva um modelo matemático cujo custo de suas soluções ótimas seja a. Dica: tente fazer um modelo de maximização.*

Exercício 66. *Considere o modelo matemático*

$$\max \quad 3x_1 + 2x_2x_3 + x_4 \tag{6.38}$$

$$s.t \quad 4x_1x_4 + 6x_2 + x_3 \leq 8 \tag{6.39}$$

$$x \in \mathbb{B}^4 \tag{6.40}$$

que claramente não é linear. Escreva um modelo linear equivalente. Dica: tente substituir x_2x_3 por uma variável p e acrescente restrições lineares para garantir que $p = x_2x_3$.

Exercício 67. *Tome duas funções lineares, ambas de $\mathbb{R}^n$ em $\mathbb{R}$, $f(x) = c^T x$ e $g(x) = d^T x$, com $c, d \in \mathbb{R}^n$. Considere o modelo matemático*

$$\max \quad \min(f(x), g(x)) \tag{6.41}$$

$$s.t \quad Ax = b \tag{6.42}$$

$$x \in \mathbb{R}^n \tag{6.43}$$

Escreva um modelo linear equivalente. Dica: apenas a função-objetivo não é linear. Tente substituí-la por uma variável q e acrescente restrições para que $q \leq \min(f(x), g(x))$. Por que isso é o bastante para termos $q = \min(f(x), g(x))$?

6.1.7 Caminho de custo mínimo

Tome um digrafo $D = (V, A)$ (A é um conjunto de arcos, e um arco é um par ordenado de vértices) Tome também uma função $c : A \to \mathbb{Z}_+$ que associa custos não-negativos aos arcos de G. Dados vértices $s, t \in V$, gostaríamos de encontrar um caminho de s a t que tenha custo mínimo, de acordo com c.

Para modelar esse problema, precisamos apenas determinar quais arcos fazem parte de um caminho de custo mínimo de s a t. Utilizamos as variáveis $x_a \in \mathbb{B}, \forall a \in A$. Para facilitar a escrita do modelo, definimos $\delta^+(u) = \{a \in A | \exists v \in V : a = (u, v)\}$ como o conjunto dos arcos partindo de u e $\delta^-(v) = \{a \in A | \exists u \in V : a = (u, v)\}$ como o conjunto dos arcos chegando em v.

Para conceber o modelo, basta percebermos que precisamos selecionar exatamente um arco partindo de s e exatamente um arco chegando em t. Além disso, todo outro vértice v ou não faz parte do caminho ou selecionamos um arco chegando em v e um arco partindo de v. Escrevemos o modelo como segue.

$$\min \quad \sum_{a \in A} c_a x_a \tag{6.44}$$

$$\text{s.t} \quad \sum_{a \in \delta^+(s)} x_a = 1 \tag{6.45}$$

$$\sum_{a \in \delta^-(t)} x_a = 1 \tag{6.46}$$

$$\sum_{a \in \delta^-(v)} x_a = \sum_{a \in \delta^+(v)} x_a, \qquad \forall v \in V \setminus \{s, t\} \tag{6.47}$$

$$x \in \mathbb{B}^{|A|} \tag{6.48}$$

Exercício 68. *Perceba que a restrição* (6.47) *permite a seleção de arcos que formam um ciclo, para além dos arcos que formam um caminho de s a t. Isso nos mostra que nem toda solução viável para o modelo representa um caminho de s a t. Prove que, se $c_a > 0, \forall a \in A$, então as soluções ótimas do modelo representam caminhos de s a t.*

O modelo acima apresentado poderia ter o domínio $\mathbb{B}^{|A|}$ de suas variáveis substitído por $[0, 1]^{|A|}$ (chamamos isso de relaxação linear) sem prejuízo para a natureza de suas soluções ótimas. Em outras palavras, poderíamos trocar as variáveis binárias por variáveis contínuas no intervalo $[0, 1]$ com a garantia de que as soluções ótimas do modelo ainda seriam valorações binárias. Quando isso ocorre, temos que o modelo representa um problema fácil de ser resolvido, como é o caso do problema do caminho de custo mínimo.

6.1.8 Árvore geradora mínima

Tome um grafo $G = (V, E)$ e $c : E \to \mathbb{Z}_+$. Um grafo $G' = (V', E')$ é dito um subgrafo de G se $V' \subseteq V$ e $E' \subseteq E$. Um subgrafo G' de G é dito gerador se G' tem o mesmo conjuto de vértices de G. Uma árvore é um grafo conexo

e acíclico. O problema da árvore geradora mínima consiste em encontrar um subgrafo gerador de G que seja uma árvore e que tenha custo mínimo, de acordo com c.

Devido às várias caracterizações de árvore, existem diversas formas de modelar esse problema. Como é sabido que uma árvore com n vértices tem $n-1$ arestas, uma árvore geradora de G tem $|V|-1$ arestas. Essa propriedade, junto com a conectividade (ausência de ciclos), é suficiente para implicar a ausência de ciclos (conectividade), e formar uma caracterização de árvore geradora.

Exercício 69. *Prove que as seguintes sentenças são equivalentes.*

- *T é uma árvore com n vértices.*
- *T é conexo e tem $n-1$ arestas.*
- *T é acíclico e tem $n-1$ arestas.*

Assim, uma das várias formas de modelar esse problema seria buscar por subgrafos acíclicos com $|V|-1$ arestas. Podemos usar as variáveis $x_e \in \mathbb{B}, \forall e \in E$, para indicar se selecionamos a aresta e para nossa solução. É fácil selecionar $|V|-1$ arestas com a restrição

$$\sum_{e \in E} x_e = |V| - 1$$

Mas como fazemos para garantir a ausência de ciclos? Perceba que um ciclo com n vértices tem n arestas. Como um grafo acíclico não tem ciclos, nenhum subconjunto S de seu conjunto de vértices pode formar um ciclo. Dessa forma, o número de arestas com ambos os extremos em S não pode exceder $|S|-1$. Definimos $E(S) = \{e \in E : |e \cap S| = 2\}$ justamente como o conjutno de arestas com ambos os extremos em S. Segue o modelo.

$$\min \quad \sum_{e \in E} c_e x_e \tag{6.49}$$

$$\text{s.t} \quad \sum_{e \in E} x_e = |V| - 1 \tag{6.50}$$

$$\sum_{e \in E(S)} x_e \leq |S| - 1, \qquad \forall S \subseteq V \tag{6.51}$$

$$x \in \mathbb{B}^{|E|} \tag{6.52}$$

Note que esse modelo apresenta um número exponencial de restrições em relação a $|V|$, o que desencoraja sua resolução direta. Além disso, esse modelo também tem a propriedade de ter as mesmas soluções ótimas de sua relaxação linear.

Exercício 70. *Escreva um modelo linear para o problema da árvore geradora mínima baseado na caracterização de que T é uma árvore geradora de G se, e somente se, T tem $|V|-1$ arestas e é conexo. Talvez a notação $\delta(S) = \{e \in E : |e \cap S| = 1\}$ seja útil.*

6.1.9 Conjunto dominante mínimo

Tome um grafo $G = (V, E)$. Um subconjunto $V' \subseteq V$ de vértices é dito dominante se, para todo $v \in V$, $v \in V'$ ou $N(v) \cap V' \neq \emptyset$, onde $N(v) = \{u \in V : \{u, v\} \in E\}$. Gostaríamos de encontrar um conjunto dominante de G cuja cardinalidade seja mínima.

Para modelar esse problema, vamos utilizar as variáveis $x_v \in \mathbb{B}, \forall v \in V$, que indicam quais vértices fazem parte do nosso conjunto dominante mínimo. O modelo segue a definição de conjunto dominante.

$$\min \quad \sum_{v \in V} x_v \tag{6.53}$$

$$\text{s.t} \quad x_v + \sum_{u \in N(v)} x_u \geq 1, \qquad \forall v \in V \tag{6.54}$$

$$x \in \mathbb{B}^{|V|} \tag{6.55}$$

Note que poderíamos facilmente ponderar os vértices de G com $c : V \to \mathbb{Z}_+$. Para isso, bastariar utilizar a função-objetivo $\sum_{v \in V} c_v x_v$.

6.1.10 Árvore dominante mínima

Tome um grafo $G = (V, E)$ e uma ponderação $w : E \to \mathbb{Z}_+$ sobre as arestas de G. Dizemos que um subgrafo $G' = (V', E')$ de G é uma árvore dominante de G se G' é uma árvore e V' é um conjunto dominante de G. Gostaríamos de encontrar uma árvore dominante de G com peso mínimo, de acordo com w.

Vamos utilizar dois conjuntos de variáveis: $x_e \in \mathbb{B}, \forall e \in E$, indicam quais arestas fazem parte da nossa árvore dominante; $y_v \in \mathbb{B}, \forall v \in V$, indicam quais vértices fazem parte da nossa árvore dominante. O modelo consiste apenas em mesclar as restrições de modelos já estudados, fazendo alguma adaptação.

$$\min \quad \sum_{e \in e} w_e x_e \tag{6.56}$$

$$\text{s.t} \quad x_{uv} \leq y_u, \qquad \forall \{u, v\} \in E \tag{6.57}$$

$$x_{uv} \leq y_v, \qquad \forall \{u, v\} \in E \tag{6.58}$$

$$\sum_{e \in E} x_e = \sum_{v \in V} y_v - 1 \tag{6.59}$$

$$\sum_{e \in E(S)} x_e \leq \sum_{v \in S} y_v - 1, \qquad \forall S \subseteq V \tag{6.60}$$

$$y_v + \sum_{u \in N(v)} y_u \geq 1, \qquad \forall v \in V \tag{6.61}$$

$$x \in \mathbb{B}^{|E|}, y \in \mathbb{B}^{|V|} \tag{6.62}$$

6.1.11 Árvore de Steiner

Tome um grafo $G = (V, E)$ e uma ponderação $w : E \to \mathbb{Z}_+$ sobre as arestas de G. Além disso, tome um certo subconjunto $U \subseteq V$. O problema da Árvore de Steiner consiste em encontrar, dentre os subgrafos $G' = (V', E')$ de G com $U \subseteq V'$ que são árvores, um que tenha peso mínimo.

Novamente, utilizamos dois conjuntos de variáveis: $x_e \in \mathbb{B}, \forall e \in E$, indicam quais arestas fazem parte da nossa árvore de Steiner; $y_v \in \mathbb{B}, \forall v \in V$, indicam quais vértices fazem parte da nossa árvore de Steiner.

$$\begin{aligned}
\min \quad & \sum_{e \in E} w_e x_e && && (6.63)\\
\text{s.t} \quad & x_{uv} \leq y_u, && \forall \{u, v\} \in E && (6.64)\\
& x_{uv} \leq y_v, && \forall \{u, v\} \in E && (6.65)\\
& \sum_{e \in E} x_e = \sum_{v \in V} y_v - 1 && && (6.66)\\
& \sum_{e \in E(S)} x_e \leq \sum_{v \in S} y_v - 1, && \forall S \subseteq V && (6.67)\\
& y_u = 1, && \forall u \in U && (6.68)\\
& x \in \mathbb{B}^{|E|}, y \in \mathbb{B}^{|V|} && && (6.69)
\end{aligned}$$

É interessante observar como o problema da Árvore de Steiner generaliza dois problemas simples: o problema do caminho mínimo e o problema da árvore geeradora mínima. Se $U = \{s, t\}$, uma árvore de Steiner ótima tem de ser um dos caminhos de custo mínimo entre s e t (perceba que nós modelamos a versão direcionada desse problema). Se $U = V$, então uma árvore de Steiner ótima tem de ser uma árvore geradora mínima de G. Mesmo assim, o problema da àrvore de Steiner é um problema difícil.

Exercício 71. *Tome $n \in \mathbb{N}$ e $\overline{x} \in \mathbb{B}^n$, além de $w : \mathbb{B}^n \setminus \{\overline{x}\} \to \mathbb{Z}_+$. Considere o modelo*

$$\begin{aligned}
\min \quad & w^T x && (6.70)\\
s.t \quad & x \in \mathbb{B}^n \setminus \{\overline{x}\} && (6.71)
\end{aligned}$$

Escreva um modelo linear equivalente.

Exercício 72. *Tome os intervalos $I_i = [a_i, b_i] \subset \mathbb{R}, i \in [n], n \in \mathbb{N}$, com $I_i \cap I_j = \emptyset, \forall i, j \in [n], i \neq j$. Tome $c \in \mathbb{Z}_+$ e considere o modelo*

$$\begin{aligned}
\min \quad & cx && (6.72)\\
s.t \quad & x \in I_1 \cup I_2 \cup \cdots \cup I_n && (6.73)\\
& x \in \mathbb{R} && (6.74)
\end{aligned}$$

Escreva um modelo linear equivalente.

6.1.12 Coloração de vértices

Tome um grafo $G = (V, E)$. Uma coloração dos vértices de G é uma atribuição $c : V \to [n]$, para algum $n \in \mathbb{N}$. Uma coloração $c : V \to [n]$ é dita própria quando $c(u) \neq c(v)$, para toda aresta $\{u, v\} \in E$. O problema da coloração de vértices consiste em determinar o menor $n \in \mathbb{N}$ tal que existe uma coloração própria $c : V \to [n]$ dos vértices de G.

Vamos modelar esse problema trocando a ideia de "pintar"um vértice por escolher um único representante para cada vértice. Um vértice pode ser um representante (nesse caso, ele representa a si mesmo) e assim pode representar outros vértices. Para fazer a analogia com o conceito de coloração própria, extremos de uma aresta têm de ter representantes distintos. Sob essa restrição, queremos escolher um representante para cada vértice, de forma a utilizar o menor número de representantes possível.

Para construir o modelo, utilizamos as variáveis $x_{uv} \in \mathbb{B}, \forall u, v \in V$, que indicam se o vértice u representa o vértice v. Caso $x_{vv} = 1$, temos que v representa a si mesmo e portanto é um representante, podendo assim representar outros vértices. Segue o modelo.

$$\min \quad \sum_{v \in V} x_{vv} \tag{6.75}$$

$$\text{s.t} \quad x_{uv} \leq x_{uu}, \qquad \forall u, v \in V, u \neq v \tag{6.76}$$

$$\sum_{u \in V} x_{uv} = 1, \qquad \forall v \in V \tag{6.77}$$

$$x_{ua} + x_{ub} \leq 1, \qquad \forall \{a, b\} \in E, \forall u \in V \tag{6.78}$$

$$x \in \mathbb{B}^{|V|^2} \tag{6.79}$$

Exercício 73. *Dado um grafo completo $G = (V, E = \binom{V}{2})$ e uma ponderação $w : E \to \mathbb{Z}_+$, o Problema do Caixeiro Viajante consiste em encontrar um ciclo hamiltoniano de G que tenha peso mínimo. Modele esse problema. Dica: observe o modelo para Árvore Geradora Mínima.*

Exercício 74. *No Problema do Caixeiro Viajante, o caixeiro quer passar por todas as cidades, voltando para a cidade inicial, percorrendo a menor distância. Agora, vamos ver o Problema do Caixeiro Medroso, onde o caixeiro quer passar por todas as cidades correndo o menor risco de ser assaltado. Dados um grafo $G = (V, E = \binom{V}{2})$ e $p : E \to [0, 1]$ uma função que determina a chance do caixeiro ser assaltado em cada deslocamento, escreva um modelo que minimiza a chance do caixeiro ser assaltado. Dica: tente transformar p antes de modelar, e tente imaginar esse problema como um de maximização.*

6.2 Forma padrão de um programa linear

Antes de aprender como resolver um programa linear, é interessante que possamos desenvolver um modo único de representar todos os nossos modelos ma-

temáticos. Aqui, vamos mostrar que é possível escrever todos os nossos modelos lineares na forma

$$\begin{aligned} \max \quad & c^T x && (6.80)\\ \text{s.t} \quad & Ax = b && (6.81)\\ & x \geq 0 && (6.82) \end{aligned}$$

com $b \geq 0$.

Primeiro, vamos lidar com a função-objetivo. Se temos uma função de maximização, nada temos a fazer. Caso tenhamos uma função de minimização, fazemos uma transformação bem intuitiva: transformamos $\min c^T x$ em $\max(-c)^T x$. É notável que as valorações $\overline{x}$ que fazem $c^T x$ assumir seu valor mínimo são as mesmas que fazem $(-c)^T x$ assumir seu valor máximo.

Exercício 75. *Prove que* $\min c^T x = c^T \overline{x}$ *se, e somente se,* $\max(-c)^T x = (-c)^T \overline{x}$. *Dica: por contradição (na ida, assuma que* $\max(-c)^T x > (-c)^T \overline{x}$, *implicando que existe* $\tilde{x}$ *tal que...)*

Agora, vamos lidar com as variáveis. Dado que nosso programa tem variáveis $x \in \mathbb{R}^n$, considere a variável $x_i, i \in [n]$: se $x_i \geq 0$, nada temos a fazer; se $x_i \leq 0$, devemos criar uma variável $y_i \geq 0$, e substituir x_i por $-y_i$ em todo o modelo; se $x_i \in \mathbb{R}$, isto é, x_i é uma variável livre, podemos criar duas variáveis $y_i^+, y_i^- \geq 0$ e fazer a substituição $x_i = y_i^+ - y_i^-$ em todo o modelo (repare que y_i^+ representa os valores positivos que x_i pode assumir, enquanto y_i^- representa os valores negativos).

Vamos agora tratar das restrições do modelo. Tome a i-ésima restrição, $i \in [m]$: para uma restrição na forma $a_i^T x = b_i$, nada temos a fazer; para uma restrição na forma $a_i^T x \leq b_i$, criamos a variável $s_i \geq 0$ e reescrevemos a restrição como $a_i^T x + s_i = b_i$ (note como s_i representa a folga da restrição); para uma restrição na forma $a_i^T x \geq b_i$, criamos a variável $s_i \geq 0$ e reescrevemos a restrição como $a_i^T x - s_i = b_i$ (note como s_i representa o excesso da restrição).

Agora que temos todas as restrições como igualdades, devemos fazer com que $b \geq 0$. Vamos tomar nossa i-ésima restrição de igualdade $a_i^T x = b_i$: se $b_i \geq 0$, nada temos a fazer; se $b_i < 0$, reescrevemos a restrição como $(-a)_i^T x = -b_i$.

Com isso, terminamos nossa transformação. Obtemos nosso modelo inicial reescrito na forma

$$\begin{aligned} \max \quad & c^T x && (6.83)\\ \text{s.t} \quad & Ax = b && (6.84)\\ & x \geq 0 && (6.85) \end{aligned}$$

com $b \geq 0$.

6.3 Resolvendo um programa linear

Agora que temos um programa linear escrito na forma padrão

$$\begin{aligned} \max \quad & c^T x && (6.86)\\ \text{s.t} \quad & Ax = b && (6.87)\\ & x \geq 0 && (6.88) \end{aligned}$$

com $A \in \mathbb{R}^{m\times n}, c \in \mathbb{R}^n, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$, sendo $b \geq 0$, vamos aprender a resolvê-lo. Mas antes de prosseguir, precisamos de alguns novos conceitos.

6.3.1 Definições

Um poliedro P é um conjunto de pontos que pode ser caracterizado como $P = \{x \in \mathbb{R}^n : Ax \leq b\}$, para alguma matriz $A \in \mathbb{R}^{m\times n}$ e um vetor $b \in \mathbb{R}^m$.

Exercício 76. *Prove que o conjunto-solução de $Ax = b$ é um poliedro. Dica: $Ax = b \iff Ax \leq b$ e $Ax \geq b$.*

Exercício 77. *Tome $P = \{x \in \mathbb{R}^n : Ax = b\}$ um poliedro. Um ponto $\overline{x}$ é dito interior a P se $\overline{x} \in P$ e, para todo $v \in \mathbb{R}^n$, existe $\epsilon > 0$ tal que $\overline{x} + \epsilon v \in P$. Prove que, dado $c \neq 0$, um ponto interior a P não pode ser solução ótima de*

$$\begin{aligned} \max \quad & c^T x && (6.89)\\ s.t \quad & Ax = b && (6.90)\\ & x \geq 0 && (6.91) \end{aligned}$$

Sabemos, pelo exercício anterior, que pontos interiores a P não nos interessam. Estamos de olho nos demais pontos de P. Como podemos caracterizar um ponto de P que não é interior? Ora, um ponto é interior a P quando, a partir dele, podemos "andar"um pouco em qualquer direção sem "sair"de P. Se um ponto de P não é interior, então ele deve se encontrar em alguma "fronteira"de P.

De fato, um ponto que não é interior a P se encontra em uma face de P. Mais ainda, temos o seguinte resultado.

Teorema 3. *Tome o programa linear*

$$\begin{aligned} \max \quad & c^T x && (6.92)\\ s.t \quad & Ax = b && (6.93)\\ & x \geq 0 && (6.94) \end{aligned}$$

O conjunto de soluções ótimas desse programa linear é uma face de $P = \{x \in \mathbb{R}^n : Ax = b\}$.

Com isso, vemos que as regiões de P que nos interessam são as faces de P. Mas será que precisamos levar em conta todos os pontos de uma face de P,

quando em busca de uma solução ótima? Vamos caracterizar as faces de P em função dos vértices de P.

Dados pontos $X = \{x_1, x_2, \ldots, x_k\} \subseteq \mathbb{R}^n$, um ponto $\overline{x} \in \mathbb{R}^n$ é dito *combinação convexa* de X se $\overline{x} = \sum_{i=1}^{k} a_i x_i$, para $a_i \in [0,1], i \in [k]$, tal que $\sum_{i=1}^{k} a_i = 1$. O conjunto de todas as combinações convexas de X é chamado de *fecho convexo* de X, e é denotado por $conv(X)$.

Exercício 78. *Tome $X \in \mathbb{R}^n$. Prove que $X \subseteq conv(X)$.*

Teorema 4. *Seja F uma face de P. Existe um conjunto X de vértices de P tal que $F = conv(X)$.*

Ora, isso quer dizer que uma face de P tem pelo menos um vértice de P ($X \subseteq conv(X)$). Com isso, percebemos que o conjunto de soluções ótimas de um programa linear inclui pelo menos um vértice de P, onde P é o poliedro que é conjunto-solução de $Ax = b$. Se levarmos em conta apenas os vértices de P, ainda seremos capazes de encontrar uma solução ótima para um programa linear. E ainda podemos caracterizar os vértices de P!

Teorema 5. *Seja P o poliedro que é conjunto-solução de $Ax = b$. Temos que $\overline{x}$ é vértice de P se, e somente se, $\overline{x}$ é solução básica de $Ax = b$.*

Com isso, vemos que podemos encontrar uma solução ótima de um programa linear em $Ax = b$ mesmo que busquemos apenas dentro do conjunto de soluções básicas de $Ax = b$.

6.3.2 O método Simplex

Tome um programa linear na forma padrão

$$\begin{aligned} \max \quad & c^T x && (6.95)\\ \text{s.t} \quad & Ax = b && (6.96)\\ & x \geq 0 && (6.97) \end{aligned}$$

com $A \in \mathbb{R}^{m \times n}, c \in \mathbb{R}^n, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$. Além disso, $b \geq 0$.

Vamos admitir que conhecemos um subconjunto B de C_A que é uma base de $\mathbb{R}^m$. Vamos chamar de A_B a matriz cujas colunas são os vetores de B e de A_N a matriz cujas colunas são os vetores de $C_A \setminus B$. Em tempo, $\mathcal{B}$ são os índices das colunas associadas a B e $\mathcal{N}$ são os índices das colunas associadas a $C_A \setminus B$.

Dada a decomposição $A = [A_B A_N]$, podemos reescrever o programa linear como

$$\begin{aligned} \max \quad & c_B^T x_B + c_N^T x_N && (6.98)\\ \text{s.t} \quad & A_B x_B + A_N x_N = b && (6.99)\\ & x_B, x_N \geq 0 && (6.100) \end{aligned}$$

Podemos isolar x_B para tornar evidente a solução básica associada a B, que ocorre quando fazemos $x_N = 0$.

$$\begin{aligned} \max \quad & c_B^T x_B + c_N^T x_N && (6.101)\\ \text{s.t} \quad & x_B = A_B^{-1} b - A_B^{-1} A_N x_N && (6.102)\\ & x_B, x_N \geq 0 && (6.103) \end{aligned}$$

Podemos substituir x_B na função-objetivo e obter

$$\begin{aligned} c_B^T x_B + c_N^T x_N = c_B^T (A_B^{-1} b - A_B^{-1} A_N x_N) + c_N^T x_N \\ = c_B^T A_B^{-1} b - c_B^T A_B^{-1} A_N x_N + c_N^T x_N = c_B^T A_B^{-1} b + (c_N^T - c_B^T A_B^{-1} A_N) x_N \\ = c_B^T A_B^{-1} b - (c_B^T A_B^{-1} A_N - c_N^T) x_N \\ = c_B^T A_B^{-1} b - ((A_B^{-1} A_N)^T c_B - c_N)^T x_N \end{aligned}$$

Com isso, podemos reescrever o programa linear na forma

$$\begin{aligned} \max \quad & c_B^T A_B^{-1} b - ((A_B^{-1} A_N)^T c_B - c_N)^T x_N && (6.104)\\ \text{s.t} \quad & x_B = A_B^{-1} b - A_B^{-1} A_N x_N && (6.105)\\ & x_B, x_N \geq 0 && (6.106) \end{aligned}$$

Após tanto algebrismo, vamos fazer algumas observações. Primeiro, note que $c_B^T A_B^{-1} b \in \mathbb{R}$. Esse é precisamente o custo da solução básica associada a B. Como podemos saber se essa é uma solução de custo máximo? Vamos chamar $z_B = c_B^T A_B^{-1} b$ e $z_N = (A_B^{-1} A_N)^T c_B - c_N$. Simplficamos a escrita do modelo, em função de B.

$$\begin{aligned} \max \quad & z_B - z_N^T x_N && (6.107)\\ \text{s.t} \quad & x_B = A_B^{-1} b - A_B^{-1} A_N x_N && (6.108)\\ & x_B, x_N \geq 0 && (6.109) \end{aligned}$$

Repare que, se $z_N > 0$, qualquer incremento nas variáveis x_N causaria uma perda no custo (como estamos lidando com a solução básica associada a B, $x_N = 0$, e como o modelo exige $x_N \geq 0$, só faz sentido pensarmos em incrementos nos valores de x_N). Nesse cenário, é melhor deixar as variáveis x_N em 0 e temos encontrado uma solução básica de custo máximo.

Caso $z_N \geq 0$, pode haver uma variável em x_N, digamos $x_{N_i}, i \in \mathcal{N}$, tal que $z_{N_i} = 0$. Nesse caso, aumentar o valor de x_{N_i} não traz perdas ao custo. Isso indica que a solução básica associada a B não é a única solução com custo z_B, e pode haver outra solução básica com custo z_B.

Por último, pode ser que $z_N \ngeq 0$, isto é, pode haver uma variável em x_N, digamos $x_{N_i}, i \in \mathcal{N}$, tal que $z_{N_i} < 0$. Nesse caso, aumentar o valor de x_{N_i} traz um ganho ao custo, sugerindo que, se x_{N_i} assumir valor positivo, podemos encontrar soluções com custo maior que z_B. Como uma solução ótima do programa linear tem de ser uma solução básica, isso indica que existe uma solução

básica com custo maior que z_B, e que podemos obter uma se considerarmos uma solução básica com x_{N_i} como variável básica.

As observações acerca da otimalidade de uma solução básica estão feitas. No entanto, apesar de estarmos interessados em soluções básicas de $Ax = b$, não podemos levar em conta todas elas. No tocante a x, o modelo não apenas exige que $Ax = b$, mas também que $x \geq 0$. Ora, quando fazemos $x_N = 0$, o que nos garante que uma base B qualquer vai fazer $x_B = A_B^{-1} x_B \geq 0$? Não temos essa garantia. Precisamos lidar apenas com bases viáveis, e precisamos saber como ir a outra base viável caso percebamos que estamos em uma que não é ótima (isso nós já sabemos identificar, certo?).

Lembra que, quando escrevemos nosso programa linear na forma padrão, fizemos com que $b \geq 0$? Isso nos garante que a base canônica sempre será uma base viável. No entanto, não é sempre que os vetores $e_i, i \in [m]$, estão em C_A.

Por um instante, vamos considerar que nosso programa linear está na forma

$$\begin{aligned} \max \quad & c^T x && (6.110) \\ \text{s.t} \quad & Ax \leq b && (6.111) \\ & x \geq 0 && (6.112) \end{aligned}$$

com $A \in \mathbb{R}^{m \times n}, c \in \mathbb{R}^n, x \in \mathbb{R}^n$ e $b \in \mathbb{R}^m$. Além disso, $b \geq 0$.

Podemos facilmente escrevê-lo na forma padrão, se utilizarmos as variáveis de folga $s \in \mathbb{R}^m$.

$$\begin{aligned} \max \quad & c^T x + 0^T s && (6.113) \\ \text{s.t} \quad & Ax + Is = b && (6.114) \\ & x, s \geq 0 && (6.115) \end{aligned}$$

Dessa forma, associadas à base canônica, temos as variáveis s como básicas e as x como não-básicas, fornecendo a solução básica $(s, x) = (b, 0)$, que é viável e tem custo $z_B = 0^T b = 0$. Será ela uma solução básica ótima? Precisamos calcular z_N para determinar isso. Como $z_N = (A_B^{-1} A_N)^T c_B - c_N$, $c_B = 0$ e $c_N = c$, temos que $z_N = -c$. Se $c \leq 0$, temos $z_N \geq 0$ e a solução básica atual é ótima. Caso $c \not\leq 0$, temos $z_N \not\geq 0$, e a solução básica atual não é ótima. Podemos encontrar uma solução básica viável com custo maior que o da atual. Como podemos fazer isso?

Como $z_N \not\geq 0$ nos diz que existe uma variável $x_{N_i}, i \in \mathcal{N}$, que, se incrementada, traz ganhos ao custo atual, precisamos construir uma solução básica onde x_{N_i} é básica (afinal, apenas as soluções básicas são nosso interesse e, nelas, somente as variáveis básicas podem ser não-nulas). Podemos encontrar, a partir de B (ques estamos tratando como a base canônica), uma outra base de $\mathbb{R}^m$ que inclua a coluna associada a x_{N_i}. Como o número de variáveis básicas é m, a entrada de x_{N_i} na nova base implica necessariamente a saída de alguma das variáveis básicas de x_B.

Lembramos agora que estamos trabalhando com B sendo a base canônica. Uma outra observação é que encontrar explicitamente a nova base é equivalente a reduzí-la por operações sobre linhas à base canônica, nossa base atual. Dessa

forma, se queremos tornar x_{N_i} em uma variável básica e, para isso, uma certa variável $x_{B_j}, j \in \mathcal{B}$, tem de deixar de ser básica, basta transformar a coluna a_{N_i} associada a x_{N_i} em e_j. A pergunta agora é: quem vai sair da base?

Precisamos escolher x_{B_j} de forma que a_{N_i} tenha sua j-ésima posição não-nula, do contrário não podemos transformar a_{N_i} em e_j por operações sobre linhas. Com isso, dado que

$$a_{N_i} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_j \\ \vdots \\ a_m \end{bmatrix}$$

representamos as operações sobre linhas como (lembrando que as operações são aplicadas da direita para a esquerda)

$$P = \prod_{k \in [m]: k \neq j} S_{jk}^{-a_k} S_j^{\frac{1}{a_j}}$$

Em palavras, estamos multiplicando a j-ésima linha da matriz por $\frac{1}{a_j}$ e, em seguida, somando múltiplos dela às demais, de forma a zerar todas as entradas de a_{N_i}, menos a j-ésima (que torna-se unitária). Repare também que, após essas operações, as colunas básicas são preservadas, com exceção de e_j, que passa a representar os multiplicadores utilizados em cada operação. Dessa forma, x_{B_j} deixa de ser básica e x_{N_i} passa a ser.

Agora que temos atualizado a solução básica, precisamos atualizar z_B e z_N (já que mudamos de base). Podemos fazer isso com mais uma operação sobre linhas. Fazemos uma operação sobre linhas envolvendo a j-ésima linha da matriz e a função-objetivo. Somamos à função-objetivo um múltiplo da j-ésima linha de forma a zerar z_{N_i}. Essa operação é representada por $S_{j0}^{-z_{N_i}}$ (tratamos a função-objetivo como uma linha de índice zero). Isso atualiza z_B para representar o custo da nova solução básica (essencialmente, acrescentamos a z_B o valor $-z_{N_i}\frac{x_{B_j}}{a_j}$) e z_N para representar a sensibilidade dela a alterações nas suas variáveis não-básicas.

Temos descrito um procedimento para mudança de solução básica. No entanto, mesmo partindo de uma solução básica viável, o procedimento não garante produzir outra. Precisamos de mais algum critério para termos a garantia de que, quando mudamos de solução básica viável, a nova solução básica também é viável.

Lembramos que, associada a uma base B, temos

$$x_B = A_B^{-1} b - A_B^{-1} A_N x_N$$

descrevendo sua solução básica com $x_N = 0$, o que faz $x_B = A_B^{-1} b$. Quando incrementamos x_{N_i} de 0 a t, é como se estivéssemos trocando x_N por te_i na

equação acima. A nova solução básica produzida seria

$$x_{B'} = x_B - A_B^{-1} A_N t e_i$$

Veja que sabemos que a entrada de x_{N_i} na base aumenta o custo atual. Quanto mais aumentarmos o valor de x_{N_i}, mais ganhamos no custo. Dessa forma, queremos determinar o valor máximo de t que preserve a viabilidade da próxima solução básica. Queremos escolher t máximo que garanta $x_{B'} \geq 0$ (lembre que $x_{N'} = 0$, de todo modo).

$$x_{B'} \geq 0 \implies x_B - A_B^{-1} A_N t e_i \geq 0 \implies A_B^{-1} A_N t e_i \leq x_B$$

Fazemos $\Delta = A_B^{-1} A_N e_i$ para simplificar (perceba que, como estamos lidando com B sendo a base canônica, $\Delta = A_N e_i$ é a coluna associada a x_{N_i})

$$\Delta t \leq x_B$$

Queremos o maior t tal que Δt não ultrapasse x_B em nenhuma componente.

$$\Delta_j t \leq x_{B_j}, \forall j \in \mathcal{B}$$

Como $x_{B_j} \geq 0, j \in \mathcal{B}$, temos que o crescimento de t é limitado por x_{B_j} apenas quando $\Delta_j \geq 0$. Dessa forma, queremos o maior t que respeite

$$t \leq \frac{x_{B_j}}{\Delta_j}, \forall j \in \mathcal{B} : \Delta_j > 0$$

Pela inequação acima, nosso t máximo é

$$t = \min_{j \in \mathcal{B}: \Delta_j > 0} \frac{x_{B_j}}{\Delta_j}$$

Quando aplicamos nosso procedimento de mudança de base, x_{N_i} assume o valor de t correspondente ao j da variável x_{B_j} que escolhemos para sair da base (lembra da operação sobre a linha da função-objetivo?). De forma a manter a viabilidade para a próxima solução básica, precisamos escolher a variável básica a deixar a base de acordo com

$$j = \arg \min_{j \in \mathcal{B}: \Delta_j > 0} \frac{x_{B_j}}{\Delta_j}$$

Como última observação, constatamos que nem sempre x_B limita os múltiplos positivos de Δ (lembre que t tem de ser positivo). Pode ser que $\Delta \leq 0$ e, como $x_B \geq 0$ (por viabilidade), $\Delta t \leq x_B$, para qualquer t positivo. Isso indica que podemos aumentar o valor de x_{N_i} sem restrições e ganhar custo de forma arbitrária. Nesse caso, o programa linear não possui solução ótima, pois é ilimitado. Em particular, Δ indica a direção (a partir da solução básica atual) em que podemos tomar soluções com custos arbitrariamente altos.

Temos descrito um procedimento de mudança de base que preserva a viabilidade. Nossa descrição, no entanto, conta com a suposição de que a base

canônica está contida em C_A. Quando isso não é verdade, podemos recorrer ao artifício de transformar

$$\begin{aligned} \max \quad & c^T x && (6.116)\\ \text{s.t} \quad & Ax = b && (6.117)\\ & x \geq 0 && (6.118) \end{aligned}$$

no programa linear, para algum $M \in \mathbb{R}_+$ arbitrariamente alto

$$\begin{aligned} \max \quad & c^T x - Ma && (6.119)\\ \text{s.t} \quad & Ax + Ia = b && (6.120)\\ & x, a \geq 0 && (6.121) \end{aligned}$$

Exercício 79. *Prove que, se o programa linear* (6.119)-(6.121) *tem solução ótima* $(\overline{x}, \overline{a})$ *com* $\overline{a} = 0$*, então* $\overline{x}$ *é solução ótima para* (6.116)-(6.118)*. Prove que, se o programa linear* (6.119)-(6.121) *tem solução ótima* $(\overline{x}, \overline{a})$ *com* $\overline{a} \neq 0$*, então* (6.116)-(6.118) *é inviável. Dica: ambas por contradição.*

Exercício 80. *Resolva o seguinte problema*

$$\begin{aligned} \max \quad & 2x_1 - x_2 + 2x_3 && (6.122)\\ s.t & && (6.123)\\ & 2x_1 + x_2 \leq 10 && (6.124)\\ & x_1 + 2x_2 - 2x_3 \leq 20 && (6.125)\\ & x_2 + 2x_3 \leq 5 && (6.126)\\ & x \geq 0 && (6.127) \end{aligned}$$

Exercício 81. *Resolva o seguinte problema*

$$\begin{aligned} \max \quad & 2x_1 + x_2 && (6.128)\\ s.t & && (6.129)\\ & x_1 - x_2 \leq 10 && (6.130)\\ & 2x_1 - x_2 \leq 40 && (6.131)\\ & x \geq 0 && (6.132) \end{aligned}$$

Exercício 82. *Resolva o seguinte problema. Após encontrar a primeira, é possível encontrar uma outra solução ótima?*

$$\begin{aligned} \max \quad & 4x_1 + 14x_2 && (6.133)\\ s.t & && (6.134)\\ & 2x_1 + 7x_2 \leq 21 && (6.135)\\ & 7x_1 + 2x_2 \leq 21 && (6.136)\\ & x \geq 0 && (6.137) \end{aligned}$$

Exercício 83. *Resolva o seguinte problema*

$$\begin{aligned} \max \quad & x_1 + x_2 && (6.138)\\ s.t \quad & && (6.139)\\ & x_1 \geq 6 && (6.140)\\ & x_2 \geq 6 && (6.141)\\ & x_1 + x_2 \leq 11 && (6.142)\\ & x \geq 0 && (6.143) \end{aligned}$$

7 Programação Inteira

Nesta seção, lidamos com as técnicas de resolução de programas lineares nos cenários em que exigimos que suas variáveis sejam inteiras. Nessas situações, não temos a garantia de que algum vértice do poliedro que contém o conjunto de soluções viáveis seja uma solução ótima, uma vez que não temos a garantia de que os vértices do poliedro são pontos inteiros.

Aprendemos a lidar com tais cenários de duas formas. Na primeira, aplicamos cortes no poliedro, sem prejuízo para as soluções viáveis. Na segunda, consideramos uma abordagem de divisão-e-conquista para forçar a integralidade das variáveis.

Observamos, antes de mais nada, que Programação Linear é bem distinta de Programação Inteira, em se tratando de complexidade. Enquanto programas lineares podem ser resolvidos em tempo polinomial, programas lineares inteiros podem ser muito difíceis de resolver.

Exercício 84. *Modele o problema 3-SAT (variáveis $x \in \mathbb{B}^n$ e cláusulas $c_i = l_{i1} \vee l_{i2} \vee l_{i3}, i \in [m], m \in \mathbb{N}$, onde $l_{ij} \in \{x_i, \neg x_i : i \in [n]\}, i \in [m], j \in [3]$) como um programa linear inteiro. Note que, enquanto 3-SAT é um problema de decisão, um programa linear inteiro fornece um valor ótimo. Faça uma função-objetivo que maximize o número de cláusulas satisfeitas. Argumente que, se for possível resolver o seu programa linear inteiro em tempo polinomial, então 3-SAT pode ser decidido em tempo polinomial.*

7.1 Cortes de Gomory-Chvátal

Tome um programa linear na forma padrão

$$\begin{aligned} \max \quad & c^T x && (7.1)\\ \text{s.t} \quad & Ax = b && (7.2)\\ & x \in \mathbb{Z}^n_+ && (7.3) \end{aligned}$$

onde $A \in \mathbb{R}^{m \times n}, c \in \mathbb{R}^n, b \in \mathbb{R}^m$ e $x \in \mathbb{Z}^n_+$. Além disso, $b \geq 0$.

Perceba que, se $P = \{x \in \mathbb{R}^n : Ax = b\}$, então o conjunto de soluções viáveis do programa (7.1)-(7.3) é descrito por $P \cap \mathbb{Z}^n_+$. Quando consideramos (7.1)-(7.3) sem as restrições de integralidade, isto é, apenas com $x \geq 0$, temos que suas

soluções viáveis estão em P, e esse caso sabemos resolver. Então, considere que, quando exigimos apenas que $x \geq 0$, uma base B de $\mathbb{R}^m$ está associada a uma solução básica ótima $(\overline{x}_B, 0)$ para o programa (7.1)-(7.3), evidenciada como

$$\begin{aligned} \max \quad & \overline{z}_B - \overline{z}_N^T 0 && (7.4) \\ \text{s.t} \quad & \overline{x}_B = A_B^{-1} b - A_B^{-1} A_N 0 && (7.5) \\ & \overline{x}_B \geq 0 && (7.6) \end{aligned}$$

Se $\overline{x}_B \in \mathbb{Z}_+^m$, então $(\overline{x}_B, 0) \in P \cap \mathbb{Z}_+^n$ e temos encontrado uma solução ótima para (7.1)-(7.3). Caso $\overline{x}_B \notin \mathbb{Z}_+^m$, então alguma das variáveis x_B está assumindo valor fracionário na solução básica associada a B, que portanto é inviável (por não ser inteira). Vamos supor que tal variável é x_{B_i}, que assume valor $\overline{x}_{B_i} \notin \mathbb{Z}_+$. Admitindo que a i-ésima componente de $A_B^{-1} b$ é d e que a i-ésima linha de $A_B^{-1} A_N$ é a^T, temos que a i-ésima igualdade de (7.4)-(7.6) nos dá

$$x_{B_i} + a^T x_N = d \tag{7.7}$$

Isso implica que

$$x_{B_i} + a^T x_N \leq d$$

Perceba que, como $\overline{x}_{B_i}$ é obtido quando fazemos $x_N = 0$, então $d \notin \mathbb{Z}_+$. Considere que, se $v \in \mathbb{R}^k$ é um vetor qualquer, o vetor $\lfloor v \rfloor$ é obtido tomando o piso de cada componente de v. Assim, podemos escrever

$$x_{B_i} + a^T x_N \leq d \implies x_{B_i} + \lfloor a \rfloor^T x_N \leq d$$

uma vez que exigimos $x_N \geq 0$. Como queremos encontrar uma solução ótima inteira, x_{B_i} e x_N são inteiros, assim como $\lfloor a \rfloor$. Dessa forma, $x_{B_i} + \lfloor a \rfloor^T x_N$ tem de ser inteiro, e podemos melhorar a desigualdade como segue.

$$x_{B_i} + \lfloor a \rfloor^T x_N \leq \lfloor d \rfloor \tag{7.8}$$

Note que, pelo desenvolvimento, $\lfloor d \rfloor < d$. Subtraímos (7.8) de (7.7) e obtemos

$$(a - \lfloor a \rfloor)^T x_N \geq d - \lfloor d \rfloor \tag{7.9}$$

A desigualdade (7.9) é chamada de corte de Gomory. Perceba que $(\overline{x}_B, 0)$ não satisfaz (7.9), uma vez que $d - \lfloor d \rfloor > 0$. Dessa forma, dizemos que (7.9) "corta" $(\overline{x}_B, 0)$. Note também que (7.9) é obtida a partir das restrições $Ax = b$ por operações que preservam a viabilidade de soluções inteiras. Com isso, percebemos que (7.9) preserva o conjunto de soluções viáveis de (7.1)-(7.3), enquanto corta uma solução fracionária que seria ótima caso exigíssemos apenas que $x \geq 0$.

Com (7.9) obtida, podemos resolver

$$\begin{aligned} \max \quad & c^T x && (7.10) \\ \text{s.t} \quad & Ax = b && (7.11) \\ & (a - \lfloor a \rfloor)^T x_N \geq d - \lfloor d \rfloor && (7.12) \\ & x \in \mathbb{Z}_+^n && (7.13) \end{aligned}$$

com a certeza de que, exigindo apenas $x \geq 0$, não obteremos a mesma solução fracionária, ao passo que nenhuma solução inteira torna-se inviável com a nova restrição. Dessa forma, podemos repetir o procedimento com o novo programa linear inteiro, caso sua solução ótima seja fracionária.

Um ponto importante é que, após um número finito de aplicações de cortes de Gomory, encontramos uma solução ótima inteira. Não poderemos demonstrar esse resultado, no entanto.

Apesar da garantia de que eventualmente obteremos uma solução inteira, temos dois problemas a considerar. O primeiro é o custo do método, aparentemente alto, uma vez que temos de resolver um programa linear por iteração. Há formas mais eficientes de se lidar com a adição de uma nova restrição a um programa linear já resolvido. Como essas utilizam dualidade, e não teremos tempo de nos aprofundar nesse tópico, vamos omití-las. O segundo problema é que ter a garantia de que uma solução inteira ótima é encontrada em um número finito de iterações não é satisfatório, já que isso nem sempre oferece limites superiores para o número de iterações.

7.2 Relaxação lagrangeana e dualidade

Vamos brevemente introduzir o conceito de dualidade através da relaxação lagrangeana. Antes, naturalmente, vamos introduzir a relaxação lagrangeana.

Suponha que queremos resolver o programa linear

$$z = \max \quad c^T x \tag{7.14}$$
$$\text{s.t} \quad Ax \leq b \tag{7.15}$$
$$x \geq 0 \tag{7.16}$$

Poderíamos pensar em, no lugar de proibir valores de x que não satisfazem $Ax \leq b$, apenas dar ao custo deles uma penalidade, tornando-os menos desejáveis que as soluções viáveis. Se $\overline{x}$ não satisfaz $Ax \leq b$, então $A\overline{x} - b \not\leq 0$. Podemos utilizar isso para criar uma penalidade, em vez de utilizar as restrições.

$$z(u) = \max \quad c^T x - u^T(Ax - b) \tag{7.17}$$
$$x \geq 0 \tag{7.18}$$

Note que, quando escolhemos os coeficientes u como $u \geq 0$, estamos diminuindo o custo de valores de x que violam algumas das restrições de $Ax \leq b$. No entanto, também introduzimos uma distorção, uma vez que qualquer solução, viável ou inviável, ganha custo pelas restrições de $Ax \leq b$ que não são violadas.

Exercício 85. *Prove que, se $u \geq 0$ e tanto z como $z(u)$ estão bem definidos, então $z \leq z(u)$.*

Como o exercício mostra, penalizar não é o mesmo que restringir. Quando escolhemos $u \geq 0$, a versão do modelo linear que considera penalidades nos fornece um limite superior para o ótimo do modelo original. Temos $z \leq z(u)$

justamente por conta das distorções de custo criadas no espaço viável (soluções viáveis estão ganhando custo simplesmente por serem viáveis). Mas podemos buscar pela escolha de u que, além de penalizar valores de x inviáveis, introduz a menor quantidade de distorção possível, fornecendo assim o melhor limite superior (*upper bound*) para z.

$$z_{UB} = \min \quad z(u) \tag{7.19}$$
$$u \geq 0 \tag{7.20}$$

Perceba que

$$z_{UB} = \min_{u \geq 0} \max_{x \geq 0} \quad c^T x - u^T (Ax - b) \tag{7.21}$$
$$z_{UB} = \min_{u \geq 0} \max_{x \geq 0} \quad c^T x - u^T Ax + u^T b \tag{7.22}$$

Como $u^T b$ é independete de x, escrevemos

$$z_{UB} = \min_{u \geq 0} \quad u^T b + \left(\max_{x \geq 0} \quad c^T x - u^T Ax \right) \tag{7.23}$$
$$z_{UB} = \min_{u \geq 0} \quad u^T b + \left(\max_{x \geq 0} \quad (c^T - u^T A)x \right) \tag{7.24}$$
$$z_{UB} = \min_{u \geq 0} \quad u^T b + \left(\max_{x \geq 0} \quad (c - A^T u)^T x \right) \tag{7.25}$$

Vamos fazer algumas observações. Queremos que z_{UB} seja o melhor limite superior para z. Note que, para valores de u com $c - A^T u \not\leq 0$, a expressão que define z_{UB} torna-se ilimitada, uma vez que podemos escolher x de forma que $(c - A^T u)^T x$ assuma valores arbitrariamente altos. Posto isso, pode ser que z_{UB} seja limitado para escolhas de u com $c - A^T u \leq 0$, e caso seja, a direção de mínimo nos garante que valores de u com $c - A^T u \not\leq 0$ não nos interessam.

$$z_{UB} = \min_{u \geq 0} \quad u^T b + \left(\max_{x \geq 0} \quad (c - A^T u)^T x \right) \tag{7.26}$$
$$\text{s.t} \quad A^T u \geq c \tag{7.27}$$

Sob a restrição $A^T u \geq c$, note que $\max_{x \geq 0} (c - A^T u)^T x = 0$. Feita essa observação, escrevemos

$$z_{UB} = \min \quad u^T b \tag{7.28}$$
$$\text{s.t} \quad A^T u \geq c \tag{7.29}$$
$$u \geq 0 \tag{7.30}$$

Exercício 86. *Prove que, se z e z_{UB} estão bem definidos, então $z \leq z_{UB}$.*

O programa linear (7.28)-(7.30) é dito o *dual* do programa (7.14)-(7.16) (chamado *primal*), e fornece o melhor limite superior para z, caso seja viável e limitado. Como estamos tratando de programas lineares, z_{UB} não poderia ser um limite superior melhor.

Exercício 87. *Desenvolva o dual do programa linear* (7.28)-(7.30). *O que você pode concluir com isso?*

Teorema 6. *Se* (7.14)-(7.16) *e* (7.28)-(7.30) *são ambos viáveis, então* $z = z_{UB}$.

Demonstração. Vamos desenvolver essa prova com base na corretude do método Simplex. Se ambos primal e dual são viáveis, o exercício 86 nos diz que um limita o outro, e portanto ambos têm soluções ótimas, o que garante que z e z_{UB} estão bem definidos. Procedemos transformando (7.14)-(7.16) para a forma padrão

$$z = \max \quad d^T x' \tag{7.31}$$
$$\text{s.t} \quad Mx' = b \tag{7.32}$$
$$x' \geq 0 \tag{7.33}$$

onde $d = (c, 0), x' = (x, s)$ e $M = [A \quad I]$. Ao final da execução do Simplex, obtemos uma solução ótima $\overline{x'} = (\overline{x'}_B, 0)$ associada a uma base B, que pode ser descrita por

$$z = \max \quad z_B - z_N^T x'_N \tag{7.34}$$
$$\text{s.t} \quad x'_B = M_B^{-1} b - M_B^{-1} M_N x'_N \tag{7.35}$$
$$x'_B, x'_N \geq 0 \tag{7.36}$$

Temos que, dado $\overline{x'}_N = 0$, $z = z_B$ e $\overline{x'}_B = M_B^{-1} b$. Pela otimalidade de $\overline{x'}$, temos $z_N \geq 0$. Como $z_B = d_B^T M_B^{-1} b$, temos $z = d_B^T M_B^{-1} b$. Vamos provar que $\overline{u} = (d_B^T M_B^{-1})^T$ é solução ótima para (7.28)-(7.30). Vamos lidar com a viabilidade de $\overline{u}$ depois, e garantir a igualdade do teorema antes.

$$z = d^T \overline{x'} = d_B^T \overline{x'}_B + d_N^T \overline{x'}_N = d_B^T \overline{x'}_B$$
$$z = d_B^T M_B^{-1} b = (d_B^T M_B^{-1}) b = ((d_B^T M_B^{-1})^T)^T b = \overline{u}^T b$$

Se $\overline{u}$ é de fato viável para o dual, temos que o ótimo do dual não pode ser maior que $\overline{u}^T b$, isto é, $z_{UB} \leq \overline{u}^T b = z$. Como o exercicio 86 nos dá $z \leq z_{UB}$, $\overline{u}$ é ótima e temos $z = z_{UB}$, desde que $\overline{u}$ seja viável para o dual.

A pendência agora é provar que $\overline{u}$ é viável para o dual, isto é, $A^T \overline{u} \geq c$ e $\overline{u} \geq 0$. Como $\overline{u} \geq 0 \implies I\overline{u} \geq 0$, as duas inequações a serem provadas podem ser escritas como

$$M^T \overline{u} \geq d$$

Perceba que

$$M^T \overline{u} = M^T (d_B^T M_B^{-1})^T = (d_B^T M_B^{-1} M)^T$$

Agora, levamos em conta a decomposição $M = [M_B \quad M_N]$ para escrever

$$\begin{bmatrix} M_B^T \\ M_N^T \end{bmatrix} \overline{u} = \begin{bmatrix} M_B^T \\ M_N^T \end{bmatrix} (d_B^T M_B^{-1})^T = \begin{bmatrix} M_B^T (d_B^T M_B^{-1})^T \\ M_N^T (d_B^T M_B^{-1})^T \end{bmatrix} = \begin{bmatrix} (d_B^T M_B^{-1} M_B)^T \\ (d_B^T M_B^{-1} M_N)^T \end{bmatrix}$$

Como $M_B^{-1} M_B = I$ e $z_N = (M_B^{-1} M_N)^T d_B - d_N$, fazemos

$$\begin{bmatrix} M_B^T \\ M_N^T \end{bmatrix} \overline{u} = \begin{bmatrix} (d_B^T)^T \\ (d_N^T + z_N^T)^T \end{bmatrix} = \begin{bmatrix} d_B \\ d_N + z_N \end{bmatrix}$$

Podemos ver que a inequação $M^T\overline{u} \geq d$ é satisfeita: na componente associada a M_B, a inequação é satisfeita na igualdade; na componente associada a M_N, segue da otimalidade de $\overline{x}$ que $z_N \geq 0$, e portanto $d_N + z_N \geq d_N$. Com isso, temos provado a viabilidade de $\overline{u}$, sustentando o resultado. □

Observamos que z_{UB} pode não estar bem-definido por dois motivos. Se seu programa linear é inviável, isso quer dizer que não existe um limite superior z_{UB} para z, sugerindo que z não pode ser limitado superiormente. Isso ocorre quando z não existe, seja por inviabilidade ou por seu programa linear ser ilimitado. Quando o programa linear de z_{UB} é ilimitado, podemos tomar um limite superior para z tão baixo quanto desejarmos. Note que nenhum valor para z é compatível com essa situação, o que sugere que o programa linear associado a z é inviável (perceba que z não pode também ser ilimitado).

Vamos recapitular o que temos feito. Desenvolvemos o conceito de relaxação lagrangeana, que penaliza soluções inviáveis para fornecer um limite superior para um programa linear de maximização. Para cada escolha de u, obtemos um limite superior. Dessa forma, procuramos pela escolha de u que nos fornece o melhor limite superior possível. O programa linear de minimização que modela essa escolha de u é o dual do programa linear original, chamado de primal.

Exercício 88. *Por meio de relaxações lagrangeanas, desenvolva os duais de programas lineares de* $\{$*maximização, minimização*$\}$*, com restrições* $\{Ax = b, Ax \geq b\}$ *e variáveis* $\{x \geq 0, x \leq 0, x \in \mathbb{R}^n\}$.

Quando lidamos com programas lineares inteiros, o programa dual pode ser bastante útil. Primeiro, dado que pode ser muito difícil resolver um programa linear inteiro, não podemos ter muita esperança de encontrar seu ótimo em uma quantidade de tempo razoável. Podemos ficar satisfeitos em determinar um intervalo $[lb, ub]$ que contém seu valor ótimo. Quando resolvemos o programa linear associado ao programa linear inteiro (sua *relaxação linear*), obtemos um valor para ub. Se conhecemos uma solução viável para o programa linear inteiro, seu custo é um valor para lb. No entanto, mesmo encontrar uma solução viável pode ser difícil (lembra que modelamos 3-SAT como um programa linear inteiro?).

O dual é interessante, dentre diversas razões, por fornecer valores para ub de uma forma bem barata. Enquanto apenas soluções ótimas para a relaxação linear de um programa inteiro fornecem valores para ub, a dualidade fraca nos garante que soluções viáveis quaisquer para o dual da relaxação linear nos fornecem valores para ub. Nesse sentido, seria preciso resolver a relaxação linear para obter um valor para ub. Para seu dual, apenas algumas iterações do Simplex (até atingir-se viabilidade) são necessárias para se obter um valor para ub.

7.3 Branch and Bound

Tome $P = \{x \in \mathbb{R}^n : Ax = b\}$. Um programa linear inteiro na forma padrão pode ser expresso como

$$z = \max \quad c^T x \tag{7.37}$$
$$\text{s.t} \quad x \in P \cap \mathbb{Z}_+^n \tag{7.38}$$

com $c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}$ e $b \in \mathbb{R}^m$. Além disso, $b \geq 0$.

Podemos abordar a resolução do programa linear inteiro a partir de sua relaxação linear. Seja $\overline{x}$ uma solução ótima para a relaxação linear. Se $\overline{x} \in \mathbb{Z}_+^n$, então $\overline{x}$ também é solução ótima para o programa linear inteiro. Caso $\overline{x} \notin \mathbb{Z}_+^n$, começamos o procedimento de *branch and bound*.

Note que $c^T\overline{x}$ é um limite superior para z. Dessa forma, temos um intervalo $[lb, ub]$ com $z \in [lb, ub]$, onde $ub = c^T\overline{x}$ e $lb = -\infty$ (ainda não conhecemos soluções viáveis para o programa inteiro).

Já que $\overline{x} \notin \mathbb{Z}_+^n$, existe $i \in [n]$ tal que $\lfloor \overline{x}_i \rfloor < \overline{x}_i < \lceil \overline{x}_i \rceil$. Note que toda solução do programa inteiro satisfaz ou $x_i \leq \lfloor \overline{x}_i \rfloor$ ou $x_i \geq \lceil \overline{x}_i \rceil$. Mais ainda, ambas as desigualdades cortam $\overline{x}$. Considere $L_i(a) = \{x \in \mathbb{R}^n : x_i \leq \lfloor a_i \rfloor\}$ e $R_i(a) = \{x \in \mathbb{R}^n : x_i \geq \lceil a_i \rceil\}$, onde $i \in [n]$ e $a \in \mathbb{R}^n$. Dessa forma, $P \cap \mathbb{Z}_+^n = ((P \cap L_i(\overline{x})) \cup (P \cap R_i(\overline{x}))) \cap \mathbb{Z}_+^n$. Com isso, podemos ver que as partes relevantes da relaxação linear são $P \cap L_i(\overline{x})$ e $P \cap R_i(\overline{x})$. Vamos fazer uma *ramificação* (*branching*) com base nessa dicotomia.

$$z_l = \max \quad c^T x \tag{7.39}$$
$$\text{s.t} \quad x \in (P \cap L_i(\overline{x})) \cap \mathbb{Z}_+^n \tag{7.40}$$

$$z_r = \max \quad c^T x \tag{7.41}$$
$$\text{s.t} \quad x \in (P \cap R_i(\overline{x})) \cap \mathbb{Z}_+^n \tag{7.42}$$

Perceba que z_l e z_r são ambos definidos por programas lineares inteiros, que podem ser resolvidos, recursivamente, pelo procedimento que estamos descrevendo. Uma vez definidos z_l e z_r, temos $z = \max(z_l, z_r)$. No caso de ambos terem seus programas inviáveis, o programa de z é inviável. No caso de z_l ou z_r ter seu programa ilimitado, z tem seu programa ilimitado.

Até agora, temos descrito uma busca exaustiva que evita soluções fracionárias por meio de cortes, mas que divide o espaço de busca em duas regiões disjuntas. Isso não parece muito esperto. Infelizmente, a NP-dificuldade da Programação Inteira não nos deixa fazer muito além disso. Podemos, ainda, tornar essa busca exaustiva um pouco mais elegante.

Dado que $z = \max(z_l, z_r)$, pode ser que não precisemos encontrar ambos z_l e z_r se tivermos garantias de que um é limitado pelo outro. Suponha que encontramos z_l, resolvendo seu programa linear inteiro. Se o ótimo da relaxação linear do programa de z_r é menor que z_l, sabemos que z_r tem de ser menor que z_l, o que nos diz $z = z_l$ sem precisarmos resolver o programa de z_r. Dizemos

que foi feita a *poda* da região viável associada a z_r com base em algum limite (*bound*) conhecido. Há a situação em que encontramos z_r primeiro, simétrica à descrita.

Pode ser que informações que encontramos durante a resolução do programa de z_l nos ajudem a diminuir o esforço necessário para a resolução do programa de z_r, e vice-versa. Como a estratégia que estamos descrevendo consiste em dividir a resolução de um programa inteiro em subproblemas, podemos estar interessados em intercalar a resolução de subproblemas do programa associado a z_l com a de subproblemas do programa associado a z_r.

Dizemos que *exploramos* um subproblema quando dividimos seu espaço viável como foi descrito. Ainda que um subproblema não tenha sido explorado, admitimos conhecer uma solução ótima para sua relaxação linear (ou saber que ela é inviável ou ilimitada). Um subproblema cuja solução ótima de sua relaxação linear é inteira não precisa ser explorado (o mesmo vale quando a relaxação linear é inviável).

Estamos interessados em encontrar um intervalo $[lb, ub]$ com $z \in [lb, ub]$. Toda vez que um subproblema apresenta uma solução ótima inteira para sua relaxação linear, temos que seu custo nos fornece um limite inferior lb' para z. Se $lb < lb'$, passamos a considerar o intervalo $[lb', ub]$. Tomamos ub como o maior custo dentre os custos ótimos das relaxações lineares dos subproblemas inexplorados. Se um subproblema inexplorado tem sua relaxação linear com custo ub', e temos $ub' < lb$, esse subproblema não precisa ser explorado (por quê?).

Pode ser que, para completar esse procedimento, precisemos explorar uma quantidade muito grande de subproblemas, o que pode ser incoveniente em situações onde temos tempo limitado. Quanto mais subproblemas explorarmos, menor tende a ser o comprimento do intervalo $[lb, ub]$, chamado de *gap*. Pode ser aceitável interromper o procedimento, mesmo havendo subproblemas inexplorados, quando se tem um gap suficientemente pequeno.

Ainda há algumas questões em aberto quanto ao procedimento de Branch and Bound. Primeiro, a solução ótima da relaxação linear de um subproblema pode ter mais de uma variável com valor fracionário. Nesse caso, qual variável devemos escolher para fazer a ramificação? Podemos também ter diversos subproblemas inexplorados. Nessa situação, qual subproblema devemos explorar?

É bem perceptível que, até agora, descrevemos apenas como diminuir o comprimento do intervalo $[lb, ub]$ encontrando valores para lb ao passo que subproblemas apresentam soluções ótimas inteiras em suas relaxações lineares. Mas podemos também diminuir o comprimento de $[lb, ub]$ encontrando melhores valores para ub. Lembra que ub é o maior custo ótimo dentre as relaxações lineares dos subproblemas inexplorados? Uma abordagem clássica para a escolha do próximo subproblema a ser explorado consiste em sempre escolher um subproblema cujo custo ótimo de sua relaxação linear seja ub. A ideia por trás dessa escolha é que, quando exploramos um subproblema, isto é, quando o substituímos por dois novos subproblemas, cada um com uma restrição a mais, esses novos subproblemas podem ter relaxações lineares com custos ótimos menores que o da relaxação linear do subproblema recém-explorado. Dessa forma, seguindo

essa escolha, a tendência é que encontremos menores valores para ub.

Além da abordagem descrita acima, existem diversas outras para a escolha de qual subproblema explorar. Podemos, por exemplo, abrir mão de melhorar ub para explorar um subproblema cuja relaxação linear esteja próxima de ser inteira (de acordo com algum critério). Dessa vez, temos a esperança de que, com as restrições adicionais, as relaxações lineares dos novos subproblemas apresentem soluções ótimas inteiras, o que pode melhorar lb.

Quanto à escolha da variável para efetuar a ramificação, existem também diversas abordagens. Uma delas consiste em, ao explorar um subproblema, tomar a variável mais fracionária (mais distante de um valor inteiro) para ramificação. Algumas outras abordagens podem usar critérios mais avançados (e caros), alguns baseados nos duais das relaxações lineares de alguns dos pares de novos subproblemas que podem ser criados.

8 Problema de Fluxo Máximo

Na última seção, nos deparamos com as técnicas de resolução de programas lineares inteiros. As constatações não foram boas. Programas lineares inteiros podem representar problemas NP-difíceis, afinal. Apesar disso, nem todo programa linear inteiro é de difícil resolução. Nesta seção, apresentamos um problema clássico cuja resolução dá-se em tempo polinomial. A motivação para apresentá-lo, em detrimento de outros, está na sua versatilidade.

8.1 Definição

Tome $D = (V, A)$ um grafo direcionado, isto é, com vértices V e arcos A, onde cada arco $a \in A$ é um par ordenado $a = (v_1, v_2) \in V^2$. Considere também uma função $c : A \to \mathbb{Z}_+$ associando aos arcos A uma capacidade inteira. Dados $s, t \in V$, o Problema de Fluxo Máximo (PFM) em (D, c) consiste em determinar a maior quantidade de fluxo que pode "ser criado"em s e "sumir"em t, "passando"pelos arcos A, tal que:

- o fluxo passando por um arco não excede sua capacidade;
- o fluxo não pode chegar em s nem sair de t;
- o fluxo passando pelos vértices $V \backslash \{s, t\}$ é conservativo, isto é, a quantidade de fluxo que chega em um vértice é a mesma que dele sai.

Lembramos que $\delta^+(u) = \{a \in A | \exists v \in V : a = (u, v)\}$ e $\delta^-(v) = \{a \in A | \exists u \in V : a = (u, v)\}$. Além disso, se não pode haver fluxo chegando em s nem saindo de t, podemos admitir que, em D, não há arcos chegando em s nem saindo de t (caso haja, basta deletá-los). Dadas variáveis $f \in \mathbb{R}^{|A|}$ representando um fluxo, podemos modelar o problema como

$$\max \quad \sum_{a \in \delta^+(s)} f_a \tag{8.1}$$

$$\text{s.t} \quad \sum_{a \in \delta^+(v)} f_a = \sum_{a \in \delta^-(v)} f_a, \qquad \forall v \in V \setminus \{s, t\} \tag{8.2}$$

$$0 \leq f_a \leq c_a, \qquad \forall a \in A \tag{8.3}$$

8.2 Propriedades

A primeira propriedade que esse problema nos apresenta é evidenciada por seu modelo. Note que, dado qualquer (D, c), o problema de fluxo máximo sempre é bem definido.

Exercício 89. *Argumente que o PFM é sempre viável. Aargumente que o PFM é sempre limitado. Dicas: apresente um fluxo trivial para qualquer (D, c); use (8.3) para obter um limite superior para $\sum_{a \in \delta^+(s)} f_a$ em termos de c.*

Exercício 90. *Pesquise o que são matrizes de incidência de grafos direcionados. Se admitirmos que os arcos associam-se a colunas e os vértices a linhas, quantas entradas não-nulas tem qualquer coluna da matriz de incidência de um grafo direcionado? Qual a soma das entradas não-nulas de cada coluna?*

Vamos introduzir brevemente o conceito de total-unimodularidade. Uma matriz quadrada com entradas inteiras é unimodular se seu determinante é ± 1. Uma matriz com entradas inteiras é dita total-unimodular (TU) se qualquer uma de suas submatrizes quadradas tem determinante 0 ou ± 1. Como essa definição pode nos ajudar? Antes, exercícios.

Exercício 91. *Prove que toda matriz unimodular é invertível.*

Exercício 92. *Prove que, se a matriz A é TU, então qualquer entrada de A é 0, 1 ou -1. Dica: matrizes em $\mathbb{R}^{1\times 1}$ são quadradas?*

Exercício 93. *Prove que, se A é TU, então qualquer matriz obtida a partir de A pela adição de uma linha ou coluna canônica também é TU.*

Exercício 94. *Prove que, se A é TU, então A^T é TU.*

Exercício 95. *Prove que, se A é TU, então $[A \quad I]$ é TU.*

Lembramos que, em programas lineares na forma padrão, nossos pontos de interesse são as soluções básicas de $Ax = b$, para $A \in \mathbb{R}^{m\times n}$. Se A é quadrada e invertível, então $Ax = b$ tem exatamente uma solução. Considere A'_j a matriz obtida a partir de A ao substituir sua j-ésima coluna por b. Cramer nos diz que

$$x_j = \frac{\det(A'_j)}{\det(A)}, \forall j \in [m]$$

Teorema 7. *Se A é unimodular e b é um vetor inteiro, então $Ax = b$ tem uma única solução, e esta é inteira.*

Demonstração. Por unimodularidade, toda entrada de A é inteira. Além disso, A é invertível (por quê?). Note que, como b é um vetor inteiro, A'_j tem todas as suas entradas inteiras, para $j \in [m]$. Como o determinante de uma matriz está em função de suas entradas, $\det(A'_j) \in \mathbb{Z}$. Como $\det(A) \in \{-1, 1\}$, temos $\frac{\det(A'_j)}{\det(A)} \in \mathbb{Z}$. A é quadrada por ser unimodular, daí $Ax = b$ tem uma única solução, e por Cramer, temos que ela é inteira. □

Teorema 8. *Se A é TU e b é um vetor inteiro, então toda solução básica de $Ax = b$ é inteira.*

Demonstração. Tome uma solução básica $\overline{x}$ de $Ax = b$, associada a uma base B. Como A é TU, temos que A_B, uma submatriz quadrada de A, é tal que $\det(A_B) \in \{-1, 0, 1\}$. Como A_B tem suas colunas linearmente independentes, temos que $\det(A_B) \in \{-1, 1\}$, e portanto A_B é unimodular. Como $\overline{x}$ é básica, temos que $\overline{x} = (\overline{x}_B, 0)$. Assim, resta provar que $\overline{x}_B$ é inteira. Note que A_B e b satisfazem a hipótese do Teorema 7, e como $\overline{x}_B$ é solução de $A_B x_B = b$, temos a integralidade de $\overline{x}$. □

Note a importância desses resultados. Se um programa linear inteiro na forma padrão tem suas restrições expressas por $Ax = b$, com A TU e b de entradas inteiras, então a solução ótima de sua relaxação linear tem de ser inteira. Isso significa que a total-unimodularidade oferece uma condição suficiente para que um programa linear inteiro possa ser resolvido em tempo hábil. O próximo resultado é uma condição suficiente para total-unimodularidade, cuja demonstração é omitida.

Teorema 9. *Uma matriz A é TU se cada uma de suas colunas tem no máximo duas entradas não-nulas e o conjunto de suas linhas pode ser particionado em C_1 e C_2 tal que:*

- *se uma coluna tem duas entradas não-nulas de mesmo sinal, as respectivas linhas estão uma em C_1 e a outra em C_2;*
- *se uma coluna tem duas entradas não-nulas de sinais opostos, as respectivas linhas estão ou ambas em C_1 ou ambas em C_2.*

Exercício 96. *Prove que a matriz de coeficientes do programa linear* (8.1)-(8.3) *é TU. Dica: prove que a matriz de incidência de um digrafo $D = (V, A)$ é TU; note que toda submatriz de uma matriz TU também é TU; note que a matriz de coeficientes de* (8.1)-(8.3) *é submatriz da matriz de incidência do digrafo associado.*

Nesse ponto, temos que o PFM, além de estar sempre bem definido, sempre tem solução ótima inteira. Isso indica que, se tivéssemos modelado o PFM com um programa linear inteiro, resolver sua relaxação linear seria suficiente para encontrar uma solução ótima inteira. No entanto, veremos que há modos mais eficientes de resolver o PFM.

8.3 Dualidade e Ford-Fulkerson

Podemos abordar o estudo de dualidade do PFM sem a utilização direta de Álgebra Linear. Vamos fazê-lo por meio de Combinatória.

Tome $D = (V, A)$ um grafo direcionado. Dados $s, t \in V$, um conjunto $A' \subseteq A$ é dito um (s,t)-corte se não há um caminho de s para t em $D' = (V, A \setminus A')$. Isso indica que, em D, todo caminho de s a t tem um arco em um (s,t)-corte.

Dada uma instância $(D = (V, A), c)$ do PFM, definimos a capacidade de um conjunto de arcos $A' \subseteq A$ como $c(A') = \sum_{a \in A'} c_a$. Note que essa definição atribui capacidades aos (s,t)-cortes de D.

Dados $D = (V, A)$, $s, t \in V$ e $c : A \to \mathbb{Z}_+$, podemos definir o Problema do Corte Mínimo (PCM) como o problema de determinar um (s,t)-corte em D com a menor capacidade possível. Vamos provar que o PFM tem o PCM como seu dual.

Antes, vamos introduzir algumas notações. Definimos a força de um fluxo $f : A \to \mathbb{R}_+$ como $|f| = \sum_{a \in \delta^+(s)} f_a$. Dado um conjunto $S \subseteq V$, definimos $\delta^+(S) = \{(u, v) \in A : u \in S, v \notin S\}$ e $\delta^-(S) = \{(u, v) \in A : u \notin S, v \in S\}$. Fazemos, dado $A' \subseteq A$, $f(A') = \sum_{a \in A'} f_a$. Além disso, definimos também a função $inc : V \times A \to \{-1, 0, 1\}$ como

$$inc(v, a) = \begin{cases} 1, & \text{se } \exists u \in V : a = (v, u) \\ -1, & \text{se } \exists u \in V : a = (u, v) \\ 0, & \text{caso contrário} \end{cases}$$

Repare que *inc* representa a matriz de incidência de D como uma função. Isso vai ser muito útil.

Exercício 97. *Tome $D = (V, A)$. Prove que, se $S \subset V$, $s \in S$ e $t \notin S$, então $\delta^+(S)$ é um (s,t)-corte de D. Sob as mesmas condições, prove que $\delta^+(S)$ é um (s,t)-corte minimal. Ainda mais, para todo (s,t)-corte minimal A', prove que existe $S \subseteq V$ com $s \in S$ e $t \notin S$ tal que $A' = \delta^+(S)$.*

Exercício 98. *No programa linear (8.1)-(8.3), a conservação de fluxo é garantida por $\sum_{a \in \delta^+(v)} f_a = \sum_{a \in \delta^-(v)} f_a$, para todo $v \in V \setminus \{s, t\}$. Argumente que essa equação é equivalente a $\sum_{a \in A} inc(v, a) f_a = 0$.*

Vamos mostrar que, para qualquer $S \subseteq V$ com $s \in S$ e $t \notin S$, temos $|f| = f(\delta^+(S)) - f(\delta^-(S))$. Notamos que

$$|f| = \sum_{a \in A} inc(s, a) f_a$$

e, como o exercício 98 nos mostra, a conservação de fluxo garante que

$$\sum_{a \in A} inc(s, a) f_a = \sum_{v \in S} \sum_{a \in A} inc(v, a) f_a = \sum_{a \in A} f_a \sum_{v \in S} inc(v, a)$$

Note que $\sum_{v \in S} inc(v, a)$ é 1 quando $a \in \delta^+(S)$ e -1 quando $a \in \delta^-(S)$ (perceba que, de outra forma, a expressão assume valor nulo). Em vista disso,

escrevemos

$$\sum_{a \in A} f_a \sum_{v \in S} inc(v,a) = \sum_{a \in \delta^+(S)} f_a - \sum_{a \in \delta^-(S)} f_a = f(\delta^+(S)) - f(\delta^-(S))$$

e chegamos a $|f| = f(\delta^+(S)) - f(\delta^-(S))$

Exercício 99. *Prove que* $\sum_{a \in \delta^+(s)} f_a = \sum_{a \in \delta^-(t)} f_a$. *Dica: tome* $S = V \setminus \{t\}$ *e tire proveito das manipulações acima.*

Teorema 10. *Dado* (D, c)*, tome* f^* *uma solução ótima para PFM e* A^* *uma solução ótima para PCM. Temos que* $|f^*| \leq c(A^*)$.

Demonstração. Se provarmos que um fluxo viável qualquer tem sua força limitada superiormente por um corte qualquer, teremos o resultado do teorema. Tome $f : A \to \mathbb{R}_+$ um fluxo viável para $(D = (V, A), c)$. Tome também $A' \subseteq A$ um (s,t)-corte em D. Existe um (s,t)-corte minimal $\delta^+(S), S \subseteq V, s \in S, t \notin S$, contido em A'. Note que $c(\delta^+(S)) \leq c(A')$. Segue que $|f| = f(\delta^+(S)) - f(\delta^-(S)) \leq c(\delta^+(S)) - c(\delta^-(S)) \leq c(\delta^+(S)) \leq c(A')$. A inequação $|f| \leq c(A')$ vale para fluxos viáveis e (s,t)-cortes quaisquer, e portanto vale para f^* e A^*. □

Temos provado a dualidade fraca entre o PFM e o PCM. Para provar a dualidade forte entre esses problemas, vamos introduzir uma técnica construtiva para a obtenção de um fluxo máximo para uma instância (D, c) do PFM.

Dada uma instância $(D = (V, A), c)$ do PFM, começamos com o fluxo trivial $f_a = 0, \forall a \in A$. Tome $P = v_1, v_2, \ldots, v_k$, com $v_i \in V, i \in [k]$, uma sequência de vértices de D com as seguintes propriedades, para $i \in [k-1]$:

(1) se $a = (v_i, v_{i+1}) \in A$, então $f_a < c_a$;

(2) se $a = (v_{i+1}, v_i) \in A$, então $f_a > 0$.

Suponha que existe uma sequência P com $v_1 = s$ e $v_k = t$. Note que, se P é tal que a propriedade (1) vale para todo $i \in [k-1]$, então P representa um caminho de s a t que suporta a passagem de um pouco mais de fluxo. Por outro lado, se a propriedade (1) não vale para algum $i' \in [k-1]$ e a propriedade (2) vale para esse i', então P certamente não representa um caminho de s a t, mas evidencia que podemos manter a viabilidade diminuindo o fluxo que passa em $(v_{i'+1}, v_{i'})$ na mesma quantidade que aumentamos o fluxo que passa por $(v_{i'-1}, v_{i'})$ e por $(v_{i'+1}, v_{i'+2})$ (note que, como admitimos s fonte e t sumidouro, $i' \neq s$ e $i' + 1 \neq t$).

Tome $\Delta_i, i \in [k-1]$, definido como

$$\Delta_i = \begin{cases} c_a - f_a, & \text{se } a = (v_i, v_{i+1}) \in A \\ f_a, & \text{se } a = (v_{i+1}, v_i) \in A \end{cases}$$

Definimos $\Delta = \min_{i \in [k-1]} \Delta_i$ (Δ é uma espécie de "gargalo"). Note que $\Delta > 0$. Criamos um fluxo f' da seguinte forma

$$f'_a = \begin{cases} f_a + \Delta, & \text{se } \exists i \in [k-1] : a = (v_i, v_{i+1}) \\ f_a - \Delta, & \text{se } \exists i \in [k-1] : a = (v_{i+1}, v_i) \\ f_a, & \text{caso contrário} \end{cases}$$

Exercício 100. *Argumente que f' é um fluxo viável, isto é, não-negativo, conservativo e limitado por c.*

Exercício 101. *Argumente que $|f'| = |f| + \Delta$.*

Notamos que, se f é um fluxo viável tal que há uma sequência P de vértices com a propriedade descrita, com $v_1 = s$ e $v_k = t$, então certamente f não é um fluxo máximo. Sabemos, inclusive, construir um fluxo de maior força a partir de f. Mas e se f é viável e não existe uma tal sequência P começando em s e terminando em t?

Dado um fluxo viável f, considere S o conjunto de todos os vértices finais de sequências P maximais em relação à propriedade descrita. Se $t \in S$, podemos aplicar o procedimento descrito e f não tem força máxima. Se $t \notin S$, então $\delta^+(S)$ é um (s,t)-corte em D. Sabemos que $|f| = f(\delta^+(S)) - f(\delta^-(S))$. Pela maximalidade das sequências que "emprestam"seus vértices finais a S, temos $f(\delta^+(S)) = c(\delta^+(S))$ (por quê?) e $f(\delta^-(S)) = 0$ (por quê?). Com isso, concluímos que $|f| = c(\delta^+(S))$. O Teorema 10 nos diz que não pode haver um fluxo de força maior que $|f|$, nem um (s,t)-corte de capacidade menor que $c(\delta^+(S))$. Isso nos permite concluir que f é um fluxo máximo e que $\delta^+(S)$ é um (s,t)-corte mínimo.

Teorema 11. *Dada (D, c), se f^* é solução ótima para o PFM e A^* é solução ótima para o PCM, então $|f^*| = c(A^*)$.*

Note que, com esse resultado, não precisamos saber que a matriz de coeficientes do programa linear (8.1)-(8.3) é TU para garantir a integralidade de sua solução ótima. Basta observar que, como as capacidades dos arcos são inteiras, a capacidade de qualquer (s,t)-corte é inteira.

Aqui, desenvolvemos um método para determinar um fluxo de força máxima. Esse método é conhecido como Ford-Fulkerson. Além de sua aplicação óbvia, o método permite provar a dualidade forte entre PFM e PCM.

Exercício 102. *Desenvolva o dual do programa linear* (8.1)-(8.3)*. Argumente que o dual desenvolvido modela o PCM.*

Exercício 103. *Dois caminhos direcionados são ditos* ***arco-disjuntos*** *se eles não tem arcos em comum. Descreva um algoritmo para para determinar a maior quantidade de caminhos direcionados arco-disjuntos de um vértice para um outro vértice de um digrafo D. Dica: talvez você já conheça um algoritmo que faz quase isso, bastando apenas dar a ele uma entrada adequada.*

*Um digrafo D é dito k-**arco-conexo** sse é preciso remover ao menos k de seus arcos para desconectá-lo. Descreva um algoritmo para determinar o menor k para o qual um digrafo D é k-arco-conexo.*

Exercício 104. *Dois caminhos direcionados, ambos de u para v, são ditos* ***disjuntos*** *sse seus únicos vértices em comum são u e v. Descreva um algoritmo para para determinar a maior quantidade de caminhos direcionados disjuntos de um vértice para um outro vértice de um digrafo D.*

*Um digrafo D é dito k-**conexo** sse é preciso remover ao menos k de seus vértices para desconectá-lo. Descreva um algoritmo para determinar o menor k para o qual um digrafo D é k-conexo.*

Exercício 105. *Um grafo $G = (V, E)$ é bipartido quando existem $V_1, V_2 \subseteq V$ com $V_1 \cup V_2 = V$ e $V_1 \cap V_2 = \emptyset$ tais que toda aresta tem um extremo em V_1 e o outro em V_2. Desenvolva um algoritmo para resolver o problema do emparelhamento máximo para grafos bipartidos.*

Exercício 106. *Um hospital tem o seguinte estoque de bolsas de sangue: 46 do tipo A, 34 do tipo B, 45 do tipo O e 45 do tipo AB. Um acidente terrível ocorreu e há diversas pessoas necessitando de transfusão: 39 tem sangue do tipo A, 38 do tipo B, 42 do tipo O e 50 do tipo AB. Cada uma dessas pessoas precisa de exatamente uma bolsa de sangue para sobreviver até haver condições de um atendimento apropriado. Determine o número mínimo de pessoas que não poderão esperar pelo atendimento apropriado, e portanto morrerão. Ah, e tem mais: pacientes tipo A recebem sangue A ou O, pacientes tipo B recebem sangue B ou O, pacientes tipo O recebem sangue O e pacientes tipo AB recebem qualquer sangue.*