Metody numeryczne

Przemysław Kiciak

wykład, III rok Informatyki

Uniwersytet Warszawski, Wydział MIM
rok akad. 2013/2014

Zasady zaliczania przedmiotu

Na zaliczenie przedmiotu składają się: zaliczenie ćwiczeń i zdanie egzaminu. Połowa ćwiczeń ma miejsce w laboratorium, pozostałe ćwiczenia są w sali przy tablicy. Na końcową ocenę składają się

- punkty, którymi prowadzący ćwiczenia ocenił prace domowe, tj. rozwiązania zadań na kartce,
- punkty za rozwiązania zadań programistycznych,
- punkty zdobyte na egzaminie pisemnym.

Po egzaminie pisemnym będą wystawione propozycje ocen, w których zadania domowe, zadania programistyczne i egzamin pisemny mają udziały odpowiednio 25%, 25% i 50%, przy czym z każdego z tych elementów trzeba zdobyć co najmniej 25% punktów, a w sumie co najmniej 50%. Otrzymaną propozycję oceny uczestnik zajęć może przyjąć, lub wystawić na ryzyko zmiany na egzaminie ustnym.

Literatura

- Kincaid D., Cheney W.: *Analiza numeryczna*, WNT, Warszawa, 2006.
- Krzyżanowski P.: *Obliczenia inżynierskie i naukowe*, PWN, Warszawa, 2011.
- Jankowska J., Jankowski M., Dryja M: *Przegląd metod i algorytmów numerycznych* cz. 1 i 2, WNT, Warszawa, 1988.
- Dahlquist G., Björck Å: *Metody numeryczne*, PWN, Warszawa, 1983.

1. Rozwiązywanie równań nieliniowych

Rozważamy zadanie znalezienia liczby x, takiej że

$$f(x) = 0,$$

mając do dyspozycji podprogram obliczający wartość funkcji f dla argumentu x podanego jako parametr. Możemy na ogół znaleźć *tylko* pewne przybliżenie rozwiązania.

Mając do czynienia z takim zadaniem, zawsze musimy wiedzieć coś więcej o funkcji f:

- Czy rozwiązanie istnieje?
- Czy istnieje więcej niż jedno? A może nieskończenie wiele?

Jeśli rozwiązań jest więcej, to czy mamy znaleźć wszystkie, kilka, czy tylko jedno, obojętnie które, albo spełniające jakiś dodatkowy warunek?

Aby wybrać algorytm rozwiązywania zadania, musimy wiedzieć też w jakim zbiorze funkcja f jest określona i czy jest ciągła, przyda się też wiedza np. czy ciągła jest jej pochodna rzędu 1, 2 i być może dalsze. W niektórych metodach oprócz podprogramu obliczającego $f(x)$ będzie też potrzebny podprogram obliczający $f'(x)$, a nawet dalsze pochodne.

Metoda Newtona

Niech A oznacza ograniczony przedział domknięty, w którym jest określona funkcja rzeczywista f klasy C^2. Chcemy znaleźć w tym przedziale miejsce zerowe funkcji f, o którym założymy, że istnieje i jest tylko jedno (*zawsze to trzeba sprawdzić*).

Napiszemy wzór Taylora:

$$f(x+h) = \frac{f(x)}{0!} + \frac{f'(x)}{1!}h + \frac{f''(\xi)}{2!}h^2.$$

Rozumiemy go tak: jeśli liczby x oraz $x+h$ należą do przedziału A, w którym funkcja f jest klasy C^2, to istnieje liczba ξ, leżąca pomiędzy x oraz $x+h$, taka że powyższa równość zachodzi.

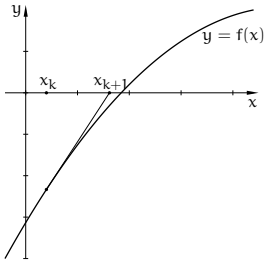
Metoda Newtona (znana też jako metoda stycznych lub metoda Newtona-Raphsona) jest następująca: wybieramy liczbę x_0, która jest przybliżeniem miejsca zerowego funkcji f, a następnie konstruujemy rekurencyjnie elementy ciągu $x_1, x_2, \ldots$, w taki sposób: mając x_k, określamy wielomian

$$w_k(h) = f(x_k) + f'(x_k)h.$$

Znajdujemy miejsce zerowe δ wielomianu w_k i przyjmujemy $x_{k+1} = x_k + \delta$. Mamy stąd formułę

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Interpretacja geometryczna: wykres funkcji f jest gładką krzywą, przechodzącą przez punkt $(x_k, f(x_k))$. Konstruujemy prostą styczną do wykresu w tym punkcie i przyjmujemy za x_{k+1} punkt przecięcia stycznej z osią x.



Znajdziemy pewne warunki wystarczające, aby ciąg $(x_k)_{k\in\mathbb{N}}$ dla dowolnego $x_0 \in A$ zbiegał do rozwiązania, które oznaczymy literą α.

Zauważamy, że w żadnym punkcie tego ciągu pochodna funkcji f nie może być zerowa. Naturalne jest założenie, że w przedziale A pochodna znaku nie zmienia, co więcej, zachodzi nierówność $|f'(x)| \geqslant K$ dla pewnej stałej $K > 0$. Ponieważ f jest klasy $C^2(A)$, istnieje stała M, taka że $|f''(x)| \leqslant M$ dla każdego $x \in A$.

Oznaczmy $\varepsilon_k = x_k - \alpha$ — jest to błąd aproksymacji rozwiązania przez k-ty element ciągu.

Na podstawie wzoru Taylora

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + \frac{1}{2}f''(\xi_k)(\alpha - x_k)^2.$$

Liczba ξ_k leży między α i x_k. Dzielimy strony przez $f'(x_k)$:

$$0 = \frac{f(x_k)}{f'(x_k)} + \alpha - x_k + \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2 =$$
$$\frac{f(x_k)}{f'(x_k)} + \alpha - x_{k+1} + x_{k+1} - x_k + \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2.$$

Ponieważ $x_{k+1} - x_k = -\frac{f(x_k)}{f'(x_k)}$, mamy stąd

$$\varepsilon_{k+1} = \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2. \qquad (*)$$

Możemy oszacować

$$|\varepsilon_{k+1}| \leqslant \frac{M}{2K}|\varepsilon_k|^2.$$

Aby zachodziła nierówność $|\varepsilon_{k+1}| < |\varepsilon_k|$, wystarczy, że $\frac{M}{2K}|\varepsilon_k| < 1$, czyli

$$|\varepsilon_k| < \frac{2K}{M}.$$

Jeśli błąd przybliżenia rozwiązania przez punkt x_0, z którego zaczynamy, spełnia tę nierówność, to każdy następny błąd ma mniejszą wartość bezwzględną niż poprzedni, co więcej, ciąg błędów zbiega do zera.

Zbadajmy szybkość zbieżności metody. Wybierzmy dowolną podstawę logarytmu większą niż 1 i oznaczmy $a_k = \log|\varepsilon_k|$, $g(k) = \log\left|\frac{f''(\xi_{k-1})}{2f'(x_{k-1})}\right|$. Na podstawie równości (*) możemy napisać równanie różnicowe

$$a_k = 2a_{k-1} + g(k).$$

Niech $G = \log\frac{M}{2K}$. Jeśli rozważymy równanie uproszczone,

$$\tilde{a}_k = 2\tilde{a}_{k-1} + G,$$

dla którego przyjmiemy $\tilde{a}_0 = a_0 < -G$, to dla każdego k mamy

$$a_k \leqslant \tilde{a}_k = (a_0 + G)\cdot 2^k - G.$$

Ciąg $(\tilde{a}_k)_{k\in\mathbb{N}}$ dąży wykładniczo do $-\infty$, a ciąg $(a_k)_{k\in\mathbb{N}}$ dąży do $-\infty$ co najmniej tak samo szybko. To zaś oznacza, że jeśli x_k jest przybliżeniem rozwiązania, które ma n cyfr dokładnych, to x_{k+1} będzie mieć w przybliżeniu 2n cyfr dokładnych. Zatem zbieżność jest bardzo szybka. Na podstawie powyższej nierówności, znając oszacowanie $|\varepsilon_0|$ i G oraz tolerancję błędu, można oszacować liczbę iteracji wystarczającą do otrzymania rozwiązania z błędem w granicach tej tolerancji.

Uwaga. Można udowodnić zbieżność metody przy słabszych założeniach, np. że funkcja f niekoniecznie jest klasy C^2, ale jej pochodna spełnia warunek Lipschitza.

Podstawowe pojęcia w numerycznym rozwiązywaniu równań

- funkcja iteracyjna
- kula zbieżności
- wykładnik zbieżności
- maksymalna graniczna dokładność

Funkcja iteracyjna jest to funkcja φ, za pomocą której konstruujemy ciąg $x_0, x_1, \ldots$, według wzoru

$$x_{k+1} = \varphi(x_k).$$

W metodzie Newtona funkcja iteracyjna jest określona wzorem

$$\varphi_N(x) = x - \frac{f(x)}{f'(x)}.$$

Funkcja iteracyjna powinna być tak skonstruowana, aby rozwiązanie α było jej punktem stałym, tj. aby było $\varphi(\alpha) = \alpha$.

Istnieje nieskończenie wiele możliwości „przerobienia” równania $f(x) = 0$ na równoważne równanie $x = \varphi(x)$. W najprostszym przypadku możemy wziąć

$$\varphi(x) = x - \tau f(x),$$

z jakimś parametrem rzeczywistym τ. Oczywiście, nie zawsze otrzymana w ten sposób funkcja φ prowadzi do otrzymania ciągu zbieżnego. Aby zbieżność miała miejsce, trzeba, by w otoczeniu rozwiązania α funkcja φ była odwzorowaniem zwężającym (może mieć np. pochodną o wartości bezwzględnej mniejszej od 1).

Funkcje iteracyjne dla pewnych metod są bardziej skomplikowane. Argumentem funkcji iteracyjnej oprócz ostatniego przybliżenia może być także jedno lub więcej poprzednich (czasami takie metody nazywa się metodami z pamięcią).

Na przykład w metodzie siecznych, o której będzie mowa dalej, potrzebne są dwa przybliżenia, które nie mogą być jednakowe. Funkcja iteracyjna ma w tej metodzie postać

$$\varphi_S(x, y) = x - \frac{f(x)}{f[x, y]}, \quad \text{gdzie} \quad f[x, y] = \frac{f(x) - f(y)}{x - y},$$

a w kolejnych iteracjach obliczamy $x_{k+1} = \varphi_S(x_k, x_{k-1})$.

Funkcja iteracyjna może też w jawny sposób zależeć od numeru iteracji, k — w tym przypadku mówimy o metodzie niestacjonarnej.

Kula zbieżności rozwiązania α jest to największa kula B o środku α (w przypadku równań z jedną niewiadomą jest to przedział symetryczny względem α), taka że jeśli wybierzemy dowolny punkt startowy x_0 wewnątrz tej kuli, to ciąg $(x_k)_{k\in\mathbb{N}}$ zbiega do α.
Znalezienie kuli zbieżności jest na ogół bardzo trudne, więc tego nie robimy, ale możemy szacować jej promień r.
Na przykład, dla metody Newtona $r \geqslant \frac{2K}{M}$.

Jeśli równanie ma kilka rozwiązań, to każde z nich ma własną kulę zbieżności i wszystkie te kule są rozłączne. Kule zbieżności pewnych rozwiązań mogą być zbiorem pustym — wtedy metoda na ogół nie jest w stanie takich rozwiązań znaleźć.

Jeśli punkt startowy nie należy do kuli zbieżności żadnego rozwiązania, to metoda może znaleźć rozwiązanie, jeśli otrzymany po pewnej liczbie iteracji punkt „wpadł” do kuli zbieżności. Tylko, że *nie należy* liczyć na taki przypadek.

W analizie metod numerycznych często przydaje się

Twierdzenie Banacha o punkcie stałym: *jeśli zbiór X z metryką ρ jest zupełną przestrzenią metryczną, a funkcja $\varphi\colon X \to X$ jest przekształceniem zwężającym (tj. istnieje stała $L < 1$, taka że $\forall_{a,b\in X}\, \rho(\varphi(a), \varphi(b)) \leqslant L\rho(a, b)$), to funkcja φ ma jeden punkt stały w zbiorze X.*

Wykazanie, że metoda działa, tj. wytwarza ciąg zbieżny do rozwiązania, często sprowadza się do znalezienia (wykazania istnienia lub oszacowania promienia) kuli X zawartej w kuli zbieżności, w której funkcja iteracyjna φ jest przekształceniem zwężającym.

Wykładnik zbieżności metody opisuje asymptotyczną szybkość zbieżności ciągu $(x_k)_{k\in\mathbb{N}}$ do rozwiązania. Przeprowadzony rachunek dla metody Newtona dowiódł, że *jeśli funkcja f spełnia uczynione założenia*, to wykładnik zbieżności jest nie mniejszy niż 2.

Formalna definicja: wykładnik zbieżności jest to największa liczba p, taka że istnieją stałe K i $C < +\infty$, takie że dla każdego $k \geqslant K$ zachodzi nierówność

$$|\varepsilon_{k+1}| \leqslant C|\varepsilon_k|^p, \quad \text{czyli} \quad \log|\varepsilon_{k+1}| \leqslant \log C + p\log|\varepsilon_k|.$$

Wykładnik zbieżności powinien być większy lub równy 1, przy czym jeśli $p = 1$, to oczywiście musi być $C < 1$.

Przykładem metody o wykładniku zbieżności 1 jest metoda bisekcji: w każdej iteracji otrzymujemy przybliżenie rozwiązania z oszacowaniem błędu mniejszym o połowę.

Również metoda Newtona ma wykładnik zbieżności 1, jeśli nie jest spełnione założenie, że pochodna funkcji f w otoczeniu rozwiązania jest niezerowa. Jeśli $p > 1$, to dla ustalonego K istnieją stałe a i b, takie że dla każdego $k > K$

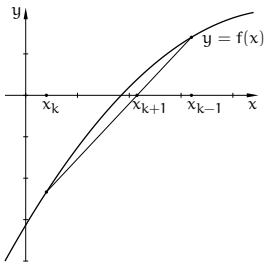
$$\log|\varepsilon_k| \leqslant a + (\log|\varepsilon_K| + b)p^{k-K}.$$

Maksymalna graniczna dokładność oznacza maksymalną dokładność osiągalną w obliczeniach.

Analiza metody Newtona była przeprowadzona przy założeniu, że nie ma błędów, tj. zarówno wartości funkcji f i pochodnej w x_k są obliczane dokładnie, jak i w końcowych działaniach obliczenia wartości funkcji iteracyjnej nie ma błędów. *Błędy jednak są* i ograniczają możliwą do uzyskania dokładność rozwiązania. Za rozwiązanie metoda może przyjąć dowolny punkt przedziału, w którym błąd obliczonej wartości funkcji f jest większy lub równy 100%. Jeśli pochodna funkcji jest bliska 0, to ten przedział może być długi.

Metoda siecznych

Wadą metody Newtona jest konieczność obliczania wartości pochodnej funkcji f. Metoda siecznych jest modyfikacją metody Newtona, w której pochodna została zastąpiona przez różnicę dzieloną (albo iloraz różnicowy, jak kto woli), czyli pewne przybliżenie pochodnej. Mając *dwa różne* przybliżenia rozwiązania, x_k i x_{k-1}, prowadzimy prostą przez punkty $(x_k, f(x_k))$ i $(x_{k-1}, f(x_{k-1}))$. Prosta ta przecina (siecze) wykres funkcji f w tych punktach, i w tym sensie jest jego sieczną.



Skonstruowana sieczna jest wykresem wielomianu pierwszego stopnia. Punkt x_{k+1} jest miejscem zerowym tego wielomianu. W metodzie siecznych należy podać dwa początkowe przybliżenia rozwiązania, x_0 i x_1, a następnie w każdej iteracji obliczać

$$x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, x_{k-1}]},$$

gdzie

$$f[x_k, x_{k-1}] = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Aby dokonać analizy metody siecznych, użyjemy pewnego uogólnienia wzoru Taylora:

$$f(z) = f(x) + f[x, y](z - x) + \frac{f''(\xi)}{2!}(z - x)(z - y).$$

Wzór ten jest szczególnym przypadkiem wzoru opisującego resztę interpolacyjną Hermite'a (będzie on udowodniony później). Podany wzór rozumiemy w ten sposób, że jeśli liczby x, y, z leżą w przedziale A, w którym funkcja f jest klasy C^2, to istnieje $\xi \in A$, takie że podana wyżej równość zachodzi (liczba ξ leży między najmniejszą i największą spośród tych trzech liczb).

Jak poprzednio, α oznacza poszukiwane rozwiązanie, zaś $\varepsilon_k = x_k - \alpha$. Liczymy

$$0 = f(\alpha) = f(x_k) + f[x_k, x_{k-1}](\alpha - x_k) + \frac{f''(\xi_k)}{2}(\alpha - x_k)(\alpha - x_{k-1})$$

i dzielimy stronami przez $f[x_k, x_{k-1}]$:

$$0 = \frac{f(x_k)}{f[x_k, x_{k-1}]} + \alpha - x_{k+1} + x_{k+1} - x_k + \frac{f''(\xi_k)}{2f[x_k, x_{k-1}]}(\alpha - x_k)(\alpha - x_{k-1}),$$

skąd, po skróceniu podkreślonych składników, otrzymujemy

$$0 = \alpha - x_{k+1} + \frac{f''(\xi_k)}{2f[x_k, x_{k-1}]}(\alpha - x_k)(\alpha - x_{k-1}).$$

Po uporządkowaniu i uwzględnieniu faktu, że istnieje liczba η_k położona między x_k i x_{k-1}, taka że $f[x_k, x_{k-1}] = f'(\eta_k)$, mamy stąd równość

$$\varepsilon_{k+1} = \frac{f''(\xi_k)}{2f'(\eta_k)}\varepsilon_k\varepsilon_{k-1}. \qquad (**)$$

Jeśli, jak poprzednio, możemy oszacować $|f'(x)| \geqslant K > 0$ i $|f''(x)| \leqslant M$ dla każdego $x \in A$, to mamy

$$|\varepsilon_{k+1}| \leqslant \frac{M}{2K}|\varepsilon_k||\varepsilon_{k-1}|.$$

Jeśli oba błędy, ε_k i ε_{k-1}, mają wartości bezwzględne mniejsze niż $\frac{2K}{M}$, to wartości bezwzględne kolejnych błędów będą coraz mniejsze — w ten sposób mamy oszacowany promień kuli zbieżności.

Aby zbadać rząd zbieżności, oznaczmy $a_k = \log|\varepsilon_k|$ oraz $g(k) = \log\left|\frac{f''(\xi_{k-1})}{f'(\eta_{k-1})}\right|$ i $G = \log\left|\frac{M}{2K}\right|$. Na podstawie (**) możemy napisać równanie różnicowe drugiego rzędu,

$$a_k = a_{k-1} + a_{k-2} + g(k),$$

i jego uproszczoną wersję

$$\tilde{a}_k = \tilde{a}_{k-1} + \tilde{a}_{k-2} + G.$$

Dla ustalonych wyrazów początkowych, $\tilde{a}_0 = a_0$ i $\tilde{a}_1 = a_1$, istnieją liczby a, b, c, takie że

$$\tilde{a}_k = a\lambda_1^k + b\lambda_2^k + c, \quad \text{gdzie} \quad \lambda_1 = \frac{1 - \sqrt{5}}{2},\ \lambda_2 = \frac{1 + \sqrt{5}}{2},$$

i jeśli liczby a_0 i a_1 są dostatecznie małe, to $b < 0$.

Składnik $b\lambda_2^k$ dominuje i dla $b < 0$ ciąg $(\tilde{a}_k)_{k\in\mathbb{N}}$ zbiega wykładniczo do $-\infty$. Zachodzi też nierówność $a_k \leqslant \tilde{a}_k$ dla każdego k, w związku z czym, dla dostatecznie dużych k, jeśli przybliżenie x_k rozwiązania α ma n cyfr dokładnych, to przybliżenie x_{k+1} będzie ich miało około $\lambda_2 n$. Wykładnik zbieżności metody siecznych, $\lambda_2 \approx 1.618$, jest ułamkiem.

Metoda siecznych ma mniejszy wykładnik zbieżności niż metoda Newtona, ale jedna jej iteracja jest tańsza — odpada obliczanie pochodnej. Okazuje się, że jeśli zadamy tolerancję ε dopuszczalnego błędu, to metoda siecznych może znaleźć dostatecznie dokładne rozwiązanie szybciej (w większej liczbie iteracji, z których każda zajmuje mniej czasu). Z tego punktu widzenia, jeśli koszt obliczania różnicy dzielonej uznamy za nieistotny, to metoda Newtona jest opłacalna, gdy koszt obliczania pochodnej nie przewyższa ok. 0.44 kosztu obliczania wartości funkcji f.

Metoda Newtona dla układu równań

Rozważamy teraz zadanie znalezienia wspólnego miejsca zerowego n rzeczywistych funkcji skalarnych, których argumentami jest n zmiennych rzeczywistych. Możemy zatem napisać układ w postaci rozwiniętej:

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

lub „zwiniętej"

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}.$$

Funkcja $\mathbf{f}$ jest określona w pewnym obszarze A przestrzeni $\mathbb{R}^n$ i ma wartości w $\mathbb{R}^n$.

Niech $\mathbf{h} = [h_1, \ldots, h_n]^T$. Dla funkcji *skalarnej* f_i klasy $C^2(A)$, możemy napisać wzór Taylora:

$$f_i(\mathbf{x}+\mathbf{h}) = \frac{1}{0!}f_i(\mathbf{x}) + \frac{1}{1!}Df_i|_{\mathbf{x}}(\mathbf{h}) + \frac{1}{2!}D^2f_i|_{\xi_i}(\mathbf{h},\mathbf{h}).$$

Rozumiemy go tak: jeśli obszar A zawiera odcinek o końcach $\mathbf{x}$ i $\mathbf{x}+\mathbf{h}$, to istnieje punkt ξ_i na tym odcinku, taki że powyższa równość zachodzi. Symbol $Df_i|_{\mathbf{x}}$ oznacza <u>różniczkę</u> funkcji f_i w punkcie $\mathbf{x}$, czyli przekształcenie liniowe, które dowolnemu wektorowi $\mathbf{h}$ przyporządkowuje liczbę

$$Df_i|_{\mathbf{x}}(\mathbf{h}) = \frac{\partial f_i}{\partial x_1}\Big|_{\mathbf{x}} h_1 + \cdots + \frac{\partial f_i}{\partial x_n}\Big|_{\mathbf{x}} h_n.$$

Wartością tego przekształcenia jest zatem iloczyn skalarny gradientu funkcji f_i w punkcie $\mathbf{x}$ i wektora $\mathbf{h}$.

Symbol $D^2f_i|_{\xi_i}$ oznacza <u>różniczkę drugiego rzędu</u>, tj. przekształcenie *dwuliniowe*, którego wartością dla pary wektorów $(\mathbf{g},\mathbf{h})$ jest liczba

$$D^2f_i|_{\xi_i}(\mathbf{g},\mathbf{h}) = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^2 f_i}{\partial x_j \partial x_k}\Big|_{\xi_i} g_j h_k.$$

Drobny kłopot (o którym *nie należy* zapominać) jest taki, że punkt ξ_i dla każdego i może być różny, dlatego nie można tak prosto zapisać odpowiedniego wzoru dla funkcji wektorowej $\mathbf{f}$. Niemniej, ze wzoru Taylora wynika, że jeśli obszar A zawiera odcinek $\overline{\mathbf{x},\mathbf{x}+\mathbf{h}}$, to dla wektorowej funkcji $\mathbf{f}$ klasy $C^2(A)$ zachodzi równość

$$\mathbf{f}(\mathbf{x}+\mathbf{h}) = \mathbf{f}(\mathbf{x}) + Df|_{\mathbf{x}}(\mathbf{h}) + \mathbf{r}, \qquad \binom{**}{*}$$

Symbol $Df|_{\mathbf{x}}$ oznacza różniczkę przekształcenia $\mathbf{f}$ w punkcie $\mathbf{x}$, a ponadto istnieje macierz B (zależna od $\mathbf{x}$ i $\mathbf{h}$) o wymiarach $n \times n$ i współczynnikach *wektorowych*

$$\mathbf{b}_{jl} = \left[\frac{\partial^2 f_1}{\partial x_j \partial x_l}\Big|_{\xi_1}, \ldots, \frac{\partial^2 f_n}{\partial x_j \partial x_l}\Big|_{\xi_n}\right]^T \in \mathbb{R}^n,$$

taka że reszta we wzorze $\binom{**}{*}$ jest równa

$$\mathbf{r} = \mathbf{h}^T B \mathbf{h} = \sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{b}_{jl} h_j h_l, \qquad \binom{**}{**}$$

i spełnia oszacowanie

$$\|\mathbf{r}\| \leqslant \frac{M}{2}\|\mathbf{h}\|^2$$

dla pewnej stałej M (stała ta jest określona przez pochodne drugiego rzędu funkcji f_i w obszarze A i przez używaną normę).

<u>Metoda Newtona</u> polega na tym, że mając przybliżenie $\mathbf{x}_k$ rozwiązania $\boldsymbol{\alpha}$, konstruujemy przekształcenie afiniczne $\mathbb{R}^n \to \mathbb{R}^n$, określone przez pierwsze dwa składniki po prawej stronie wzoru $\binom{**}{*}$, a następnie przyjmujemy za $\mathbf{x}_{k+1}$ miejsce zerowe tego przekształcenia. Czyli

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (Df|_{\mathbf{x}_k})^{-1}\big(\mathbf{f}(\mathbf{x}_k)\big).$$

Aby obliczyć $\mathbf{x}_{k+1}$, należy obliczyć wektor $\mathbf{f}_k = \mathbf{f}(\mathbf{x}_k)$ oraz macierz pochodnych cząstkowych pierwszego rzędu

$$J_k = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}\Big|_{\mathbf{x}_k} & \cdots & \frac{\partial f_1}{\partial x_n}\Big|_{\mathbf{x}_k} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}\Big|_{\mathbf{x}_k} & \cdots & \frac{\partial f_n}{\partial x_n}\Big|_{\mathbf{x}_k} \end{bmatrix}$$

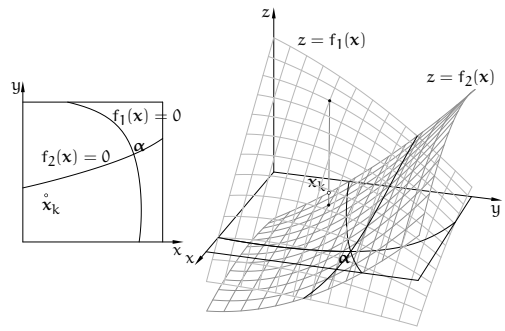
zwaną <u>jakobianem</u>, która reprezentuje różniczkę funkcji $\mathbf{f}$ w punkcie $\mathbf{x}_k$,

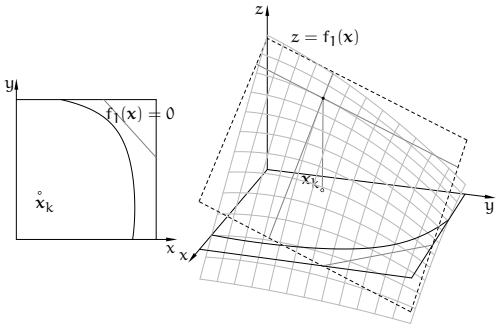
a następnie rozwiązać układ równań liniowych

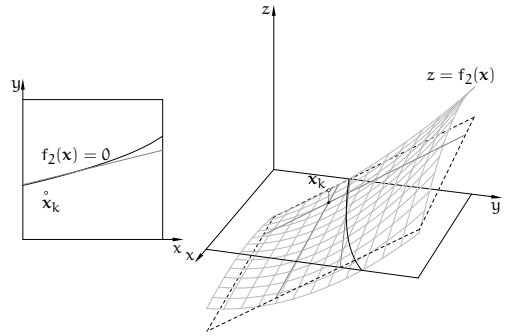
$$J_k\boldsymbol{\delta} = -\mathbf{f}_k$$

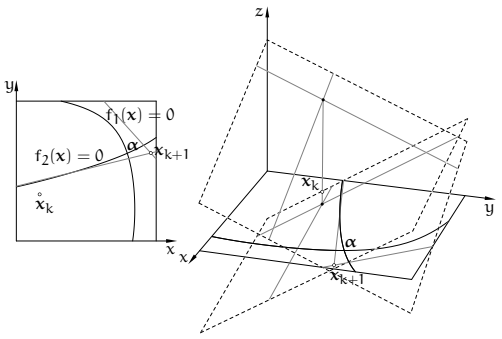
i obliczyć $\mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\delta}$. Oczywiście, aby to obliczenie było wykonalne, macierz J_k musi być nieosobliwa.

Ilustrację kroku metody Newtona dla układu dwóch równań przedstawia seria obrazków.









Aby znaleźć wykładnik zbieżności przyjmiemy założenie, że istnieje taka stała K, że dla każdego punktu $\mathbf{x}$ w rozpatrywanym obszarze A różniczka przekształcenia $\mathbf{f}$ spełnia warunek $\|(D\mathbf{f}|_{\mathbf{x}})^{-1}\| \leqslant K^{-1}$. Zatem, dla $\mathbf{x}_k \in A$ jest $\|J_k^{-1}\| \leqslant K^{-1}$. Na podstawie wzorów $(\overset{*}{*})$ i $(\overset{**}{**})$, mamy

$$\mathbf{0} = \mathbf{f}(\boldsymbol{\alpha}) = \mathbf{f}(\mathbf{x}_k) + J_k(\boldsymbol{\alpha} - \mathbf{x}_k) + (\boldsymbol{\alpha} - \mathbf{x}_k)^T B_k(\boldsymbol{\alpha} - \mathbf{x}_k),$$

Dalej postępujemy identycznie, jak w przypadku skalarnym. Oznaczamy $\boldsymbol{\varepsilon}_k = \mathbf{x}_k - \boldsymbol{\alpha}$. Strony równości mnożymy przez J_k^{-1}, oraz odejmujemy i dodajemy $\mathbf{x}_{k+1}$ i skracamy:

$$\mathbf{0} = \underbrace{J_k^{-1}\mathbf{f}(\mathbf{x}_k)} + \boldsymbol{\alpha} - \mathbf{x}_{k+1} + \underbrace{\mathbf{x}_{k+1} - \mathbf{x}_k} + J_k^{-1}\left(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\right) = \boldsymbol{\alpha} - \mathbf{x}_{k+1} + J_k^{-1}\left(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\right).$$

Stąd wielkość błędu kolejnego przybliżenia rozwiązania,

$$\boldsymbol{\varepsilon}_{k+1} = J_k^{-1}\left(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\right),$$

możemy oszacować tak:

$$\|\boldsymbol{\varepsilon}_{k+1}\| \leqslant \frac{M}{2K}\|\boldsymbol{\varepsilon}_k\|^2.$$

Jeśli funkcja $\mathbf{f}$ spełnia przyjęte założenia, to wykładnik zbieżności metody Newtona jest równy 2 — końcowy rachunek (z rozwiązywaniem równania różnicowego) jest identyczny jak dla równania z jedną niewiadomą.

Modyfikacje

Metoda Newtona dla układu równań może być dość kosztowna: oprócz wartości funkcji $\mathbf{f}$, składającej się z n liczb, trzeba obliczyć macierz J_k, tj. w ogólności n^2 liczb, a następnie rozwiązać układ równań, co może wymagać wykonania $\Theta(n^3)$ działań zmiennopozycyjnych. Ze wzrostem liczby równań i niewiadomych koszty te mogą stać się zaporowe. Dla bardzo dużych n często macierz J_k jest *rzadka*, tj. ma znacznie mniej niż n^2 współczynników niezerowych. W takim przypadku należy po pierwsze obliczać tylko współczynniki niezerowe (ich rozmieszczenie w macierzy należy wyznaczyć zawczasu), a ponadto użyć metody rozwiązywania układu równań liniowych dostosowanej do macierzy rzadkiej.

Często stosuje się rozmaite modyfikacje metody Newtona. Po pierwsze, zamiast obliczać współczynniki macierzy J_k na podstawie dokładnych wzorów, które mogą być znacznie bardziej skomplikowane (czyli droższe) niż wzory opisujące funkcje f_i, można obliczać różnice dzielone; w tym celu trzeba obliczyć wartości funkcji $\mathbf{f}$ w $n+1$ punktach.

Jeśli punkty $\mathbf{x}_{k-n}, \ldots, \mathbf{x}_k$ są w położeniu ogólnym, tj. wektory $\mathbf{x}_j - \mathbf{x}_k$ dla $j = k-n, \ldots, k-1$ są liniowo niezależne, to można obliczyć przybliżenie $\tilde{J}_k$ macierzy J_k na podstawie wartości funkcji $\mathbf{f}$ w tych punktach. W ten sposób powstaje wielowymiarowa metoda siecznych. Różniczka przekształcenia afinicznego $\tilde{\mathbf{f}}\colon \mathbb{R}^n \to \mathbb{R}^n$, które w punktach $\mathbf{x}_{k-n}, \ldots, \mathbf{x}_k$ przyjmuje wartości $\mathbf{f}_{k-n}, \ldots, \mathbf{f}_k$, jest taka sama w każdym punkcie przestrzeni i spełnia warunek

$$D\mathbf{f}(\mathbf{x} - \mathbf{x}_k) = \mathbf{f}(\mathbf{x}) - \mathbf{f}_k,$$

z którego wynika równość

$$\tilde{J}_k X = F,$$

gdzie $\tilde{J}_k$ oznacza jakobian przekształcenia $\tilde{\mathbf{f}}$, zaś

$$X = [\mathbf{x}_{k-n} - \mathbf{x}_k, \ldots, \mathbf{x}_{k-1} - \mathbf{x}_k], \quad F = [\mathbf{f}_{k-n} - \mathbf{f}_k, \ldots, \mathbf{f}_{k-1} - \mathbf{f}_k].$$

Jeśli więc macierze X i F są nieosobliwe, to mamy $\tilde{J} = FX^{-1}$ oraz $\tilde{J}^{-1} = XF^{-1}$. W $k+$ pierwszym kroku metody siecznych rozwiązujemy układ równań

$$F\boldsymbol{\beta} = -\mathbf{f}_k,$$

po czym obliczamy

$$\boldsymbol{\delta} = X\boldsymbol{\beta} \quad \text{i} \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\delta}.$$

Koszt tego obliczenia w ogólnym przypadku jest rzędu n^3 operacji. Wadą wielowymiarowej metody siecznych jest bardzo mały wykładnik zbieżności (bliski 1) dla dużych n.

Kolejna modyfikacja polega na wykorzystaniu macierzy J_k w kilku kolejnych iteracjach. To również obniża wykładnik zbieżności, ale dodatkowe iteracje z tą samą macierzą są bardzo tanie: nie trzeba obliczać pochodnych i można skorzystać z „gotowych” czynników (np. trójkątnych) rozkładu macierzy. Koszt rzędu n^3 w rozwiązywaniu układów równań liniowych jest związany z rozkładaniem macierzy na te czynniki, mając je, można rozwiązać układ kosztem $\Theta(n^2)$ działań.

Istnieją modyfikacje metody Newtona, mające na celu „powiększenie” kuli zbieżności poszukiwanych rozwiązań. Dla nie dość dobrego punktu $\mathbf{x}_k$ często zdarza się, że przyrost δ, otrzymany przez rozwiązanie układu równań $J_k\delta = -\mathbf{f}_k$ jest za duży. Wtedy można przyjąć $\mathbf{x}_{k+1} = \mathbf{x}_k + \beta\delta$, dla odpowiednio wybranego parametru $\beta \in (0, 1)$. Metoda skuteczniejsza, choć bardziej kosztowna, polega na wyznaczeniu przyrostu przez rozwiązanie układu równań

$$(J_k + \lambda I)\delta = -\mathbf{f}_k,$$

z odpowiednio wybranym parametrem λ. Metoda ta może być też skuteczna w pewnych przypadkach, gdy macierz J_k jest osobliwa. Parametr λ dobieramy tak, aby otrzymać jak najmniejsze residuum układu, tj. aby zminimalizować normę wektora $\mathbf{f}_{k+1}$. Po pewnej liczbie iteracji możemy otrzymać przybliżenie rozwiązania należące do kuli zbieżności metody Newtona i od tej chwili przyjmować $\lambda = 0$.

Kryteria stopu

Ważną decyzją w obliczeniach jest, kiedy je przerwać. Na przykład wykonywanie kolejnych iteracji po osiągnięciu maksymalnej granicznej dokładności jest stratą czasu. Dlatego w pętli, realizującej iteracje, musi się pojawić jedna lub więcej instrukcji przerywających obliczenia po spełnieniu pewnego warunku.

Po pierwsze, można dać limit liczby iteracji, np. określony przez parametr procedury. W *wielu* typowych zastosowaniach, jeśli metoda Newtona nie znalazła rozwiązania (z graniczną dokładnością) po siedmiu iteracjach, to już nie znajdzie (bo funkcja nie spełnia warunków koniecznych działania metody, zaczęliśmy od złego przybliżenia startowego, lub w ogóle nie ma rozwiązania).

Drugie kryterium stopu jest residualne. Residuum równania w punkcie x_k jest to liczba $f(x_k)$ (lub wektor $\mathbf{f}(\mathbf{x}_k)$). Jeśli residuum ma dostatecznie małą wartość bezwzględną (lub normę, dla układu równań), na przykład porównywalną z oszacowaniem błędu, z jakim obliczamy wartości funkcji f, to przerywamy obliczenia.

Wreszcie jest kryterium przyrostowe. Obliczenia przerywamy, gdy wartość bezwzględna (lub norma) przyrostu $\delta = x_{k+1} - x_k$ jest mniejsza niż pewna wielkość progowa. Dla wielu metod długość przyrostu w danym kroku jest górnym oszacowaniem błędu rozwiazania przybliżonego x_k (ale to zależy także od funkcji f).

2. Arytmetyka zmiennopozycyjna

Liczb rzeczywistych jest nieskończenie (a nawet nieprzeliczalnie) wiele, a pamięć choćby największego komputera jest skończona. Dlatego w obliczeniach numerycznych musimy się zadowolić poruszaniem się w pewnym skończonym zbiorze, którego elementy tylko przybliżają wszelkie liczby rzeczywiste, jakie mogłyby się pojawić w tych obliczeniach.

W rozmaitych zastosowaniach istotny jest błąd względny przetwarzanych liczb.

Reprezentacja zmiennopozycyjna

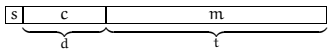
Powszechnie używana reprezentacja zmiennopozycyjna liczb rzeczywistych jest kompromisem między dokładnością i złożonością czasową i pamięciową. Jej głównym celem jest masowe przetwarzanie liczb, czemu służy stosunkowo mała ilość miejsca zajmowanego przez tę reprezentację i możliwość szybkiego wykonywania działań przez specjalnie opracowane w tym celu podukłady procesorów. Błędy tej reprezentacji są dostatecznie małe na potrzeby znakomitej większości zastosowań. Istnieją inne reprezentacje, umożliwiające prowadzenie obliczeń ze znacznie większą dokładnością, ale znacznie wolniej i w większej pamięci. Te inne reprezentacje są poza zakresem tego wykładu. Jeszcze jedno: reprezentacje zmiennopozycyjne mają powszechnie przyjęty standard, który ułatwia m.in. wymianę danych. Reprezentacje niestandardowe tak fajnie nie mają.

Idea reprezentacji zmiennopozycyjnej wiąże się z tzw. półlogarytmicznym zapisem liczby. Każdą dodatnią liczbę rzeczywistą możemy przedstawić za pomocą liczby z przedziału $[1, 10)$ i całkowitej potęgi liczby 10, na przykład

$$27182818 = 2.7182818 \cdot 10^7.$$

W komputerach zamiast podstawy 10 i dziesięciu różnych cyfr, wygodniej jest używać podstawy 2 i bitów.

Podstawowa reprezentacja określona przez standard IEEE-754 (opracowany w 1985 r.) składa się z bitu znaku, s, po którym następuje cecha c i mantysa m:



Mantysa jest liczbą rzeczywistą; jeśli reprezentuje ją ciąg bitów $b_{t-1}b_{t-2}\ldots b_1b_0$, to $m = \sum_{k=0}^{t-1} b_k 2^{k-t}$, a zatem zawsze $0 \leqslant m < 1$. Cecha jest liczbą całkowitą (bez znaku), reprezentowaną za pomocą d bitów, która wpływa na sposób interpretacji całego ciągu bitów. Liczba reprezentowana przez taki ciąg, w zależności od cechy, jest równa

$$\begin{aligned}
&x = (-1)^s\, 2^{c-b}\,(1+m) && \text{dla } 0 < c < 2^d - 1,\\
&x = (-1)^s\, 2^{1-b}\, m && \text{dla } c = 0,\\
&x = (-1)^s\, \infty && \text{dla } c = 2^d - 1,\ m = 0,\\
&x = \mathrm{NaN}\ (\text{„nie-liczba”}) && \text{dla } c = 2^d - 1,\ m \neq 0.
\end{aligned}$$

Liczby d, t i b są ustalone dla konkretnej reprezentacji. Cechą charakterystyczną reprezentacji z użyciem pierwszego wzoru jest tzw. normalizacja. Mając dowolną liczbę rzeczywistą $x \neq 0$, przedstawioną w układzie dwójkowym, dobieramy cechę c (czyli równoważnie czynnik 2^{c-b}) tak, że czynnik $(1 + m)$ w wyrażeniu opisującym x jest liczbą z przedziału $[1, 2)$. Jeśli otrzymana w ten sposób cecha jest za duża (większa lub równa $2^d - 1$), to mamy nadmiar zmiennopozycyjny (ang. *floating point overflow*), czyli niewykonalne zadanie reprezentowania liczby o za dużej wartości bezwzględnej, zwykle będące powodem do przerwania obliczeń. Jeśli nie ma nadmiaru, to pierwszy wzór opisuje liczbę w ten sposób, że najbardziej znacząca jedynka w rozwinięciu dwójkowym nie jest jawnie pamiętana — właśnie to jest normalizacja. Dzięki niej każdy ciąg bitów reprezentuje inną liczbę, co m.in. umożliwia optymalne wykorzystanie bitów do zmniejszenia błędów.

Niech x oznacza dowolną liczbę rzeczywistą. Jej reprezentację, tj. położoną najbliżej niej liczbę zmiennopozycyjną, oznaczymy symbolem $\mathrm{rd}(x)$ (z ang. *rounding*). Jeśli liczbę x możemy przedstawić w postaci

$$x = (-1)^s 2^{c-b}(1+f),$$

dobierając cechę c tak, aby mieć $f \in [0,1)$ oraz $0 < c < 2^d - 1$, to (z jednym rzadkim wyjątkiem, gdy f trzeba zaokrąglić w górę do jedynki) będziemy mieli

$$\mathrm{rd}(x) = (-1)^s 2^{c-b}(1+m),$$

przy czym $|f-m| \leqslant 2^{-t-1}$. Błąd względny reprezentacji spełnia nierówność

$$\frac{|x-\mathrm{rd}(x)|}{|x|} = \frac{|(-1)^s 2^{c-b}(1+f) - (-1)^s 2^{c-b}(1+m)|}{|(-1)^s 2^{c-b}(1+f)|} \leqslant |f-m| \leqslant 2^{-t-1}.$$

Co ciekawe, nierówność ta jest spełniona też w specjalnym przypadku wspomnianym wcześniej (bo w mianowniku $1+f \approx 2$). Zatem, maksymalny błąd względny reprezentacji zmiennopozycyjnej, jeśli nie ma niedomiaru ani nadmiaru, jest na poziomie 2^{-t-1}, gdzie t jest liczbą bitów mantysy. Jeśli kierunek zaokrąglania wybieramy mniej starannie (np. zawsze obcinamy w kierunku zera), to błąd względny może być dwa razy większy, czyli rzędu $\nu = 2^{-t}$.

Bardziej skomplikowana sytuacja zdarza się w przypadku, gdy cecha jest za mała (tj. gdy w pierwszym wzorze należałoby przyjąć $c \leqslant 0$). Wtedy korzystamy z drugiego wzoru, w którym występuje czynnik m (przypominam, że $m \in [0,1)$). Jeśli $c = m = 0$, to mamy reprezentację zera; liczba 0 jako jedyna ma dwie reprezentacje, różniące się bitem znaku. Jeśli $c = 0$ i $m \neq 0$, to mamy do czynienia z niedomiarem zmiennopozycyjnym, czyli reprezentowaniem liczby x za pomocą mantysy o mniejszej liczbie bitów istotnych (jeśli w użyciu jest pierwszy wzór, to istotne są wszystkie bity mantysy, jeśli drugi, to tylko bity od pozycji najmniej znaczącej, do najbardziej znaczącej pozycji, na której jest jedynka).

Najdokładniejszą reprezentacją liczb o bardzo małej wartości bezwzględnej (mniejszej niż 2^{-b-t}) jest 0. Niedomiar wiąże się zatem ze (stopniową) utratą dokładności reprezentacji. Dla $x \to 0$ błąd względny reprezentacji dąży do 100%, a błąd bezwzględny jest ograniczony. W analizie błędów najczęściej nie bierzemy tego przypadku pod uwagę.

Reprezentacja umożliwia używanie nieskończoności, także w rachunkach (np. wynik dzielenia dowolnej liczby przez nieskończoność jest równy 0).

Nie-liczby są wykorzystywane do sygnalizowania błędów, np. próby obliczenia pierwiastka kwadratowego z liczby ujemnej. Można je też wykorzystać do odpluskwiania programu, np. nadając zmiennym takie wartości początkowe, a następnie śledząc, czy nie ma do nich odwołań przed przypisaniem właściwej wartości liczbowej.

W standardzie IEEE-754 są zdefiniowane formaty liczb pojedynczej i podwójnej precyzji, a także liczb pojedynczej i podwójnej rozszerzonej precyzji. Liczby pojedynczej rozszerzonej precyzji się nie przyjęły, procesory w komputerach PC ich nie obsługują.

Dane na temat standardowych formatów są w tabelce:

	B	d	t	b	M	S	ν	μ
pojedyncza, `float`	32	8	23	127	10^{38}	10^{-38}	10^{-7}	10^{-45}
pojed. rozszerzona —	44	11	31	1023	10^{308}	10^{-308}	10^{-10}	10^{-317}
podwójna `double`	64	11	52	1023	10^{308}	10^{-308}	10^{-15}	10^{-323}
podw. rozszerzona `long double`	80 (96, 128)	15	63	16383	10^{4932}	10^{-4932}	10^{-19}	10^{-4951}

Oznaczenia: B — całkowita liczba bitów, d — liczba bitów cechy, t — liczba bitów mantysy, b — stała odejmowana od cechy w celu otrzymania wykładnika. Stała b jest równa $2^{d-1}-1$, dzięki czemu jeśli liczba x ma reprezentację znormalizowaną, to $1/x$ na ogół też.

Liczby $M = 2^{2^d-b-2}(2-2^{-t})$ — największa liczba zmiennopozycyjna, $S = 2^{1-b}$ — najmniejsza dodatnia liczba reprezentowana w postaci znormalizowanej (tj. bez niedomiaru), $\nu = 2^{-t}$ — oszacowanie maksymalnego błędu względnego reprezentacji znormalizowanej, oraz $\mu = 2^{1-b-t}$ — najmniejsza zmiennopozycyjna liczba dodatnia, są podane w przybliżeniu (tylko rząd wielkości).

Reprezentacje rozszerzonej precyzji nie wymuszają normalizacji (mantysa ma $t+1$ bitów i jest liczbą z przedziału $[0,2)$, jej najbardziej znaczący bit ma wartość 1), ale wyniki działań, jeśli nie ma niedomiaru, są normalizowane przez procesor.

Jeszcze jedno: w 32-bitowych systemach operacyjnych zmienna rozszerzonej podwójnej precyzji zajmuje 12 bajtów, z których 2 są nieużywane. W systemach 64-bitowych taka zmienna zajmuje 16 bajtów, z których 6 jest nieużywanych. To utrudnia m.in. przenoszenie danych między komputerami w postaci binarnej. Jeśli nie ma istotnego powodu, to najlepiej nie używać tej reprezentacji liczb.

Oprócz standardu IEEE-754 istnieje też standard IEEE-854, który definiuje reprezentacje liczb zmiennopozycyjnych z podstawami 2 i 10. Standard ten służy do wymiany danych między komputerami, natomiast określone przezeń reprezentacje nie są przetwarzane bezpośrednio przez jednostki zmiennopozycyjne procesorów (w każdym razie znanych mi). Jeśli nie ma ważnych powodów do używania reprezentacji określonych w tym standardzie, to można się nim nie przejmować.

Reprezentacje niestandardowe: istnieje dość rzadko spotykany format poczwórnej precyzji, w którym reprezentacja liczby zajmuje 128 bitów (cecha ma w nim 15 bitów, mantysa 112). Nie słyszałem o procesorach z rejestrami zmiennopozycyjnymi o takiej długości, zatem działania na takich liczbach muszą być wykonywane przez odpowiednie podprogramy. Z drugiej strony, reprezentacje 16- 11- i 10-bitowe (bit znaku może być nieobecny, cecha ma 5 bitów, a mantysa 10, 6 albo 5) są używane przez niektóre karty graficzne podczas wykonywania obrazów, gdy dokładność ma małe znaczenie, zaś najważniejsza jest szybkość obliczeń i oszczędność miejsca. Wspomniane karty graficzne mają specjalizowane podukłady do wykonywania działań na takich liczbach.

Arytmetyka i błędy zaokrągleń

Na potrzeby analizy błędów działanie procesora podczas wykonywania operacji arytmetycznych można sobie wyobrazić tak: dokładny wynik działania jest poddawany normalizacji (tj. dobierana jest cecha), a następnie zaokrągleniu — nieskończony ciąg bitów mantysy jest obcinany i ewentualnie zaokrąglany w górę. Nie wyznacza się oczywiście nieskończonego ciągu bitów mantysy, zamiast tego wykorzystuje się trzy bity dodatkowe („wystające" poza format), z których pierwsze dwa są zwykłe, a trzeci „lepki" — bit ten otrzymuje wartość 1, jeśli dowolny dalszy bit nieskończenie długiej mantysy jest niezerowy. Te trzy bity zawsze wystarczą do poprawnego zaokrąglenia liczby.

Wyboru kierunku zaokrąglania można dokonać, ustawiając odpowiednie bity w rejestrze sterującym procesora (zwykle zostawiamy domyślne zaokrąglanie do najbliższej liczby zmiennopozycyjnej).

Istotne jest, że oprócz reprezentacji liczb, standard IEEE-754 określa własności działań, w tym wymagania dotyczące dokładności wyników — dotyczy to czterech działań arytmetycznych, pierwiastka kwadratowego, oraz konwersji reprezentacji całkowitej i zmiennopozycyjnej. Istnieją procesory, które wprawdzie przetwarzają liczby w standardowym formacie, ale realizowane przez nie działania *nie spełniają* wszystkich warunków określonych w standardzie.

Najbardziej rozpowszechnionym sprzętem tego rodzaju są karty graficzne, które mogą m.in. nie obsługiwać liczb nieznormalizowanych (tj. zapisanych przy użyciu drugiego wzoru podanego w opisie formatu; w razie niedomiaru wynikiem działania jest zero) lub zaokrąglać wyniki działań w arbitralnie określony sposób (standard nakazuje umożliwiać dokonanie wyboru). Powinien o tym pamiętać każdy, kto zajmuje się tzw. GPGPU (*general programming on graphics processing unit*).

Jeśli x jest liczbą rzeczywistą, a $\mathrm{rd}(x)$ jest jej znormalizowanym zmiennopozycyjnym przybliżeniem (bez nadmiaru i niedomiaru), to mamy $|x - \mathrm{rd}(x)| \leqslant |x|2^{-1-t}$, skąd wynika, że istnieje liczba ε, taka że

$$\mathrm{rd}(x) = x(1+\varepsilon) \quad \text{oraz} \quad |\varepsilon| \leqslant 2^{-1-t}.$$

Sposób zaokrąglania (do najbliższej liczby zmiennopozycyjnej, zawsze w stronę zera, zawsze w przeciwną stronę, zawsze w górę albo zawsze w dół) może być ustawiony różnie, przez co błąd względny może być dwa razy większy. Jeśli zatem $\diamond$ oznacza dowolne z czterech działań arytmetycznych, to zamiast wyniku $x = a \diamond b$, po zaokrągleniu, otrzymamy liczbę

$$\tilde{x} = \mathrm{fl}(a \diamond b) = (a \diamond b)(1+\varepsilon),$$

dla pewnego $\varepsilon \in (-\nu, \nu)$ (piszemy $\mathrm{fl}(a \diamond b)$ zamiast $\mathrm{rd}(a \diamond b)$, bo ten ostatni symbol oznacza u nas wynik zaokrąglenia do najbliższej liczby zmiennopozycyjnej).

Wyniki działań są najczęściej argumentami dalszych działań, zatem podczas obliczeń numerycznych ma miejsce zjawisko zwane kumulacją błędów. W szczególnych przypadkach może ono doprowadzić do otrzymania bardzo niedokładnych wyników końcowych, mimo że poszczególne błędy zaokrągleń są małe. Ponadto wskutek zaokrągleń zbiór liczb zmiennopozycyjnych z działaniami dodawania i mnożenia *nie jest* ciałem (z punktu widzenia algebry). Przede wszystkim, nie jest zamknięty ze względu na działania (bo może wystąpić nadmiar) i są w nim dzielniki zera (np. jeśli liczba $|x| \neq 0$ jest dostatecznie mała, to $\mathrm{fl}(x * x) = 0$). Po drugie, dodawanie i mnożenie nie są działaniami łącznymi i dodawanie nie jest rozdzielne względem mnożenia.

W konsekwencji, algorytmy oparte na różnych wzorach algebraicznie równoważnych (w ciele $\mathbb{R}$), mogą produkować *różne* wyniki (czasem bardzo od siebie odległe). Analiza algorytmów ma na celu między innymi badanie, na jaką dokładność wyników obliczeń wykonywanych z błędami zaokrągleń można liczyć (i może się przydać do wybrania najlepszego algorytmu, albo przynajmniej do odrzucenia najgorszego).

Arytmetyka zmiennopozycyjna zespolona

W różnych zadaniach występują liczby zespolone. W obliczeniach ich części rzeczywiste i urojone są reprezentowane w postaci zmiennopozycyjnej. Jeśli zatem zamiast liczby $z = (a, b) \neq 0$ mamy liczbę $\tilde{z} = (\tilde{a}, \tilde{b}) = (a(1+\varepsilon_a), b(1+\varepsilon_b))$, gdzie $|\varepsilon_a|, |\varepsilon_b| < \nu$, to liczbę z reprezentujemy z błędem względnym

$$\frac{|z - \tilde{z}|}{|z|} = \frac{\sqrt{a^2\varepsilon_a^2 + b^2\varepsilon_b^2}}{\sqrt{a^2+b^2}} < \frac{\sqrt{a^2\nu^2 + b^2\nu^2}}{\sqrt{a^2+b^2}} = \nu.$$

Zatem reprezentacja zmiennopozycyjna liczby zespolonej zapewnia równie mały błąd, jak reprezentacja liczby rzeczywistej.

Dodawanie i odejmowanie liczb zespolonych wykonujemy na podstawie wzorów będących definicją tych działań, w związku z czym, jeśli nie ma nadmiaru ani niedomiaru, otrzymamy

$$\mathrm{fl}(z_1 \pm z_2) = (z_1 \pm z_2)(1+\varepsilon), \quad \text{gdzie} \quad |\varepsilon| < \nu.$$

Mnożenie też wykonuje się na podstawie definicji:

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1a_2 - b_1b_2, a_1b_2 + a_2b_1).$$

Zamiast dokładnego wyniku otrzymamy

$$\begin{aligned}\mathrm{fl}((a_1, b_1) \cdot (a_2, b_2)) = \\ ((a_1a_2(1+\varepsilon_1) - b_1b_2(1+\varepsilon_2))(1+\varepsilon_3), \\ (a_1b_2(1+\varepsilon_4) + a_2b_1(1+\varepsilon_5))(1+\varepsilon_6)),\end{aligned}$$

przy czym, jeśli w żadnym działaniu nie wystąpił nadmiar ani niedomiar, to wszystkie epsilony mają wartości bezwzględne mniejsze niż ν. Można udowodnić, że

$$(a_1, b_1) \cdot (a_2, b_2) \cdot (1+\xi),$$

gdzie ξ jest pewną liczbą zespoloną, taką że $|\xi| < (1+\sqrt{2})\nu$.

Dzielenie zespolone jest bardziej kłopotliwe, bo algorytm musi unikać nadmiaru i niedomiaru (zwróćmy uwagę, że nawet w przypadku mnożenia, wynik działania może mieć reprezentację, zaś wyniki pośrednie mogą jej nie mieć z powodu nadmiaru — w dzieleniu ten problem też występuje). Algorytm dzielenia:

```
if ( fabs ( a2 ) >= fabs ( b2 ) ) {
  p = b2/a2;
  q = a2+b2*p;
  wynik = ((a1+b1*p)/q, (b1-a1*p)/q);
}
else {
  p = a2/b2;
  q = a2*p+b2;
  wynik = ((a1*p+b1)/q, (b1*p-a1)/q);
}
```

Jeśli nie ma nadmiaru ani niedomiaru, to względny błąd zaokrąglenia wyniku nie jest większy niż $(4+\sqrt{2})\nu$.

3. Błędy w obliczeniach

W obliczeniach numerycznych występują błędy pięciu rodzajów.

- Błędy modelu,
- Błędy danych wejściowych,
- Błędy aproksymacji,
- Błędy zaokrągleń,
- Błędy grube.

Błędy modelu. Model matematyczny dowolnego zjawiska (przyrodniczego, ekonomicznego i w ogóle każdego) jest tego zjawiska uproszczeniem. Na przebieg zjawiska ma wpływ wiele różnych czynników, z których jedne są ignorowane (bo ich wpływ został uznany za pomijalny), a inne nie są znane dostatecznie dokładnie, aby można było napisać całkowicie poprawny wzór. Jeśli model znacznie odbiega od zjawiska, to i wyniki obliczeń mogą bardzo się różnić od tego, co można zaobserwować w rzeczywistości.

Błędy danych wejściowych. Dane wejściowe trzeba zapisać w postaci liczb zmiennopozycyjnych, co powoduje ich zaburzenie. Jeśli wynik od danych zależy (a zwykle tak jest), to nawet gdyby nie było innych błędów, wynik obliczeń może się różnić od wyniku doświadczenia. Ponadto, na ogół dane otrzymujemy z pomiarów, których niedokładności mogą być znacznie większe niż błąd reprezentacji zmiennopozycyjnej. Najdokładniejsze pomiary w fizyce dają kilkanaście cyfr dokładnych, często znamy dane z dokładnością rzędu 1%, a czasami błędy są na poziomie kilkudziesięciu procent. Sygnały lub obrazy mogą być zniekształcone z powodu szumu i bardzo niewyraźne. To wszystko ma bardzo duży wpływ na wynik (albo jego brak, jeśli algorytm nie poradzi sobie z niedokładnymi danymi).

Błędy aproksymacji. W obliczeniach numerycznych stosuje się przybliżenia funkcji, których dokładne obliczenie jest niewykonalne lub zbyt kosztowne. Na przykład, zamiast granicy nieskończonego ciągu zbieżnego, bierze się pewien element tego ciągu. Zamiast sumy szeregu nieskończonego oblicza się sumę iluś początkowych składników. Zamiast całki oblicza się kwadraturę. Równania różniczkowe często zastępuje się równaniami różnicowymi; można podać wiele dalszych przykładów.

Błędy aproksymacji granicy ciągu nieskończonego przez pewien element tego ciągu, lub sumy nieskończonego szeregu przez pewną sumę częściową są często nazywane błędami obcięcia.

Błędy zaokrągleń. Wynik każdego działania wykonanego przez komputer podlega zaokrągleniu. Skutki bardzo często są małe w porównaniu ze skutkami innych błędów, ale czasem mogą zupełnie zmienić wynik.

Błędy grube. To są skutki wszelkich pomyłek, awarii, oraz błędów popełnionych w procesie pozyskiwania danych lub w implementacji algorytmu. Z innych przyczyn można tu też wymienić sabotaż (np. uprawiany przez producentów wirusów komputerowych i przez nierzetelnych autorów oprogramowania).

Uwarunkowanie zadania

Większość zadań numerycznych polega na obliczeniu wartości pewnej funkcji f, której dziedziną jest pewien obszar $D \subset \mathbb{R}^n$. Wynik obliczenia jest wektorem w $\mathbb{R}^m$, przy czym m może być określone przez konkretny argument $\mathbf{x} \in D$ — na przykład, gdy trzeba znaleźć wszystkie rzeczywiste miejsca zerowe wielomianu, którego współczynniki są współrzędnymi wektora $\mathbf{x}$. Załóżmy jednak, że m jest ustalone (i znane) dla wszystkich $\mathbf{x} \in D$, a funkcja f jest ciągła. Zanim zaczniemy rozpatrywać jakiekolwiek algorytmy obliczania wyniku, zajmiemy się wpływem, jaki zaburzenia danych (które mogą pochodzić z niedokładnych pomiarów i które trzeba zastąpić liczbami zmiennopozycyjnymi) mają na wynik.

Pojęcie numerycznego uwarunkowania zadania określa wrażliwość wyniku na zaburzenia danych; dla zadania dobrze uwarunkowanego niewielkie zaburzenie danych powoduje niewielką zmianę wyniku. Zadanie jest źle uwarunkowane, jeśli po małej zmianie danych otrzymujemy zupełnie inny wynik. W związku ze sposobem reprezentowania liczb (który zapewnia mały błąd względny), bierzemy pod uwagę *względne* zaburzenia danych i spowodowane przez nie zmiany wyniku.

Liczbowa miara uwarunkowania nazywa się wskaźnikiem uwarunkowania zadania. Określa się go wzorem

$$\operatorname{cond}_{f(\mathbf{x})} \mathbf{x} = \sup_{\|\tilde{\mathbf{x}}-\mathbf{x}\|<\varepsilon\|\mathbf{x}\|} \left(\frac{\|f(\tilde{\mathbf{x}}) - f(\mathbf{x})\|}{\|f(\mathbf{x})\|} \Big/ \frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \right).$$

Symbol cond pochodzi od angielskiego *condition number*; napis po lewej stronie czytamy: „wskaźnik uwarunkowania zadania obliczenia $f(\mathbf{x})$ dla danych $\mathbf{x}$”. W określeniu wskaźnika uwarunkowania używamy jakichś norm (zależnie od zadania) i określamy największą dopuszczalną zmianę (zaburzenie względne) ε danych $\mathbf{x}$. Następnie badamy iloraz względnego zaburzenia wyniku i powodującej to zaburzenie względnej zmiany danych.

Jeśli dane znamy z błędem względnym nie większym niż ε, to błąd względny wyniku (uwaga: *dokładnego* wyniku dla danych $\tilde{\mathbf{x}}$, jakimi dysponujemy, w porównaniu z wynikiem dla nieznanych nam danych dokładnych $\mathbf{x}$) nie jest większy niż $\varepsilon \operatorname{cond}_{f(\mathbf{x})} x$. Na przykład, jeśli wskaźnik uwarunkowania jest równy 100 (to jeszcze nie jest dużo), a dane reprezentujemy w formacie pojedynczej precyzji, tj. z błędem nie większym niż $\nu \approx 10^{-7}$ (i poza zaokrągleniem nie ma innych błędów), to wiemy, że jesteśmy w stanie otrzymać wynik z pięcioma cyframi dokładnymi. Jeśli jednak pomiar danych ma błąd rzędu 1%, to otrzymany wynik może mieć błąd 100%; na ogół taki wynik jest bezwartościowy. Albo należy wtedy zdobyć dokładniejsze dane, albo zająć się innym zadaniem (być może można jakoś przeformułować problem). Pamiętajmy przy tym, że założyliśmy brak błędów w algorytmie, który może dodatkowo zepsuć wynik.

Często przyjmuje się, że zaburzenia danych są bardzo małe (bo względne błędy reprezentacji zmiennopozycyjnej są bardzo małe), więc dla uproszczenia oblicza się wartość graniczną wskaźnika uwarunkowania, dla $\varepsilon \to 0$ (co ma sens, jeśli wskaźnik jest ciągły w otoczeniu $\mathbf{x}$). Jeśli zadanie polega na obliczeniu wartości skalarnej funkcji f, która ma skalarny argument x, przy czym funkcja f ma pochodną, to mamy wtedy

$$\operatorname{cond}_{f(x)} x = \left| \frac{x}{f(x)} f'(x) \right|.$$

Błędy reprezentacji wektorów

Niech $\mathbf{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ i niech $\tilde{\mathbf{x}} = [\tilde{x}_1, \ldots, \tilde{x}_n]^T \in \mathbb{R}^n$, przy czym $\tilde{x}_i = x_i(1 + \varepsilon_i)$ dla każdego i. Zamiast rozpatrywać osobno błędy poszczególnych składowych wektora, co mogłoby zbyt być pracochłonne, często błąd opisuje się jedną liczbą, za pomocą jakiejś normy. Najczęściej wykorzystywane są tzw. normy Höldera, określone wzorem

$$\|\mathbf{x}\|_p = \left(|x_1|^p + \cdots + |x_n|^p \right)^{1/p},$$

dla pewnego $p \geqslant 1$. Często stosuje się normę — przypadek graniczny, dla $p \to \infty$:

$$\|\mathbf{x}\|_\infty = \max_{i \in \{1,\ldots,n\}} |x_i|.$$

Za miarę błędu bezwzględnego możemy przyjąć liczbę $\|\mathbf{x} - \tilde{\mathbf{x}}\|_p$. Jeśli $\mathbf{x} \neq 0$ i dla każdego i jest $|\varepsilon_i| \leqslant \nu$, to miara błędu względnego spełnia nierówność

$$\begin{aligned} \frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|_p}{\|\mathbf{x}\|_p} &= \frac{\left(|x_1\varepsilon_1|^p + \cdots + |x_n\varepsilon_n|^p \right)^{1/p}}{\|\mathbf{x}\|_p} \\ &\leqslant \frac{\left(|x_1\nu|^p + \cdots + |x_n\nu|^p \right)^{1/p}}{\|\mathbf{x}\|_p} = \frac{\|\mathbf{x}\|_p \nu}{\|\mathbf{x}\|_p} = \nu. \end{aligned}$$

Zatem, błąd względny reprezentacji wektora, którego współrzędne zostały zokrąglone do najbliższych liczb zmiennopozycyjnych, mierzony za pomocą dowolnej normy Höldera (także $\|\cdot\|_\infty$), jest na poziomie błędu reprezentacji pojedynczej liczby.

Uwaga: Należy pamiętać, że z nierówności $\frac{\|\mathbf{x}-\tilde{\mathbf{x}}\|_p}{\|\mathbf{x}\|_p} \leqslant \varepsilon > 0$ *nie wynika*, że błędy względne poszczególnych składowych są małe. Jeśli pewna składowa jest równa 0, to dowolne niezerowe jej zaburzenie daje nieograniczony błąd względny. Tak więc, wykonując odpowiednie rachunki, nie należy wyciągać pochopnych wniosków.

Numeryczna poprawność algorytmu

Skutki błędów zaokrągleń w obliczeniach czasem można zinterpretować jako skutki takiego zaburzenia danych, że otrzymany wynik jest dla tych zaburzonych danych dokładny. Jeśli takie hipotetyczne zaburzenie danych jest małe, to mówimy, że algorytm jest numerycznie poprawny. Pewne algorytmy są numerycznie poprawne, inne nie są. W zasadzie numeryczna poprawność „to jest to” — w praktyce niczego lepszego po algorytmach numerycznych spodziewać się nie można.

Tak, jak uwarunkowanie zadania, numeryczną poprawność można mierzyć, badając tzw. stałe kumulacji algorytmu. Algorytm jest tym lepszy, im te stałe są mniejsze. Aby je zdefiniować, wprowadzimy potrzebne oznaczenia. Niech A oznacza algorytm. Zatem, niech $A(\mathbf{x})$ oznacza wynik obliczenia, który powinien być jak najbliższy „prawdziwemu" rozwiązaniu zadania, $f(\mathbf{x})$. Obliczony wynik składa się z liczb zmiennopozycyjnych, zatem możemy dopuścić do rozważań jego błąd reprezentacji. Przypuśćmy zatem, że istnieją liczby K_d i K_w, takie że dla każdego $\mathbf{x} \in D$ istnieją dane zaburzone $\tilde{\mathbf{x}}$, dla których spełnione są nierówności

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant K_d \nu, \quad \text{oraz} \quad \frac{\|f(\tilde{\mathbf{x}}) - A(\mathbf{x})\|}{\|f(\tilde{\mathbf{x}})\|} \leqslant K_w \nu.$$

Mówimy wtedy, że algorytm A jest numerycznie poprawnym algorytmem obliczania wartości funkcji f w dziedzinie (klasie zadań) D, ze stałymi kumulacji (danych) K_d i (wyniku) K_w.

Trzeba podkreślić, że w analizie algorytmu często występuje swoboda wybierania danych lub wyniku, do których „doczepiamy" błędy; z jednej strony to utrudnia analizę, a z drugiej stwarza możliwości pewnej „gimnastyki", wskutek czego pewne oszacowania mogą być poprawione — nieraz jest tak, że algorytm w praktyce działa bardzo dobrze, tj. wytwarza bardzo dokładne wyniki, zaś analiza tego nie potwierdza, bo na przykład daje bardzo grube oszacowania stałych kumulacji. Wspomniana „gimnastyka" czasem pomaga. W analizie błędu zwykle zakłada się, że błędy w poszczególnych działaniach są niezależne (i nieskorelowane), a ich wartości bezwzględne sumują się, tymczasem poszczególne błędy względne mogą być mniejsze niż ν, mogą się też znosić. Czasem analiza błędu pozwala wykryć newralgiczne miejsca i pomaga przeprojektować wzory.

Jeśli stała K_d jest równa 0, to znaczy, że niezależnie od uwarunkowania zadania otrzymany wynik jest bardzo dokładny, tj. otrzymany z dokładnością na poziomie błędu reprezentacji (tj. błąd wyniku jest co najwyżej K_w razy większy). Taka sytuacja występuje w praktyce nadzwyczaj rzadko. Częściej „winę" za niedokładność wyniku można „zwalić" na dane. Takie postępowanie, tj. znalezienie i oszacowanie zaburzenia danych, które prowadzi do otrzymanego wyniku, nazywa się analizą wstecz; jej twórcą był Wilkinson. Jeśli zadanie jest dobrze uwarunkowane i stałe kumulacji są nieduże, to stąd wynika, że obliczony wynik jest dobrym przybliżeniem wyniku poszukiwanego.

Numeryczna stabilność algorytmu

Często się nie udaje udowodnienie numerycznej poprawności algorytmu, tj. znalezienie stałych kumulacji niezależnych od danych w ustalonej dziedzinie D. Wówczas można spróbować zbadać, czy jest on numerycznie stabilny — ta własność jest pewnego rodzaju „minimum przyzwoitości" algorytmu. Aby ją zdefiniować, zbadajmy, jak duży byłby błąd wyniku, gdyby dane zostały zaburzone na poziomie błędu reprezentacji (co musi mieć miejsce — dane do obliczeń są liczbami zmiennopozycyjnymi) i wynik też należałoby zaokrąglić (bo też go reprezentujemy w ten sposób), ale poza zaokrągleniem końcowego wyniku wszystkie obliczenia byłyby wykonywane dokładnie.

Błąd (bezwzględny) wyniku spełniający wymienione warunki można oszacować przez liczbę, zwaną optymalnym poziomem błędu:

$$\|f(\mathbf{x})\|(\mathrm{cond}_{\mathbf{w}}\, \mathbf{d} + 1)\nu.$$

Względny błąd danych, na poziomie ν, przenosi się na wynik z czynnikiem $\mathrm{cond}_{\mathbf{w}}\, \mathbf{d}$; do tego wynik trzeba jeszcze zaokrąglić, stąd do wskaźnika uwarunkowania została dodana jedynka.

Mówimy, że algorytm A jest numerycznie stabilnym algorytmem obliczania funkcji f, jeśli istnieje liczba K (stała kumulacji), taka że dla dowolnych danych $\mathbf{d} \in D$ spełniona jest nierówność

$$\|f(\mathbf{x}) - A(\mathbf{x})\| \leqslant K\|f(\mathbf{x})\|(\mathrm{cond}_{\mathbf{w}}\, \mathbf{d} + 1)\nu.$$

Ważne jest też, aby stała kumulacji nie była bardzo duża.

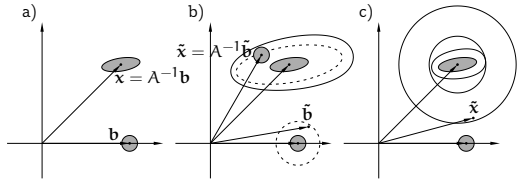
W tym ujęciu analizy błędów nie zajmujemy się tym, czy istnieją takie dane, bliskie danych $\mathbf{x}$, dla których otrzymujemy (ewentualnie zaburzony na poziomie błędu reprezentacji) wynik. Dane takie mogą więc nie istnieć — możemy na przykład otrzymać sinus pewnego kąta rzeczywistego większy niż 1. Istotne jest to, że mając algorytm numerycznie stabilny, możemy dowolnie zmniejszyć skutki błędów zaokrągleń, wykorzystując w obliczeniach dostatecznie dokładną arytmetykę (czyli taką o dostatecznie długiej mantysie: przypominam, że $\nu = 2^{-t}$). Oczywiście, dla zadań źle uwarunkowanych arytmetyki standardowe mogą nie wystarczyć, ale wtedy czy na pewno znamy dane aż tak dokładnie?

Jeśli funkcja f, której wartość należy obliczyć, spełnia warunek Lipschitza, tj. istnieje stała L, taka że

$$\forall_{\mathbf{x},\mathbf{y} \in D}\ \|f(\mathbf{x}) - f(\mathbf{y})\| \leqslant L\|\mathbf{x} - \mathbf{y}\|,$$

to każdy algorytm numerycznie poprawny jest też numerycznie stabilny, ale numeryczna stabilność nie gwarantuje numerycznej poprawności.

Ilustracją opisanych pojęć może być rysunek, będący ilustracją zadania rozwiązywania układu dwóch równań liniowych $A\mathbf{x} = \mathbf{b}$, z nieosobliwą macierzą A. Rozwiązaniem zadania jest wektor $\mathbf{x} = A^{-1}\mathbf{b}$, przy czym ten wzór jest pożyteczny w teoretycznej analizie zadania i algorytmów jego rozwiązywania, ale nie jest dobrym algorytmem numerycznym (i proszę go *nie używać* w tym charakterze). Danymi są współczynniki macierzy A i wektora prawej strony $\mathbf{b}$. Dla ilustracji pojęć rozpatrujemy tylko zaburzenia wektora prawej strony. Wielkość tych zaburzeń jest taka, jak gdyby mantysa miała mniej więcej trzy bity.



Na rysunku a) mamy ilustrację uwarunkowania zadania. Zaznaczona kula (tj. koło) o środku $\mathbf{b}$ ma promień $\nu\|\mathbf{b}\|$. Zaburzenie danych polega na zastąpieniu wektora $\mathbf{b}$ przez jakiś element tej kuli. Obrazem tej kuli jest elipsoida (elipsa) o środku $\mathbf{x}$. Wskaźnik uwarunkowania zadania (ze względu na zaburzanie wektora $\mathbf{b}$, ale także macierzy A, co będziemy badać na jednym z dalszych wykładów) jest ilorazem długości najdłuższej i najkrótszej osi elipsoidy.

Numeryczna poprawność jest zilustrowana na rysunku b). Algorytm wyprodukował pewien wektor $\tilde{\mathbf{x}}$. Niech $\tilde{\mathbf{b}} = A\mathbf{x}$. Przypuśćmy, że stała kumulacji $K_w = 0$. Wtedy

$$\frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|\nu} \leqslant K_d,$$

i mamy gwarancję, że dla otrzymanego wyniku $\tilde{\mathbf{x}}$, który leży w obrębie narysowanej linią przerywaną elipsy, istnieją dane $\tilde{\mathbf{b}}$, które leżą w narysowanym linią przerywaną kole (promień tego koła jest K_d razy większy niż $\|\mathbf{b}\|\nu$). Jeśli zaś weźmiemy $K_w > 0$, to dopuszczamy dodatkowe zaburzenie wyniku; leży on w nieco większym obszarze ograniczonym przez krzywą zobrazowaną przez linię ciągłą (ta krzywa nie jest elipsą). Dla takiego wyniku istnieje bliski punkt leżący w obszarze ograniczonym elipsą, który jest dokładnym wynikiem dla pewnych danych położonych w większym kole o środku $\mathbf{b}$.

Numeryczna stabilność jest przedstawiona na rysunku c). Rozważamy zaburzenia danych $\mathbf{b}$ na poziomie błędu reprezentacji. Dla tak zaburzonych danych wynik leży w obszarze zacienionym, ograniczonym przez elipsę. Ten obszar rozszerzamy, aby uwzględnić błąd reprezentacji wyniku, a następnie opisujemy koło. Promień tego koła jest optymalnym poziomem błędu. Wynik jest punktem koła o promieniu K razy większym. Dla pewnych punktów tego koła, położonych daleko od elipsy, nie istnieją dane $\tilde{\mathbf{b}}$, leżące blisko danych $\mathbf{b}$ i takie, że mamy dokładny wynik dla danych $\tilde{\mathbf{b}}$.

4. Rozwiązywanie układów równań liniowych

Zajmujemy się rozwiązywaniem układu równań liniowych

$$A\mathbf{x} = \mathbf{b},$$

w którym dane są: nieosobliwa macierz A o wymiarach $n \times n$ i wektor $\mathbf{b} \in \mathbb{R}^n$. Układ ten ma jednoznaczne rozwiązanie, $\mathbf{x} = A^{-1}\mathbf{b}$, ale ten wzór, poza bardzo szczególnymi przypadkami, nie nadaje się do numerycznego rozwiązywania naszego zadania (ale w rachunkach symbolicznych *nie zawahamy się go użyć*).

Uwarunkowanie układu równań liniowych

Zbadamy, jak zmieni się rozwiązanie układu, jeśli dane, tj. macierz A lub wektor $\mathbf{b}$ zaburzymy. Dla układu równań

$$A\mathbf{x}' = \mathbf{b} + \delta\mathbf{b}$$

otrzymujemy rozwiązanie

$$\mathbf{x}' = A^{-1}\mathbf{b} + A^{-1}\delta\mathbf{b} = \mathbf{x} + A^{-1}\delta\mathbf{b},$$

skąd wynika, że

$$\begin{aligned}\|\mathbf{x}' - \mathbf{x}\| &\leqslant \|A^{-1}\|\|\delta\mathbf{b}\| = \|A^{-1}\|\|\mathbf{b}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \|A^{-1}\|\|A\mathbf{x}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \\ &\leqslant \|A^{-1}\|\|A\|\|\mathbf{x}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|},\end{aligned}$$

i ostatecznie

$$\frac{\|\mathbf{x}' - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|A\|\|A^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Zaburzymy teraz macierz A, tj. będziemy rozwiązywać układ $(A + \delta A)\mathbf{x}'' = \mathbf{b}$. Mamy

$$A(I + A^{-1}\delta A)\mathbf{x}'' = \mathbf{b}.$$

Musimy założyć, że zaburzenie macierzy A jest na tyle małe, że macierz $(I + A^{-1}\delta A)$ jest nieosobliwa, dzięki czemu możemy ją odwrócić i użyć wzoru przybliżonego

$$(I + A^{-1}\delta A)^{-1} \approx I - A^{-1}\delta A.$$

Dostaniemy wtedy

$$\mathbf{x}'' \approx (I - A^{-1}\delta A)A^{-1}\mathbf{b} = A^{-1}\mathbf{b} - A^{-1}\delta A A^{-1}\mathbf{b} = \mathbf{x} - A^{-1}\delta A\mathbf{x},$$

skąd wynika przybliżona nierówność

$$\frac{\|\mathbf{x}'' - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|A\|\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}.$$

Zatem, oba zaburzenia względne danych, tj. wektora $\mathbf{b}$ i macierzy A, mogą przenieść się na wynik z czynnikiem co najwyżej $\|A\|\|A^{-1}\|$. Ten czynnik jest wskaźnikiem uwarunkowania zadania rozwiązywania układu równań $A\mathbf{x} = \mathbf{b}$ i bywa też nazywany wskaźnikiem uwarunkowania macierzy A. Jeśli przyjmiemy normę p-tą indukowaną, to mamy wskaźnik uwarunkowania macierzy A w normie p-tej, który oznaczamy symbolem $\mathrm{cond}_p A$ ($\mathrm{cond}_p A = \|A\|_p\|A^{-1}\|_p$).

Normy indukowane $\|\cdot\|_1$ i $\|\cdot\|_\infty$ macierzy A są łatwe do znalezienia. Ponieważ na ogół nie znamy (i nie tracimy czasu na znajdowanie) macierzy A^{-1}, jej normę możemy zwykle tylko oszacować. Jeśli dysponujemy dodatkową informacją o zadaniu, z którego wziął się nasz układ równań, to warto z takiej informacji skorzystać w tym celu. Szacowanie normy macierzy A^{-1} jest też w zasadzie możliwe na podstawie czynników rozkładu znalezionych podczas rozwiązywania układu jedną z metod bezpośrednich.

Metody bezpośrednie

Metody bezpośrednie możemy stosować wtedy, gdy liczba równań i niewiadomych jest mała (co najwyżej rzędu 10^3) lub gdy macierz układu jest „szczególnie łatwa", np. trójdiagonalna. Metody te, gdyby nie było błędów zaokrągleń, dawałyby dokładny wynik po wykonaniu skończenie wielu działań. Błędy zaokrągleń oczywiście to psują.

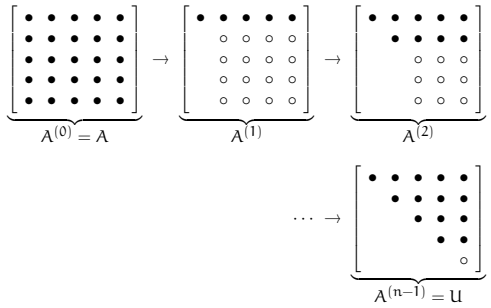
Metoda eliminacji Gaussa

Metoda eliminacji Gaussa jest najprostszym i chyba najczęściej używanym algorytmem rozwiązywania układów równań liniowych. Składa się on z dwóch etapów. W pierwszym układ jest przekształcany tak, aby powstał równoważny danemu układ równań liniowych z macierzą trójkątną górną. W etapie drugim, na podstawie kolejnych równań (od końca) obliczamy kolejne niewiadome (też od końca) — w każdym równaniu występuje tylko jedna niewiadoma, której wartość nie została obliczona wcześniej.

W pierwszym etapie (który jest właściwą eliminacją Gaussa), konstruujemy ciąg macierzy $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-1)} = U$, takich że macierz $A^{(k)}$ ma w kolumnach $1, \ldots, k$ współczynniki poniżej diagonali równe 0. Mając macierz $A^{(k-1)} = [a_{ij}^{(k-1)}]_{i,j}$, obliczamy współczynniki macierzy $A^{(k)}$:

$$\left.\begin{array}{rcl} l_{ik} & = & a_{ik}^{(k-1)}/a_{kk}^{(k-1)}, \\ a_{ij}^{(k)} & = & a_{ij}^{(k-1)} - l_{ik}a_{kj}^{(k-1)}, \quad \text{dla } j = k+1, \ldots, n \end{array}\right\} \text{dla } i = k+1, \ldots, n$$

Ponadto $a_{ij}^{(k)} = a_{ij}^{(k-1)}$ dla $i \leqslant k$, oraz $a_{ik}^{(k)} = 0$ dla $i > k$.



Przekształcanie wektora prawej strony polega na skonstruowaniu ciągu wektorów $\mathbf{b}^{(0)} = \mathbf{b}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(n-1)} = \mathbf{y}$. W k-tym kroku eliminacji obliczamy współrzędne wektora $\mathbf{b}^{(k)}$:

$$b_i^{(k)} = b_i^{(k-1)} - l_{ik}b_k^{(k-1)}, \quad \text{dla } i = k+1, \ldots, n,$$

zaś dla $i = 1, \ldots, k$ mamy $b_i^{(k)} = b_i^{(k-1)}$.

W wyniku eliminacji otrzymujemy macierz trójkątną

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ 0 & u_{22} & u_{23} & \ldots & u_{2n} \\ 0 & 0 & u_{33} & \ldots & u_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & u_{nn} \end{bmatrix},$$

i wektor $\mathbf{y}$, takie że układ $U\mathbf{x} = \mathbf{y}$ jest równoważny układowi danemu.

Niech

$$L = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ l_{21} & 1 & 0 & \ldots & 0 \\ l_{31} & l_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{n1} & l_{n2} & \ldots & l_{n,n-1} & 1 \end{bmatrix},$$

Okazuje się, że zachodzi równość $A = LU$. Przekształcanie wektora prawej strony jest równoważne rozwiązywaniu układu równań $L\mathbf{y} = \mathbf{b}$. Zatem możemy najpierw wyznaczyć tylko macierze L i U, a przetwarzanie wektora prawej strony przenieść do drugiego etapu, w którym trzeba rozwiązać kolejno układy równań z macierzami trójkątnymi, $L\mathbf{y} = \mathbf{b}$ i $U\mathbf{x} = \mathbf{y}$.

Eliminację można wykonać *w miejscu* (po łacinie *in situ*). Po obliczeniu współczynnika l_{ik}, można go zapamiętać na miejscu współczynnika $a_{ik}^{(k-1)}$ (czyli na miejscu zajmowanym początkowo przez a_{ik}). Obliczone współczynniki $a_{ij}^{(k)}$ dla $i \leqslant j$ wpisujemy w miejsce $a_{ij}^{(k-1)}$. W ten sposób otrzymamy tablicę z liczbami

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \ldots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \ldots & u_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \ldots & l_{n,n-1} & u_{nn} \end{bmatrix},$$

do której może sięgać podprogram rozwiązujący układy trójkątne. Podprogram eliminacji *in situ* „psuje" początkową zawartość tablicy. Jeśli oryginalna macierz A jest potrzebna (często jest), należy ją skopiować i „zepsuć" kopię.

Opisany wyżej algorytm jest zawodny; nieosobliwość macierzy A nie gwarantuje wykonalności dzielenia przez współczynnik $a_{kk}^{(k)}$, który może być zerem. Co więcej, jeśli współczynnik ten ma małą wartość bezwzględną, to skutki błędów zaokrągleń mogą prowadzić do otrzymania bardzo niedokładnych wyników. Dlatego stosuje się wybór elementu głównego (ang. *pivoting*). Najczęściej stosowany wybór częściowy w kolumnie polega na wyszukaniu w zbiorze $\{a_{kk}^{(k-1)}, \ldots, a_{nk}^{(k-1)}\}$ współczynnika $a_{lk}^{(k-1)}$ o największej wartości bezwzględnej, a następnie (jeśli $l \neq k$) przestawieniu równań l i k.

Jeśli macierz A jest nieosobliwa, to któraś z tych liczb nie jest zerem i dzielenie przez nią jest wykonalne. Ponieważ dzielimy przez liczbę o największej wartości bezwzględnej, współczynniki l_{ik} mają wartości bezwzględne nie większe niż 1.

Można dowieść, że skutki takiego przestawiania są takie same, jak gdyby równania zostały poprzestawiane *przed* przystąpieniem do eliminacji. Zatem, po zastosowaniu częściowego wyboru elementu głównego, otrzymamy macierze L i U, takie że $LU = PA$, gdzie P oznacza macierz dokonanej permutacji równań.

Macierz P trzeba jakoś reprezentować, aby można było odpowiednio poprzestawiać współrzędne wektora prawej strony, ponieważ trzeba będzie rozwiązać układy równań $L\mathbf{y} = P\mathbf{b}$ i $U\mathbf{x} = \mathbf{y}$. Najprostszy sposób polega na użyciu tablicy liczb całkowitych o długości n; pozycji k-tej przypisujemy indeks l wiersza, który został przestawiony z k-tym (albo k, jeśli nie było przestawienia). Przestawianie liczb zmiennopozycyjnych w tablicy jest operacją wolną od błędów zaokrągleń.

Istnieje też wybór pełny elementu głównego; przestawiamy w nim wiersze i kolumny tak, aby współczynnik $a_{kk}^{(k-1)}$, przez który będziemy dzielić, miał największą wartość bezwzględną w prawej dolnej podmacierzy $n+1-k \times n+1-k$. W ten sposób otrzymujemy rozkład macierzy $LU = PAQ^T$. Do rozwiązania mamy układy $L\mathbf{y} = P\mathbf{b}$ i $U\mathbf{z} = \mathbf{y}$, a następnie trzeba obliczyć $\mathbf{x} = Q^T\mathbf{z}$, czyli odpowiednio poprzestawiać współrzędne rozwiązania. Macierz permutacji Q można reprezentować w taki sam sposób jak P. Pełny wybór elementu głównego jest dosyć kosztowny i *bardzo rzadko* zdarza się sytuacja, gdy dokładność wyniku otrzymanego z wyborem częściowym jest za mała, a wybór pełny daje dostatecznie mały błąd.

Zwróćmy uwagę, że aby rozwiązać zadanie, rozwiazujemy numerycznie dwa podzadania, tj. układy z macierzami trójkątnymi. Dla każdego p iloczyn wskaźników uwarunkowania tych podzadań, $\mathrm{cond}_p L$ i $\mathrm{cond}_p U$, jest *zawsze* większy lub równy wskaźnikowi uwarunkowania całego zadania, $\mathrm{cond}_p A$. Ponadto dla dowolnych permutacji reprezentowanych przez macierze P i Q mamy $\mathrm{cond}_p A = \mathrm{cond}_p PAQ^T$. Wybór elementu głównego można interpretować jak dążenie do tego, aby iloczyn wskaźników uwarunkowania czynników rozkładu macierzy PA (lub PAQ^T) był możliwie mały.

Metoda eliminacji Gaussa z wyborem elementu głównego jest algorytmem numerycznie poprawnym, tj. istnieje macierz $\tilde{A}$, taka że zachodzi równość $P\tilde{A}Q^T = LU$ dla *obliczonych* macierzy L i U, oraz

$$\frac{\|\tilde{A} - A\|}{\|A\|} \leqslant F_n(A)\nu,$$

przy czym w tym wzorze jest użyta norma indukowana $\|\cdot\|_1$ lub $\|\cdot\|_\infty$. Liczba $F_n(A)$ jest stałą kumulacji; jeśli jest stosowany wybór częściowy, to ma ona duże oszacowanie ($F_n(A) \leqslant 3 \cdot 2^n - 5$), ale w praktyce stała ta jest prawie zawsze znacznie mniejsza; jej dokładniejsze oszacowanie zależy od wartości bezwzględnych obliczonych współczynników $a_{ij}^{(k)}$, a wybór elementu głównego jest pewną metodą przeciwdziałania wystąpieniu w obliczeniach wielkich liczb. Obliczenie wektora $\mathbf{x}$ przez rozwiązanie układów równań z macierzami L i U jest również numerycznie poprawnym algorytmem rozwiązywania układu równań liniowych. Rachunki w dowodzie tego twierdzenia są dosyć długie i żmudne.

Jeśli macierz A jest pełna, to wyznaczanie czynników trójkątnych ma koszt $(n^3 - n)/3$ operacji zmiennopozycyjnych (za jedną operację uznamy dzielenie lub mnożenie z dodawaniem), natomiast rozwiązywanie układów trójkątnych kosztuje n^2 takich operacji. W wielu zastosowaniach mamy do czynienia z macierzami rzadkimi, tj. mającymi dużo zerowych współczynników. W takich przypadkach *czynem karalnym* jest użycie ogólnego algorytmu, odpowiedniego dla macierzy pełnych. Jeśli np. macierz jest wstęgowa, tj. istnieje $k \ll n$, takie że $a_{ij} = 0$ dla $|i - j| > k$, to odpowiedni wariant metody eliminacji Gaussa może znaleźć czynniki trójkątne kosztem rzędu $k^2 n$ operacji, a koszt rozwiązywania układów równań z tymi czynnikami jest rzędu kn. I tego wariantu należy użyć.

Metoda odbić Householdera

Przypomnijmy własności odbić symetrycznych w $\mathbb{R}^n$. Niech $\boldsymbol{\nu}$ oznacza dowolny wektor jednostkowy (w sensie normy drugiej, tj. taki że $\|\boldsymbol{\nu}\|_2 = 1$). Odbicie symetryczne względem hiperpłaszczyzny prostopadłej do wektora $\boldsymbol{\nu}$ jest określone wzorem

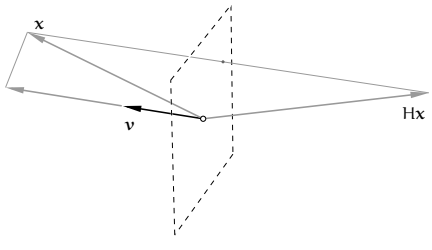
$$H\mathbf{x} = \mathbf{x} - 2\boldsymbol{\nu}\boldsymbol{\nu}^T\mathbf{x}.$$

Macierz odbicia jest więc równa

$$H = I - 2\boldsymbol{\nu}\boldsymbol{\nu}^T.$$

Odbicie jest inwolucją, tj. swoją własną odwrotnością:

$$H^2 = (I - 2\boldsymbol{\nu}\boldsymbol{\nu}^T)(I - 2\boldsymbol{\nu}\boldsymbol{\nu}^T) = I - 4\boldsymbol{\nu}\boldsymbol{\nu}^T + 4\boldsymbol{\nu}\underbrace{\boldsymbol{\nu}^T\boldsymbol{\nu}}_{=1}\boldsymbol{\nu}^T = I.$$



Macierz odbicia jest symetryczna, a zatem $H^T H = H^2 = I$, czyli macierz H jest ortogonalna. Tak więc odbicie jest izometrią; dla dowolnych wektorów $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ zachodzi równość

$$\langle H\mathbf{x}, H\mathbf{y}\rangle = \mathbf{y}^T H^T H \mathbf{x} = \mathbf{y}^T \mathbf{x} = \langle \mathbf{x}, \mathbf{y}\rangle,$$

skąd dalej wynika, że dla dowolnego wektora $\mathbf{x}$ jest $\|H\mathbf{x}\|_2 = \|\mathbf{x}\|_2$. Jeśli wektor $\boldsymbol{\nu}$ jest niezerowy, ale niekoniecznie jednostkowy, to macierz odbicia symetrycznego względem hiperpłaszczyzny prostopadłej do niego jest określona wzorem

$$H = I - \boldsymbol{\nu}\frac{2}{\boldsymbol{\nu}^T\boldsymbol{\nu}}\boldsymbol{\nu}^T.$$

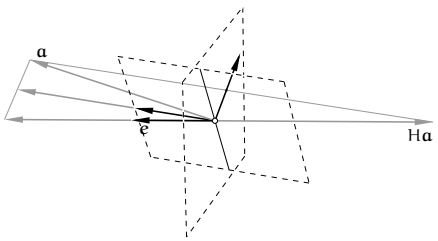
W algorytmach numerycznych, jeśli to nie jest wynikiem, który koniecznie musimy otrzymać, nigdy nie wyznaczamy jawnie macierzy H. Mając wektor $\boldsymbol{\nu}$, możemy obliczyć liczbę $\gamma = \frac{2}{\boldsymbol{\nu}^T\boldsymbol{\nu}}$, a następnie, chcąc obliczyć obraz $\mathbf{y}$ dowolnego wektora $\mathbf{x}$ w odbiciu, obliczamy kolejno

$s = \boldsymbol{\nu}^T\mathbf{x}$,
$t = \gamma s$,
$\mathbf{y} = \mathbf{x} - t\boldsymbol{\nu}$.

W tym obliczeniu należy wykonać $2n + 1$ operacji (mnożeń z dodawaniami), podczas gdy mnożenie wektora $\mathbf{x}$ przez macierz H ma koszt n^2 operacji; ponadto metoda z macierzą H reprezentowaną jawnie jest znacznie gorsza ze względu na skutki błędów zaokrągleń.

Zauważmy jeszcze jedną własność macierzy odbicia: jeśli k-ta współrzędna wektora $\boldsymbol{\nu}$ jest równa 0, to k-ty wiersz i k-ta kolumna macierzy H są takie, jak w macierzy jednostkowej. Wtedy k-ta współrzędna wektora $H\mathbf{x}$ jest identyczna, jak k-ta współrzędna wektora $\mathbf{x}$. Zatem każda zerowa współrzędna wektora $\boldsymbol{\nu}$ (jeśli wiemy, które to są) umożliwia zaoszczędzenie dwóch operacji w powyższym obliczeniu.

Niech $\mathbf{a}$ oznacza pewien wektor w $\mathbb{R}^n$. Niech $\boldsymbol{e}$ oznacza pewien ustalony wektor jednostkowy (tj. $\|\boldsymbol{e}\|_2 = 1$). Odbicie Householdera jest to odbicie H skonstruowane w taki sposób, aby wektor $H\mathbf{a}$ miał kierunek wektora $\boldsymbol{e}$. Musi być zatem $H\mathbf{a} = \pm\|\mathbf{a}\|_2\boldsymbol{e}$. To zaś oznacza, że wektor normalny hiperpłaszczyzny odbicia, $\boldsymbol{\nu}$, musi mieć kierunek wektora $\mathbf{a} - \|\mathbf{a}\|_2\boldsymbol{e}$, albo $\mathbf{a} + \|\mathbf{a}\|_2\boldsymbol{e}$. Chcąc zmniejszyć skutki błędów zaokrągleń, należy zawsze wybierać dłuższy z tych dwóch wektorów.



Zastosujmy teraz odbicia do przekształcania układu równań liniowych $A\mathbf{x} = \mathbf{b}$. W pierwszym kroku odbijemy kolumny $\mathbf{a}_1, \ldots, \mathbf{a}_n$ macierzy A i wektor prawej strony $\mathbf{b}$ tak, aby obraz $H_1\mathbf{a}_1$ pierwszej kolumny miał kierunek wektora $\boldsymbol{e}_1$. Powstanie układ $H_1A\mathbf{x} = H_1\mathbf{b}$, którego macierz ma zera w pierwszej kolumnie pod diagonalą (tj. wszystkie współczynniki oprócz pierwszego są zerowe). Jeśli teraz odrzucimy pierwsze równanie, to otrzymamy podukład, w którym nie występuje niewiadoma x_1. Podukład ten możemy dalej przekształcać w podobny sposób.

Na początku k-tego kroku mamy równoważny wyjściowemu układ równań $A^{(k-1)}\mathbf{x} = \mathbf{b}^{(k-1)}$, którego macierz $A^{(k-1)}$ ma zerowe współczynniki poniżej diagonali w pierwszych $k-1$ kolumnach. Przekształcamy macierz $\tilde{A}^{(k-1)}$, która jest prawym dolnym blokiem macierzy $A^{(k-1)}$ o wymiarach $n+1-k \times n+1-k$, oraz blok $\tilde{\mathbf{b}}^{(k-1)}$ wektora $\mathbf{b}^{(k-1)}$ złożony z jego ostatnich $n+1-k$ współczynników.

W tym celu konstruujemy wektor $\tilde{\boldsymbol{\nu}}^{(k)} \in \mathbb{R}^{n+1-k}$, dany wzorem

$$\tilde{\boldsymbol{\nu}}^{(k)} = \tilde{\mathbf{a}}_1^{(k-1)} \mp \|\tilde{\mathbf{a}}_1^{(k-1)}\|_2\boldsymbol{e}_1,$$

w którym $\tilde{\mathbf{a}}_1^{(k-1)}$ oznacza pierwszą kolumnę macierzy $\tilde{A}^{(k-1)}$ (czyli „dolną część" k-tej kolumny macierzy $A^{(k-1)}$). Pierwsza współrzędna wektora $\boldsymbol{e}_1$ jest jedynką, pozostałe $n-k$ to zera. Aby wektor $\tilde{\boldsymbol{\nu}}^{(k)}$ był jak najdłuższy, wybieramy znak „+" jeśli pierwsza współrzędna wektora $\tilde{\mathbf{a}}_1^{(k-1)}$ jest dodatnia, a „−" w przeciwnym razie. Następnie obliczamy liczbę $\gamma_k = \frac{2}{\tilde{\boldsymbol{\nu}}^{(k)T}\tilde{\boldsymbol{\nu}}^{(k)}}$ i poddajemy kolumny macierzy $\tilde{A}^{(k-1)}$ i wektor $\tilde{\mathbf{b}}^{(k-1)}$ odbiciu. Nie ma przy tym potrzeby stosowania ogólnego wzoru do odbijania wektora $\tilde{\mathbf{a}}_1^{(k-1)}$, bo skądinąd wiemy, co z tego wyjdzie.

Przekształcenie kolumn bloku $\tilde{A}^{(k-1)}$ jest równoważne odbiciu kolumn macierzy $A^{(k-1)}$ względem hiperpłaszczyzny, której wektor normalny $\boldsymbol{\nu}^{(k)}$ składa się z $k-1$ zer i z wektora $\tilde{\boldsymbol{\nu}}^{(k)}$. Oznaczmy macierz tego odbicia literą $H^{(k)}$. Po wykonaniu $n-1$ odbić mamy układ równań liniowych

$$R\mathbf{x} = Q^T\mathbf{b},$$

którego macierz

$$R = H^{(n-1)}\ldots H^{(1)}A = Q^TA$$

jest trójkątna górna. Mamy równość

$$A = QR, \quad \text{gdzie} \quad Q = H^{(1)}\ldots H^{(n-1)},$$

przy czym macierz Q, jako iloczyn macierzy ortogonalnych, jest ortogonalna. W ten sposób, za pomocą odbić symetrycznych, znaleźliśmy rozkład ortogonalno-trójkątny macierzy A.

Podobnie, jak w eliminacji Gaussa, przekształcanie prawej strony możemy wykonać później, ale w tym celu trzeba zapamiętać wektory $\tilde{\boldsymbol{\nu}}^{(k)}$ (i, aby nie obliczać ich ponownie, co kosztuje, liczby $\gamma^{(k)}$). W tym celu możemy użyć miejsc w tablicy początkowo zawierającej współczynniki macierzy A, ale potrzebujemy dla każdego wektora odbicia dwóch dodatkowych miejsc. Jeden z możliwych sposobów przechowywania wyników obliczeń jest taki:

$$\left[\begin{array}{ccccc|c}
r_{11} & r_{12} & r_{13} & \cdots & r_{1n} & \\
\nu_2^{(1)} & r_{22} & r_{23} & \cdots & r_{2n} & \\
\nu_3^{(1)} & \nu_3^{(2)} & r_{33} & \cdots & r_{3n} & \\
\vdots & \vdots & \ddots & \ddots & \vdots & \\
\nu_n^{(1)} & \nu_n^{(2)} & \cdots & \nu_n^{(n-1)} & r_{nn} & \\
\hline
\nu_1^{(1)} & \nu_2^{(2)} & \cdots & \nu_{n-1}^{(n-1)} & \bullet & \\
\hline
\gamma^{(1)} & \gamma^{(2)} & \cdots & \gamma^{(n-1)} & \bullet &
\end{array}\right]$$

Symbole r_{ij} oznaczają tu współczynniki macierzy R, zaś $v_i^{(k)}$ oznaczają współrzędne wektora $\boldsymbol{v}^{(k)}$. Jest też możliwe zmieszczenie wyników obliczenia z wykorzystaniem tylko jednej dodatkowej zmiennej dla każdej kolumny, po przeskalowaniu wektorów $\boldsymbol{v}^{(k)}$. Jak poprzednio, wykonanie obliczeń *in situ* oznacza „zepsucie" tablicy współczynników macierzy A, zatem najlepiej, aby takiemu „zepsuciu" poddać kopię.

Złożoność obliczeniowa wyznaczania rozkładu QR macierzy $n \times n$ jest równa $\frac{2}{3}n^3 + O(n^2)$, jest zatem w przybliżeniu dwukrotnie większa niż eliminacja Gaussa. Z drugiej strony, odpadają koszty wybierania elementu głównego, zresztą decydujący wpływ na czas obliczeń ma efektywność wykorzystania pamięci podręcznej (*cache*'a) procesora przez implementację algorytmu, dlatego nie można powiedzieć z góry, że eliminacja Gaussa działa dwukrotnie szybciej. Natomiast użycie izometrii (tj. przekształceń reprezentowanych przez macierze ortogonalne) daje bardzo dobre własności numeryczne algorytmu. Zauważmy, że oryginalne zadanie zastępujemy dwoma podzadaniami — układami równań $Q\mathbf{y} = \mathbf{b}$ i $R\mathbf{x} = \mathbf{y}$. Wskaźnik uwarunkowania w normie drugiej macierzy ortogonalnej Q jest równy 1 (bo $\|Q\|_2 = \|Q^{-1}\|_2 = 1$), zaś $\operatorname{cond}_2 R = \operatorname{cond}_2 A$.

Metoda Choleskiego

W wielu zastosowaniach należy rozwiązać układ $A\mathbf{x} = \mathbf{b}$, którego macierz A jest symetryczna i dodatnio określona. Dla takich macierzy można stosować eliminację Gaussa, przy czym okazuje się, że wybór elementu głównego jest niepotrzebny. Jednak symetria macierzy to okazja do zmniejszenia kosztu obliczeń o połowę. Takich okazji nie wypada marnować.

Metoda Choleskiego polega na rozłożeniu macierzy A na czynniki trójkątne, z których każdy jest transpozycją drugiego: $A = LL^T$, gdzie macierz L jest trójkątna dolna. Po znalezieniu takiego rozkładu można rozwiązać kolejno układy równań $L\mathbf{y} = \mathbf{b}$ i $L^T\mathbf{x} = \mathbf{y}$.

Macierz L można znaleźć, traktując równość $LL^T = A$ jak układ równań. Symetryczna macierz A ma $\frac{1}{2}n(n+1)$ danych niezależnych współczynników. Tyle samo współczynników na diagonali i pod nią ma poszukiwana macierz L. Zatem, dla i, j takich że $1 \leqslant j \leqslant i \leqslant n$ zachodzą równości

$$a_{ij} = \sum_{k=1}^{n} l_{ik}l_{jk} = \sum_{k=1}^{j} l_{ik}l_{jk} = \sum_{k=1}^{j-1} l_{ik}l_{jk} + l_{ij}l_{jj}.$$

Wyodrębniony składnik sumy powyżej umożliwia obliczenie współczynnika l_{ij}, jeśli znamy wszystkie pozostałe współczynniki macierzy L występujące w sumowanych iloczynach. Mianowicie, możemy obliczać kolejno

$$\left.\begin{array}{l} l_{ij} = \dfrac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}} \quad \text{dla } j = 1, \ldots, i-1, \\ l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}, \end{array}\right\} \quad \text{dla } i = 1, \ldots, n.$$

Obliczenie można wykonać *in situ*, wpisując liczby l_{ij} natychmiast po obliczeniu w miejsce a_{ij} (a zatem, „psując" daną macierz A lub jej kopię). Trzeba podkreślić, że macierz A musi być nie tylko symetryczna, ale także dodatnio określona, aby powyższy algorytm był wykonalny, tj. aby wyrażenia, z których należy obliczać pierwiastki kwadratowe, miały dodatnie wartości.

Jeśli początkowych k współczynników w i-tym wierszu (dla $k < i$) to zera, to również macierz L ma na początku i-tego wiersza k zerowych współczynników. Można to wykorzystać do efektywnego wykorzystania miejsca w pamięci i do zmniejszenia kosztu znajdowania rozkładu (np. jeśli macierz A jest wstęgowa).

Jeśli macierz A jest pełna, to też można przechowywać tylko dolny jej trójkąt w tablicy o długości $\frac{1}{2}n(n+1)$. Dla macierzy pełnej znalezienie rozkładu wymaga wykonania ok. $\frac{1}{6}n^3$ operacji mnożenia z dodawaniem lub dzielenia i obliczenia n pierwiastków kwadratowych.

Układy i algorytmy blokowe

Wiele układów równań liniowych rozwiązywanych w praktycznych zastosowaniach ma macierze o specjalnych własnościach, które można wykorzystać do zmniejszenia kosztu rozwiązywania. Bardzo często macierz w naturalny sposób dzieli się na wyróżniające się jakoś bloki. W najprostszej sytuacji, niech

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}.$$

Przypuśćmy, że macierz $A \in \mathbb{R}^{n,n}$ jest nieosobliwa i blok $B \in \mathbb{R}^{k,k}$ dla pewnego $k \in \{1, \ldots, n-1\}$ też jest nieosobliwy. Podzielmy również prawą stronę i wektor niewiadomy na bloki:

$$\mathbf{b} = \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}.$$

Układ $A\mathbf{x} = \mathbf{b}$ podzieliliśmy w ten sposób na dwa podukłady:

$$\begin{cases} B\mathbf{y} + C\mathbf{z} = \mathbf{p}, \\ D\mathbf{y} + E\mathbf{z} = \mathbf{q}. \end{cases}$$

Znając $\mathbf{z}$, moglibyśmy rozwiązać pierwszy podukład; jego rozwiązanie wyraża się wzorem

$$\mathbf{y} = B^{-1}(\mathbf{p} - C\mathbf{z}).$$

Wstawiamy to wyrażenie do drugiego podukładu; mamy

$$DB^{-1}(\mathbf{p} - C\mathbf{z}) + E\mathbf{z} = \mathbf{q},$$

czyli

$$(E - DB^{-1}C)\mathbf{z} = \mathbf{q} - DB^{-1}\mathbf{p}.$$

Macierz $S = E - DB^{-1}C$ nazywa się macierzą Schura; jeśli macierze A i B są nieosobliwe, to również macierz S jest nieosobliwa.

Powiedzmy, że mamy macierz permutacji P i macierze trójkątne L i U, takie że $PB = LU$. Wtedy możemy wykonać następujący algorytm:

1. Korzystając z macierzy P, L, U, rozwiąż (macierzowy) układ równań $BF = C$, a następnie oblicz macierz $S = E - DF$,
2. Rozwiąż układ równań $B\boldsymbol{v} = \mathbf{p}$, a następnie oblicz wektor $\boldsymbol{w} = \mathbf{q} - D\boldsymbol{v}$,
3. Rozwiąż układ równań $S\mathbf{z} = \boldsymbol{w}$; w tym celu należy wyznaczyć i wykorzystać jakiś użyteczny rozkład macierzy S,
4. Oblicz wektor $\mathbf{u} = \mathbf{p} - C\mathbf{z}$, a następnie, korzystając z macierzy P, L, U, rozwiąż układ równań $B\mathbf{y} = \mathbf{u}$.
   Alternatywnie, oblicz wektor $\mathbf{t} = C\mathbf{z}$, a następnie rozwiąż układ $B\mathbf{s} = \mathbf{t}$ i oblicz $\mathbf{y} = \boldsymbol{v} - \mathbf{s}$.

Z wyjątkiem ograniczenia możliwości wyboru elementu głównego, nie mamy tu żadnych zmian (w szczególności kosztu) w porównaniu ze zwykłą eliminacją Gaussa (choć pierwsze dwa kroki można wykonać równolegle). Jeśli jednak blok B jest macierzą symetryczną dodatnio określoną, to zamiast eliminacji Gaussa możemy użyć dwukrotnie tańszej metody Choleskiego. Jeśli zaś blok B jest na przykład macierzą diagonalną lub ortogonalną, to niezależnie od tego, jakie są pozostałe bloki macierzy A, układy równań z macierzą B możemy rozwiązywać znacznie mniejszym kosztem. Ponadto, gdyby blok B był macierzą odbicia symetrycznego, reprezentowaną przez wektor normalny hiperpłaszczyzny odbicia (lub iloczynem takich macierzy, reprezentowanych przez odpowiednie wektory), to jawne wyznaczanie współczynników macierzy B, po to by następnie rozwiązać układ równań z tą macierzą, byłoby przejawem *skrajnego niedołęstwa*. Dlatego *najpierw* należy się dowiedzieć jak najwięcej o zadaniu, a *potem* dobierać algorytm.

Szacowanie błędu i poprawianie rozwiązania

Oznaczmy symbolem $\boldsymbol{\alpha}$ *dokładne* rozwiązanie układu równań $A\mathbf{x} = \mathbf{b}$ (czyli $\boldsymbol{\alpha} = A^{-1}\mathbf{b}$), i niech symbol $\tilde{\mathbf{x}}$ oznacza wynik *numerycznego* rozwiązywania tego układu (jakimś algorytmem z błędami zaokrągleń). Residuum rozwiązania $\tilde{\mathbf{x}}$, tj. wektor $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$, jest równe $\mathbf{0}$ wtedy i tylko wtedy, gdy $\tilde{\mathbf{x}} = \boldsymbol{\alpha}$. Możemy napisać

$$\boldsymbol{\alpha} - \tilde{\mathbf{x}} = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = A^{-1}\mathbf{r}.$$

Z równości $\boldsymbol{\alpha} - \tilde{\mathbf{x}} = A^{-1}\mathbf{r}$ oraz $A(\boldsymbol{\alpha} - \tilde{\mathbf{x}}) = \mathbf{r}$ wynikają nierówności

$$\frac{1}{\|A\|_p}\|\mathbf{r}\|_p \leqslant \|\boldsymbol{\alpha} - \tilde{\mathbf{x}}\|_p \leqslant \|A^{-1}\|_p\|\mathbf{r}\|_p.$$

Możemy użyć tych nierówności do oszacowania wielkości błędu rozwiązania, pod warunkiem, że

1. umiemy oszacować normę macierzy A^{-1} (dla $p = 1$ lub $p = \infty$ obliczenie $\|A\|_p$ jest łatwe, ale nie chcemy jawnie wyznaczać macierzy A^{-1}),
2. umiemy obliczyć wektor $\mathbf{r}$.

Jeśli do rozwiązania układu użyliśmy metody eliminacji Gaussa, to mamy znalezione trójkątne czynniki L, U rozkładu macierzy A (lub PA albo PAQ^T, zależnie od użytego wariantu wyboru elementu głównego). Istnieje algorytm, który na podstawie tych czynników znajduje, kosztem $O(n^2)$ działań, normę macierzy A^{-1} z dokładnością rzędu 50%, co w zastosowaniu do sprawdzania dokładności otrzymanego rozwiązania wystarczy.

Znacznie gorzej wygląda kwestia obliczenia residuum — robiąc to przy użyciu arytmetyki zmiennopozycyjnej, dostaniemy inny wektor, $\tilde{\mathbf{r}}$, przy czym w tym obliczeniu występuje silne znoszenie się składników, wskutek czego znaleziona potem liczba $\|\tilde{\mathbf{r}}\|_p$ może mieć bardzo niewiele wspólnego z $\|\mathbf{r}\|_p$.

Z tego powodu podczas obliczania residuum trzeba zadbać o dokładność. Najprostszym sposobem jest użycie *silniejszej arytmetyki*, jeśli na przykład domyślnie używamy pojedynczej precyzji, to residuum powinniśmy obliczyć w precyzji podwójnej. Jeśli do rozwiązania układu użyliśmy precyzji podwójnej, to można sięgnąć po precyzję rozszerzoną, lub użyć algorytmu Kahana. Działania w wyższej precyzji mogą zajmować więcej czasu, ale całe to obliczenie ma złożoność $\Theta(n^2)$, co jest mało istotne w porównaniu ze złożonością eliminacji Gaussa, rzędu n^3.

Jeśli otrzymane oszacowanie błędu jest za duże, to rozwiązanie można poprawić (w praktyce rzadko się to robi, ale *każdy informatyk powinien wiedzieć*, jak to zrobić, jeśli pojawi się taka konieczność). W tym celu wystarczy rozwiązać układ równań

$$A\boldsymbol{\delta} = \tilde{\mathbf{r}},$$

a następnie obliczyć poprawione rozwiązanie

$$\hat{\mathbf{x}} = \tilde{\mathbf{x}} + \boldsymbol{\delta}.$$

Koszt tego postępowania jest rzędu n^2, ponieważ do rozwiązania układu równań z wektorem prawej strony $\tilde{\mathbf{r}}$ wykorzystujemy trójkątne czynniki rozkładu znalezione wcześniej (jeśli do rozwiązania układu używamy innego rozkładu macierzy A, np. ortogonalno-trójkątnego wyznaczonego metodą odbić Householdera, to też otrzymamy koszt poprawiania rzędu n^2).

Rozwinięciem tego postępowania jest iteracyjne poprawianie rozwiązania, w którym po obliczeniu nowego przybliżenia, $\hat{\mathbf{x}}$, obliczamy jego residuum i w razie potrzeby poprawiamy je dalej. Iteracje przerywamy, jeśli norma residuum jest dostatecznie mała, lub jeśli nie jest jest istotnie mniejsza od normy residuum poprzedniego przybliżenia — to oznacza osiągnięcie maksymalnej granicznej dokładności. Kluczowym dla dokładności elementem tego postępowania jest dokładność obliczania wektorów residuum.

Metody iteracyjne

Jeśli liczba n równań i niewiadomych jest wielka, to koszt metod bezpośrednich rozwiązywania takich układów jest zbyt duży. Często w zastosowaniach pojawiają się układy z n rzędu tysięcy lub milionów. Bardzo często macierze układów w takich zastosowaniach są rzadkie, np. mają tylko $O(n)$ niezerowych współczynników, ale rozmieszczenie tych współczynników uniemożliwia stosowanie metod bezpośrednich (np. znalezienie metodą różnic skończonych rozwiązania przybliżonego równania różniczkowego Poissona w kwadracie sprowadza się do rozwiązania układu równań liniowych, którego macierz jest wstęgowa, przy czym wstęga ma szerokość rzędu $\sqrt{n}$, a w każdym wierszu jest co najwyżej 5 współczynników niezerowych).

Do rozwiązywania wielkich układów równań liniowych stosuje się metody iteracyjne. Mając początkowe przybliżenie $\mathbf{x}_0$ rozwiązania $\boldsymbol{\alpha}$, metoda konstruuje elementy ciągu, $\mathbf{x}_1, \mathbf{x}_2, \ldots$, zbieżnego do $\boldsymbol{\alpha}$. Obliczenia przerywa się na podstawie kryteriów stopu podobnych, jak dla równań nieliniowych, tj. ustalonego limitu liczby iteracji, kryterium residualnego (residuum jest w tym przypadku wektor $\mathbf{b} - A\mathbf{x}$) lub kryterium przyrostowego (opartego na badaniu wartości wyrażenia $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$). Podstawową operacją, wykonywaną w każdej iteracji, jest mnożenie pewnego wektora przez macierz układu A, lub inną macierz, skonstruowaną na podstawie A. Dzięki takiemu ograniczeniu można korzystać z bardzo oszczędnych reprezentacji macierzy, które są w istocie wykazami (tablicami lub listami) miejsc, w których są niezerowe współczynniki. Koszt mnożenia wektora przez macierz jest wtedy proporcjonalny do liczby tych współczynników.

Metody iteracji prostej

Metody iteracji prostej polegają na tym, że na podstawie macierzy A i wektora prawej strony $\mathbf{b}$ konstruuje się pewną macierz B i wektor $\mathbf{t}$, przyjmuje początkowe przybliżenie rozwiązania, $\mathbf{x}_0$, i w kolejnych iteracjach oblicza

$$\mathbf{x}_{k+1} = B\mathbf{x}_k + \mathbf{t}.$$

Macierz B i wektor $\mathbf{t}$ muszą być tak dobrane, aby zachodziła równość $\boldsymbol{\alpha} = B\boldsymbol{\alpha} + \mathbf{t}$, tj. aby rozwiązanie było punktem stałym funkcji $\boldsymbol{\varphi}(\mathbf{x}) = B\mathbf{x} + \mathbf{t}$. Ponadto, aby ciąg wektorów $\mathbf{x}_k$ był zbieżny do $\boldsymbol{\alpha}$ dla dowolnego punktu startowego $\mathbf{x}_0$, funkcja $\boldsymbol{\varphi}$ musi być przekształceniem zwężającym, a zatem pewna (dowolna) norma indukowana macierzy B musi być mniejsza od 1.

Warunkiem koniecznym i dostatecznym istnienia takiej normy, która dla macierzy B przyjmuje wartość mniejszą od 1, jest nierówność $\rho(B) < 1$, gdzie $\rho(B)$ jest to promień spektralny, tj. największa wartość bezwzględna wartości własnej macierzy B. Aby zbieżność była szybka, $\rho(B)$ musi być jak najmniejsze, ale możliwość skonstruowania macierzy B o małym promieniu spektralnym zależy od macierzy A (i w szczególności od jej wskaźnika uwarunkowania).

Metoda Jacobiego: macierz A przedstawiamy w postaci sumy $A = L + D + U$, gdzie macierz L powstaje z A przez zastąpienie zerami współczynników na i nad diagonalą, macierz D jest diagonalna, a macierz U ma zerowe współczynniki na i pod diagonalą. Układ $A\mathbf{x} = \mathbf{b}$ przepisujemy w postaci

$$(L + D + U)\mathbf{x} = \mathbf{b}, \quad \text{czyli} \quad D\mathbf{x} = \mathbf{b} - (L + U)\mathbf{x},$$

skąd otrzymujemy metodę:

$$\mathbf{x}_{k+1} = D^{-1}\big(\mathbf{b} - (L + U)\mathbf{x}_k\big).$$

W metodzie Jacobiego mamy zatem $B_J = -D^{-1}(L + U)$ oraz $\mathbf{t}_J = D^{-1}\mathbf{b}$.

W obliczeniach nie wyznaczamy macierzy D^{-1}, choć to jest łatwe; zamiast tego rozwiązujemy układ równań z macierzą diagonalną D. Warunek $\rho(B_J) < 1$ jest spełniony dla wielu macierzy A; łatwym do sprawdzenia przypadkiem jest macierz diagonalnie dominująca, tj. taka że $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ dla każdego i.

Metoda Gaussa-Seidela: jak poprzednio, bierzemy $A = L + D + U$, po czym piszemy układ

$$(L + D)\mathbf{x} = \mathbf{b} - U\mathbf{x},$$

skąd otrzymujemy układ równań do rozwiązania w każdej iteracji:

$$(L + D)\mathbf{x}_{k+1} = \mathbf{b} - U\mathbf{x}_k.$$

W metodzie Gaussa-Seidela mamy zatem $B_{GS} = -(L + D)^{-1}U$ i $\mathbf{t}_{GS} = (L + D)^{-1}\mathbf{b}$. Podobnie jak w metodzie Jacobiego, jeśli macierz A jest diagonalnie dominująca, to ciąg $(\mathbf{x}_k)_{k \in \mathbb{N}}$ jest zbieżny do $\boldsymbol{\alpha}$ dla każdego $\mathbf{x}_0 \in \mathbb{R}^n$, przy czym zbieżność metody Gaussa-Seidela jest zwykle szybsza.

Metoda Richardsona: Wprowadzamy parametr τ i piszemy układ równoważny układowi $A\mathbf{x} = \mathbf{b}$:

$$\mathbf{x} + \tau A\mathbf{x} = \mathbf{x} + \tau\mathbf{b},$$

skąd mamy

$$\mathbf{x} = (I - \tau A)\mathbf{x} + \tau\mathbf{b} = \mathbf{x} + \tau(\mathbf{b} - A\mathbf{x}).$$

Metoda Richardsona polega na obliczaniu wektorów

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau(\mathbf{b} - A\mathbf{x}_k).$$

W tej metodzie mamy $B_R = I - \tau A$, $\mathbf{t}_R = \tau\mathbf{b}$. Podstawowym problemem jest dobranie parametru τ tak, aby ciąg $(\mathbf{x}_k)_k$ był zbieżny i aby ta zbieżność była jak najszybsza. Jeszcze do tego wrócimy.

Metoda sprzężonych gradientów

Metoda sprzężonych gradientów (ang. *conjugate gradient method*, w skrócie *CG*) służy do rozwiązywania układu n równań liniowych $A\mathbf{x} = \mathbf{b}$ z macierzą A symetryczną i dodatnio określoną. Co ciekawe, z punktu widzenia algebry metoda ta jest metodą bezpośrednią, mianowicie startując z dowolnego punktu $\mathbf{x}_0$, wytwarza skończony ciąg $\mathbf{x}_1, \ldots, \mathbf{x}_m$ dla pewnego $m \leqslant n$; gdyby nie było błędów zaokrągleń, to ostatni element tego ciągu byłby rozwiązaniem $\boldsymbol{\alpha}$.

Błędy zaokrągleń dla układów z macierzami wielkimi i źle uwarunkowanymi (jakie często występują w zastosowaniach) likwidują tę własność — wytworzony ciąg punktów początkowo zbliża się do rozwiązania, a następnie oddala. Dlatego zwykle metodę CG wykorzystuje się jak metodę iteracyjną, która wytwarza dostatecznie dokładne przybliżenie $\mathbf{x}_k$ rozwiązania dla pewnego k znacznie mniejszego niż n.

Podstawowym krokiem metody CG jest minimalizacja wielomianu kwadratowego wzdłuż pewnej prostej. Funkcja określona wzorem

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}$$

jest wielomianem drugiego stopnia zmiennych $x_1, \ldots, x_n$.

Ponieważ macierz A jest dodatnio określona, funkcja f ma minimum; jej gradient jest równy

$$\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b},$$

a więc wektor residuum, $\mathbf{r} = \mathbf{b} - A\mathbf{x} = -\nabla f(\mathbf{x})$, określa kierunek najszybszego spadku funkcji f w punkcie $\mathbf{x}$.
Poszukiwane rozwiązanie $\boldsymbol{\alpha}$ jest właśnie tym punktem przestrzeni $\mathbb{R}^n$, w którym funkcja f przyjmuje wartość minimalną.

Niech $\mathbf{x}_k$ oznacza bieżące przybliżenie rozwiązania, zaś $\mathbf{v}_k$ oznacza pewien niezerowy wektor. Zbiór $L = \{\mathbf{x} = \mathbf{x}_k + t\mathbf{v}_k \colon t \in \mathbb{R}\}$ jest prostą w $\mathbb{R}^n$. Podstawiając jako argument funkcji f wyrażenie $\mathbf{x}_k + t\mathbf{v}_k$, otrzymujemy wielomian jednej zmiennej,

$$\begin{aligned} g(t) &= f(\mathbf{x}_k + t\mathbf{v}_k) \\ &= \frac{1}{2}\big(\mathbf{v}_k^T A\mathbf{v}_k t^2 - 2\mathbf{v}_k^T(\mathbf{b} - A\mathbf{x}_k)t + \mathbf{x}_k^T(A\mathbf{x}_k - 2\mathbf{b})\big). \end{aligned}$$

Trójmian kwadratowy $g(t) = at^2 - 2bt + c$ ze współczynnikiem $a > 0$ przyjmuje wartość minimalną dla $t = b/a$. Zatem, obliczając

$$\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k, \qquad t_k = \frac{\mathbf{v}_k^T \mathbf{r}_k}{\mathbf{v}_k^T A\mathbf{v}_k}, \qquad \mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{v}_k,$$

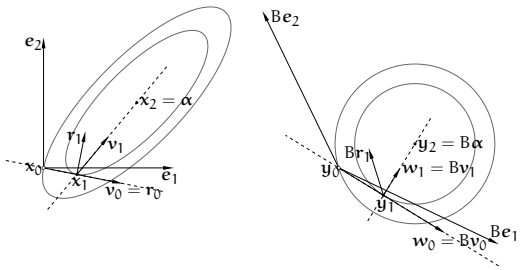
otrzymamy punkt $\mathbf{x}_{k+1}$ prostej L, w którym wartość funkcji f jest najmniejsza. Możemy sprawdzić, że następny wektor residuum, $\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{r}_k - t_k A\mathbf{v}_k$, spełnia warunek $\mathbf{v}_k^T \mathbf{r}_{k+1} = 0$.

Istnieje rodzina metod iteracyjnych, w których powyższe obliczenie łączy się z pewną strategią wyboru wektorów $\mathbf{v}_k$, wyznaczających kierunki odpowiednich prostych w kolejnych iteracjach; oczywiście, istotne są tylko *kierunki* tych wektorów. W metodzie CG, która należy do tej rodziny, wektory $\mathbf{v}_k$ mają kierunki sprzężone względem macierzy A, tzn. spełniają równości $\mathbf{v}_i^T A\mathbf{v}_k = 0$ dla $i \neq k$. Dla dowolnej symetrycznej i dodatnio określonej macierzy A istnieje (jednoznacznie określona) macierz B symetryczna i dodatnio określona taka że $A = B^2$. Niech $\mathbf{w}_i = B\mathbf{v}_i$ dla każdego i. Dla $i \neq k$ mamy $\langle \mathbf{w}_k, \mathbf{w}_i \rangle = \mathbf{w}_i^T \mathbf{w}_k = 0$. Po podstawieniu $\mathbf{x} = B^{-1}\mathbf{y}$ jako argumentu funkcji f, otrzymujemy wielomian

$$\begin{aligned} h(\mathbf{y}) = f(B^{-1}\mathbf{y}) &= \frac{1}{2}\mathbf{y}^T B^{-T} A B^{-1}\mathbf{y} - \mathbf{y}^T B^{-T}\mathbf{b} \\ &= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T B^{-T}\mathbf{b}, \end{aligned}$$

którego część kwadratowa jest równa $\frac{1}{2}(y_1^2 + \cdots + y_n^2)$.

Poszukiwanie minimum funkcji $f(\mathbf{x})$ wzdłuż prostych o kierunkach $\mathbf{v}_0, \ldots, \mathbf{v}_k$, zaczynając od punktu $\mathbf{x}_0$, jest równoważne minimalizacji funkcji $h(\mathbf{y})$ wzdłuż prostych o wzajemnie prostopadłych kierunkach $\mathbf{w}_0, \ldots, \mathbf{w}_k$, zaczynając od punktu $\mathbf{y}_0 = B\mathbf{x}_0$.



Na rysunku wektor $\mathbf{v}_0 = \mathbf{r}_0$ jest prostopadły do przechodzącej przez $\mathbf{x}_0$ warstwicy funkcji f, która jest elipsą o środku $\boldsymbol{\alpha}$. Punkt $\mathbf{x}_1$ leży na kolejnej warstwicy; residuum w tym punkcie, $\mathbf{r}_1$, ma kierunek najszybszego spadku funkcji f. Obrazy $\mathbf{w}_0$ i $\mathbf{w}_1$ wektorów $\mathbf{v}_0$ i $\mathbf{v}_1$ w przekształceniu liniowym określonym przez macierz B są prostopadłe do siebie; obrazami warstwic funkcji f, tj. warstwicami funkcji h, są okręgi. Prosta przechodząca przez punkt $\mathbf{y}_1$ o kierunku wektora $\mathbf{w}_1$ przechodzi przez środek $B\boldsymbol{\alpha}$ tych okręgów, zatem wektor $\mathbf{v}_1$ wyznacza kierunek prostej łączącej punkt $\mathbf{x}_1$ i rozwiązanie $\boldsymbol{\alpha}$.

W ogólności do znalezienia w $\mathbb{R}^n$ minimum funkcji kwadratowej h, której warstwice są sferami, wystarczy wykonać co najwyżej n kroków minimalizacji wzdłuż prostych wzajemnie prostopadłych.

W metodzie CG przyjmujemy $\mathbf{v}_0 = \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ oraz

$$\mathbf{v}_{k+1} = \mathbf{r}_{k+1} + s_k\mathbf{v}_k, \text{ gdzie } s_k = -\frac{\mathbf{v}_k^T A\mathbf{r}_{k+1}}{\mathbf{v}_k^T A\mathbf{v}_k},$$
$$\text{dla } k \geqslant 0, \text{ takiego że } \mathbf{v}_k \neq \mathbf{0}.$$

Twierdzenie. *Otrzymane w ten sposób wektory $\mathbf{v}_k$ mają kierunki sprzężone względem macierzy A i metoda znajduje rozwiązanie po co najwyżej n iteracjach.*

Szkic dowodu: Zauważamy, że dla każdego $j \leqslant k+1$ wektor $\mathbf{v}_j$ jest kombinacją liniową wektorów $\mathbf{r}_0, \ldots, \mathbf{r}_j$, a wektor $\mathbf{r}_j$ jest kombinacją liniową wektorów $\mathbf{v}_0, \ldots, \mathbf{v}_j$. Niech K_j oznacza podprzestrzeń przestrzeni $\mathbb{R}^n$ rozpiętą przez te wektory. Przed obliczeniem wektora $\mathbf{v}_{k+1}$ były wyznaczone wektory $\mathbf{r}_0, \ldots, \mathbf{r}_{k+1}$ (oraz $\mathbf{v}_0, \ldots \mathbf{v}_k$), które określają przestrzenie $K_0 \subset \cdots \subset K_{k+1}$. Przyjmujemy założenie indukcyjne, że dla każdego wektora $\mathbf{u} \in K_i$, gdzie $0 \leqslant i < j \leqslant k$, jest $\mathbf{u}^T A\mathbf{v}_j = \mathbf{u}^T\mathbf{r}_j = 0$. Równoważnie, dla $0 \leqslant i < j \leqslant k$ są spełnione równości $\mathbf{v}_i^T A\mathbf{v}_j = \mathbf{r}_i^T A\mathbf{v}_j = \mathbf{v}_i^T\mathbf{r}_j = \mathbf{r}_i^T\mathbf{r}_j = 0$.

Wiemy, że residuum w punkcie $\mathbf{x}_{k+1}$ spełnia warunek $\mathbf{v}_k^T\mathbf{r}_{k+1} = 0$. Mamy też

$$\mathbf{v}_k^T A\mathbf{v}_{k+1} = \mathbf{v}_k^T A(\mathbf{r}_{k+1} + s_k\mathbf{v}_k) = \mathbf{v}_k^T A\mathbf{r}_{k+1} - \frac{\mathbf{v}_k^T A\mathbf{r}_{k+1}}{\mathbf{v}_k^T A\mathbf{v}_k}\mathbf{v}_k^T A\mathbf{v}_k = 0.$$

Niech $i < k$. Jeśli $\mathbf{u} \in K_i$, to

$$\mathbf{u}^T\mathbf{r}_{k+1} = \mathbf{u}^T(\mathbf{r}_k - t_k A\mathbf{v}_k) = \mathbf{u}^T\mathbf{r}_k - t_k\mathbf{u}^T A\mathbf{v}_k = 0.$$

Jeśli zatem $\mathbf{u} \in K_k$, to $\mathbf{u}^T\mathbf{r}_{k+1} = 0$, bo wektor $\mathbf{u}$ jest kombinacją liniową wektora $\mathbf{v}_k$ i pewnego wektora $\mathbf{w} \in K_{k-1}$.
Ponieważ $\mathbf{v}_{k+1} = \mathbf{r}_{k+1} + s_k\mathbf{v}_k$ oraz $t_i A\mathbf{v}_i = \mathbf{r}_i - \mathbf{r}_{i+1} \in K_{i+1} \subset K_k$, mamy również

$$\begin{aligned} \mathbf{v}_i^T A\mathbf{v}_{k+1} &= \mathbf{v}_i^T A(\mathbf{r}_{k+1} + s_k\mathbf{v}_k) \\ &= \mathbf{v}_i^T A\mathbf{r}_{k+1} + s_k \underbrace{\mathbf{v}_i^T A\mathbf{v}_k}_{=0} = \frac{1}{t_i}(\mathbf{r}_i - \mathbf{r}_{i+1})^T\mathbf{r}_{k+1} = 0. \end{aligned}$$

Z założenia indukcyjnego i wykazanych wyżej równości wynika, że $\mathbf{u}^T A\mathbf{v}_{k+1} = \mathbf{u}^T\mathbf{r}_{k+1} = 0$ dla każdego wektora $\mathbf{u} \in K_k$.

Z powyższego rachunku wynika, że wektory $\mathbf{r}_0, \mathbf{r}_1, \ldots$ są do siebie nawzajem prostopadłe. Ale ciąg ortogonalny w $\mathbb{R}^n$ może składać się z co najwyżej n niezerowych wektorów, zatem residuum w którymś punkcie $\mathbf{x}_k$, otrzymanym po co najwyżej n iteracjach metody CG, musi być zerowe. □

Dzięki temu, że $\mathbf{v}_k^T\mathbf{r}_k = \mathbf{r}_k^T\mathbf{r}_k$, wyrażenia opisujące t_k i s_k można przekształcić tak, aby zmniejszyć koszty ich obliczania:

$$\begin{aligned} t_k &= \frac{\mathbf{v}_k^T\mathbf{r}_k}{\mathbf{v}_k^T A\mathbf{v}_k} = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{v}_k^T A\mathbf{v}_k}, \\ s_k &= -\frac{\mathbf{v}_k^T A\mathbf{r}_{k+1}}{\mathbf{v}_k^T A\mathbf{v}_k} = -\frac{1}{t_k}\frac{(\mathbf{r}_k - \mathbf{r}_{k+1})^T\mathbf{r}_{k+1}}{\mathbf{v}_k^T A\mathbf{v}_k} = \frac{\mathbf{v}_k^T A\mathbf{v}_k}{\mathbf{r}_k^T\mathbf{r}_k}\frac{\mathbf{r}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{v}_k^T A\mathbf{v}_k} \\ &= \frac{\mathbf{r}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{r}_k^T\mathbf{r}_k}. \end{aligned}$$

Teoretycznie obliczenia należy zakończyć po otrzymaniu $\mathbf{r}_k = \mathbf{0}$. W implementacji korzystającej z arytmetyki zmiennopozycyjnej obliczenia są przerywane, gdy wektor $\mathbf{v}_k$ lub $\mathbf{r}_{k+1}$ ma dostatecznie małą normę drugą.

Metodę CG realizuje następujący podprogram:

```
r = b − Ax;  /* r = r_0 */
v = r;       /* v = v_0 */
c = r^T r;
for ( k = 0; k < n; k++ ) {
  if ( v^T v < δ^2 ) return;
  z = Av;
  t = c/(v^T z);  /* t = t_k */
  x = x + tv;  /* x = x_{k+1} */
  r = r − tz;  /* r = r_{k+1} */
  d = r^T r;
  if ( d < ε^2 ) return;
  v = r + (d/c)v;  /* v = v_{k+1} */
  c = d;
}
```


Tablica x początkowo zawiera współrzędne wektora x_0. Obliczone przybliżenie rozwiązania jest końcową zawartością tej tablicy; oprócz niej podprogram używa jeszcze trzech tablic o długości n. Mnożenie wektora przez macierz A może być realizowane przez podprogram podany jako parametr i będący „czarną skrzynką” dla implementacji metody. Parametry δ i ε określają kryteria stopu.

Poprawianie uwarunkowania

Rozważmy metodę Richardsona. Jeśli macierz A jest symetryczna i dodatnio określona, to jej wartości własne są rzeczywiste i dodatnie, zawarte w przedziale $[\lambda_{min}, \lambda_{max}]$. Macierz B_R o najmniejszym promieniu spektralnym otrzymamy, przyjmując

$$\tau = \tau_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}}.$$

Często mamy pewne informacje na temat wartości własnych macierzy A, co umożliwia dobranie optymalnego lub prawie optymalnego parametru. Przyjrzyjmy się jednak szybkości zbieżności. Promień spektralny macierzy $B_R = I - \tau_{opt}A$ (który dla macierzy symetrycznej jest równy jej normie drugiej indukowanej) jest równy

$$\rho(B_R) = 1 - \tau_{opt}\lambda_{min} = 1 - \frac{2\lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

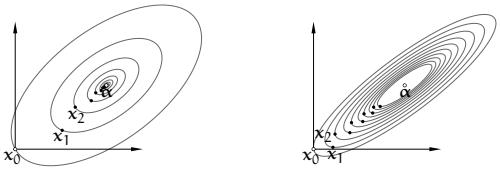
Dla macierzy symetrycznej i dodatnio określonej jest $\mathrm{cond}_2 A = \lambda_{max}/\lambda_{min}$. Korzystając z tej formuły, otrzymujemy

$$\rho(B_R) = \frac{\mathrm{cond}_2 A - 1}{\mathrm{cond}_2 A + 1}.$$

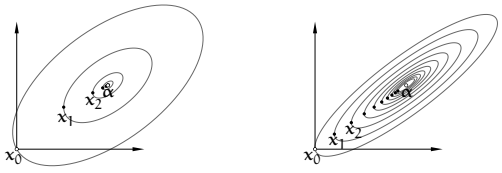
Tak więc, nawet jeśli wybierzemy optymalnie parametr τ, jeśli macierz A ma wielki wskaźnik uwarunkowania, zbieżność konstruowanego przez metodę Richardsona ciągu $(x_k)_{k\in\mathbb{N}}$ do rozwiązania jest bardzo wolna. Podobne spostrzeżenie dotyczy także innych metod iteracyjnych. W szczególności dla metody CG błąd $\varepsilon_k = x_k - \alpha$ ma oszacowanie (z macierzą $B = B^T$, taką że $B^2 = A$)

$$\|B\varepsilon_k\|_2 \leqslant 2\left(\frac{\sqrt{\mathrm{cond}_2 A} - 1}{\sqrt{\mathrm{cond}_2 A} + 1}\right)^k \|B\varepsilon_0\|_2.$$

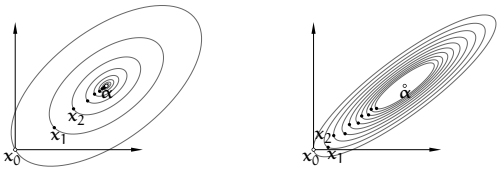
metoda Jacobiego:



metoda Gaussa-Seidela:



metoda Richardsona:



Rysunki przedstawiają eksperyment, w którym były rozwiązywane dwa układy równań o tym samym rozwiązaniu, których macierze symetryczne i dodatnio określone mają wskaźniki uwarunkowania odpowiednio 4 i 16. Pokazane są ciągi przybliżeń rozwiązania wygenerowane przez opisane wcześniej metody (dla metody Richardsona przyjęty został optymalny parametr τ) i warstwice wielomianu kwadratowego f rozważanego w opisie metody CG, przechodzące przez kolejne punkty x_k.

Zbieżność metod iteracyjnych można przyspieszyć, zastępując układ dany układem równoważnym, którego macierz ma mniejszy wskaźnik uwarunkowania. Cel ten można osiągnąć za pomocą macierzy S o następujących własnościach: układy równań z macierzą S są łatwe do rozwiązania, i zachodzi nierówność $\operatorname{cond}(S^{-1}A) \ll \operatorname{cond} A$; aby tak było, macierz S musi w jakiś sposób przybliżać macierz A. Metodę iteracyjną stosujemy do układu $S^{-1}A\mathbf{x} = S^{-1}\mathbf{b}$, przy czym nigdy nie wyznaczamy macierzy $S^{-1}A$; zamiast tego, mając obliczyć wektor $\mathbf{u} = S^{-1}A\boldsymbol{v}$, obliczamy $\boldsymbol{w} = A\boldsymbol{v}$ i rozwiązujemy układ $S\mathbf{u} = \boldsymbol{w}$.

Nawet jeśli macierze A i S są symetryczne, macierz $S^{-1}A$ na ogół nie jest taka. Niektóre metody, na przykład CG, wymagają, aby macierz układu była symetryczna. Aby zachować symetrię, wybieramy taką macierz C, aby układy równań z nią były łatwe do rozwiązania i aby było $\operatorname{cond}(C^{-T}AC^{-1}) \ll \operatorname{cond} A$. Układ dany zastępujemy układem równań

$$C^{-T}AC^{-1}\hat{\mathbf{x}} = C^{-T}\mathbf{b}, \qquad (*)$$

przy czym mamy równość $\mathbf{x} = C^{-1}\hat{\mathbf{x}}$. Mniejszy wskaźnik uwarunkowania macierzy $C^{-T}AC^{-1}$ przekłada się na szybszą zbieżność metod iteracyjnych.

Metoda CG z poprawianiem uwarunkowania
(ang. *preconditioned conjugate gradient method*, *PCG*)

Wprowadzimy kilka oznaczeń: symbole $\mathbf{x}_k$ i $\mathbf{r}_k$ oznaczają kolejne przybliżenie rozwiązania układu $A\mathbf{x} = \mathbf{b}$ i wektor residuum, zaś $\hat{\mathbf{x}}_k = C\mathbf{x}_k$ i $\hat{\mathbf{r}}_k$ to odpowiednio kolejne przybliżenie rozwiązania i residuum układu (*). Wektor $\boldsymbol{v}_k$ wyznacza kierunek prostej, na której leży punkt $\mathbf{x}_k$ i kolejny punkt $\mathbf{x}_{k+1}$, zaś symbolem $\hat{\boldsymbol{v}}_k$ oznaczymy wektor kierunkowy prostej łączącej punkty $\hat{\mathbf{x}}_k$ i $\hat{\mathbf{x}}_{k+1}$. Dla każdego k jest

$$\hat{\mathbf{r}}_k = C^{-T}\mathbf{b} - C^{-T}AC^{-1}\hat{\mathbf{x}}_k = C^{-T}(\mathbf{b} - A\mathbf{x}_k) = C^{-T}\mathbf{r}_k,$$

czyli

$$\mathbf{r}_k = C^T\hat{\mathbf{r}}_k.$$

Oznaczmy $S = C^TC$. Mając wektor $\hat{\boldsymbol{v}}_k = C\boldsymbol{v}_k$, obliczamy

$$t_k = \frac{\hat{\mathbf{r}}_k^T\hat{\mathbf{r}}_k}{\hat{\boldsymbol{v}}_k^TC^{-T}AC^{-1}\hat{\boldsymbol{v}}_k} = \frac{(C^{-T}\mathbf{r}_k)^TC^{-T}\mathbf{r}_k}{\boldsymbol{v}_k^TA\boldsymbol{v}_k} = \frac{\mathbf{r}_k^TS^{-1}\mathbf{r}_k}{\boldsymbol{v}_k^TA\boldsymbol{v}_k}.$$

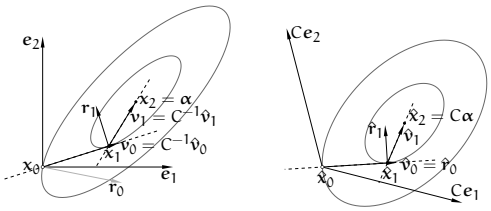
Następnie moglibyśmy obliczyć $\hat{\mathbf{x}}_{k+1} = \hat{\mathbf{x}}_k + t_k\hat{\boldsymbol{v}}_k$ itd., ale zamiast tego obliczamy

$$\mathbf{x}_{k+1} = C^{-1}\hat{\mathbf{x}}_{k+1} = C^{-1}\hat{\mathbf{x}}_k + t_kC^{-1}\hat{\boldsymbol{v}}_k = \mathbf{x}_k + t_k\boldsymbol{v}_k,$$

$$\mathbf{r}_{k+1} = C^T\hat{\mathbf{r}}_{k+1} = C^T\hat{\mathbf{r}}_k - C^Tt_kC^{-T}AC^{-1}\hat{\boldsymbol{v}}_k = \mathbf{r}_k - t_kA\boldsymbol{v}_k,$$

$$s_k = \frac{\hat{\mathbf{r}}_{k+1}^T\hat{\mathbf{r}}_{k+1}}{\hat{\mathbf{r}}_k^T\hat{\mathbf{r}}_k} = \frac{(C^{-T}\mathbf{r}_{k+1})^TC^{-T}\mathbf{r}_{k+1}}{(C^{-T}\mathbf{r}_k)^TC^{-T}\mathbf{r}_k} = \frac{\mathbf{r}_{k+1}^TS^{-1}\mathbf{r}_{k+1}}{\mathbf{r}_k^TS^{-1}\mathbf{r}_k},$$

$$\boldsymbol{v}_{k+1} = C^{-1}\hat{\boldsymbol{v}}_{k+1} = C^{-1}\hat{\mathbf{r}}_{k+1} + s_kC^{-1}\hat{\boldsymbol{v}}_k = S^{-1}\mathbf{r}_{k+1} + s_k\boldsymbol{v}_k.$$



Rysunek przedstawia przykład. Macierz C jest tu symetryczna, dodatnio określona i taka, że $A = C^4$, dzięki czemu $\operatorname{cond}_2(C^{-T}AC^{-1}) = \sqrt{\operatorname{cond}_2 A}$. Oczywiście, niezależnie od uwarunkowania, dla układu dwóch równań metoda znajduje rozwiązanie $\boldsymbol{\alpha}$ już po drugiej iteracji. Ale możemy zauważyć, że przybliżenie otrzymane już w pierwszej iteracji, $\mathbf{x}_1$, leży bliżej rozwiązania niż wtedy, gdy metodę stosujemy do oryginalnego układu równań $A\mathbf{x} = \mathbf{b}$.

Mówiąc niezbyt ściśle, im lepiej uwarunkowany jest układ równań, tym lepiej wektory kierunkowe prostych, wzdłuż których minimalizujemy funkcję kwadratową, „celują” w punkt $\boldsymbol{\alpha}$. To ma ogromne znaczenie praktyczne, bo dla *wielkich* układów równań liniowych metodę CG stosujemy jako metodę iteracyjną i chcemy wykonać *dużo mniej* iteracji niż n.

Nigdzie w końcowych wzorach nie występuje macierz C. Zamiast niej mamy macierz S, a dokładniej S^{-1}.

Dwa parametry procedury, która realizuje metodę PCG, powinny być wskaźnikami do procedur, z których pierwsza ma za zadanie obliczenie iloczynu $A\boldsymbol{v}$, a zadaniem drugiej jest obliczenie iloczynu $S^{-1}\mathbf{r}$, gdzie wektory $\boldsymbol{v}$ i $\mathbf{r}$ są przekazywane jako parametry. To drugie obliczenie może (powinno) polegać na rozwiązaniu układu równań $S\mathbf{z} = \mathbf{r}$ jakimś szybkim sposobem dobranym do postaci macierzy S. W pewnych zastosowaniach macierzy S nie znamy jawnie; dysponujemy tylko macierzą S^{-1} lub jakimś szybkim algorytmem mnożenia wektora przez tę macierz.


```
r = b − Ax;  /* r = r_0 */
z = S^{-1}r;  /* rozwiązywanie układu Sz = r */
v = z;        /* v = v_0 */
c = z^T r;
for ( k = 0; k < n; k++ ) {
  if ( v^T v < δ^2 ) return;
  z = Av;
  t = c/(v^T z);  /* t = t_k */
  x = x + tv;  /* x = x_{k+1} */
  r = r − tz;  /* r = r_{k+1} */
  z = S^{-1}r;  /* rozwiązywanie układu Sz = r */
  d = z^T r;
  if ( d < ε^2 ) {
    if ( r^T r < ε^2 ) /* kończymy, jeśli residuum */
      return; /* oryginalnego układu jest dostatecznie małe */
  }
  v = z + (d/c)v;  /* v = v_{k+1} */
  c = d;
}
```

Zastępowanie układu równań układem o macierzy lepiej uwarunkowanej ma angielską nazwę *preconditioning*, a używana do tego macierz C to tzw. *preconditioner*; terminy te nie mają powszechnie przyjętych polskich odpowiedników.

Metody znajdowania odpowiednich macierzy są silnie związane ze specyfiką zadania i daleko wykraczają poza ten wykład. Niemniej, warto wiedzieć, że metody iteracyjne z poprawianiem uwarunkowania są w zasadzie jedynymi skutecznymi metodami rozwiązywania *naprawdę wielkich* układów równań liniowych (które w praktyce często są *naprawdę źle* uwarunkowane).

5. Liniowe zadania najmniejszych kwadratów

Rozważamy układ równań liniowych $A\mathbf{x} = \mathbf{b}$ z macierzą $A \in \mathbb{R}^{m,n}$ i wektorem $\mathbf{b} \in \mathbb{R}^m$. Układ ten może (choć nie musi) być sprzeczny. Liniowe zadanie najmniejszych kwadratów (LZNK) polega na znalezieniu wektora $\mathbf{x}^\star$, takiego że norma druga wektora residuum, $\mathbf{b} - A\mathbf{x}^\star$, jest najmniejsza. Jeśli układ jest niesprzeczny, to rozwiązanie LZNK jest zwykłym rozwiązaniem tego układu.

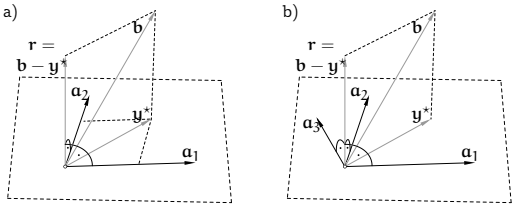
Twierdzenie. *LZNK ma rozwiązanie; jest nim taki wektor $\mathbf{x}^\star$, że wektor residuum jest prostopadły (w sensie iloczynu skalarnego $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{v}^T\mathbf{u}$) do przestrzeni liniowej (podprzestrzeni $\mathbb{R}^m$) rozpiętej przez kolumny $\mathbf{a}_1, \ldots, \mathbf{a}_n$ macierzy A.*

Dowód. Rozważmy wektor $\mathbf{y}^\star$, który jest rzutem prostopadłym wektora $\mathbf{b}$ na tę podprzestrzeń. Zatem istnieje wektor $\mathbf{x}^\star$, taki że $\mathbf{y}^\star = A\mathbf{x}^\star$ i wektor $\mathbf{b} - \mathbf{y}^\star = \mathbf{b} - A\mathbf{x}^\star$ (czyli residuum) jest prostopadły do tej podprzestrzeni. Jeśli weźmiemy dowolny wektor $\mathbf{x} \in \mathbb{R}^n$ i obliczymy $\mathbf{y} = A\mathbf{x}$, to wektor $\mathbf{y} - \mathbf{y}^\star = A(\mathbf{x} - \mathbf{x}^\star)$ jest prostopadły do wektora $\mathbf{b} - \mathbf{y}^\star$. Ale wtedy, na podstawie twierdzenia Pitagorasa, mamy

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|_2^2 &= \|\mathbf{b} - \mathbf{y}\|_2^2 = \|\mathbf{b} - \mathbf{y}^\star\|_2^2 + \|\mathbf{y}^\star - \mathbf{y}\|_2^2 \\ &\geqslant \|\mathbf{b} - \mathbf{y}^\star\|_2^2 = \|\mathbf{b} - A\mathbf{x}^\star\|_2^2. \end{aligned}$$

Dla $\mathbf{y} \neq \mathbf{y}^\star$ powyższa nierówność jest ostra. □

LZNK ma rozwiązanie jednoznaczne wtedy i tylko wtedy, gdy kolumny macierzy A są liniowo niezależne (co jest możliwe tylko dla $m \geqslant n$). Zadania z takimi macierzami to tzw. regularne liniowe zadania najmniejszych kwadratów (RLZNK). Jeśli macierz ma kolumny liniowo zależne, to zadanie (nieregularne, NLZNK) ma wiele rozwiązań, ich zbiór jest warstwą przestrzeni $\mathbb{R}^n$ o wymiarze $n - r$ (gdzie r oznacza rząd macierzy A).



Ilustracje liniowych zadań najmniejszych kwadratów mamy na rysunku. Rysunek a) przedstawia zadanie regularne dla układu z macierzą 3×2. Kolumny macierzy $A = [\mathbf{a}_1, \mathbf{a}_2]$ rozpinają dwuwymiarową podprzestrzeń przestrzeni $\mathbb{R}^3$, nie zawierającą wektora prawej strony $\mathbf{b}$.

Ponieważ kolumny te są liniowo niezależne, rzut prostopadły $\mathbf{y}^\star$ wektora $\mathbf{b}$ na tę podprzestrzeń jest ich kombinacją liniową o jednoznacznie określonych współczynnikach — współrzędnych wektora $\mathbf{x}^\star$, który jest jedynym rozwiązaniem tego zadania.

Na rysunku b) jest pokazane zadanie nieregularne, z macierzą 3×3 o liniowo zależnych kolumnach. Kolumny te rozpinają przestrzeń dwuwymiarową, której elementem wektor $\mathbf{b}$ nie jest. Jego rzut prostopadły $\mathbf{y}^\star$ na tę podprzestrzeń jest jednoznacznie określony, ale można go wyrazić jako kombinację liniową kolumn $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ na nieskończenie wiele sposobów i właśnie tyle rozwiązań ma zadanie.

Regularne LZNK

Prostopadłość dowolnego wektora w $\mathbb{R}^m$ do podprzestrzeni jest równoważna prostopadłości tego wektora do wszystkich elementów dowolnej bazy tej podprzestrzeni. Zatem, mając układ równań $A\mathbf{x} = \mathbf{b}$, możemy pomnożyć skalarnie residuum przez kolumny macierzy A i przyrównać do zera:

$$\langle \mathbf{b} - A\mathbf{x}, \mathbf{a}_i \rangle = 0, \quad i = 1, \ldots, n.$$

Można to zapisać w postaci macierzowej, po prostych przekształceniach otrzymując tzw. układ równań normalnych:

$$A^TA\mathbf{x} = A^T\mathbf{b}.$$

Jeśli kolumny $\mathbf{a}_1, \ldots, \mathbf{a}_n$ są liniowo niezależne, to ich zbiór jest bazą odpowiedniej podprzestrzeni; wtedy macierz symetryczna $M = A^TA$ jest dodatnio określona (skąd wynika, że nieosobliwa) i układ ma jednoznaczne rozwiązanie — rozwiązanie RLZNK (jeśli kolumny są liniowo zależne, to układ równań normalnych jest niesprzeczny, ale ma nieskończenie wiele rozwiązań, którymi są wszystkie rozwiązania NLZNK).

Algorytm równań normalnych jest najprostszą i najtańszą metodą numeryczną rozwiązywania RLZNK. Polega on na obliczeniu macierzy $M = A^TA$ i wektora $\mathbf{d} = A^T\mathbf{b}$, a następnie rozwiązaniu układu równań $M\mathbf{x} = \mathbf{d}$. Ponieważ macierz M jest symetryczna, obliczenie jej współczynników może być wykonane kosztem $mn(n+1)/2$ działań (mnożeń i dodawań zmiennopozycyjnych). Układ równań z macierzą M może być rozwiązany metodą Choleskiego.

Większą dokładność rozwiązania można osiągnąć, korzystając z rozkładu ortogonalno-trójkątnego macierzy A. Dla ustalonej macierzy $A \in \mathbb{R}^{m,n}$ istnieje macierz ortogonalna $Q \in \mathbb{R}^{m,m}$ i macierz $R \in \mathbb{R}^{m,n}$, której współczynniki poniżej diagonali są zerowe, przy czym jeśli macierz A ma liniowo niezależne kolumny, to macierz R również (zatem ma niezerowe współczynniki diagonalne). Dla $m \geqslant n$ pierwsze n kolumn macierzy Q i wierszy macierzy R są określone jednoznacznie z dokładnością do zwrotów.

Macierze Q i R podzielimy na bloki, odpowiednio

$$Q = [Q_1, Q_2], \qquad R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix},$$

takie że $Q_1 \in \mathbb{R}^{m,n}$ i $R_1 \in \mathbb{R}^{n,n}$. Ponieważ blok R_2 jest zerowy, mamy $A = Q_1 R_1$. Podstawmy to do układu równań normalnych:

$$R_1^T Q_1^T Q_1 R_1 x = R_1^T Q_1^T b.$$

Macierz $Q_1^T Q_1$ jest macierzą jednostkową $n \times n$, a ponieważ macierz R_1 jest nieosobliwa, mamy układ równoważny układowi równań normalnych:

$$R_1 x = Q_1^T b,$$

z nieosobliwą macierzą trójkątną górną R_1.

Jeśli dany układ równań, $Ax = b$, dla którego stawiamy LZNK, pomnożymy stronami przez Q^T, to możemy w nim wyróżnić dwa podukłady:

$$\begin{cases} R_1 x = Q_1^T b, \\ 0x = Q_2^T b. \end{cases}$$

Układ dany jest niesprzeczny wtedy i tylko wtedy, gdy wektor $Q_2^T b = 0$. Co więcej, ponieważ pierwszy podukład ma rozwiązanie jednoznaczne (jest nim rozwiązanie LZNK), a macierz drugiego podukładu jest zerowa, długość wektora $Q_2^T b$ jest najmniejszą osiągalną normą residuum, $b - Ax$.

Istnieje wiele metod rozkładania macierzy A na czynniki Q i R albo Q_1 i R_1. Jedną z nich jest zastosowanie odbić Householdera.
Za pomocą n odbić, konstruowanych tak samo, jak w zastosowaniu do układu równań liniowych z nieosobliwą macierzą $n \times n$, macierz A przekształcamy na macierz R. Macierz ortogonalną Q reprezentujemy za pomocą wektorów normalnych hiperpłaszczyzn kolejnych odbić (które możemy przechowywać w tablicy początkowo zawierającej współczynniki macierzy A); mamy

$$Q^T = H_n H_{n-1} \dots H_1, \quad Q = H_1 \dots H_{n-1} H_n,$$

gdzie $H_i = I - \gamma_i v v^T$. Macierzy Q *nie wyznaczamy* w postaci jawnej.

Algorytm rozwiązywania RLZNK za pomocą odbić składa się z następujących kroków:

1. Znajdź rozkład macierzy A, tj. reprezentację macierzy Q w postaci wektorów odbić i macierz R.
2. Oblicz wektor $y = Q^T b = H_n \dots H_1 b$.
3. Wybierz pierwsze n wierszy macierzy R i wektora y, tj. macierz $R_1 = Q_1^T A$ i wektor $y_1 = Q_1^T b$, i rozwiąż układ $R_1 x = y_1$.

Rozkładu macierzy A na czynniki Q_1 i R_1 możemy dokonać za pomocą ortonormalizacji Grama-Schmidta. W tak zwanym algorytmie modyfikowanym (MGS) konstruujemy macierze $A^{(0)} = A, \dots, A^{(n)} = Q_1$. Kolumny macierzy $A^{(k)}$ oznaczymy $a_1^{(k)}, \dots, a_n^{(k)}$. Obliczamy

```
for ( k = 1; k ⩽ n; k++ ) {
  r_kk = √(a_k^(k−1)T a_k^(k−1));
  a_k^(k) = (1/r_kk) a_k^(k−1);
  for ( i = k + 1; i ⩽ n; i++ ) {
    r_ki = a_k^(k)T a_i^(k−1);
    a_i^(k) = a_i^(k−1) − r_ki a_k^(k);
  }
}
```

Wynikiem obliczenia są kolumny $q_i = a_i^{(n)}$ macierzy Q_1 i współczynniki r_{ki} na i powyżej diagonali macierzy R_1.

Do rozwiązania RLZNK za pomocą ortonormalizacji służy następujący algorytm:

1. Za pomocą ortonormalizacji Grama-Schmidta znajdź macierze Q_1 i R_1.
2. Oblicz wektor $y_1 = Q_1^T b$.
3. Rozwiąż układ równań $R_1 x = y_1$.

Przyczyna, dla której algorytmy korzystające z rozkładu ortonormalno-trójkątnego dają dokładniejsze wyniki niż algorytm równań normalnych jest taka, że wyjściowe zadanie jest zwykle znacznie lepiej uwarunkowane niż układ równań normalnych. Dlatego błędy zaokrągleń popełnione podczas obliczania macierzy M i jej rozkładania na czynniki trójkątne przenoszą się na wynik ze znacznie większym czynnikiem. Tymczasem uwarunkowanie układu równań $R_1 x = Q_1^T b$ (w normie drugiej) jest takie samo, jak uwarunkowanie zadania wyjściowego.

Dualne LZNK

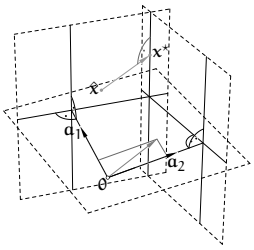
Inny rodzaj liniowego zadania najmniejszych kwadratów możemy postawić, gdy dany układ współrzędnych, $Ax = b$, jest niesprzeczny i nieokreślony.
W dualnym liniowym zadaniu najmniejszych kwadratów (DLZNK) celem jest wybranie jednego elementu ze zbioru rozwiązań układu $Ax = b$. Należy wybrać rozwiązanie $x^\star$ najkrótsze (o najmniejszej normie drugiej), lub takie, aby dla ustalonego wektora $\hat{x} \in \mathbb{R}^n$ wektor $x^\star - \hat{x}$ był najkrótszy; pierwsza sytuacja jest szczególnym przypadkiem drugiej.

Twierdzenie. *DLZNK ma rozwiązanie. Niech* $A^T = [a_1, \dots, a_m]$*, tj. niech wektory* $a_1, \dots, a_m \in \mathbb{R}^n$ *będą transponowanymi wierszami macierzy* A*. Rozwiązaniem DLZNK jest taki wektor* $x^\star$*, że* $Ax^\star = b$ *i różnica* $x^\star - \hat{x}$ *jest kombinacją liniową wektorów* $a_1, \dots, a_m$.

Dowód. Zbiór rozwiązań równania liniowego $a_i^T x = b_i$ jest warstwą równoległą do podprzestrzeni o wymiarze $n - 1$ prostopadłej do wektora a_i. Zbiór rozwiązań całego układu równań $Ax = b$ jest przecięciem tych warstw, i jest to warstwa przestrzeni $\mathbb{R}^n$ równoległa do podprzestrzeni, do której należą wszystkie wektory prostopadłe do wektorów $a_1, \dots, a_m$. Jeśli zatem wektor x jest dowolnym rozwiązaniem układu $Ax = b$, to wektor $x - x^\star$ jest prostopadły do wektorów $a_1, \dots, a_m$, a więc także do ich kombinacji liniowej $x^\star - \hat{x}$, i z twierdzenia Pitagorasa mamy

$$\|x - \hat{x}\|_2^2 = \|x - x^\star\|_2^2 + \|x^\star - \hat{x}\|_2^2 \geqslant \|x^\star - \hat{x}\|_2^2.$$

Jeśli $x \neq x^\star$, to nierówność jest ostra. □



Ilustrację DLZNK dla układu dwóch równań z trzema niewiadomymi mamy na rysunku. Zbiór rozwiązań układu jest prostą prostopadłą do płaszczyzny rozpiętej przez wektory $\mathbf{a}_1$, $\mathbf{a}_2$, tj. przecięciem dwóch płaszczyzn prostopadłych do tych wektorów. Wektor $\mathbf{x}^\star - \hat{\mathbf{x}}$ jest prostopadły do tej prostej.

Jeśli więc wektor $\mathbf{x}$ jest rozwiązaniem DLZNK, to wektor $\mathbf{x} - \hat{\mathbf{x}}$ musi być kombinacją liniową transponowanych wierszy macierzy A, a zatem istnieje wektor $\mathbf{y} \in \mathbb{R}^m$, taki że $A^T\mathbf{y} = \mathbf{x} - \hat{\mathbf{x}}$. Jeśli tę równość pomnożymy przez macierz A, to otrzymamy

$$AA^T\mathbf{y} = A\mathbf{x} - A\hat{\mathbf{x}}.$$

Po podstawieniu $A\mathbf{x} = \mathbf{b}$ mamy stąd układ równań z niewiadomym wektorem $\mathbf{y}$

$$AA^T\mathbf{y} = \mathbf{b} - A\hat{\mathbf{x}},$$

zwany dualnym układem równań normalnych. Macierz AA^T jest symetryczna i jeśli wiersze macierzy A są liniowo niezależne, to jest dodatnio określona. Aby tak było, musi być $n \geqslant m$. Jeśli wiersze macierzy A są liniowo zależne, to nie mamy gwarancji, że układ równań $A\mathbf{x} = \mathbf{b}$ jest niesprzeczny, i mamy do czynienia z zadaniem nieregularnym.

Algorytm dualnych równań normalnych polega na obliczaniu macierzy $M = AA^T$ i wektora $\mathbf{d} = \mathbf{b} - A\hat{\mathbf{x}}$, a następnie rozwiązaniu układu $M\mathbf{y} = \mathbf{d}$ (do czego można użyć metody Choleskiego) i obliczeniu rozwiązania $\mathbf{x} = \hat{\mathbf{x}} + A^T\mathbf{y}$. Jeśli ma być znalezione rozwiązanie o najmniejszej normie drugiej, to $\hat{\mathbf{x}} = \mathbf{0}$; można wtedy pominąć niektóre obliczenia.

Większą dokładność można uzyskać, korzystając z rozkładu trójkątno-ortogonalnego macierzy A. Istnieje macierz $L \in \mathbb{R}^{m,n}$, która ma zera za współczynnikiem diagonalnym w każdym wierszu, i macierz ortogonalna $Q \in \mathbb{R}^{n,n}$, takie że $A = LQ^T$; macierze te można otrzymać, stosując do macierzy A^T (kolumnowo regularnej) te same algorytmy wyznaczania rozkładu ortogonalno-trójkątnego, których użycie do rozwiązania RLZNK było opisane wcześniej. Otrzymujemy macierze $L = [L_1, L_2]$ i $Q = [Q_1, Q_2]$, w których blok $L_1 \in \mathbb{R}^{m,m}$ jest nieosobliwą macierzą trójkątną dolną, blok L_2 jest zerowy, i macierze L_1 i Q_1 są dane jednoznacznie z dokładnością do zwrotów kolumn. Zachodzi równość $A = L_1Q_1^T$.

Po podstawieniu czynników rozkładu do dualnego układu równań normalnych mamy

$$L_1Q_1^TQ_1L_1^T\mathbf{y} = \mathbf{b} - L_1Q_1^T\hat{\mathbf{x}},$$

a ponieważ $Q_1^TQ_1 = I$ i macierz L_1 jest nieosobliwa, mamy układ równoważny

$$L_1^T\mathbf{y} = L_1^{-1}\mathbf{b} - Q_1^T\hat{\mathbf{x}}.$$

Rozwiązując powyższy układ równań, można by obliczyć wektor $\mathbf{y}$, a następnie obliczyć $\mathbf{x} = \hat{\mathbf{x}} + A^T\mathbf{y}$, ale ponieważ poza tym wektor $\mathbf{y}$ *nie jest do niczego potrzebny*, lepszym rozwiązaniem po znalezieniu czynników rozkładu macierzy A jest użycie *tylko* tych czynników.

Oznaczając $\boldsymbol{w} = L_1^{-1}\mathbf{b} - Q_1^T\hat{\mathbf{x}}$ i podstawiając $\mathbf{y} = L_1^{-T}\boldsymbol{w}$, otrzymamy $A^T\mathbf{y} = Q_1L_1^TL_1^{-T}\boldsymbol{w} = Q_1\boldsymbol{w}$. Stąd otrzymujemy algorytm rozwiązywania DLZNK:

1. Za pomocą ortonormalizacji Grama-Schmidta znajdź macierze trójkątną dolną L_1 i kolumnowo-ortogonalną Q_1, takie że $A = L_1Q_1^T$.
2. Rozwiąż układ równań liniowych $L_1\boldsymbol{z} = \mathbf{b}$ i oblicz wektor $\boldsymbol{w} = \boldsymbol{z} - Q_1^T\hat{\mathbf{x}}$.
3. Oblicz $\mathbf{x} = \hat{\mathbf{x}} + Q_1\boldsymbol{w}$.

Powyższy algorytm można zrealizować również za pomocą odbić Householdera, bez jawnego wyznaczania macierzy Q_1.

Inny algorytm rozwiązywania DLZNK korzystający z odbić możemy otrzymać w taki sposób: Niech $\mathbf{s} = Q^T\mathbf{x}$ i $\hat{\mathbf{s}} = Q^T\hat{\mathbf{x}}$. Podstawiając nowe wyrażenie do układu $LQ^T\mathbf{x} = \mathbf{b}$, otrzymujemy układ równań $L\mathbf{s} = \mathbf{b}$, który możemy przedstawić w postaci $L_1\mathbf{s}_1 + L_2\mathbf{s}_2 = \mathbf{b}$. Ponieważ blok L_2 jest zerowy, wektor $\mathbf{s}_1$ musi być rozwiązaniem układu równań $L_1\mathbf{s}_1 = \mathbf{b}$, zaś wektor $\mathbf{s}_2$ trzeba zatem wybrać tak, aby wektor $\mathbf{x} - \hat{\mathbf{x}} = Q(\mathbf{s} - \hat{\mathbf{s}})$ miał najmniejszą normę drugą. Ale jest ona równa normie drugiej wektora $\mathbf{s} - \hat{\mathbf{s}}$. Zatem, jeśli wektor $\hat{\mathbf{s}}$ podzielimy (w tym samym miejscu co $\mathbf{s}$) na bloki $\hat{\mathbf{s}}_1 = Q_1^T\hat{\mathbf{x}}$ i $\hat{\mathbf{s}}_2 = Q_2^T\hat{\mathbf{x}}$, to aby zminimalizować normę drugą wektora $\mathbf{s} - \hat{\mathbf{s}}$, musimy przyjąć $\mathbf{s}_2 = \hat{\mathbf{s}}_2$. Mamy stąd taki algorytm:

1. Znajdź macierz trójkątną dolną L i wektory odbić reprezentujące macierz Q, takie że $A = LQ^T$. Wybierz blok L_1 macierzy L.
2. Oblicz $\hat{\mathbf{s}} = Q^T\hat{\mathbf{x}}$, stosując odpowiednie odbicia.
3. Rozwiąż układ $L_1\mathbf{s}_1 = \mathbf{b}$ i złóż wektor $\mathbf{s}$ z bloków $\mathbf{s}_1$ i $\mathbf{s}_2 = \hat{\mathbf{s}}_2$.
4. Oblicz $\mathbf{x} = Q\mathbf{s}$, stosując odpowiednie odbicia.

Nieregularne LZNK

Jeśli rząd r macierzy A jest mniejszy zarówno od liczby kolumn n, jak i od liczby wierszy m, to liniowe zadanie najmniejszych kwadratów dla układu $A\mathbf{x} = \mathbf{b}$ jest nieregularne. Zbiór rozwiązań takiego zadania jest nieskończony; jest on warstwą $n - r$-wymiarową (przestrzeni $\mathbb{R}^n$), której elementami są takie wektory $\mathbf{x}$, że wektor $\mathbf{y}^\star = A\mathbf{x}$ jest rzutem prostopadłym wektora $\mathbf{b}$ na podprzestrzeń rozpiętą przez kolumny macierzy A (tj. residuum, $\mathbf{b} - \mathbf{y}^\star$, jest wektorem prostopadłym do tej podprzestrzeni). Dokładnie jeden element tej warstwy ma najmniejszą normę drugą; co więcej, dla dowolnego wektora $\hat{\mathbf{x}} \in \mathbb{R}^n$ istnieje dokładnie jeden element $\mathbf{x}^\star$ tej warstwy, taki że norma druga różnicy $\mathbf{x}^\star - \hat{\mathbf{x}}$ jest najmniejsza. Rozwiązanie NLZNK zwykle polega na znalezieniu tego wektora $\mathbf{x}^\star$.

Rozumowanie podobne do przeprowadzonego wcześniej dla DLZNK uzasadnia stwierdzenie, że wektor $x^* - \hat{x}$ jest kombinacją liniową transponowanych wierszy macierzy A.

NLZNK *są trudne* do numerycznego rozwiązania. Jest tak dlatego, że rozwiązanie zadania zależy od danych w sposób *paskudnie nieciągły*. NLZNK jest szczególnie trudne, jeśli nie znamy rzędu macierzy A i dopiero mamy go na podstawie obliczeń numerycznych ustalić.

Najodporniejsze numeryczne algorytmy rozwiązywania NLZNK korzystają z rozkładu względem wartości szczególnych (ang. *singular value decomposition*, *SVD*) macierzy A. Dowodzi się, że dla dowolnej macierzy $A \in \mathbb{R}^{m,n}$ istnieją macierze ortogonalne $U \in \mathbb{R}^{m,m}$ i $V \in \mathbb{R}^{n,n}$ oraz macierz diagonalna $\Sigma \in \mathbb{R}^{m,n}$, takie że $A = U\Sigma V^T$. Współczynniki diagonalne macierzy Σ, $\sigma_1, \ldots, \sigma_l$, gdzie $l = \min\{m, n\}$, nazywają się wartościami szczególnymi macierzy A i są nieujemne. Rozkład w ogólności *nie jest* jednoznaczny, ale same wartości szczególne i liczby ich wystąpień (krotności) są określone przez macierz A jednoznacznie. Zwykle rozkładu dokonuje się w taki sposób, aby wartości szczególne były uporządkowane nierosnąco na diagonali macierzy Σ: $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > \sigma_{r+1} = \cdots = \sigma_l = 0$. Liczba niezerowych wartości szczególnych jest rzędem macierzy A.

Wyznaczanie rozkładu SVD wiąże się z rozwiązywaniem algebraicznego zagadnienia własnego i dla macierzy o więcej niż czterech wierszach i kolumnach może być dokonane tylko jakąś metodą iteracyjną. Opis algorytmu Goluba, który dokonuje rozkładu (otrzymując reprezentacje macierzy U i V w postaci ciągu wektorów odbić Householdera i tzw. obrotów Givensa), pominiemy (kto będzie potrzebował, ten go znajdzie). Natomiast przyjrzyjmy się zastosowaniu tego rozkładu do rozwiązania zadania.

NLZNK dla układu równań $Ax = b$ i wektora $\hat{x}$ można zastąpić zadaniem równoważnym dla układu równań $\Sigma y = d$, gdzie $d = U^T b$, i wektora $\hat{y} = V^T\hat{x}$ (po rozwiązaniu tego zadania możemy obliczyć $x = Vy$). Załóżmy dla uproszczenia, że $\hat{x} = 0$, czyli $\hat{y} = 0$. Wtedy rozwiązaniem NLZNK dla układu $\Sigma y = d$ jest wektor o współrzędnych

$$y_i = \begin{cases} d_i/\sigma_i & \text{dla } i \leqslant r, \\ 0 & \text{dla } i > r. \end{cases}$$

Mamy stąd wyjaśnienie trudności zadania: niewielkie zaburzenie macierzy A może spowodować pewne niegroźne zmiany macierzy U i V, oraz zaburzenie macierzy Σ: jeśli dowolna zerowa wartość szczególna zmieni się na niezerową (czyli skutkiem zaburzenia będzie zwiększenie rzędu macierzy A) i $d_i \neq 0$, to trzeba będzie przyjąć $y_i = d_i/\sigma_i$, zamiast zera, dla pewnego $i > r$. Tak więc, *im mniej* zaburzymy macierz A (w sposób zmieniający σ_i), *tym większa* będzie zmiana wyniku.

Jeśli znamy rząd r macierzy A, to po znalezieniu rozkładu SVD możemy zamienić na zera obliczone numerycznie wartości szczególne σ_i dla $i > r$ — obliczone wartości niezerowe są skutkiem błędów zaokrągleń i aproksymacji popełnionych podczas rozkładania. Jeśli rzędu nie znamy, to możemy przyjąć pewien próg i zamienić na zera znalezione wartości szczególne mniejsze od tego progu; to postępowanie nazywa się regularyzacją dyskretną.

Inne podejście to tzw. regularyzacja ciągła — do *wszystkich* wartości szczególnych dodajemy pewną liczbę $s > 0$, otrzymując zadanie z macierzą pełnego rzędu, tj. RLZNK, jeśli $m > n$, układ równań z macierzą kwadratową nieosobliwą, jeśli $n = m$, albo DLZNK, jeśli $m < n$. Wybór metody regularyzacji zależy od zastosowania.

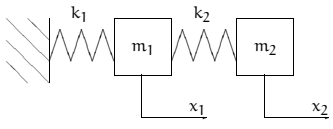
6. Algebraiczne zagadnienie własne

Niech $A \in \mathbb{R}^{n,n}$. Jeśli wektor $x \neq 0$ spełnia równanie $Ax = \lambda x$ dla pewnej liczby λ, to mówimy, że jest to wektor własny macierzy A, zaś liczba λ jest to wartość własna tej macierzy; parę (x, λ) nazywamy parą własną macierzy A.

Algebraiczne zagadnienie własne polega na znalezieniu, dla danej macierzy A, jej (wszystkich, kilku lub jednej) wartości własnych albo par własnych. Algebraiczne zagadnienia własne występują w różnych zastosowaniach, np. w mechanice, mają też związek z innymi zadaniami numerycznej algebry liniowej, np. rozwiązywaniem układów równań lub liniowych zadań najmniejszych kwadratów.

Równanie $Ax = \lambda x$ można przepisać w postaci $(A - \lambda I)x = 0$. Z tej postaci natychmiast wynika, że para (x, λ), w której wektor $x \neq 0$, może spełniać to równanie (czyli być parą własną) wtedy i tylko wtedy, gdy macierz $A - \lambda I$ jest osobliwa. To oznacza, że jej wyznacznik jest zerowy. Wyrażenie $\det(A - \lambda I)$ jest wielomianem stopnia n zmiennej λ. Na podstawie zasadniczego twierdzenia algebry (Gauss, 1799 r.), równanie charakterystyczne $\det(A - \lambda I) = 0$ ma rozwiązanie, które jest liczbą rzeczywistą albo zespoloną. Tak więc każda macierz ma jakąś wartość własną. Zbiór (w ogólności zespolonych) wartości własnych dowolnej macierzy A nazywa się widmem tej macierzy; oznaczamy je symbolem $\operatorname{spect} A$.

Przykład zastosowania:



Rozważmy układ złożony z dwóch ciężarków połączonych ze sobą i z nieruchomym podłożem sprężynkami. Jeśli ciężarki potrącimy, to będą one drgać, przy czym drgania, które są skutkiem tylko początkowego wytrącenia z położenia równowagi, są nazywane drganiami własnymi układu.

Ciężarki mają masy odpowiednio m_1 i m_2; ich odchylenia od położeń równowagi oznaczymy symbolami x_1 i x_2. Każda ze sprężynek działa z siłą proporcjonalną do jej odkształcenia, przy czym współczynniki proporcjonalności oznaczymy odpowiednio k_1 i k_2.

Na pierwszy ciężarek działa siła

$$-k_1x_1 + k_2(x_2 - x_1) = -(k_1 + k_2)x_1 + k_2x_2$$

(uwaga na zwrot; siła jest dodatnia jeśli ma ten sam zwrot co dodatnie przemieszczenie). Na drugi ciężarek działa siła $k_2(x_1 - x_2)$. Każda z tych sił jest równoważona przez bezwładność ciężarka proporcjonalną do jego masy, zatem ruch ciężarków jest opisany przez taki układ równań różniczkowych zwyczajnych:

$$\begin{cases} m_1\ddot{x}_1 = -(k_1 + k_2)x_1 + k_2x_2, \\ m_2\ddot{x}_2 = k_2x_1 - k_2x_2 \end{cases}$$

(każda kropka oznacza tu jednokrotne różniczkowanie względem czasu).

Możemy to zapisać w postaci macierzowej:

$$\begin{bmatrix} -(k_1 + k_2) & k_2 \\ k_2 & -k_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} m_1\ddot{x}_1 \\ m_2\ddot{x}_2 \end{bmatrix}.$$

Można (dokonując zamiany zmiennych) przekształcić ten układ tak, aby utrzymać symetrię macierzy z prawej strony, ale to zaniedbamy; zamiast tego weźmy

$$\begin{bmatrix} -(k_1 + k_2)/m_1 & k_2/m_1 \\ k_2/m_2 & -k_2/m_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix}.$$

Przypuśćmy, że $x_1(t) = a_1 \sin \omega t$ oraz $x_2(t) = a_2 \sin \omega t$. Wtedy $\ddot{x}_1(t) = -a_1\omega^2 \sin \omega t$ oraz $\ddot{x}_2(t) = -a_2\omega^2 \sin \omega t$. Po podstawieniu i podzieleniu przez $-\sin \omega t$ dostaniemy równanie

$$\begin{bmatrix} (k_1 + k_2)/m_1 & -k_2/m_1 \\ -k_2/m_2 & k_2/m_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \omega^2 \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},$$

czyli algebraiczne zagadnienie własne z macierzą 2×2. Wektor własny występującej w nim macierzy opisuje amplitudy a_1 i a_2 drgań własnych, zaś odpowiadająca mu wartość własna jest kwadratem prędkości fazowej ω.

Można dowieść, że w zagadnieniach własnych utworzonych dla układów ciężarków podobnych do rozpatrywanego wyżej wszystkie wartości własne są rzeczywiste i dodatnie, a więc hipoteza, że drgania mogą być opisane za pomocą funkcji sinus, znajduje potwierdzenie.

Dla ustalonego λ układ równań $(A - \lambda I)\mathbf{x} = \mathbf{0}$ jest jednorodny; jeśli zatem $\lambda \in \mathbb{R}$ jest wartością własną macierzy A, to zbiór rozwiązań jest podprzestrzenią liniową przestrzeni $\mathbb{R}^n$. Jest to tzw. <u>podprzestrzeń własna macierzy A przynależna do wartości własnej λ</u>. Wymiar tej podprzestrzeni jest nazywany <u>krotnością geometryczną</u> wartości własnej λ. Z kolei, wielomian charakterystyczny można przedstawić w postaci

$$\det(A - \lambda I) = (\lambda_1 - \lambda) \cdot \ldots \cdot (\lambda_n - \lambda).$$

Liczby $\lambda_1, \ldots, \lambda_n$ to wartości własne, które mogą się powtarzać. Liczba wystąpień wartości własnej λ_i w tym rozkładzie jest zwana jej <u>krotnością algebraiczną</u>. Krotność algebraiczna dowolnej wartości własnej jest większa lub równa krotności geometrycznej tej wartości własnej.

O macierzach A i B, dla których istnieje nieosobliwa macierz C, taka że $B = C^{-1}AC$ mówimy, że to są <u>macierze podobne</u>. Podobieństwo macierzy jest oczywiście relacją równoważności. Można udowodnić, że jeśli macierze są podobne, to mają identyczne wartości własne, o identycznych krotnościach algebraicznych i geometrycznych.

Wektory własne przynależne do różnych wartości własnych dowolnej danej macierzy są liniowo niezależne. Jeśli krotność algebraiczna każdej wartości własnej macierzy A jest równa krotności geometrycznej, to suma baz wszystkich podprzestrzeni własnych składa się z n niezależnych liniowo wektorów własnych macierzy A. Ustawmy te wektory w macierz $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$; macierz ta jest nieosobliwa. Wtedy

$$AX = [A\mathbf{x}_1, \ldots, A\mathbf{x}_n] = [\lambda_1\mathbf{x}_1, \ldots, \lambda_n\mathbf{x}_n] = X\Lambda,$$

gdzie macierz Λ jest diagonalna; jej współczynniki diagonalne są wartościami własnymi macierzy A. Możemy napisać równości

$$X^{-1}AX = \Lambda \quad \text{i} \quad X\Lambda X^{-1} = A.$$

Taka macierz A jest zatem podobna do macierzy diagonalnej, mówimy też, że jest <u>diagonalizowalna</u>. Macierz nie jest diagonalizowalna, jeśli co najmniej jedna jej wartość własna ma krotność algebraiczną różną (większą) od geometrycznej.

<u>Przykłady</u>: Macierz

$$\begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 7 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}$$

jest diagonalizowalna. Macierz

$$\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}$$

też jest diagonalizowalna, ale jej wartości własne są liczbami zespolonymi, $\lambda_1 = (3, -4)$, $\lambda_2 = (3, 4)$, zatem wektory własne — kolumny odpowiedniej macierzy X — mają co najmniej jedną współrzędną zespoloną. Natomiast macierz

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

nie jest diagonalizowalna; krotność algebraiczna wartości własnej 1 jest równa 2, a krotność geometryczna jest równa 1.

Niech $w(x) = a_kx^k + \cdots + a_1x + a_0$ będzie dowolnym wielomianem. Możemy użyć macierzy A jako argumentu, tj. napisać

$$w(A) = a_kA^k + \cdots + a_1A + a_0I.$$

Łatwo jest sprawdzić, że jeśli liczba λ jest wartością własną macierzy A, to liczba $w(\lambda)$ jest wartością własną macierzy $w(A)$. To samo stwierdzenie dotyczy funkcji wymiernych, tj. ilorazów wielomianów. Dzielenie licznika przez mianownik (podczas obliczania skalarnej wartości funkcji dla danego x) zastępujemy przez mnożenie macierzy — licznika — przez odwrotność macierzy — mianownika.

Jeśli macierz $A \in \mathbb{R}^{n,n}$ jest symetryczna, to jest diagonalizowalna, co więcej, jej wszystkie wartości własne są liczbami rzeczywistymi i istnieje baza ortonormalna przestrzeni $\mathbb{R}^n$ złożona z wektorów własnych tej macierzy. Zatem, istnieje macierz ortogonalna X, taka że $X^{-1}AX = X^TAX = \Lambda$ jest macierzą diagonalną. W wielu zastosowaniach pojawia się potrzeba rozwiązania algebraicznego zagadnienia własnego z macierzą symetryczną — jest to przypadek prostszy do numerycznego rozwiązywania niż przypadek ogólny i głównie na nim się dalej skupimy.

Kilka uwag na temat uwarunkowania zadania: jeśli macierz A jest diagonalizowalna, to na podstawie twierdzenia Bauera-Fikego mamy następujące oszacowanie: niech δA oznacza zaburzenie macierzy A. Niech μ oznacza (dowolną) wartość własną macierzy $A + \delta A$ i niech λ_i oznacza wartość własną macierzy A, taką że różnica $\mu - \lambda_i$ ma najmniejszą wartość bezwzględną. Zachodzi nierówność

$$|\mu - \lambda_i| \leqslant \operatorname{cond} X \cdot \|\delta A\|,$$

gdzie X oznacza macierz, której kolumny są wektorami własnymi macierzy A (norma indukowana może tu być dowolna). Jeśli macierz A jest symetryczna, to istnieje odpowiednia macierz ortogonalna X, i wtedy $\operatorname{cond}_2 X = 1$. Dla macierzy diagonalizowalnej niesymetrycznej żadna macierz X zbudowana z wektorów własnych nie jest ortogonalna i dlatego $\operatorname{cond}_2 X > 1$. Jeśli macierz A nie jest diagonalizowalna, to zmiany wartości własnych zależą od powodujących je zaburzeń macierzy A w sposób ciągły, ale nie lipschitzowski. Numeryczne obliczanie wartości własnych takich macierzy jest kłopotliwe.

Jeśli macierze A i $A + \delta A$ są symetryczne i symbolami $\boldsymbol{\lambda}$ i $\boldsymbol{\mu}$ oznaczymy wektory zbudowane odpowiednio z tak samo (np. malejąco) uporządkowanych wartości własnych tych macierzy, to na podstawie twierdzenia Wielandta-Hoffmana zachodzi nierówność

$$\|\boldsymbol{\mu} - \boldsymbol{\lambda}\|_2 \leqslant \|\delta A\|_F.$$

Zadanie wyznaczania wektora $\boldsymbol{\lambda}$ jest zatem bardzo dobrze uwarunkowane, choć jeśli pewne wartości własne mają bardzo małe wartości bezwzględne, to ich zaburzenia względne spowodowane dodaniem małego zaburzenia δA mogą być duże.

Uwarunkowanie zadania wyznaczania wektorów własnych zależy od odległości między wartościami własnymi, i jest tym gorsze, im mniej odpowiednie wartości własne się różnią. Zauważmy, że jeśli pewna wartość własna ma krotność geometryczną $k > 1$, to istnieje nieskończenie wiele baz odpowiedniej podprzestrzeni własnej, złożonych z wektorów jednostkowych. Macierz zaburzona może mieć zamiast tej wartości własnej k różnych wartości własnych (jednokrotnych) i dlatego w tym przypadku rozwiązanie zależy od zaburzenia w sposób nieciągły (jest to możliwe nawet, jeśli macierz A jest symetryczna).

Uwaga: Nie jest dobrym pomysłem obliczanie współczynników wielomianu charakterystycznego $\det(A - \lambda I)$, np. w bazie potęgowej, a następnie znajdowanie jego miejsc zerowych. Nawet jeśli zadanie wyjściowe jest dobrze uwarunkowane, zadanie znalezienia miejsc zerowych wielomianu na podstawie jego współczynników jest zwykle *bardzo źle* uwarunkowane.

Metoda potęgowa

Przypuśćmy, że jedna z wartości własnych macierzy A dominuje, tj. jej wartość bezwzględna jest większa niż wartości bezwzględne wszystkich pozostałych wartości własnych, i przypuśćmy, że mamy wyznaczyć parę własną z właśnie tą wartością własną. Załóżymy, że dominująca wartość własna jest liczbą rzeczywistą (możemy, jeśli macierz A jest symetryczna) i chwilowo przyjmiemy, że jej krotność jest równa 1. Niech będzie to wartość własna λ_1.

Wybieramy niezerowy wektor $\mathbf{x}^{(0)} \in \mathbb{R}^n$, a następnie dla $k > 0$ określamy wektory $\mathbf{x}^{(k)}$, wzorem $\mathbf{x}^{(k)} = A\mathbf{x}^{(k-1)}$ (czyli $\mathbf{x}^{(k)} = A^k\mathbf{x}^{(0)}$). Jeśli macierz A jest diagonalizowalna, to istnieją liczby $c_1, \ldots, c_n$, takie że

$$\mathbf{x}^{(0)} = \sum_{i=1}^{n} c_i \mathbf{x}_i,$$

gdzie $\mathbf{x}_i$ to wektory własne macierzy A. Wtedy mamy

$$\mathbf{x}^{(k)} = \sum_{i=1}^{n} c_i \lambda_i^k \mathbf{x}_i = \lambda_1^k \sum_{i=1}^{n} c_i \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathbf{x}_i.$$

Jeśli $|\lambda_i| < |\lambda_1|$, to dla $k \to \infty$ ciąg liczb $(\lambda_i/\lambda_1)^k$ dąży do zera. To oznacza, że jeśli $c_1 \neq 0$, to ciąg kierunków wektorów $\mathbf{x}^{(k)}$ dąży do kierunku wektora własnego $\mathbf{x}_1$, przynależnego do dominującej wartości własnej. Po wykonaniu dostatecznie wielu iteracji możemy w ten sposób znaleźć wektor bliski wektora własnego $\mathbf{x}_1$.

Podane rozumowanie jest podstawą metody potęgowej rozwiązywania algebraicznego zagadnienia własnego, a dokładniej wyznaczania pary własnej $(\mathbf{x}_1, \lambda_1)$ z dominującą wartością własną. Jeśli krotność geometryczna tej wartości własnej jest większa niż 1, to kierunki otrzymanego ciągu wektorów zbiegają do kierunku *pewnego* wektora własnego związanego z dominującą wartością własną. Opisane postępowanie jest jednak niepraktyczne, ponieważ jeśli $|\lambda_1| \neq 1$, to długości wektorów $\mathbf{x}^{(k)}$ maleją do zera lub rosną nieograniczenie. Dlatego należy stosować normalizację, tj. dzielić kolejne otrzymane wektory przez ich długości — pamiętamy, że istotne są tylko kierunki tych wektorów. Mamy stąd algorytm:

1. Przyjmij $\mathbf{z}^{(0)} \neq \mathbf{0}$,
2. Dla $k = 1, 2, \ldots$ obliczaj

$$\mathbf{y}^{(k)} = A\mathbf{z}^{(k-1)}, \qquad \mathbf{z}^{(k)} = \frac{1}{\|\mathbf{y}^{(k)}\|_2}\mathbf{y}^{(k)}.$$

Jeśli pewien wektor $\mathbf{z}$ jest wektorem własnym macierzy A, to spełnia równanie $A\mathbf{z} = \lambda\mathbf{z}$. Możemy je potraktować jak układ n równań z jedną niewiadomą, którą jest wartość własna λ; macierz tego układu jest kolumnowa, jest nią wektor $\mathbf{z}$. Dla takiego układu stawiamy RLZNK. Układ równań normalnych ma postać

$$\mathbf{z}^T\mathbf{z}\lambda = \mathbf{z}^TA\mathbf{z},$$

aby go rozwiązać, obliczamy tzw. iloraz Rayleigha

$$\lambda = \frac{\mathbf{z}^TA\mathbf{z}}{\mathbf{z}^T\mathbf{z}}.$$

Jeśli wektor $\mathbf{z}$ *nie jest* wektorem własnym, to oczywiście układ $A\mathbf{z} = \lambda\mathbf{z}$ jest sprzeczny, ale jeśli wektor $\mathbf{z}$ jest przybliżeniem wektora własnego $\mathbf{x}_i$, to iloraz Rayleigha jest przybliżeniem wartości własnej λ_i. Ale jeśli $\|\mathbf{z}\|_2 = 1$, to mianownik ilorazu Rayleigha jest równy 1. Zatem, po obliczeniu wektora $\mathbf{z}^{(k)}$ obliczamy liczbę $\rho_{k-1} = \mathbf{z}^{(k)T}\mathbf{y}^{(k-1)}$. Podczas gdy ciąg wektorów jednostkowych $\mathbf{z}^{(k)}$ zbiega do wektora własnego, ciąg liczb ρ_k zbiega do dominującej wartości własnej λ_1.

Jeśli macierz A jest symetryczna, λ_2 jest drugą co do wartości bezwzględnej wartością własną, i symbolem t_k oznaczymy tangens najmniejszego kąta między wektorem $z^{(k)}$ i wektorem x_1 należącym do podprzestrzeni własnej przynależnej do wartości własnej λ_1, to można udowodnić, że

$$|t_k| \leqslant \left|\frac{\lambda_2}{\lambda_1}\right|^k |t_0|, \quad \text{oraz} \quad |\rho_k - \lambda_1| \leqslant 2\|A\||t_k|^2 = O\Big(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\Big).$$

Szybkość zbieżności zależy więc od tego, „jak bardzo dominuje" wartość własna λ_1. Zbieżność nie ma miejsca, jeśli dwie wartości własne dominują, tj. $\lambda_2 = -\lambda_1$. W takim przypadku „prosta" metoda potęgowa nie wystarczy do rozwiązania zadania.

Jeśli liczba c_1 dla przyjętego wektora $z^{(0)}$ jest zerem, to teoretycznie ciąg $(z^{(k)})_{k\in\mathbb{N}}$ zbiega do wektora własnego związanego z którąś z pozostałych wartości własnych. Ale w obliczeniach numerycznych występują błędy zaokrągleń, których skutki w tym przypadku *mogą być dobroczynne*: zaburzenie spowodowane zaokrągleniem zwykle doprowadza do pojawienia się odpowiedniej składowej o kierunku wektora własnego związanego z wartością własną λ_1, po czym kolejne iteracje „wzmacniają" tę składową, jednocześnie „wygaszając" pozostałe.

Odwrotna metoda potęgowa

Jeśli liczba λ jest wartością własną macierzy A, to dla dowolnego $\alpha \notin \operatorname{spect} A$ liczba $1/(\lambda - \alpha)$ jest wartością własną macierzy $(A - \alpha I)^{-1}$. Zauważmy, że jeśli liczba α jest najbliżej wartości własnej λ_i macierzy A (tj. $|\lambda_i - \alpha| < |\lambda_j - \alpha|$ dla każdego $j \neq i$), to wartość własna $1/(\lambda_i - \alpha)$ macierzy $(A - \alpha I)^{-1}$ dominuje; co więcej, im lepsze przybliżenie α wartości własnej λ_i wybierzemy, tym szybsza jest zbieżność metody potęgowej zastosowanej do macierzy $(A - \alpha I)^{-1}$.

Otrzymana na podstawie powyższego spostrzeżenia odwrotna metoda potęgowa, zwana też metodą Wielandta, umożliwia obliczenie dowolnej wartości własnej macierzy A (a nie tylko dominującej), a poza tym umożliwia otrzymanie szybkiej zbieżności.

Algorytm jest taki:

1. Oblicz macierz $B = A - \alpha I$ i rozłóż ją (np. na czynniki trójkątne, za pomocą eliminacji Gaussa).
2. Przyjmij $z^{(0)} \neq 0$,
3. Dla $k = 1, 2, \ldots$ obliczaj

$$y^{(k)} = B^{-1}z^{(k-1)}, \quad \text{rozwiązując układ równań} \quad By^{(k)} = z^{(k-1)},$$
$$z^{(k)} = \frac{1}{\|y^{(k)}\|_2} y^{(k)}.$$

Ciąg wektorów $(z^{(k)})_{k\in\mathbb{N}}$ dąży do wektora własnego macierzy B^{-1}, który jest także wektorem własnym macierzy A. Na podstawie ilorazu Rayleigha $\rho_k = z^{(k)T}y^{(k-1)}$ można obliczyć przybliżenie $\lambda_i \approx 1/\rho_k + \alpha$.

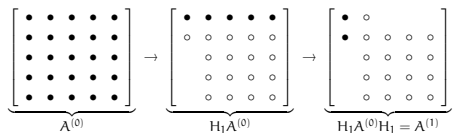
Jeśli macierz A jest pełna, to koszt jej rozłożenia w kroku pierwszym jest rzędu n^3, zaś koszt rozwiązywania układu równań w każdej iteracji jest rzędu n^2, czyli taki sam jak koszt jednej iteracji zwykłej metody potęgowej. Koszt jednej iteracji można zmniejszyć, dokonując wstępnego przekształcenia macierzy, co będzie opisane dalej.

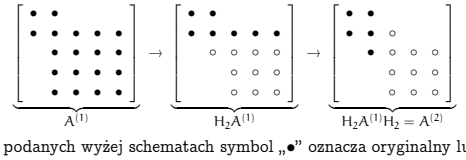
Przybliżenie wartości własnej otrzymane na podstawie ilorazu Rayleigha po wykonaniu pewnej liczby iteracji umożliwia (znaczne) przyspieszenie zbieżności, kosztem ponownego rozkładania na czynniki macierzy $A - \alpha' I$. Macierz ta jest źle uwarunkowana (tym gorzej, im lepszym przybliżeniem wartości własnej macierzy A jest liczba α'), ale ponieważ prawa strona rozwiązywanego układu równań jest przybliżeniem wektora własnego przynależnego do dominującej wartości własnej macierzy $(A - \alpha' I)^{-1}$, okazuje się, że skutki błędów zaokrągleń nie są groźne dla dokładności obliczeń.

Sprowadzanie macierzy symetrycznej do postaci trójdiagonalnej

Wprawdzie (dla macierzy $n \times n$, gdzie $n > 4$) na ogół *nie można* w skończenie wielu krokach skonstruować macierzy X, takiej że macierz $\Lambda = X^{-1}AX$ jest diagonalna, ale dla macierzy symetrycznej *można* skonstruować macierz ortogonalną U, taką że macierz $T = U^{-1}AU$ jest trójdiagonalna. Koszt tego obliczenia jest (dla macierzy pełnej) rzędu n^3, ale można je wykonać jednorazowo, a następnie rozwiązać zagadnienie własne dla macierzy T; ma ona te same wartości własne, co macierz A, jeśli zaś wektor y jest wektorem własnym macierzy T, to wektor $x = Uy$ jest wektorem własnym macierzy A. Zarówno koszt obliczania iloczynu $y^{(k)} = Tz^{(k-1)}$, jak i koszt rozwiązywania układu równań $(T - \alpha I)y^{(k)} = z^{(k-1)}$, jest rzędu n. Wstępne przekształcenie macierzy do postaci trójdiagonalnej jest też wstępnym krokiem wielu innych algorytmów rozwiązywania algebraicznego zagadnienia własnego.

Opiszemy algorytm Ortegi-Householdera. Otrzymana w nim macierz U jest iloczynem macierzy $n - 2$ odbić Householdera; jak zwykle, nie wyznaczamy jej w postaci jawnej, tylko zapamiętujemy odpowiedni ciąg wektorów normalnych hiperpłaszczyzn odbić. Obliczenie polega na skonstruowaniu ciągu macierzy symetrycznych, $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-2)} = T$. Współczynniki macierzy $A^{(k)}$ spełniają warunek $a^{(k)}_{ij} = a^{(k)}_{ji} = 0$ dla $j \leqslant k$ oraz $i > j + 1$. Ponadto, jeśli $i < k$ lub $j < k$, to $a^{(k)}_{ij} = a^{(k-1)}_{ij}$.





W podanych wyżej schematach symbol „•" oznacza oryginalny lub niezmieniony współczynnik macierzy, zaś „◦" oznacza współczynnik, który wskutek odbicia uległ zmianie. Puste miejsca oznaczają (wytworzone lub zachowane) zera.

Pierwsza współrzędna wektora v_1, określającego odbicie reprezentowane przez macierz $H_1 = I - \gamma_1 v_1 v_1^T$, jest równa 0. Dla takiego odbicia macierze $A^{(0)}$ i $H_1 A^{(0)}$ mają taki sam pierwszy wiersz. Odbicie konstruujemy w taki sposób, aby w pierwszej kolumnie macierzy $H_1 A^{(0)}$ w wierszach $3, \ldots, n$ otrzymać zera. Mnożenie przez macierz odbicia z prawej strony zachowuje pierwszą kolumnę macierzy $H_1 A^{(0)}$, w tym jej zerowe współczynniki. Wykonane przekształcenie $A^{(0)} \to A^{(1)}$ jest podobieństwem macierzy, ponieważ macierz H_1 jest symetryczna i ortogonalna. Ponadto przekształcenie to zachowuje symetrię, a zatem w pierwszym wierszu macierzy $A^{(1)}$, w kolumnach $3, \ldots, n$ też mamy zera.

Wektor v_2 ma dwie pierwsze współrzędne równe zero, czego konsekwencją jest zachowanie pierwszego wiersza i pierwszej kolumny macierzy $A^{(1)}$.

Teraz implementacja. W k-tym kroku mamy obliczyć macierz

$$\begin{aligned} A^{(k)} &= H_k A^{(k-1)} H_k = (I - \gamma_k v_k v_k^T) A^{(k-1)} (I - \gamma_k v_k v_k^T) \\ &= A^{(k-1)} - \gamma_k v_k v_k^T A^{(k-1)} - \gamma_k A^{(k-1)} v_k v_k^T + \\ &\qquad \gamma_k^2 v_k v_k^T A^{(k-1)} v_k v_k^T. \end{aligned}$$

Oznaczmy $w = \gamma_k A^{(k-1)} v_k$. Wtedy

$$A^{(k)} = A^{(k-1)} - v_k w^T - w v_k^T + v_k (\gamma_k v_k^T w) v_k^T.$$

Niech $p = w - v_k (v_k^T w) \gamma_k / 2$. Możemy sprawdzić, że

$$A^{(k)} = A^{(k-1)} - (v_k p^T + p v_k^T).$$

Właśnie tego wzoru używamy w obliczeniach. Zauważmy, że wektory w i p obliczone w k-tym kroku mają $k-1$ początkowych współrzędnych równych 0. Dzięki symetrii można obliczać tylko współczynniki na i pod (albo na i nad) diagonalą, dla zmniejszenia kosztu.

Algorytm QR

Niech A będzie nieosobliwą macierzą symetryczną i niech Z_{k-1} będzie dowolną macierzą nieosobliwą $n \times n$. Kolumny macierzy $Y_k = A Z_{k-1}$, zgodnie ze spostrzeżeniami, na których opiera się metoda potęgowa, mają „kierunki bliższe" kierunku wektora własnego x_1, przynależnego do dominującej wartości własnej, λ_1. Ale gdybyśmy układ wektorów $y_1^{(k)}, \ldots, y_n^{(k)}$, tj. kolumn macierzy Y_k poddali ortonormalizacji Grama-Schmidta, to otrzymalibyśmy układ wektorów $z_1^{(k)}, \ldots, z_n^{(k)}$, z których każdy ma „kierunek bliższy" kierunku wektora przynależnego do kolejnej wartości własnej (zakładamy, że wartości własne są ponumerowane w taki sposób, że $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, nierówności nieostre dopuścimy później). Jest tak dlatego, bo ortonormalizacja „likwiduje" składowe wektora $y_i^{(k)}$ w kierunkach wektorów $z_1^{(k)}, \ldots, z_{i-1}^{(k)}$, które są przybliżeniami wektorów własnych $x_1, \ldots, x_{i-1}$ macierzy A. Stąd wynika przypuszczenie, że dla każdego $i \in \{1, \ldots, n\}$ ciąg wektorów $(z_i^{(k)})_{k \in \mathbb{N}}$ dąży do wektora własnego x_i przynależnego do wartości własnej λ_i.

Macierz $Z_k = [z_1^{(k)}, \ldots, z_n^{(k)}]$ jest ortogonalna, a ponadto istnieje macierz trójkątna górna R_k, taka że $Y_k = Z_k R_k$. Przyjmijmy $Z_0 = I$ i oznaczmy

$$A_k \stackrel{\text{def}}{=} Z_k^T A Z_k$$

(czyli w szczególności $A_0 = A$, ponadto wszystkie macierze A_k są podobne do A i symetryczne). Wtedy

$$A_k = Z_k^T Y_{k+1} = Z_k^T Z_{k+1} R_{k+1}.$$

Niech $Q_{k+1} = Z_k^T Z_{k+1}$. Stąd $Z_{k+1} = Z_k Q_{k+1}$, a przez indukcję mamy stąd

$$Z_k = Z_0 Q_1 \ldots Q_k = Q_1 \ldots Q_k.$$

Na tej podstawie

$$A_k = Q_k^T \ldots Q_1^T A Q_1 \ldots Q_k = Q_k^T A_{k-1} Q_k = R_k Q_k.$$

Ten rachunek jest podstawą dla następującego algorytmu:

1. Przyjmij $A_0 = A$,
2. Dla $k = 1, 2, \ldots$

   znajdź macierze ortogonalną Q_k i trójkątną górną R_k,
   takie że $A_{k-1} = Q_k R_k$,
   oblicz $A_k = R_k Q_k$.

Jeśli ciąg macierzy $(Z_k)_{k \in \mathbb{N}}$ zbiega do macierzy X, której kolumny są wektorami własnymi macierzy A, to ciąg macierzy $(A_k)_{k \in \mathbb{N}}$ zbiega do macierzy diagonalnej Λ, której znalezienie jest równoznaczne z obliczeniem wszystkich wartości własnych. Zbieżność może jednak nie mieć miejsca (z tego samego powodu, dla którego metoda potęgowa może nie być zbieżna — wystarczy, że dwie wartości własne mają tę samą wartość bezwzględną i przeciwne znaki).

Zanim zajmiemy się zbieżnością, dokonajmy pewnego spostrzeżenia, które ma wpływ na koszt algorytmu. Jeśli macierz A_{k-1} jest trójdiagonalna, to macierz A_k też jest taka. Dlatego pierwszym etapem obliczeń powinno być przekształcenie danej macierzy do postaci trójdiagonalnej (przy użyciu algorytmu Ortegi-Householdera), co kosztuje $O(n^3)$ działań i jest równoważne zastąpieniu macierzy jednostkowej Z_0 w procesie iteracyjnym rozważanym wyżej przez macierz ortogonalną $Q_0 = H_1 \ldots H_{n-2}$, która reprezentuje złożenie wykonanych odbić.

Rozkładanie macierzy trójdiagonalnej na czynniki Q_k i R_k, a następnie obliczanie A_k jest wykonywane kosztem $O(n)$ działań. Zamiast ortonormalizacji Grama-Schmidta (która zawiedzie, jeśli macierz jest osobliwa), lepiej jest tu użyć innej metody; zwykle korzysta się z tzw. obrotów Givensa. Można by też użyć odbić Householdera, ale do rozkładania macierzy trójdiagonalnej są one mniej wygodne.

Aby osiągnąć zbieżność i sprawić, by była jak najszybsza, w kolejnych iteracjach dobiera się parametr a_k (tzw. przesunięcie) i znajduje czynniki rozkładu macierzy $A_{k-1} - a_k I = Q_k R_k$, a nastęnie oblicza się macierz $A_k = R_k Q_k + a_k I$. Zauważmy, że

$$A_k = Q_k^T (A_{k-1} - a_k I) Q_k + a_k I = Q_k^T A_{k-1} Q_k,$$

a więc dla dowolnego przesunięcia macierze A_{k-1} i A_k są podobne. Mamy też $A_k = Z_k^T A Z_k$ oraz $Q_k = Z_{k-1}^T Z_k$, tak samo jak w przypadku bez przesunięć.

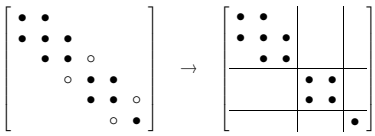
Gdyby przesunięcie a_k było równe pewnej wartości własnej λ, to wszystkie kolumny iloczynu $Y_k = (A - a_k I) Z_{k-1}$ byłyby prostopadłe do wszystkich wektorów własnych x przynależnych do tej wartości własnej (oczywiście macierz Y_k byłaby osobliwa). Przypuśćmy, że krotność wartości własnej λ jest równa 1 i wektory $y_1, \ldots, y_{n-1}$ (początkowe kolumny Y_k) są liniowo niezależne. Otrzymane z nich metodą Grama-Schmidta wektory $z_1, \ldots, z_{n-1}$ są prostopadłe do x. Macierz ortogonalna Z_k, której to są początkowe kolumny, ma kolumnę z_n do nich prostopadłą, ale to znaczy, że ta kolumna ma kierunek wektora x, czyli *jest jednostkowym wektorem własnym* przynależnym do wartości własnej λ macierzy A. Łatwo jest sprawdzić, że wtedy współczynnik macierzy $A_k = Z_k^T A Z_k$ na ostatnim miejscu diagonali byłby równy λ, a pozostałe współczynniki w ostatnim wierszu i kolumnie byłyby równe 0.

Jeśli zatem $a_k = \lambda$, to *w jednym kroku* dostaniemy macierz A_k ze współczynnikiem $a^{(k)}_{nn} = \lambda$. Jeśli przesunięcie a_k jest tylko przybliżeniem λ, a dokładniej, są spełnione nierówności $|a_k - \lambda| < |a_k - \lambda_i|$ dla każdej wartości własnej $\lambda_i \neq \lambda$, to ciąg współczynników $(a^{(k)}_{nn})_{k\in\mathbb{N}}$ będzie zbieżny do λ, tym szybciej, im lepiej parametr przesunięcia przybliża tę wartość własną. Aby zbieżność była jeszcze szybsza, w każdej iteracji wybiera się nowe przesunięcie.

Istnieją różne sposoby wybierania przesunięcia; jego wartość powinna przybliżać pewną wartość własną macierzy A. Najprostszy (i skuteczny) wybór to $a_k = a^{(k-1)}_{nn}$.

Inny sposób (tzw. przesunięcie Wilkinsona) polega na przyjęciu parametru a_k równego jednej z wartości własnych bloku 2×2 wybranego z dwóch ostatnich wierszy i kolumn macierzy A_{k-1} (w tym celu trzeba rozwiązać równanie kwadratowe).

Współczynniki diagonalne kolejnych macierzy A_k dążą (z różnymi szybkościami) do wartości własnych, zaś współczynniki kodiagonalne (tj. sąsiadujące z diagonalą) dążą do zera. Dla odpowiednio dobranych przesunięć najszybciej zbiegają współczynniki w ostatnim wierszu i kolumnie. Jeśli wartość bezwzględna pewnego współczynnika na kodiagonali jest dostatecznie mała, tj. na poziomie błędów zaokrągleń, to współczynnik ten zastępuje się zerem, ale wtedy powstaje macierz blokowo-diagonalna z trójdiagonalnymi blokami:



i obliczenia można kontynuować dla tych bloków niezależnie, dobierając niezależnie przesunięcia.

Przejście od zadania postawionego dla całej macierzy do zadań w mniejszych blokach nazywa się deflacją.

Algorytm QR ze wstępnym przekształceniem do postaci trójdiagonalnej, przesunięciami i rekurencyjną deflacją jest najefektywniejszym znanym algorytmem znajdowania wszystkich wartości własnych macierzy symetrycznej.

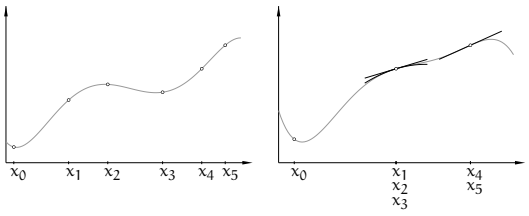
7. Interpolacja wielomianowa

Zadania interpolacyjne Lagrange'a i Hermite'a

Niech $x_0, \ldots, x_n$ będą danymi liczbami, z których każde dwie są różne i niech $y_0, \ldots, y_n$ będą liczbami dowolnymi.
Zadanie interpolacyjne Lagrange'a polega na skonstruowaniu wielomianu $h(x)$ stopnia co najwyżej n, takiego że $h(x_i) = y_i$ dla $i = 0, \ldots, n$.

Wymaganie, aby liczby x_i, zwane węzłami interpolacyjnymi, były parami różne, jest oczywiste; nie można zadawać dwóch różnych wartości funkcji w tym samym punkcie. Ale możemy dopuścić, aby węzły powtarzały się, jeśli dla każdego dodatkowego „egzemplarza" węzła określimy inny warunek interpolacyjny.

Jeśli warunek ten polega na podaniu wartości pochodnej kolejnego rzędu, to mamy ogólniejsze zadanie interpolacyjne Hermite'a: dla każdego węzła określamy jego krotność — jest to liczba jego wystąpień w danym ciągu węzłów. Dla węzła x_i o krotności $r > 1$ zadajemy wartość funkcji, $h(x_i)$, pochodnej, $h'(x_i)$, i pochodnych do rzędu $r - 1$ włącznie.



Twierdzenie. *Zadanie interpolacyjne Hermite'a i jego przypadek szczególny — zadanie interpolacyjne Lagrange'a — ma jednoznaczne rozwiązanie.*

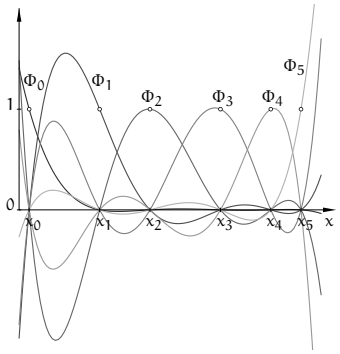
Dowód. Poszukiwany wielomian przedstawimy jako kombinację liniową elementów dowolnej bazy przestrzeni $\mathbb{R}[x]_n$. Warunki interpolacyjne możemy zapisać w postaci układu równań liniowych, z niewiadomymi współczynnikami w wybranej bazie. Wymiar przestrzeni, czyli liczba niewiadomych, jest równy $n + 1$, tj. taki sam jak liczba równań.

Przypuśćmy, że wszystkie zadane wartości funkcji i pochodnych są równe 0. Wtedy układ ma rozwiązanie — wektor zerowy, który reprezentuje wielomian zerowy. Gdyby istniał niezerowy wielomian $h(x)$ stopnia co najwyżej n spełniający te same warunki interpolacyjne, to musiałby być podzielny przez wielomian $p_{n+1}(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$, ale to oznacza, że stopień wielomianu h musiałby być co najmniej $n + 1$. Jednoznaczność rozwiązania układu równań opisującego jednorodne warunki interpolacyjne oznacza, że macierz tego układu jest nieosobliwa, a więc dla dowolnej prawej strony układ ma jednoznaczne rozwiązanie. □

Rozwiązanie zadania interpolacyjnego Lagrange'a można przedstawić wzorem

$$h(x) = \sum_{i=0}^{n} y_i \Phi_i(x), \quad \text{gdzie} \quad \Phi_i(x) = \prod_{j\in\{0,\ldots,n\}\setminus\{i\}} \frac{x - x_j}{x_i - x_j},$$

ale wzór ten nie jest praktyczny w obliczeniach numerycznych (należy go raczej traktować jako dowód istnienia rozwiązania zadania, czasem przydaje się też w rachunkach symbolicznych i w rozważaniach teoretycznych).



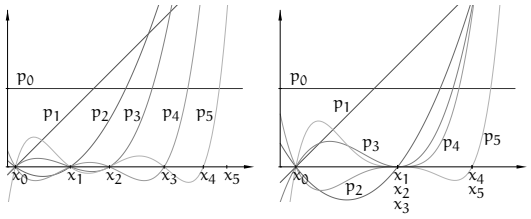
Wykresy wielomianów $\Phi_0, \ldots, \Phi_n$ dla przykładowego ciągu węzłów (z $n = 5$) są pokazane na rysunku. Warto zauważyć, że niektóre z tych wielomianów przyjmują między węzłami wartości bezwzględne sporo większe niż 1. Maksymalne wartości bezwzględne wielomianów bazowych Lagrange'a między węzłami zależą od liczby węzłów i od ich rozmieszczenia, i jeśli stopień jest duży, to mogą być bardzo duże. W konsekwencji rozwiązanie zadania interpolacji może przyjmować między węzłami interpolacyjnymi wartości leżące daleko poza przedziałem, w którym leżą dane wartości funkcji w tych węzłach.

Bazy Newtona

Niech $x_0, \ldots, x_n$ będą liczbami danymi. Możemy określić wielomiany

$$\begin{aligned}
p_0(x) &= 1,\\
p_1(x) &= x - x_0,\\
p_2(x) &= (x - x_0)(x - x_1),\\
&\vdots\\
p_n(x) &= (x - x_0) \cdot \ldots \cdot (x - x_{n-1}),\\
p_{n+1}(x) &= (x - x_0) \cdot \ldots \cdot (x - x_{n-1})(x - x_n).
\end{aligned}$$

Zbiór wielomianów $\{p_0, \ldots, p_k\}$ jest bazą przestrzeni $\mathbb{R}[x]_k$, której elementami są wszystkie wielomiany stopnia co najwyżej k. Ta tzw. baza Newtona jest wygodniejsza od bazy potęgowej w zastosowaniu do zadań interpolacji wielomianowej.



W szczególności, mając współczynniki $b_0, \ldots, b_n$ wielomianu stopnia co najwyżej n, możemy obliczyć wartość wielomianu $w(x) = \sum_{i=0}^{n} b_i p_i(x)$ za pomocą odpowiednio uogólnionego schematu Hornera:

```
w = b_n;
for ( i = n − 1; i ⩾ 0; i-- )
  w = w*(x − x_i) + b_i;
```


Aby rozwiązać zadanie interpolacyjne Lagrange'a, możemy dla wybranej bazy $\{f_0, \ldots, f_n\}$ przestrzeni $\mathbb{R}[x]_n$ utworzyć macierz $A \in \mathbb{R}^{n+1,n+1}$, taką że jej współczynnik $a_{ij} = f_i(x_j)$ (numerujemy tu wiersze i kolumny od 0 do n). Rozwiązanie zadania sprowadza się do rozwiązania układu równań z tą macierzą. Dla bazy potęgowej mamy układ równań z macierzą pełną

$$\begin{bmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix},$$

którego rozwiązaniem jest wektor współczynników wielomianu $h(x) = \sum_{k=0}^{n} a_k x^k$.

Dla bazy Newtona określonej za pomocą węzłów interpolacyjnych mamy układ z macierzą trójkątną dolną:

$$\begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & p_1(x_1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & p_1(x_n) & \ldots & p_n(x_n) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Możemy obliczyć współczynniki tej macierzy i rozwiązać układ kosztem tylko $\Theta(n^2)$ operacji (dalej poznamy inny algorytm obliczania współczynników wielomianu interpolacyjnego w bazie Newtona). W razie potrzeby, możemy następnie kosztem $\Theta(n^2)$ operacji przejść do bazy potęgowej, ale jeśli nie jest to konieczne, to nie warto tego robić.

Różnice dzielone

Niech f oznacza pewną funkcję $A \subset \mathbb{R} \to \mathbb{R}$. Dla ustalonych liczb $x_i \in A$ (węzłów interpolacyjnych) określamy różnice dzielone rzędu 0:

$$f[x_i] \stackrel{\text{def}}{=} f(x_i).$$

Zakładajac, że węzły są jednokrotne (czyli parami różne), możemy następnie określić dla $k > 0$ różnice dzielone rzędu k wzorem

$$f[x_i, \ldots, x_{i+k}] \stackrel{\text{def}}{=} \frac{f[x_i, \ldots, x_{i+k-1}] - f[x_{i+1}, \ldots, x_{i+k}]}{x_i - x_{i+k}}. \qquad (*)$$

Różnicę dzieloną można postrzegać na dwa sposoby:

1. Dla ustalonych węzłów $x_i, \ldots, x_{i+k}$ jest to kombinacja liniowa wartości funkcji f w tych węzłach, a zatem jest to funkcjonał liniowy w przestrzeni funkcji o ustalonej dziedzinie A, do której należą te węzły,
2. Dla ustalonej funkcji f jest to funkcja $k + 1$ zmiennych. Łatwo jest dowieść, że jest to funkcja symetryczna, tj. dowolne przestawienie jej argumentów (węzłów) nie zmienia jej wartości.

Jak wiemy, jeśli wyrażenie $\lim_{h\to 0} f[x, x+h]$ ma określoną (i skończoną) wartość, to jest to pochodna funkcji f w punkcie x. Możemy zatem rozszerzyć definicję różnicy dzielonej na przypadek, gdy pewne (lub nawet wszystkie) węzły mają krotności większe niż 1, wykorzystując przejście do granicy. Okazuje się, że jeśli funkcja f jest klasy C^k, to różnica dzielona, widziana jako funkcja, której argumentami są węzły, jest ciągła i zachodzi równość

$$\lim_{x_{i+1},\ldots,x_{i+k}\to x_i} f[x_i, \ldots, x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

Na tej podstawie możemy zdefiniować różnicę dzieloną rzędu k w przypadku, gdy $x_i = \cdots = x_{i+k}$, wzorem

$$f[\underbrace{x_i, \ldots, x_i}_{k+1}] \stackrel{\text{def}}{=} \frac{f^{(k)}(x_i)}{k!}, \qquad (**)$$

natomiast w przypadku, gdy pewne węzły mają krotność większą niż 1, ale nie wszystkie węzły są jednakowe, możemy (dzięki symetrii) uporządkować je tak, aby było $x_i \neq x_{i+k}$, i użyć wzoru (*).

W przypadku ogólnym różnica dzielona rzędu k, $f[x_i, \ldots, x_{i+k}]$, jest kombinacją liniową wartości funkcji f i jej pochodnych w węzłach, przy czym jeśli pewien węzeł ma krotność r, to kombinacja obejmuje pochodne funkcji f w tym węźle do rzędu $r-1$.

Algorytm różnic dzielonych

Przypuśćmy, że węzły $x_0, \ldots, x_n$ są parami różne. Obliczmy różnicę dzieloną wielomianu $p_k(x)$ należącego do bazy Newtona określonej dla tych węzłów:

$$\begin{aligned} p_k[x, x_0] &= \frac{(x-x_0)\cdot\ldots\cdot(x-x_{k-1}) - (x_0-x_0)\cdot\ldots\cdot(x_0-x_{k-1})}{x-x_0} \\ &= (x-x_1)\cdot\ldots\cdot(x-x_{k-1}). \end{aligned}$$

Otrzymaliśmy wielomian stopnia $k-1$. Obliczając różnice dzielone coraz wyższych rzędów, dostaniemy wielomiany coraz niższych stopni:

$$\begin{aligned} p_k[x, x_0, x_1] &= (x-x_2)\cdot\ldots\cdot(x-x_{k-1}), \\ &\vdots \\ p_k[x, x_0, \ldots, x_{k-2}] &= (x-x_{k-1}), \\ p_k[x, x_0, \ldots, x_{k-2}, x_{k-1}] &= 1. \end{aligned}$$

Po ostatnim kroku możemy oczywiście podstawić $x = x_k$, co nie zmieni wartości otrzymanego wielomianu stopnia 0. Różnice dzielone rzędów wyższych niż k są równe 0. Biorąc pod uwagę zbiór miejsc zerowych wielomianu p_k, mamy

$$p_k[x_0, \ldots, x_i] = \begin{cases} 0 & \text{dla } i \neq k, \\ 1 & \text{dla } i = k. \end{cases}$$

Podany wyżej rachunek „przechodzi” też na przypadek węzłów powtarzających się — wystarczy użyć indukcji i w kroku indukcyjnym dokonać odpowiedniego przejścia do granicy.

Niech $h(x)$ będzie rozwiązaniem zadania interpolacyjnego Lagrange'a dla węzłów $x_0, \ldots, x_n$. Wielomian h możemy przedstawić jako kombinację liniową elementów bazy Newtona: $h(x) = \sum_{k=0}^{n} b_k p_k(x)$. Z tego, że dla ustalonych węzłów różnice dzielone są funkcjonałami liniowymi, i z przeprowadzonego wyżej rachunku wynika, że zachodzą równości

$$h[x_0, \ldots, x_i] = \sum_{k=0}^{n} b_k p_k[x_0, \ldots, x_i] = b_i.$$

Znamy wartości wielomianu h w węzłach interpolacyjnych, są nimi liczby $y_0, \ldots, y_n$, a zatem możemy obliczyć współczynniki $b_0, \ldots, b_n$ wielomianu h w bazie Newtona. Wygodnie jest przedstawić ich obliczenie za pomocą schematu

$$\begin{array}{c|ccccc} x_0 & y_0 = b_0 & & & & \\ & \searrow & & & & \\ x_1 & y_1 & \to & h[x_0,x_1] = b_1 & & \\ & \searrow & & \searrow & & \\ x_2 & y_2 & \to & h[x_1,x_2] & \to & h[x_0,x_1,x_2] = b_2 \\ \vdots & \vdots & & \vdots & & \vdots \\ & \searrow & & \searrow & & \searrow \\ x_n & y_n & \to & h[x_{n-1},x_n] & \to & h[x_{n-2},x_{n-1},x_n] \ldots \to h[x_0,\ldots,x_n] = b_n \end{array}$$

Podprogram realizujący to obliczenie zastępuje w tablicy y dane wartości funkcji przez współczynniki $b_0, \ldots, b_n$:

```
for ( j = 1; j <= n; j++ )
  for ( i = n; i >= j; i-- )
    y[i] = (y[i] - y[i-1])/(x[i] - x[i-j]);
```


Aby rozwiązać zadanie interpolacyjne Hermite'a, należy zmodyfikować ten algorytm. Istotne jest uporządkowanie danych; wymagamy, aby w tablicy x wszystkie „egzemplarze” węzła krotnego występowały obok siebie. W tablicy y zadaną wartość funkcji podajemy w miejscu odpowiadającym pierwszemu wystąpieniu odpowiedniego węzła, a na kolejnych miejscach podajemy wartości kolejnych pochodnych. Algorytm można zrealizować w taki sposób:

```
k[0] = 0;
for ( i = 1; i <= n; i++ )
  k[i] = x[i] == x[i-1] ? k[i-1] + 1 : 0;
for ( j = 1; j <= n; j++ )
  for ( i = n; i >= j; i-- )
    if ( k[i] == 0 )
      y[i] = (y[i] - y[i-1-k[i-1]])/(x[i] - x[i-j]);
    else { y[i] /= j; k[i]--; }
```

W pierwszej pętli w miejscu i-tym pomocniczej tablicy k zapisujemy informację, którego rzędu pochodnej wartością jest dana liczba y[i]. W drugiej pętli używamy tej informacji do wybrania odpowiedniej instrukcji: obliczenia różnicy dzielonej za pomocą wzoru (*) lub podzielenia y[i] przez odpowiednią liczbę całkowitą, co prowadzi do otrzymania silni w mianowniku wzoru (**).

Dowodzi się, że jeśli węzły są monotonicznie uporządkowane, to algorytm różnic dzielonych jest numerycznie poprawny, tj. obliczone współczynniki w bazie Newtona reprezentują wielomian interpolacyjny dla odpowiednio mało zaburzonych danych wartości funkcji (i jej pochodnych).

Reszta interpolacyjna

Z uwagi na liczne zastosowania zadań interpolacyjnych Lagrange'a i Hermite'a w aproksymacji funkcji i w konstrukcji różnych metod numerycznych (np. rozwiązywania równań nieliniowych i obliczania całek), duże znaczenie ma wzór opisujący resztę interpolacyjną.

Twierdzenie. *Jeśli funkcja* f *jest klasy* C^{n+1} *w przedziale* $A \subset \mathbb{R}$ *i* $h(x)$ *oznacza wielomian interpolacyjny Hermite'a funkcji* f *dla węzłów* $x_0, \ldots, x_n \in A$*, to dla każdego* $x \in A$ *istnieje liczba* $\xi \in A$*, taka że*

$$f(x) - h(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} p_{n+1}(x).$$

Dowód. Jeśli $x = x_i$ dla pewnego $i \in \{0, \ldots, n\}$, to $p_{n+1}(x) = 0$ i dowodzona równość jest oczywista (z dowolnym $\xi \in A$). Dla ustalonego $x \in A \setminus \{x_0, \ldots, x_n\}$ uporządkujmy ciąg $x_0, \ldots, x_n, x$ tak, aby otrzymać ciąg niemalejący $x_0^{(0)} \leqslant \cdots \leqslant x_{n+1}^{(0)}$. Określamy funkcję

$$g_x(s) \stackrel{\text{def}}{=} f(s) - h(s) - z p_{n+1}(s),$$

z parametrem $z = z(x)$, który dobierzemy za chwilę. Korzystając z tego, że $p_{n+1}(x) \neq 0$, bierzemy

$$z = \frac{f(x) - h(x)}{p_{n+1}(x)},$$

i w ten sposób dostajemy $g_x(x) = 0$. Tak określona funkcja spełnia warunek $g_x(x_i^{(0)}) = 0$ dla $i = 0, \ldots, n+1$, tzn. ma co najmniej $n+2$ miejsca zerowe.

Funkcja g_x jest klasy $C^{n+1}(A)$. Jej pochodna rzędu $k \leqslant n+1$ ma co najmniej $n+2-k$ miejsca zerowe, które tworzą ciąg niemalejący $x_0^{(k)} \leqslant \ldots \leqslant x_{n+1-k}^{(k)}$. Istotnie, jeśli $x_i^{(0)} < x_{i+1}^{(0)}$, to (z twierdzenia Rolle'a) funkcja g_x, która na końcach przedziału $[x_i^{(0)}, x_{i+1}^{(0)}]$ przyjmuje tę samą wartość 0, osiąga w tym przedziale maksimum lub minimum, w punkcie $x_i^{(1)}$ będącym miejscem zerowym funkcji g_x'. Jeśli zaś funkcja g_x ma miejsce zerowe o krotności $r > 1$ (w węźle $x_i^{(0)} = \cdots = x_{i+r-1}^{(0)}$), to jej pochodna ma w tym punkcie miejsce zerowe o krotności $r-1$ (zatem mamy podciąg $x_i^{(1)} = \cdots = x_{i+r-2}^{(1)}$).

Korzystając z indukcji, stosujemy to rozumowanie do kolejnych pochodnych. Wynika z niego, że pochodna rzędu $n+1$ funkcji g_x ma w przedziale A co najmniej jedno miejsce zerowe, $\xi = x_0^{(n+1)}$.

Podstawiając $s = \xi$, dostajemy

$$0 = g_x^{(n+1)}(\xi) = f^{(n+1)}(\xi) - h^{(n+1)}(\xi) - z p_{n+1}^{(n+1)}(\xi).$$

Pochodna rzędu $n+1$ wielomianu $h(s)$ (stopnia n) jest równa 0, zaś pochodna wielomianu $p_{n+1}(s)$ (stopnia $n+1$), którego współczynnik (w bazie potęgowej) przy s^{n+1} jest równy 1, jest dla każdego s równa $(n+1)!$. Zatem

$$z = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Dowód zakończymy, wstawiając to do definicji funkcji g_x i biorąc $s = x$. □

Przypuśćmy, że węzły $x_0, \ldots, x_n$ są parami różne i $x \notin \{x_0, \ldots, x_n\}$. Rozważmy wielomian interpolacyjny Lagrange'a $h(s)$ funkcji f dla węzłów $x_0, \ldots, x_n$ i wielomian $h_{n+1}(s)$ stopnia co najwyżej $n+1$, taki że $h_{n+1}(s) = f(s)$ dla każdego $s \in \{x_0, \ldots, x_n, x\}$. Wtedy

$$h_{n+1}(s) = h(s) + f[x_0, \ldots, x_n, x] p_{n+1}(s),$$

gdzie $p_{n+1}(s) = (s - x_0) \cdot \ldots \cdot (s - x_n)$. Ta równość ma miejsce dla *dowolnej* funkcji f określonej *w punktach* $x_0, \ldots, x_n, x$. Dla $s = x$

$$h_{n+1}(x) = f(x) = h(x) + f[x_0, \ldots, x_n, x] p_{n+1}(x).$$

Jeśli funkcja f jest klasy C^{n+1} *w przedziale* zawierającym wszystkie te punkty, to dla pewnego ξ należącego do tego przedziału

$$f(x) = h(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!} p_{n+1}(x),$$

a ponieważ $p_{n+1}(x) \neq 0$, zachodzi równość

$$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Wniosek. *Jeśli funkcja* f *jest klasy* C^k *w otoczeniu punktu* x_i*, to*

$$\lim_{x_{i+1}, \ldots, x_{i+k} \to x_i} f[x_i, \ldots, x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

Dowód. Wystarczy dla każdego układu liczb $x_i, \ldots, x_{i+k}$ należących do dostatecznie małego otoczenia punktu x_i przyjąć najkrótszy przedział zawierający te liczby i zauważyć, że potrzebną liczbę ξ możemy znaleźć w tym przedziale. □

Fakt ten jest podstawą definicji różnic dzielonych także dla węzłów o krotnościach większych niż 1.

Wzór na resztę interpolacyjną w szczególnym przypadku, gdy $x_0 = \cdots = x_n$, jest wzorem Taylora (z resztą w postaci Lagrange'a). Inny przypadek szczególny, dla dwóch węzłów jednokrotnych, wykorzystaliśmy już w analizie metody siecznych. Dalsze zastosowania nastąpią.

8. Interpolacja funkcjami sklejanymi

Motywacja dla stosowania funkcji sklejanych

Funkcje sklejane są to funkcje określone w ten sposób, że pewien przedział $[a, b] \subset \mathbb{R}$ (albo, jeśli jest taka potrzeba, cały zbiór liczb rzeczywistych) dzielimy na podprzedziały, wybierając węzły. W każdym przedziale, którego końcami są węzły, funkcja sklejana stopnia n jest wielomianem stopnia co najwyżej n.

W wielu zastosowaniach funkcje sklejane są wygodniejsze niż funkcje wielomianowe. W szczególności, kształt wykresu funkcji sklejanej nawet niskiego stopnia może być dowolnie skomplikowany, łatwo jest więc aproksymować różne funkcje z dobrą dokładnością. Zastosowanie funkcji sklejanych w interpolacji ma również przewagę nad wielomianami. Występujący we wzorze opisującym rozwiązanie zadania interpolacji Lagrange'a wielomian

$$\Phi_i(x) = \prod_{j\in\{0,\dots,n\}\setminus\{i\}} \frac{x - x_j}{x_i - x_j},$$

który przyjmuje wartość 1 dla $x = x_i$ oraz 0 dla $x = x_j \neq x_i$, między swoimi miejscami zerowymi oscyluje i (zależnie od n i od liczb $x_0, \dots, x_n$) może przyjmować wartości wychodzące daleko poza przedział $[0, 1]$. Funkcje sklejane nie mają tej wady, dzięki czemu np. wykres funkcji sklejanej przechodzącej przez zadane punkty wydaje się zwykle zgodny z oczekiwaniami.

Obcięte funkcje potęgowe

Węzły funkcji sklejanej oznaczymy symbolami u_i; przyjmiemy, że tworzą one ciąg rosnący (na razie pominiemy przypadek węzłów krotnych, tj. tworzących ciąg niemalejący, ale niekoniecznie różnowartościowy). Założymy, że wielomiany p_{i-1} i p_i stopnia co najwyżej n, opisujące funkcję sklejaną s odpowiednio w przedziałach (u_{i-1}, u_i) oraz (u_i, u_{i+1}), przyjmują w punkcie u_i tę samą wartość, a ponadto mają takie same pochodne rzędu $1, \dots, n-1$ (uwaga: założenie, że również pochodna rzędu n obu wielomianów jest taka sama oznacza, że to jest ten sam wielomian).

Różnica wielomianów p_i i p_{i-1} musi być zatem wielomianem stopnia co najwyżej n, który w punkcie u_i ma miejsce zerowe o krotności n, ale każdy taki wielomian ma postać $c(x - u_i)^n$ dla pewnej stałej c. Jednym z wielu sposobów określania funkcji sklejanych jest użycie tzw. obciętych funkcji potęgowych, określonych wzorem

$$(x - u_i)_+^n \stackrel{\text{def}}{=} \begin{cases} (x - u_i)^n & \text{dla } x \geqslant u_i, \\ 0 & \text{dla } x < u_i. \end{cases}$$

Mając węzły np. $u_0, \dots, u_N$, możemy wybrać dowolny wielomian p_{-1} (stopnia co najwyżej n) oraz liczby $c_0, \dots, c_N$, i określić funkcję sklejaną stopnia n wzorem

$$s(x) = p_{-1}(x) + \sum_{i=0}^{N} c_i (x - u_i)_+^n.$$

Takie przedstawienie funkcji sklejanej daje pewne intuicje (np. „od razu widać", że funkcja s jest klasy $C^{n-1}(\mathbb{R})$), ale obcięte funkcje potęgowe są niewygodne w zastosowaniach i oparta na nich reprezentacja jest podatna na błędy zaokrągleń. Dlatego do reprezentowania funkcji sklejanych i przetwarzania ich w obliczeniach numerycznych stosuje się inne sposoby. Chyba *najmniej wygodnym* z tych sposobów jest stosowanie bazy potęgowej.

Uwaga. W tym wykładzie posługuję się pojęciem stopnia funkcji sklejanej, tj. największego dopuszczalnego przez reprezentację stopnia wielomianu opisującego tę funkcję w pewnym przedziale, ale w wielu publikacjach i bibliotekach podprogramów jest w użyciu tzw. rząd funkcji sklejanej, tj. liczba o 1 większa od stopnia. Zatem, np. funkcje sklejane pierwszego rzędu to funkcje stopnia 0, czyli kawałkami stałe, wykresem funkcji sklejanej rzędu 2, czyli stopnia 1, jest łamana.

Z wielu różnych względów w zastosowaniach dominują funkcje sklejane trzeciego stopnia (tzw. kubiczne) i teraz na nich skupimy uwagę.

Reprezentacja Hermite'a funkcji sklejanych trzeciego stopnia

Określamy cztery wielomiany:

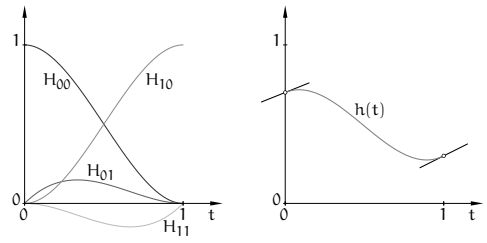
$$H_{00}(t) = 2t^3 - 3t^2 + 1, \qquad H_{10}(t) = -2t^3 + 3t^2,$$
$$H_{01}(t) = t^3 - 2t^2 + t, \qquad H_{11}(t) = t^3 - t^2.$$

Wielomiany te są rozwiązaniami zadań interpolacyjnych Hermite'a dla dwóch dwukrotnych węzłów, 0 i 1. Mianowicie, zachodzą równości

$$H_{00}(0) = 1, \quad H_{00}(1) = H'_{00}(0) = H'_{00}(1) = 0,$$
$$H_{10}(1) = 1, \quad H_{10}(0) = H'_{10}(0) = H'_{10}(1) = 0,$$
$$H'_{01}(0) = 1, \quad H_{01}(0) = H_{01}(1) = H'_{01}(1) = 0,$$
$$H'_{11}(1) = 1, \quad H_{11}(0) = H_{11}(1) = H'_{11}(0) = 0.$$

Dzięki temu rozwiązanie zadania interpolacyjnego Hermite'a z tymi węzłami dla dowolnej funkcji f możemy zapisać wzorem

$$h(t) = f(0)H_{00}(t) + f'(0)H_{01}(t) + f(1)H_{10}(t) + f'(1)H_{11}(t).$$



Przez zamianę zmiennej możemy znaleźć rozwiązanie zadania interpolacyjnego dla dowolnych dwóch węzłów interpolacyjnych o krotności 2. Jeśli węzłami tymi są liczby u_i i u_{i+1}, i oznaczymy $h_i = u_{i+1} - u_i$ (zakładamy, że $h_i > 0$), to mamy stąd wzór

$$h(x) = f(u_i)H_{i,00}(x) + f'(u_i)H_{i,01}(x) + f(u_{i+1})H_{i,10}(x) + f'(u_{i+1})H_{i,11}(x),$$

w którym użyliśmy funkcji

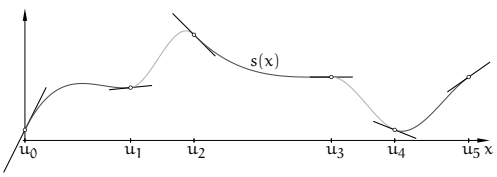
$$H_{i,00}(x) = H_{00}(t), \qquad H_{i,01}(x) = h_i H_{01}(t),$$
$$H_{i,10}(x) = H_{10}(t), \qquad H_{i,11}(x) = h_i H_{11}(t),$$

przy czym $t = (x - u_i)/h_i$.

Dla ustalonych węzłów $u_0, \dots, u_N$ (tworzących ciąg rosnący), mając dowolne liczby $a_0, \dots, a_N$ oraz $b_0, \dots, b_N$, możemy określić funkcję kawałkami wielomianową w przedziale $[u_0, u_N]$ wzorem

$$s(x) = p_i(x) = a_i H_{i,00}(x) + b_i H_{i,01}(x) + a_{i+1} H_{i,10}(x) + b_{i+1} H_{i,11}(x) \qquad \text{dla } x \in [u_i, u_{i+1}].$$

Jest oczywiste, że funkcja ta jest klasy $C^1[u_0, u_N]$, i w węźle u_i ma wartość a_i, a jej pochodna w u_i jest równa b_i, dla $i = 0, \dots, N$.



Funkcja s jest funkcją sklejaną, ale w istocie opartą na ciągu węzłów, z których każdy należy liczyć dwukrotnie. Funkcja ta jest (sklejanym) rozwiązaniem zadania interpolacyjnego Hermite'a, w którym dla każdego węzła zadajemy *dwa* warunki interpolacyjne: wartość funkcji i pochodnej. Funkcja ta jest skonstruowana za pomocą wielomianów spełniających warunki interpolacyjne Hermite'a dla dwóch węzłów o krotności 2 (końców każdego przedziału (u_i, u_{i+1})) i dlatego jej reprezentacja w tej postaci jest nazywana reprezentacją Hermite'a.

Kubiczne interpolacyjne funkcje sklejane

Aby określić funkcję sklejaną trzeciego stopnia, która jest rozwiązaniem zadania interpolacyjnego Lagrange'a (tj. dla każdego węzła u_i chcemy podać tylko jedną liczbę, a_i, będącą wartością funkcji), skorzystamy z warunku ciągłości pochodnej drugiego rzędu. Zatem, pochodne drugiego rzędu wielomianów $p_{i-1}(x)$ i $p_i(x)$, opisujących poszukiwaną funkcję s w przedziałach (u_{i-1}, u_i) oraz (u_i, u_{i+1}), mają przyjmować w punkcie u_i tę samą wartość (reprezentacja Hermite'a gwarantuje, że będą miały w tym punkcie identyczne wartości i pochodne pierwszego rzędu). Na tej podstawie możemy obliczyć liczby b_i, tj. wartości pochodnej pierwszego rzędu funkcji s w węzłach interpolacyjnych.

Aby wyprowadzić odpowiednie równania, należy obliczyć pochodne wielomianów, za pomocą których przedstawiamy rozwiązanie:

$$H''_{i-1,00}(u_i) = \frac{6}{h_{i-1}^2}, \quad H''_{i,00}(u_i) = \frac{-6}{h_i^2},$$
$$H''_{i-1,10}(u_i) = \frac{-6}{h_{i-1}^2}, \quad H''_{i,10}(u_i) = \frac{6}{h_i^2},$$
$$H''_{i-1,01}(u_i) = \frac{2}{h_{i-1}}, \quad H''_{i,01}(u_i) = \frac{-4}{h_i},$$
$$H''_{i-1,11}(u_i) = \frac{4}{h_{i-1}}, \quad H''_{i,11}(u_i) = \frac{-2}{h_i}.$$

Zatem, warunek ciągłości drugiej pochodnej w punkcie u_i, $p''_{i-1}(u_i) = p''_i(u_i)$, ma postać

$$\frac{6}{h_{i-1}^2}a_{i-1} - \frac{6}{h_{i-1}^2}a_i + \frac{2}{h_{i-1}}b_{i-1} + \frac{4}{h_{i-1}}b_i = -\frac{6}{h_i^2}a_i + \frac{6}{h_i^2}a_{i+1} - \frac{4}{h_i}b_i - \frac{2}{h_i}b_{i+1}.$$

Po pomnożeniu stron przez $h_{i-1}h_i/2$ i uporządkowaniu, dostajemy równanie

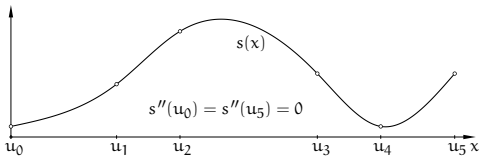
$$h_i b_{i-1} + 2(h_{i-1} + h_i)b_i + h_{i-1}b_{i+1} = 3\left(\frac{h_i}{h_{i-1}}(a_i - a_{i-1}) + \frac{h_{i-1}}{h_i}(a_{i+1} - a_i)\right). \quad (*)$$

W ten sposób otrzymaliśmy równania ciągłości pochodnych drugiego rzędu funkcji s w „wewnętrznych" węzłach $u_1, \ldots, u_{N-1}$; dla ustalonych liczb $a_0, \ldots, a_N$ musimy znaleźć liczby $b_0, \ldots, b_N$ spełniające te równania. Zauważamy, że liczba niewiadomych jest o 2 większa niż liczba równań. Aby mieć rozwiązanie jednoznaczne, trzeba dołożyć dodatkowe dwa równania.

Dodatkowe równania w konstrukcji sklejanych funkcji interpolacyjnych opisują zwykle tzw. warunki brzegowe, tj. pewne warunki narzucone na pochodne funkcji s w skrajnych węzłach u_0 i u_N. Najprostszy sposób, to arbitralne określenie wartości pochodnych pierwszego rzędu, tj. liczb b_0 i b_N. To jednak może być kłopotliwe dla użytkownika programu. Często stosowanym rozwiązaniem jest żądanie, aby pochodna drugiego rzędu funkcji s była w punktach u_0 i u_N równa 0. Powstaje wtedy tzw. naturalna funkcja sklejana. Na podstawie wypisanych wcześniej wartości funkcji $H_{i,jk}$, równania $p''_0(u_0) = 0$ i $p''_{N-1}(u_N) = 0$ możemy przedstawić w postaci

$$2h_0b_0 + h_0b_1 = 3(a_1 - a_0),$$
$$h_{N-1}b_{N-1} + 2h_{N-1}b_N = 3(a_N - a_{N-1}).$$

Po dołączeniu tych równań otrzymujemy układ równań liniowych z macierzą trójdiagonalną. Możemy zauważyć, że dla dowolnego rozmieszczenia węzłów, tj. dla dowolnych dodatnich liczb $h_0, \ldots, h_{N-1}$, macierz ta jest diagonalnie dominująca. Zatem, mamy układ o jednoznacznym rozwiązaniu, które możemy znaleźć za pomocą eliminacji Gaussa kosztem $O(N)$ działań arytmetycznych.



Istnieje wiele innych sposobów określania warunków brzegowych, np. można zażądać, aby pochodna trzeciego rzędu wielomianów p_0 i p_{N-1} była równa 0 (zatem, aby były to wielomiany drugiego stopnia), lub też, aby wielomian p_0 był identyczny z p_1, a wielomian p_{N-1} z p_{N-2} (a więc, aby węzły interpolacyjne u_1 i u_{N-1} *nie były* węzłami funkcji sklejanej — po angielsku nazywa się to warunek *not a knot*).

Jeszcze inna możliwość, to przyjęcie $b_N = b_0$ i wymaganie, aby pochodna drugiego rzędu funkcji s w węzłach u_0 i u_N była taka sama. Wtedy, jeśli $a_0 = a_N$, tj. funkcja s ma tę samą wartość w punktach u_0 i u_N, skonstruujemy tzw. okresową funkcję sklejaną — nakładając warunek, że dla każdego $x \in \mathbb{R}$ $s(x + T) = s(x)$, gdzie $T = u_N - u_0$, otrzymujemy okresową funkcję klasy $C^2(\mathbb{R})$. W układzie równań (*) zastępujemy niewiadomą b_N przez b_0 i dołączamy równanie

$$h_0b_{N-1} + 2(h_{N-1} + h_0)b_0 + h_{N-1}b_1 = 3\left(\frac{h_0}{h_{N-1}}(a_0 - a_{N-1}) + \frac{h_{N-1}}{h_0}(a_1 - a_0)\right),$$

otrzymując w ten sposób układ równań z macierzą cykliczną trójdiagonalną, diagonalnie dominującą.

Twierdzenie Holladaya

Dla dużej liczby węzłów wielomiany interpolacyjne Lagrange'a „źle się zachowują", tj. ich wartości między sąsiednimi węzłami mogą wystawać daleko poza przedział, którego końcami są wartości wielomianu w tych węzłach. Udowodnimy twierdzenie, które można zinterpretować w ten sposób, że pod tym względem najlepiej, jak tylko się da, zachowuje się naturalna kubiczna interpolacyjna funkcja sklejana. W tym celu określimy funkcjonał, który nazwiemy energią, i który przyjmiemy za miarę „powyginania" wykresów funkcji klasy $C^2[u_0, u_N]$:

$$E(f) \stackrel{\text{def}}{=} \int_{u_0}^{u_N} \big(f''(x)\big)^2\,dx.$$

Twierdzenie Holladaya. *W zbiorze funkcji klasy $C^2[u_0, u_N]$ spełniających warunki interpolacyjne Lagrange'a określone w węzłach $u_0, \ldots, u_N$ najmniejszą energię ma naturalna kubiczna funkcja sklejana.*

Dowód. Niech s oznacza naturalną kubiczną funkcję sklejaną spełniającą zadane warunki interpolacyjne, i niech f oznacza dowolną inną funkcję klasy C^2, przyjmującą w węzłach te same wartości. Mamy $f = (f - s) + s$, zatem

$$E(f) = \int_{u_0}^{u_N} \big(f''(x) - s''(x)\big)^2\,dx + 2\int_{u_0}^{u_N} \big(f''(x) - s''(x)\big)s''(x)\,dx + \int_{u_0}^{u_N} \big(s''(x)\big)^2\,dx.$$

Obliczymy drugą całkę w powyższym wzorze, całkując przez części:

$$\int_{u_0}^{u_N} \big(f''(x) - s''(x)\big)s''(x)\,dx = \big(f'(x) - s'(x)\big)s''(x)\Big|_{u_0}^{u_N} - \int_{u_0}^{u_N} \big(f'(x) - s'(x)\big)s'''(x)\,dx.$$

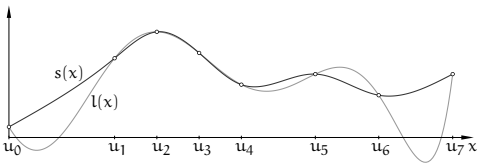
Dla naturalnej funkcji sklejanej s jest $s''(u_0) = s''(u_N) = 0$, ponadto w każdym przedziale (u_i, u_{i+1}) pochodna trzeciego rzędu funkcji s jest stała; oznaczmy ją symbolem s_i. Rozpatrywana całka jest więc równa

$$-\sum_{i=0}^{N-1} \int_{u_i}^{u_{i+1}} \big(f'(x) - s'(x)\big)s_i\,dx = -\sum_{i=0}^{N-1} s_i\big(f(x) - s(x)\big)\Big|_{u_i}^{u_{i+1}} = 0,$$

bo dla każdego i jest $f(u_i) = s(u_i)$. Mamy stąd

$$E(f) = \int_{u_0}^{u_N} \big(f''(x) - s''(x)\big)^2\,dx + \int_{u_0}^{u_N} \big(s''(x)\big)^2\,dx.$$

Jeśli funkcje f i s mają taką samą pochodną drugiego rzędu w przedziale $[u_0, u_N]$ i są różne, to ich różnica jest wielomianem stopnia 0 lub 1, ale wtedy nie mogą przyjmować tych samych wartości we wszystkich węzłach. Zatem, jeśli przyjmują i są różne, to zachodzi nierówność $E(f) > E(s)$, co pragnęliśmy udowodnić. □



Funkcje B-sklejane

Funkcje sklejane trzeciego stopnia są odpowiednie w większości zastosowań praktycznych, ale w pewnych przypadkach potrzebne są też funkcje sklejane innych stopni. Reprezentacja Hermite'a, opisana wcześniej, jest w miarę wygodna dla funkcji stopnia nieparzystego, ale jeszcze wygodniejsza w zastosowaniach, dla dowolnego stopnia, jest reprezentacja B-sklejana. Funkcję sklejaną stopnia n z węzłami $u_0, \ldots, u_N$ (ogólnie w literaturze jest wiele sposobów numerowania i oznaczania węzłów i funkcji B-sklejanych) reprezentuje się w postaci

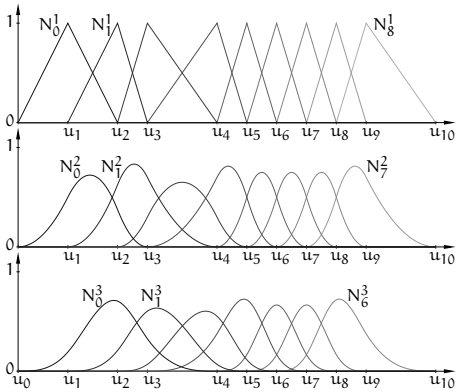
$$s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x),$$

tj. jako kombinację liniową tzw. funkcji B-sklejanych N_i^n, określonych wzorem Mansfielda-de Boora-Coxa:

$$N_i^0(x) = \begin{cases} 1 & \text{dla } x \in [u_i, u_{i+1}) \\ 0 & \text{w przeciwnym razie} \end{cases}$$

$$N_i^n(x) = \frac{x - u_i}{u_{i+n} - u_i} N_i^{n-1}(x) + \frac{u_{i+n+1} - x}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x).$$

Określenie funkcji B-sklejanych dopuszcza węzły o krotności większej niż 1, przy czym dla każdego i powinna zachodzić nierówność $u_i < u_{i+n+1}$, bo w przeciwnym razie funkcja N_i^n byłaby funkcją zerową.



Najważniejsze własności tych funkcji (bez dowodu):

- Funkcje N_i^n są nieujemne.
- Funkcja N_i^n jest różna od zera tylko w przedziale $[u_i, u_{i+n+1})$ i między kolejnymi węzłami w tym przedziale jest opisana za pomocą wielomianów stopnia n.
- Funkcja N_i^n ta ma tylko jedno maksimum (stąd nazwa B-sklejane, ang. *B-spline*, od kształtu wykresu, który trochę przypomina przekrój dzwonu).
- Suma funkcji stopnia n określonych za pomocą węzłów $u_0, \ldots, u_N$ w przedziale $[u_n, u_{N-n}]$ jest równa 1. Zwykle w zastosowaniach za dziedzinę funkcji s przyjmuje się ten przedział (ewentualnie domknięty, tj. $[u_n, u_{N-n}]$).

- Pochodna funkcji B-sklejanej stopnia n wyraża się wzorem
$$\frac{d}{dx}N_i^n(x) = \frac{n}{u_{i+n}-u_i}N_i^{n-1}(x) - \frac{n}{u_{i+n+1}-u_{i+1}}N_{i+1}^{n-1}(x).$$
- W otoczeniu dowolnego węzła o krotności r funkcje B-sklejane (a zatem i każda funkcja s, która jest ich kombinacją liniową) są ciągłe razem z pochodnymi rzędu $1,\dots,n-r$.
- Zachodzi równość
$$\int_{\mathbb{R}} N_i^n(x)\,dx = \int_{u_i}^{u_{i+n+1}} N_i^n(x)\,dx = \frac{1}{n+1}(u_{i+n+1}-u_i).$$

Zależnie od potrzeb, możemy dobrać stopień i węzły tak, aby otrzymać funkcje sklejane odpowiedniej do zastosowania klasy. Prawie zawsze wystarczy $n < 10$, najczęściej bierze się $n = 3$.

W przedziale $[u_k, u_{k+1}) \subset [u_n, u_{N-n})$ niezerowe wartości przyjmuje $n+1$ funkcji B-sklejanych stopnia n, mianowicie funkcje $N_{k-n}^n, \dots, N_k^n$. Wielomiany opisujące te funkcje w przedziale $[u_k, u_{k+1})$ są liniowo niezależne. Wartości tych wielomianów dla ustalonego x można obliczyć za pomocą algorytmu de Boora, który realizuje obliczenie na podstawie wzoru Mansfielda-de Boora-Coxa.

Przed wykonaniem podanej niżej procedury należy znaleźć przedział $[u_k, u_{k+1})$, do którego należy liczba x; można w tym celu zastosować wyszukiwanie binarne, lub, jeśli węzły $u_n, \dots, u_{N-n}$ są równoodległe, posłużyć się dzieleniem (tj. obliczyć $k = n + \lfloor (x-u_n)/(u_{n+1}-u_n) \rfloor$).


```
/* x ∈ [u_k, u_{k+1}) ⊂ [u_n, u_{N-n}) */
b[k] = 1;
for ( j = 1; j ⩽ n; j++ ) {
  β = (u_{k+1} − x)/(u_{k+1} − u_{k−j+1});
  b[k − j] = β ∗ b[k − j + 1];
  for ( i = k − j + 1; i < k; i++ ) {
    α = 1 − β;
    β = (u_{i+j+1} − x)/(u_{i+j+1} − u_{i+1});
    b[i] = α ∗ b[i] + β ∗ b[i + 1];
  }
  b[k] ∗= (1 − β);
}
/* b[i] = N_i^n(x) dla i = k − n, ..., k */
```


Niech $x \in [u_k, u_{k+1}) \subset [u_n, u_{N-n})$. Wtedy mamy
$$s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x) = \sum_{i=k-n}^{k} d_i N_i^n(x),$$
Jeśli $n > 0$, to
$$\begin{aligned} s(x) &= \sum_{i=k-n}^{k} d_i \Bigg(\underbrace{\frac{x-u_i}{u_{i+n}-u_i}}_{\alpha_i^{(1)}} N_i^{n-1}(x) + \underbrace{\frac{u_{i+n+1}-x}{u_{i+n+1}-u_{i+1}}}_{1-\alpha_{i+1}^{(1)}} N_{i+1}^{n-1}(x) \Bigg) \\ &= \sum_{i=k-n+1}^{k} \alpha_i^{(1)} d_i N_i^{n-1}(x) + \sum_{i=k-n}^{k-1} (1-\alpha_{i+1}^{(1)}) d_i N_{i+1}^{n-1}(x) \\ &= \sum_{i=k-n+1}^{k} \alpha_i^{(1)} d_i N_i^{n-1}(x) + \sum_{i=k-n+1}^{k} (1-\alpha_i^{(1)}) d_{i-1} N_i^{n-1}(x) \\ &= \sum_{i=k-n+1}^{k} \left(\alpha_i^{(1)} d_i + (1-\alpha_i^{(1)}) d_{i-1} \right) N_i^{n-1}(x). \end{aligned}$$

Oznaczmy $d_i^{(1)} = \alpha_i^{(1)} d_i + (1-\alpha_i^{(1)}) d_{i-1}$. Jeśli $n > 1$, to powyższe przekształcenie możemy zastosować rekurencyjnie, otrzymując na końcu
$$s(x) = \sum_{i=k}^{k} d_i^{(n)} N_i^0(x) = d_k^{(n)}.$$
Na tym rachunku opiera się algorytm de Boora obliczania wartości funkcji sklejanej danej za pomocą współczynników w bazie B-sklejanej stopnia n:

```
/* d_i^(0) = d_i dla i = k − n, ..., k, x ∈ [u_k, u_{k+1}) */
for ( j = 1; j ⩽ n j++ )
  for ( i = k − n + j; i ⩽ k; i++ ) {
    α = (x − u_i)/(u_{i+n+1−j} − u_i); /* α = α_i^(j) */
    d_i^(j) = (1 − α) ∗ d_{i−1}^(j−1) + α ∗ d_i^(j−1);
  }
/* d_k^(n) = s(x) */
```


Koszty obu algorytmów de Boora są rzędu n^2; dla funkcji niskich stopni, zazwyczaj stosowanych w praktyce, to są małe koszty. Ponadto można wykazać, że oba algorytmy mają bardzo dobre własności numeryczne (tj. niedokładności wyników będące skutkiem błędów zaokrągleń w implementacjach korzystających z arytmetyki zmiennopozycyjnej są małe). Jest to konsekwencją faktu, że dla $x \in [u_k, u_{k+1})$ wszystkie wartości przypisywane zmiennym α i β są liczbami z przedziału $[0, 1]$.

Kubiczne funkcje interpolacyjne w reprezentacji B-sklejanej

Konstrukcja B-sklejanej reprezentacji kubicznej funkcji interpolacyjnej, tj. obliczenie współczynników d_i, polega na rozwiązaniu układu równań liniowych. W tym przypadku przyjmujemy, że wartości a_i funkcji są zadane w węzłach $u_n, \dots, u_{N-n}$, tworzących ciąg rosnący. Ponieważ w każdym z tych węzłów tylko trzy funkcje B-sklejane są niezerowe, mamy równania
$$N_{i-3}^3(u_i) d_{i-3} + N_{i-2}^3(u_i) d_{i-2} + N_{i-1}^3(u_i) d_{i-1} = a_i.$$
Wartości funkcji B-sklejanych w węzłach możemy obliczyć za pomocą algorytmu de Boora. Do tych równań (dla $i = 3, \dots, N-3$) należy dołączyć jeszcze dwa równania opisujące warunki brzegowe (np. prowadzące do otrzymania naturalnej funkcji sklejanej s, tj. spełniającej warunek $s''(u_n) = s''(u_{N-n}) = 0$). Dla dowolnych sensownych warunków brzegowych równania można przekształcić tak, aby otrzymać układ równoważny z macierzą trójdiagonalną.

Można też skonstruować kubiczną sklejaną funkcję okresową klasy C^2. Ciąg węzłów musi być rosnący i jeśli $T = u_{N-3} - u_3$, to trzeba przyjąć $u_{N-6+i} = u_i + T$ dla $i = 1, \dots, 5$. Okresowa funkcja sklejana dla takiego ciągu ma współczynniki $d_0 = d_{N-6}$, $d_1 = d_{N-5}$, $d_2 = d_{N-4}$. Warunki interpolacyjne $s(u_i) = a_i$ zadajemy w węzłach $u_3, \dots, u_{N-3}$, przy czym $a_3 = a_{N-3}$. Na tej podstawie otrzymuje się układ $N-6$ równań liniowych, którego rozwiązaniem są współczynniki $d_0, \dots, d_{N-7}$ okresowej sklejanej funkcji interpolacyjnej, z macierzą cykliczną trójdiagonalną.

Twierdzenie Schoenberga-Whitney

Przypomnijmy, że słowo „węzły" było już używane w dwóch znaczeniach. Po pierwsze, w znaczeniu węzły interpolacyjne, czyli punkty, w których zadajemy wartości funkcji. Drugie znaczenie to węzły funkcji sklejanej, czyli punkty rozgraniczające przedziały, w których funkcja jest (a dokładniej, może być) opisana za pomocą różnych wielomianów. Do tej pory wybieraliśmy węzły interpolacyjne pokrywające się z węzłami funkcji sklejanej, ale możemy dopuścić inny ich wybór. Trzeba jednak wiedzieć, jaki wybór jest dopuszczalny, aby rozwiązanie zadania interpolacyjnego Lagrange'a istniało.

Oznaczmy symbolami $u_0, \ldots, u_N$ węzły funkcji sklejanej, a konkretniej niemalejący ciąg węzłów, których użyjemy do określenia funkcji B-sklejanych stopnia n. Liczba tych funkcji to $N - n$. Węzły interpolacyjne oznaczymy symbolami $v_0, \ldots, v_{N-n-1}$. Zatem, liczba warunków interpolacyjnych, które nakładamy, jest równa wymiarowi przestrzeni funkcji sklejanych rozpiętej przez nasze funkcje B-sklejane, dzięki czemu warunki brzegowe są zbędne. Założymy, że węzły interpolacyjne są ponumerowane tak, aby tworzyły ciąg rosnący (jest jasne, że węzły interpolacyjne muszą być parami różne).

Twierdzenie Schoenberga-Whitney. *Funkcja sklejana stopnia n, oparta na ciągu węzłów $u_0, \ldots, u_N$ i przyjmująca zadane wartości $a_0, \ldots, a_{N-n-1}$ odpowiednio w punktach $v_0, \ldots, v_{N-n-1}$, takich że $v_0 < v_1 < \cdots < v_{N-n-1}$, istnieje i jest jednoznacznie określona wtedy i tylko wtedy, gdy $N_i^n(v_i) \neq 0$ dla $i = 0, \ldots, N - n - 1$.*

Dowód tego twierdzenia jest żmudny i polega na wykazaniu że odpowiednia macierz Vandermonde'a (tj. macierz V o współczynnikach $a_{ij} = N_j^n(v_i)$) jest nieosobliwa wtedy i tylko wtedy, gdy podany w twierdzeniu warunek jest spełniony. Twierdzenie rozstrzyga problem z punktu widzenia algebry, ale nie gwarantuje, że macierz V jest dobrze uwarunkowana. Aby tak było, dla każdego $i \in \{0, \ldots, N - n - 1\}$ węzeł interpolacyjny v_i powinien leżeć w pobliżu punktu, w którym odpowiadająca mu funkcja B-sklejana N_i^n osiąga wartość maksymalną.

Przypuśćmy, że dane są węzły interpolacyjne, $v_0, \ldots, v_{N-n-1}$, ustawione w ciąg rosnący. Jeśli n jest nieparzyste, to dobrym (ale *nie jedynym dobrym*) wyborem jest przyjęcie węzłów funkcji sklejanej $u_0 = \cdots = u_n = v_0$, oraz $u_i = v_{i-(n+1)/2}$ dla $i = n + 1, \ldots, N - n - 1$, i $u_{N-n} = \cdots = u_N = v_{N-n-1}$.

Dla parzystego n można wybrać $u_0 = \cdots = u_n = v_0$, i $u_i = (v_{i-n/2-1} + v_{i-n/2})/2$ dla $i = n + 1, \ldots, N - n - 1$, oraz $u_{N-n} = \cdots = u_N = v_{N-n-1}$.

Wspomniana wyżej macierz V jest wstęgowa, a dokładniej, ma w każdym wierszu co najwyżej $n + 1$ niezerowych współczynników. Dzięki temu znalezienie sklejanej funkcji interpolacyjnej może być bardzo mało kosztowne (koszt eliminacji Gaussa w tym przypadku to $O(Nn^2)$ operacji).

9. Interpolacja trygonometryczna

Def. Wielomian trygonometryczny stopnia n jest to funkcja o postaci

$$w(t) = a_0 + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt). \qquad (*)$$

Wielomiany trygonometryczne występują w różnych zastosowaniach, zwłaszcza takich, w których pojawiają się funkcje okresowe. Często powstają z obcięcia szeregów Fouriera, tj. szeregów opisanych wzorem podobnym do wzoru (*), w którym zamiast n jest ∞.

Wzór (*) opisuje funkcję o okresie 2π. Aby otrzymać funkcję o dowolnym okresie T, można dokonać zamiany zmiennych:

$$f(x) = a_0 + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt),$$

biorąc $t = 2\pi(x - x_0)/T$, dla dowolnie wybranego x_0.

Trygonometryczne zadanie interpolacyjne Lagrange'a polega na znalezieniu wielomianu trygonometrycznego stopnia n, którego wartości $f_0, \ldots, f_{2n}$ są określone w $2n + 1$ węzłach interpolacyjnych, $x_0, \ldots, x_{2n} \in \mathbb{R}$.

Twierdzenie. *Warunek konieczny i dostateczny istnienia i jednoznaczności rozwiązania tego zadania dla dowolnych liczb $f_0, \ldots, f_{2n} \in \mathbb{R}$ jest $(x_j - x_k)/T \notin \mathbb{Z}$ dla $j \neq k$.*

Dowód. Z uwagi na to, że rozwiązanie jest funkcją okresową, konieczność tego warunku jest oczywista. To, że ten warunek jest dostateczny, wystarczy udowodnić dla przypadku szczególnego $T = 2\pi$.

Dla węzłów interpolacyjnych x_j i wartości funkcji f_j podanych dla tych węzłów, określamy liczby zespolone $z_j = e^{ix_j}$ oraz $h_j = z_j^n f_j$. Jeśli węzły $x_0, \ldots, x_{2n}$ spełniają rozważany warunek, to liczby z_j są parami różne. Jak wiemy, zadanie interpolacyjne Lagrange'a, tj. wyznaczenie wielomianu $h(z)$ stopnia co najwyżej $2n$, takiego że $h(z_j) = h_j$ dla $j = 0, \ldots, 2n$, ma rozwiązanie, $h(z) = \sum_{k=0}^{2n} c_{k-n} z^k$.

Zespolona funkcja wymierna

$$g(z) \stackrel{\text{def}}{=} \sum_{k=-n}^{n} c_k z^k$$

w węzłach z_j przyjmuje wartości f_j, i jest tylko jedna taka funkcja o tej postaci (bo liczby $c_{-n}, \ldots, c_n$ są określone jednoznacznie przez warunki interpolacyjne nałożone na wielomian h).

Niech $\hat{g}(z) \stackrel{\text{def}}{=} \overline{g(z)}$. Dla każdego $z \in \mathbb{C}$ takiego że $|z| = 1$, w tym dla każdego z_j, jest $\overline{z} = \frac{1}{z}$ i stąd

$$\hat{g}(z) = \overline{g(z)} = \sum_{k=-n}^{n} \overline{c}_k \overline{z}^k = \sum_{k=-n}^{n} \overline{c}_k z^{-k} = \sum_{k=-n}^{n} \overline{c}_{-k} z^k.$$

Ponieważ liczby f_j są rzeczywiste, zachodzą równości $\hat{g}(z_j) = f_j = g(z_j)$ dla każdego j. Spełniająca te warunki interpolacyjne funkcja $\hat{g}(z)$ też jest tylko jedna (tj. liczby $\overline{c}_{-n}, \ldots, \overline{c}_n$ są jednoznacznie określone przez te warunki).

Stąd wynika, że w zbiorze $\{ z \in \mathbb{C} \colon |z| = 1 \}$ funkcje $g(z)$ i $\hat{g}(z)$ są identyczne, a stąd wynika, że $c_{-k} = \overline{c}_k$ dla $k = -n, \ldots, n$.

Zatem, funkcja $w(t) = g(e^{it}) = \overline{g(e^{it})}$ ma dla każdego $t \in \mathbb{R}$ wartość rzeczywistą i spełnia warunki $w(x_j) = f_j$ dla $j = 0, \ldots, 2n$. Tak więc

$$\begin{aligned} w(t) = \sum_{k=-n}^{n} c_k e^{ikt} &= c_0 + \sum_{k=1}^{n} (c_k e^{ikt} + \overline{c}_k e^{-ikt}) = \\ c_0 + \sum_{k=1}^{n} \Big(c_k(\cos kt + i \sin kt) + \overline{c}_k(\cos kt - i \sin kt)\Big) &= \\ c_0 + \sum_{k=1}^{n} \Big((c_k + \overline{c}_k)\cos kt + i(c_k - \overline{c}_k)\sin kt\Big) &= \\ c_0 + \sum_{k=1}^{n} (2 \operatorname{Re} c_k \cos kt - 2 \operatorname{Im} c_k \sin kt). & \end{aligned}$$

Mamy stąd współczynniki wielomianu trygonometrycznego (*): $a_0 = \operatorname{Re} c_0$ (jest $\operatorname{Im} c_0 = 0$), oraz $a_k = 2 \operatorname{Re} c_k$ i $b_k = -2 \operatorname{Im} c_k$ dla $k > 0$. □

W praktycznych zastosowaniach najczęściej wybiera się węzły $x_0, \ldots, x_{2n}$, które dzielą przedział $[x_0, x_0 + T)$ (o długości okresu T interpolowanej funkcji) na części o jednakowych długościach. Zamiast bezpośrednio rozwiązywać zadania interpolacji trygonometrycznej, zwykle sprowadza się problem do konstrukcji zespolonego wielomianu algebraicznego, w sposób podobny do użytego w powyższym dowodzie.

Dyskretna transformata Fouriera

Def. Dyskretną transformatą Fouriera ciągu zespolonego $(a_k)_{k\in\mathbb{Z}}$ o okresie n (tj. spełniającego warunek $a_{k+n} = a_k$ dla każdego $k \in \mathbb{Z}$) jest ciąg zespolony $(b_j)_{j\in\mathbb{Z}}$ określony wzorem

$$b_j = \sum_{k=0}^{n-1} a_k e^{-2\pi ijk/n}.$$

Odwrotną dyskretną transformatą Fouriera ciągu $(a_k)_{k\in\mathbb{Z}}$ nazywamy ciąg $(c_j)_{j\in\mathbb{Z}}$ określony wzorem

$$c_j = \frac{1}{n} \sum_{k=0}^{n-1} a_k e^{2\pi ijk/n}.$$

Ciągi $(b_j)_{j\in\mathbb{Z}}$ i $(c_j)_{j\in\mathbb{Z}}$ są okresowe o okresie n. Oba przekształcenia zdefiniowane wyżej są liniowe i każde z nich jest jest odwrotnością tego drugiego, co uzasadnia nazwę. Mamy bowiem

$$d_l = \frac{1}{n} \sum_{j=0}^{n-1} \Big(\sum_{k=0}^{n-1} a_k e^{-2\pi ijk/n}\Big) e^{2\pi ilj/n} = \frac{1}{n} \sum_{k=0}^{n-1} a_k \sum_{j=0}^{n-1} e^{2\pi i(l-k)j/n}.$$

Jeśli $k = l$, to $e^{2\pi i(l-k)j/n} = e^0 = 1$, zaś jeśli $k \neq l$, to liczby $e^{2\pi i(l-k)j/n}$ są pierwiastkami zespolonymi z 1; ich suma dla $j \in \{0, \ldots, n-1\}$ jest równa 0. Stąd wynika, że $d_l = a_l$.

Interpolacja trygonometryczna i dyskretna transformata Fouriera występuje w wielu problemach związanych z analizą, transmisją i przetwarzaniem sygnałów (np. akustycznych lub obrazów), a także w rozwiązywaniu równań różniczkowych.

Algorytm FFT

Możemy zauważyć, że ciąg $(b_j)_{j\in\mathbb{Z}}$, który jest transformatą ciągu $(a_k)_{k\in\mathbb{Z}}$, składa się z wartości wielomianu stopnia $n-1$ o współczynnikach $a_0, \ldots, a_{n-1}$ w punktach $e^{-2\pi ij/n}$, $j = 0, \ldots, n-1$. Dyskretną transformatę Fouriera można wyznaczyć za pomocą schematu Hornera; wyznaczenie pełnej transformaty kosztowałoby wtedy $n^2 - n$ mnożeń i dodawań zespolonych. Okazuje się, że można to zadanie rozwiązać kosztem $\Theta(n(p_1 + \cdots + p_r))$ działań, gdzie $p_1, \ldots, p_r$ są liczbami pierwszymi, takimi że $n = p_1 \cdot \ldots \cdot p_r$. Odkrycia tego dokonali w 1952 r. Cooley i Tukey.

Zauważmy, że ciąg okresowy $(a_k)_{k\in\mathbb{Z}}$ o okresie 1 jest ciągiem stałym i jest on identyczny ze swoją dyskretną transformatą Fouriera. Dalej, przypuśćmy, że liczba n jest podzielna przez $p > 1$. Oznaczmy $w_j = e^{-2\pi ij/n}$. Wtedy wzór definiujący dyskretną transformatę Fouriera można przedstawić w postaci

$$\begin{aligned} b_j = &\sum_{k=0}^{n/p-1} a_{pk} w_j^{pk} + w_j \sum_{k=0}^{n/p-1} a_{pk+1} w_j^{pk} + \cdots \\ &+ w_j^{p-1} \sum_{k=0}^{n/p-1} a_{pk+p-1} w_j^{pk}. \end{aligned}$$

Podzieliliśmy tu ciąg $p_0, \ldots, p_{n-1}$ na podciągi n/p-elementowe, wybierając do każdego z nich co p-ty element. Możemy dalej zauważyć, że sumy mnożone przez kolejne potęgi liczby w_j są wyrażeniami opisującymi transformaty tych podciągów, a dokładniej ich obustronnie nieskończonych rozszerzeń o okresie n/p.

Obliczenie dyskretnej transformaty Fouriera dla ciągu o okresie n może być zatem wykonane przez następujący algorytm rekurencyjny:

- Jeśli $n = 1$, to przyjmij $b_0 = a_0$ (dla $n = 1$ przekształceniu poddajemy ciąg stały, którego obrazem jest ten sam ciąg).
- Jeśli n jest liczbą pierwszą, to zastosuj wzór podany jako definicja dyskretnej transformaty Fouriera i użyj schematu Hornera.
- Jeśli $n > 1$ jest podzielne przez liczbę pierwszą $p < n$, to podziel ciąg na p podciągów (zgodnie z opisem wyżej), oblicz transformaty tych podciągów i „scal" je, stosując wzór podany wyżej i schemat Hornera.

Wzór opisujący transformatę odwrotną może być przekształcony podobnie; zamiast $w_j = (\cos\frac{2\pi j}{n}, -\sin\frac{2\pi j}{n})$ występuje w nim liczba $\overline{w_j} = (\cos\frac{2\pi j}{n}, \sin\frac{2\pi j}{n})$. Możemy zatem użyć takiego samego algorytmu, zostawiając mnożenie wyniku działania procedury rekurencyjnej przez czynnik $\frac{1}{n}$ na sam koniec. Koszt algorytmu w istotny sposób zależy od możliwości rozłożenia liczby n na czynniki.

Algorytm jest najbardziej efektywny, jeśli liczba n jest potęgą liczby 2 i często określenie FFT (od angielskiego *Fast Fourier Transform*) dotyczy takiego wariantu algorytmu. Zbadamy go dokładniej. Dla parzystej liczby n transformatę otrzymamy przez „scalenie" transformat dwóch podciągów, złożonych odpowiednio z elementów parzystych i nieparzystych ciągu danego. Oznaczymy je symbolami $(p_j)_{j\in\mathbb{Z}}$ i $(q_j)_{j\in\mathbb{Z}}$.

Transformaty te są ciągami obustronnie nieskończonymi, o okresie $n/2$, reprezentowanymi przez podciągi $p_0, \ldots, p_{n/2-1}$ i $q_0, \ldots, q_{n/2-1}$. Możemy napisać

$$b_j = \sum_{k=0}^{n/2-1} a_{2k} w_j^{2k} + w_j \sum_{k=0}^{n/2-1} a_{2k+1} w_j^{2k} = p_j + w_j q_j.$$

Podstawiając $j + n/2$ w miejsce j, i biorąc pod uwagę, że $w_{j+n/2} = e^{-2\pi i(j+n/2)/n} = e^{-2\pi ij/n} e^{-2\pi in/(2n)} = -w_j$ oraz $w_{j+n/2}^{2k} = w_j^{2k}$, dostajemy

$$\begin{aligned} b_{j+n/2} &= \sum_{k=0}^{n/2-1} a_{2k} w_{j+n/2}^{2k} + w_{j+n/2} \sum_{k=0}^{n/2-1} a_{2k+1} w_{j+n/2}^{2k} \\ &= p_j - w_j q_j. \end{aligned}$$

Implementacja algorytmu FFT w postaci procedury rekurencyjnej:

```
void rFFT ( int n, complex a[] )
{
  complex *p, *q, u, w, t; int j;

  if ( n > 1 ) {
    p = malloc ( n*sizeof(complex) );
    q = &p[n/2];
    for ( j = 0; j < n/2; j++ ) {
      p[j] = a[2*j];
      q[j] = a[2*j+1];
    }
    rFFT ( n/2, p );
    rFFT ( n/2, q );
```

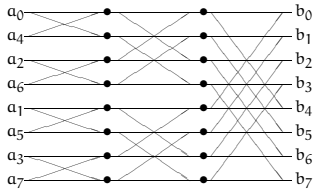


```
    u = 1;
    w = e^{-2πi/n};
    for ( j = 0; j < n/2; j++ ) {
      t = u*q[j];
      a[j] = p[j] + t;
      a[j+n/2] = p[j] - t;
      u = u*w;
    }
    free ( p );
  }
} /*rFFT*/
```

Widzimy, że choć procedura umieszcza transformatę w tej samej tablicy, w której są początkowo dane liczby $a_0, \ldots, a_{n-1}$, potrzebuje ona sporo pamięci dodatkowej (w rzeczywistości potrzeba dodatkowych tablic o sumarycznej długości $2n$).

Można zaprojektować taką implementację, która wszystkie obliczenia wykonuje „w miejscu”, tj. która oprócz tablicy z danym ciągiem, który należy zastąpić przez jego transformatę, potrzebuje tylko niewielkiej ustalonej liczby zmiennych prostych. Aby otrzymać taką procedurę, nierekurencyjną i dodatkowo oszczędzającą pewne działania, przyjrzymy się „przepływowi danych”, to znaczy zbadamy, od których współczynników zależą transformaty obliczane „po drodze”.

Dla $n = 8$ „przepływ danych” jest przedstawiony na rysunku (najlepiej go oglądać od prawej do lewej strony).



Krawędzie łączą dane z wynikami, tj. każda liczba (z wyjątkiem danych) jest obliczana na podstawie liczb znajdujących się w kolumnie na lewo od niej, połączonych z nią kreskami. Widzimy, że w każdym przypadku obliczenie polega na zastąpieniu pary liczb przez inną parę, obliczoną tylko na jej podstawie (i dana para liczb nie jest do niczego innego potrzebna). Jeśli zatem ustawimy dane wejściowe w odpowiedniej kolejności, to można całe obliczenie wykonać bez potrzeby rezerwowania dodatkowej tablicy.

Ostatnia transformata powstaje z transformat podciągów „parzystego” i „nieparzystego”. Każda z tych dwóch transformat jest obliczana na podstawie transformat „parzystego” i „nieparzystego” podciągu odpowiedniego podciągu itd.; zatem ogólna reguła porządkowania danych wejściowych polega na ustawieniu ich *w kolejności odwróconych bitów.* Jeśli indeks j danego współczynnika a_j przedstawimy w układzie dwójkowym, przy użyciu $l = \log_2 n$ cyfr dwójkowych (bitów), to indeks miejsca w tablicy, na którym ma się on znaleźć, otrzymamy wypisując te bity w odwrotnej kolejności.

Procedura FFT, która realizuje to obliczenie, ma postać

```
void FFT ( int n, complex a[] )
{
  complex t, u, w;
  int i, j, k, l, m, p;

  l = log2n; m = n/2;
        /* przestawianie danych w tablicy */
  for ( i = 1, j = m; i < n-1; i++ ) {
    if ( i < j ) przestaw ( &a[i], &a[j] );
    k = m;
    while ( k <= j ) { j -= k; k /= 2; }
    j += k;
  }
```



```
        /* obliczanie transformaty */
  for ( k = 1; k <= l; k++ ) {
    m = 2^k; p = m/2;
    u = 1; w = e^{-πi/p};
    for ( j = 0; j < p; j++ ) {
      i = j;
      do {
        t = a[i+p]*u;
        a[i+p] = a[i]-t; a[i] = a[i]+t;
        i += m;
      } while ( i <= n );
      u *= w;
    }
  }
} /*FFT*/
```

Algorytm ten opublikowali w 1965 r. Cooley, Lewis i Welch. Pierwsza pętla, `for ( i = ... ) ...`, dokonuje przestawienia elementów w tablicy zgodnie z kolejnością odwróconych bitów. Kolejne przebiegi drugiej pętli, `for ( k = ... ) ...`, mają na celu obliczenie $n/2$ transformat podciągów o okresie 2, $n/4$ transformat podciągów o okresie 4, itd. Liczba $e^{-2\pi i/n}$ (wartość zmiennej `w`) i jej potęgi, czyli liczby w_j (kolejne wartości zmiennej `u`) są obliczane tylko raz dla wszystkich transformat podciągów o tym samym okresie. W każdym przebiegu pętli `for ( j = ... ) ...` obliczane są pary współczynników o numerach j oraz $j+p$ we wszystkich transformatach podciągów o okresie $m = 2p$, ponieważ pętla najbardziej wewnętrzna (`do...while`) przebiega przez wszystkie te transformaty.

Można udowodnić, że algorytm FFT, także w wersji ogólnej (dla dowolnego n), jest numerycznie stabilny, tj. istnieje stała K (zależna od n), taka że współczynniki $\tilde{b}_j$ obliczone przy użyciu arytmetyki zmiennopozycyjnej przybliżają dokładne współczynniki b_j dyskretnej transformaty Fouriera z błędem spełniającym nierówność

$$\max_j |\tilde{b}_j - b_j| \leqslant K\nu \max_j |b_j|,$$

gdzie $\nu = 2^{-t}$. Dla liczby n będącej potęgą 2 można przyjąć

$$K = \left(\sqrt{2}\,\log_2 n + (\log_2 n - 1)(3 + 2\varepsilon)\right)\sqrt{n},$$

gdzie ε jest oszacowaniem błędu bezwzględnego obliczonych kosinusów i sinusów, tj. części rzeczywistych i urojonych liczb w_j.

Szybkie mnożenie wielomianów

Zajmiemy się następującym zadaniem: dane są współczynniki $a_0, \ldots, a_n$ i $b_0, \ldots, b_m$ wielomianów $a(x) = \sum_{k=0}^{n} a_k x^k$ i $b(x) = \sum_{k=0}^{m} b_k x^k$. Należy obliczyć współczynniki $c_0, \ldots, c_{n+m}$ wielomianu $c(x) = \sum_{k=0}^{n+m} c_k x^k = a(x)b(x)$. „Zwykły” algorytm mnożenia wielomianów można zrealizować za pomocą podprogramu

```
for ( k = 0; k <= n+m; k++ ) c[k] = 0;
for ( i = 0; i <= n; i++ )
  for ( j = 0; j <= m; j++ ) c[i+j] += a[i]*b[j];
```

Operacją dominującą w tym algorytmie jest mnożenie współczynników; operacji tych należy wykonać $(n+1)(m+1)$; jeśli $m \approx n$, to złożoność obliczeniowa ma rząd $\Theta(n^2)$, choć zarówno danych, jak i wyników jest $\Theta(n)$.

Alternatywny sposób rozwiązywania tego zadania polega na wybraniu liczb $x_0, \ldots, x_{n+m}$, obliczeniu wartości wielomianów a i b, obliczeniu wartości $c(x_j) = a(x_j)b(x_j)$ wielomianu c i znalezieniu jego współczynników w bazie potęgowej, przez rozwiązanie zadania interpolacyjnego Lagrange'a. Mnożenie wielomianów — w postaci mnożenia ich wartości w wybranych punktach — wymaga wykonania tylko $n + m + 1$ mnożeń. Trzeba tylko umieć szybko obliczyć wartości wielomianów a i b i szybko rozwiązać zadanie interpolacyjne.

Do tego celu możemy użyć algorytmu FFT; jeśli przyjmiemy, że $x_j = e^{-2\pi ij/N}$, gdzie liczba N jest najmniejszą całkowitą potęgą liczby 2 większą niż $n + m$, to ciąg wartości wielomianu a w tych punktach jest dyskretną transformatą Fouriera ciągu współczynników $a_0, \ldots, a_n, 0, \ldots, 0$ o długości (a raczej okresie) N. Mając wartości wielomianu c w punktach x_j, możemy obliczyć jego współczynniki w bazie potęgowej, wyznaczając odwrotną dyskretną transformatę Fouriera. Całe to obliczenie jest wykonalne za pomocą $\Theta(N \log N) = \Theta((n+m)\log(n+m))$ działań zmiennopozycyjnych.

10. Aproksymacja funkcji

Niech f oznacza funkcję określoną w przedziale $[a, b]$. Definicja tej funkcji może nie być wygodnym algorytmem obliczania wartości tej funkcji (np. funkcja f może być granicą nieskończonego ciągu), ewentualnie możemy mieć tylko „czarną skrzynkę” w postaci podprogramu obliczającego wartość funkcji f, przy czym koszt „sięgnięcia do tej skrzynki” może być bardzo duży, jeśli na przykład obliczenie wartości funkcji f w punkcie x polega na przeprowadzeniu eksperymentu fizycznego z parametrem x i dokonaniu pomiaru.

Zadanie aproksymacji polega na znalezieniu w ustalonej przestrzeni liniowej V, której elementy są funkcjami określonymi na przedziale $[a, b]$, funkcji g przybliżającej funkcję f (która w ogólności *nie jest* elementem przestrzeni V). Oczywiście, przestrzeń V wybieramy tak, aby koszt obliczania wartości należących do niej funkcji był mały, bo w zamierzeniu będziemy wielokrotnie obliczać wartości funkcji g, której chcemy używać zamiast f w jakimś celu.

Aby funkcja g mogła skutecznie „udawać” funkcję f, *musi* być skonstruowana w oparciu o dodatkową wiedzę na temat własności funkcji f i na temat *zamierzonej jakości aproksymacji*. Jeśli na przykład funkcja f ma ciągłą pochodną, to możemy chcieć, aby nie tylko wartości funkcji g były bliskie wartości funkcji f, ale także aby pochodna funkcji g przybliżała pochodną funkcji f (samo przybliżanie wartości funkcji f tego *nie zapewnia*). Przyjmiemy, że funkcja f jest elementem pewnej przestrzeni liniowej U, na przykład przestrzeni funkcji klasy $C^k[a, b]$ dla pewnego k. Będziemy rozpatrywać algorytmy dobierania funkcji z przestrzeni $V \subset U$. Algorytm będzie skuteczny, jeśli konkretna funkcja f, której przybliżenie chcemy znaleźć, jest istotnie elementem przestrzeni U (uwaga: to jest nietrywialne, jeśli funkcję f znamy tylko na podstawie pomiarów).

Błąd aproksymacji będziemy mierzyć za pomocą pewnej normy określonej w przestrzeni U. Zwykle normy określa się za pomocą całek i bardzo często bierze się normy Höldera; wtedy miarą błędu przybliżenia funkcji f przez g jest wyrażenie (dla ustalonego $p \geqslant 1$)

$$\|f-g\|_p = \left(\int_a^b |f(x)-g(x)|^p\,dx\right)^{1/p}.$$

Dwa szczególnie ważne przypadki to $p=2$ (mówimy wtedy o aproksymacji średniokwadratowej) oraz przypadek graniczny dla $p \to \infty$, gdy błąd jest określony wzorem

$$\|f-g\|_\infty = \max_{x\in[a,b]} |f(x)-g(x)|.$$

Ten przypadek nazywa się aproksymacją jednostajną.

Gdybyśmy byli zainteresowani także przybliżaniem pochodnej funkcji f, to mierzylibyśmy błąd innymi sposobami, np. obliczając wyrażenie

$$\max\{\|f-g\|_\infty, c\|f'-g'\|_\infty\},$$

z jakoś wybraną zawczasu stałą $c>0$.

Aproksymacja jednostajna

Z analizy znamy twierdzenie aproksymacyjne Weierstrassa: *Jeśli funkcja f jest ciągła na przedziale $[a,b]$, to dla każdego $\varepsilon>0$ istnieje wielomian p_n pewnego stopnia n, taki że $\|f-p_n\|_\infty \leqslant \varepsilon$.*

Twierdzenie Weierstrassa ma konstruktywny dowód (Bernstein, 1912 r.), ale konstrukcja użyta w tym dowodzie nie nadaje się do praktycznego stosowania, bo nawet dla „łatwych" funkcji i niezbyt małego ε wynikające z dowodu oszacowanie liczby n może być rzędu wielu tysięcy, podczas gdy wystarczy n mniejsze niż 10. Jedną z przyczyn tak słabych wyników konstrukcji jest to, że poza ciągłością o funkcji f niczego się nie zakłada.

Jeśli funkcja f jest klasy $C^{n+1}[a,b]$, to zadanie aproksymacji możemy rozwiązać przez skonstruowanie wielomianu interpolacyjnego Lagrange'a lub Hermite'a. W tym celu wybieramy węzły interpolacyjne $x_i \in [a,b]$ dla $i=0,\ldots,n$, obliczamy wartości funkcji f (i ewentualnie pochodnych, jeśli są krotne węzły) i stosujemy algorytm różnic dzielonych. Dla tak skonstruowanego wielomianu $h_n(x)$, na podstawie wzoru opisującego resztę, mamy

$$\|f-h_n\|_\infty = \max_{x\in[a,b]} \frac{|f^{(n+1)}(\xi(x))|}{(n+1)!}|p_{n+1}(x)|,$$

gdzie $p_{n+1}(x) = (x-x_0)\cdot\ldots\cdot(x-x_n)$. Mamy zatem problem, jak dobrać węzły, aby opisany powyższym wzorem błąd aproksymacji był jak najmniejszy.

Wielomiany i węzły Czebyszewa

Możemy ustalić $\varepsilon>0$, a następnie starać się dobrać węzły interpolacyjne w przedziale $[a,b]$ w dowolny sposób zapewniający, że błąd aproksymacji jest mniejszy niż ε. Jeśli się to uda, to nie przejmujemy się tym, że inny wybór mógłby dać jeszcze mniejszy błąd. Jeśli $\max_{x\in[a,b]} |f^{(n+1)}(x)| \leqslant M_{n+1}$, to

$$\|f-h_n\|_\infty \leqslant \frac{M_{n+1}}{(n+1)!}\|p_{n+1}\|_\infty. \qquad (*)$$

Możemy wybierać węzły tak, aby zminimalizować czynnik $\|p_{n+1}\|_\infty$. Aby to zrobić, zbadamy tzw. wielomiany Czebyszewa, zdefiniowane za pomocą wzorów

$$T_0(u) = 1,$$
$$T_1(u) = u,$$
$$T_k(u) = 2uT_{k-1}(u) - T_{k-2}(u) \quad \text{dla } k>1.$$

Jest jasne, że funkcja $T_k(u)$ jest wielomianem stopnia k. Wzór rekurencyjny dla $k>1$ to tak zwana formuła trójczłonowa, która umożliwia m.in. numeryczne obliczanie wartości tych wielomianów i ich kombinacji liniowych dla ustalonego u. Wielomiany Czebyszewa można określić także innymi sposobami, z których nam się przyda taki:

$$T_k(u) = \cos(k \arccos u) \quad \text{dla } u\in[-1,1].$$

Sprawdźmy, że to jest równoważna definicja: oznaczmy $u=\cos t$. Wtedy $T_0(u)=\cos 0 = 1$ oraz $T_1(u)=\cos t = u$, zaś dla $k>1$, podstawiając $\alpha = kt$ i $\beta=(k-2)t$ do tożsamości trygonometrycznej

$$\cos\alpha+\cos\beta = 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2},$$

otrzymujemy równość

$$\cos kt = 2\cos(k-1)t\cos t - \cos(k-2)t,$$

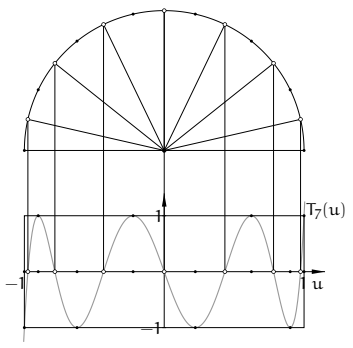
czyli formułę trójczłonową.

Na podstawie trygonometrycznego wzoru określającego wielomiany Czebyszewa możemy stwierdzić, że k miejsc zerowych wielomianu T_k (czyli wszystkie) znajduje się w przedziale $[-1,1]$, mianowicie są nimi liczby

$$z_j = \cos\frac{2j+1}{2k}\pi \quad \text{dla } j=0,\ldots,k-1,$$

a ponadto wielomian T_k w przedziale $[-1,1]$ przyjmuje wartości ekstremalne, na przemian $+1$ i -1, w punktach

$$y_j = \cos\frac{j}{k}\pi \quad \text{dla } j=0,\ldots,k.$$



Mając ustalony przedział $[a, b]$ oraz liczbę $k > 0$, możemy określić wielomian

$$q_k(x) = \frac{(b-a)^k}{2^{2k-1}} T_k(u),$$

gdzie $u = 2(x-a)/(b-a) - 1$, czyli $x = \frac{b+a}{2} + \frac{b-a}{2}u$. Z formuły trójczłonowej łatwo można wywnioskować, że wielomian $T_k(u)$ jest sumą wyrażenia $2^{k-1}u^k$ i pewnego wielomianu stopnia mniejszego niż k. Zatem współczynnik w bazie potęgowej przy x^k, czyli współczynnik wiodący wielomianu $q_k(x)$ jest równy 1. Wielomian q_k ma k miejsc zerowych w przedziale $[a, b]$ i w $k+1$ punktach tego przedziału, w tym w obu jego końcach, przyjmuje wartości ekstremalne, równe $\pm(b-a)^k/2^{2k-1}$.

Udowodnimy, że żaden wielomian stopnia k ze współczynnikiem wiodącym równym 1 nie może mieć mniejszych co do modułu wartości w całym przedziale $[a, b]$. Istotnie, gdyby taki wielomian, $w(x)$, istniał, to wielomian $r(x) = q_k(x) - w(x)$ miałby stopień mniejszy niż k, ale musiałby mieć co najmniej k miejsc zerowych w przedziale $[a, b]$, bo wykres wielomianu w przecinałby wykres wielomianu q_k co najmniej raz między każdymi jego sąsiednimi punktami ekstremalnymi (sąsiednie ekstrema mają tę samą wartość bezwzględną i przeciwne znaki, a wielomian w ma mieć w przedziale $[a, b]$ mniejsze wartości bezwzględne). Zatem, taki wielomian w nie istnieje. □

Dla dowolnych węzłów interpolacyjnych wielomian p_{n+1} występujący w oszacowaniu błędu ma współczynnik wiodący równy 1. Mamy zatem narzędzie do rozwiązywania zadania aproksymacji: aby przybliżyć funkcję klasy C^{n+1} w przedziale $[a, b]$, wybieramy tzw. węzły Czebyszewa, określone wzorem

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} \cos \frac{2j+1}{2n+2}\pi \quad \text{dla } j = 0, \ldots, n,$$

i konstruujemy wielomian interpolacyjny Lagrange'a h_n stopnia n z tymi węzłami. Wtedy otrzymamy $p_{n+1} = q_{n+1}$ i

$$\|f - h_n\|_\infty \leqslant \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Wyrażenie po prawej stronie tej nierówności możemy porównać z przyjętym progiem ε, aby sprawdzić, czy błąd jest dostatecznie mały. Jeśli nie, ale funkcja f ma ciągłe pochodne wyższych rzędów (i umiemy znaleźć ich oszacowania), to możemy spróbować szczęścia z wielomianem interpolacyjnym wyższego stopnia.

Alternans i algorytm Remeza

Teraz zajmiemy się następującym problemem: dla ustalonej funkcji rzeczywistej f należy dobrać taki wielomian $g^\star$ stopnia co najwyżej n, aby błąd aproksymacji w normie maksimum w przedziale $[a, b]$ był najmniejszy. Nieco uogólniając zadanie, rozważymy problem aproksymacji przez określone w przedziale $[a, b]$ funkcje, które są elementami ustalonej przestrzeni V o wymiarze k; zatem, mając taką przestrzeń, chcemy w niej znaleźć element najlepiej przybliżający daną funkcję f, o której założymy, że jest ciągła.

Def. Przestrzeń liniowa V o wymiarze k, której elementami są rzeczywiste funkcje ciągłe określone w przedziale $[a, b]$, spełnia warunek Haara (albo: ma własność Haara), jeśli z faktu, że funkcja $g \in V$ ma k różnych miejsc zerowych w przedziale $[a, b]$ wynika, że jest to funkcja zerowa.

Własność Haara, dla dowolnie wybranego przedziału $[a, b]$, ma zatem przestrzeń liniowa $\mathbb{R}[x]_n$, której elementy są wielomianami stopnia co najwyżej n, ale nie tylko: weźmy przestrzeń wielomianów trygonometrycznych stopnia co najwyżej n i ustalmy dowolny przedział $[a, b]$ krótszy niż 2π (tj. krótszy niż okres wszystkich tych funkcji). Przestrzeń ta ma wymiar $2n+1$, i jak wiemy, zadanie interpolacji Lagrange'a dla $2n+1$ dowolnie wybranych w przedziale $[a, b]$ (parami różnych) węzłów ma w tej przestrzeni jednoznaczne rozwiązanie. Jeśli więc pewien wielomian trygonometryczny stopnia n ma $2n+1$ miejsc zerowych w przedziale $[a, b]$, to jest on funkcją zerową. Natomiast *nie mają* własności Haara przestrzenie, których elementami są funkcje sklejane: istnieją niezerowe funkcje sklejane, które mają nieskończenie wiele miejsc zerowych.

Twierdzenie Czebyszewa o alternansie: *Jeśli przestrzeń V o wymiarze k spełnia warunek Haara, to dla dowolnej funkcji ciągłej f zadanie aproksymacji jednostajnej ma w przestrzeni V jednoznaczne rozwiązanie, $g^\star$. Funkcja $f - g^\star$, opisująca błąd aproksymacji, ma w przedziale $[a, b]$ co najmniej $k+1$ punktów, w których przyjmuje maksymalną wartość bezwzględną, przy czym znaki wartości funkcji $f - g^\star$ w kolejnych punktach z tego zbioru są przeciwne.*

Dowód twierdzenia Czebyszewa, który pominiemy, jest podobny do przeprowadzonego wcześniej dowodu stwierdzenia, że wielomian q_k ma najmniejszą normę $\|\cdot\|_\infty$ dla przedziału $[a, b]$ wśród wszystkich wielomianów stopnia k o współczynniku wiodącym 1 (i jest to jedyny taki wielomian).

Zbiór punktów, w których funkcja $f - g^\star$ przyjmuje na przemian minimalną i maksymalną wartość (wszystkie o tej samej wartości bezwzględnej $\|f - g^\star\|_\infty$) nazywany jest alternansem. Rozwiązanie zadania aproksymacji polega na znalezieniu takiego wielomianu $g^\star$ stopnia co najwyżej n, aby funkcja $f - g^\star$ przyjmowała w $n+2$ punktach przedziału $[a, b]$ wartości ekstremalne o zmieniających się znakach. Jeśli funkcja f jest wypukła albo wklęsła i poszukujemy optymalnego wielomianu stopnia 1, to alternans składa się z trzech punktów, z których dwa są końcami przedziału $[a, b]$, dzięki czemu zadanie jest dosyć łatwe.

Jeśli poszukujemy optymalnego wielomianu wyższego stopnia, to możemy użyć opisanego niżej algorytmu Remeza, w którym konstruuje się pewien ciąg wielomianów $(g^{(j)})_{j \in \mathbb{N}}$ stopnia n, zbieżny do poszukiwanego wielomianu $g^\star$.

Za $g^{(0)}$ można przyjąć wielomian interpolacyjny Lagrange'a z $n+1$ węzłami Czebyszewa w przedziale $[a, b]$. Istotne jest, aby funkcja $f - g^{(0)}$ miała w przedziale $[a, b]$ co najmniej $n+2$ lokalne minima i maksima, rozmieszczone na przemian (wartości bezwzględne tych ekstremów mogą być różne), i taki wybór funkcji $g^{(0)}$ to zapewnia: funkcja $f - g^{(0)}$ ma minimum lub maksimum między każdymi dwoma węzłami interpolacyjnymi, a także przed pierwszym i za ostatnim węzłem, a jeśli znaki kolejnych ekstremów są takie same (o co jest bardzo trudno), to zamiast jednego z nich można przyjąć węzeł między nimi.

Na podstawie wielomianu $g^{(j-1)}$ należy skonstruować $g^{(j)}$. W tym celu trzeba znaleźć *wszystkie* ekstrema funkcji $f - g^{(j-1)}$ w przedziale $[a, b]$. To może być bardzo trudnym zadaniem obliczeniowym. Mając pewne informacje o funkcji f, możemy ustalić gęstość, z jaką wystarczy stablicować tę funkcję i wielomian $g^{(j-1)}$ w przedziale $[a, b]$, aby nie „zgubić" żadnego ekstremum (to może być np. 100, 1000, lub nawet więcej punktów), potem trzeba zastosować jakąś metodę numeryczną znajdowania punktów ekstremalnych z dużą dokładnością. Następnie tworzymy j-te przybliżenie alternansu: wybieramy $n + 2$ punkty w przedziale $[a, b]$, w których funkcja $f - g^{(j-1)}$ przyjmuje wartości ekstremalne, przy czym jeśli lokalnych ekstremów jest więcej niż $n + 2$, to trzeba wybrać punkty, w których ekstrema mają największe wartości bezwzględne, z zachowaniem warunku zmieniających się znaków. Oznaczmy wybrane punkty symbolami $y_0^{(j)}, \ldots, y_{n+1}^{(j)}$ (lub lepiej w skrócie $y_0, \ldots, y_{n+1}$). Założymy, że są one uporządkowane monotonicznie.

Wielomian $g^{(j)}$ ma spełniać następujący warunek: dla $i = 0, \ldots, n + 1$ ma być $f(y_i) - g^{(j)}(y_i) = (-1)^i r_j$, gdzie r_j jest niewiadomą liczbą. Zatem, zachodzi równość $f(x) - g^{(j)}(x) = r_j h^{(j)}(x)$ dla pewnej funkcji $h^{(j)}$, takiej że $h^{(j)}(y_i) = (-1)^i$ dla $i = 0, \ldots, n + 1$. Obliczając różnicę dzieloną rzędu $n + 1$, otrzymamy

$$f[y_0, \ldots, y_{n+1}] = r_j h^{(j)}[y_0, \ldots, y_{n+1}],$$

bo różnica dzielona rzędu $n + 1$ wielomianu $g^{(j)}$ stopnia n jest zerem. Ale stąd możemy obliczyć

$$r_j = \frac{f[y_0, \ldots, y_{n+1}]}{h^{(j)}[y_0, \ldots, y_{n+1}]},$$

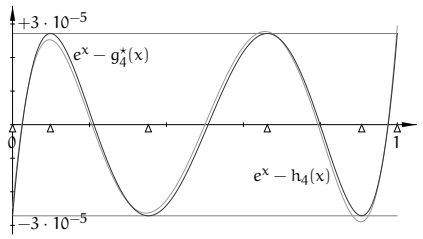
a następnie użyć r_j do obliczenia wartości wielomianu $g^{(j)}$ w punktach y_i i znaleźć ten wielomian przez rozwiązanie zadania interpolacyjnego Lagrange'a (mamy tu o 1 węzeł i warunek interpolacyjny za dużo, ale to nie szkodzi).

Ciąg wielomianów $g^{(j)}$ zwykle dość szybko zbiega do wielomianu $g^\star$, który przybliża funkcję f z najmniejszym błędem w przestrzeni $\mathbb{R}[x]_n$, przy czym ciąg liczb $|r_j|$ zbiega do normy błędu, tj. maksymalnej wartości bezwzględnej różnicy $f(x) - g^\star(x)$ w przedziale $[a, b]$.

Jak widać z opisu (który jest dosyć uproszczony), to jest kosztowny algorytm, którego stosowanie może być opłacalne tylko wtedy, gdy wartości wielomianu $g^\star$ mają być obliczane *bardzo wiele razy*.

Przykład 1. Przybliżamy funkcję $f(x) = e^x$ w przedziale $[0, 1]$. Symbolem h_n oznaczymy wielomian interpolacyjny stopnia n oparty na węzłach Czebyszewa, a symbolem $g_n^\star$ wielomian optymalny znaleziony przy użyciu algorytmu Remeza. W ostatniej kolumnie tabeli podana jest liczba wykonanych iteracji (w każdej iteracji zostało znalezione nowe przybliżenie alternansu); punktem początkowym w każdym przypadku był wielomian h_n.

n	$\|f - h_n\|_\infty$	$\|f - g_n^\star\|_\infty$	k
1	$1.24 \cdot 10^{-1}$	$1.06 \cdot 10^{-1}$	2
2	$9.87 \cdot 10^{-3}$	$8.76 \cdot 10^{-3}$	2
3	$6.00 \cdot 10^{-4}$	$5.45 \cdot 10^{-4}$	2
4	$2.95 \cdot 10^{-5}$	$2.72 \cdot 10^{-5}$	2
5	$1.21 \cdot 10^{-6}$	$1.13 \cdot 10^{-6}$	2
6	$4.28 \cdot 10^{-8}$	$4.03 \cdot 10^{-8}$	2



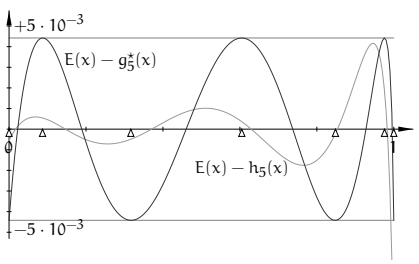
Przykład 2. W przedziale $[0, 1]$ przybliżamy funkcję

$$E(x) \stackrel{\text{def}}{=} \int_0^1 \sqrt{\frac{1 - x^2 t^2}{1 - t^2}}\, dt.$$

Jest to tak zwana zupełna całka eliptyczna drugiego rodzaju; występuje ona w różnych zastosowaniach (m.in. w mechanice). Funkcja ta maleje monotonicznie w przedziale $[0, 1]$, przyjmując na jego końcach wartości $E(0) = \frac{\pi}{2}$, $E(1) = 1$. Podobnie jak dla funkcji wykładniczej, *nie istnieje* wzór umożliwiający obliczanie $E(x)$ dla danego $x \in (0, 1)$ przy użyciu *skończenie wielu* działań arytmetycznych, co więcej, całek eliptycznych nie można wyrazić za pomocą funkcji wykładniczych i trygonometrycznych i ich odwrotności. Mając dane x, można konstruować rozmaite ciągi nieskończone, których granicą jest $E(x)$. W eksperymencie został użyty podprogram obliczający pewien wyraz takiego ciągu, przybliżający wartość funkcji E z błędem mniejszym niż 10^{-6}.

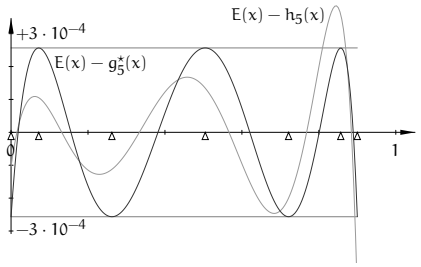
Możemy zauważyć, że funkcja E jest znacznie trudniejsza do aproksymacji — błędy przybliżeń znacznie wolniej maleją ze wzrostem stopnia niż w przypadku funkcji e^x. Powód jest taki, że funkcja E ma w przedziale $[0, 1)$ nieograniczoną pochodną; jest $\lim_{x \to 1} E'(x) = -\infty$.

n	$\|E - h_n\|_\infty$	$\|E - g_n^\star\|_\infty$	k
1	$1.54 \cdot 10^{-1}$	$9.49 \cdot 10^{-2}$	2
2	$5.43 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$	3
3	$3.00 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$	3
4	$1.88 \cdot 10^{-2}$	$6.81 \cdot 10^{-3}$	3
5	$1.30 \cdot 10^{-2}$	$4.42 \cdot 10^{-3}$	3
6	$9.48 \cdot 10^{-3}$	$3.09 \cdot 10^{-3}$	4



Wielomiany interpolacyjne funkcji E z węzłami Czebyszewa i wielomiany optymalne w przedziale $[0, 0.9]$ przybliżają funkcję E z błędami pokazanymi w następnej tabelce i (dla $n = 5$) na rysunku.

n	$\|E - h_n\|_\infty$	$\|E - g_n^\star\|_\infty$	k
1	$7.82 \cdot 10^{-2}$	$5.92 \cdot 10^{-2}$	2
2	$1.33 \cdot 10^{-2}$	$7.85 \cdot 10^{-3}$	3
3	$4.10 \cdot 10^{-3}$	$2.34 \cdot 10^{-3}$	3
4	$1.34 \cdot 10^{-3}$	$7.23 \cdot 10^{-4}$	3
5	$4.90 \cdot 10^{-4}$	$2.57 \cdot 10^{-4}$	3
6	$1.88 \cdot 10^{-4}$	$9.60 \cdot 10^{-5}$	3



Praktyczne wnioski z eksperymentów podobnych do powyższych dwóch są takie: jeśli funkcja, której przybliżenie wielomianowe należy skonstruować, ma pochodne wspólnie ograniczone, to zwykle nie sprawia kłopotów, ale zastąpienie wielomianu interpolacyjnego z węzłami Czebyszewa przez wynik działania algorytmu Remeza niewiele poprawia aproksymację; skuteczniejszym sposobem zmniejszenia błędu jest zwykle znalezienie wielomianu interpolacyjnego wyższego stopnia (z węzłami Czebyszewa). Jeśli funkcja ma w rozpatrywanym przedziale osobliwość, to żaden z tych sposobów nie jest dobry.

Dobrym sposobem na pokonanie trudności jest zwykle podzielenie przedziału na krótsze podprzedziały i poszukiwanie wielomianów aproksymacyjnych w tych podprzedziałach. Wynikiem takiego postępowania jest aproksymacyjna funkcja sklejana. Jeśli jednak w pewnym podprzedziale jest osobliwość (np. nieciągła pochodna), to warto się zastanowić nad innym sposobem przybliżania funkcji, niż za pomocą wielomianów. Jedna z możliwości to użycie funkcji wymiernych. Aby skutecznie aproksymować, *zawsze* należy wiedzieć, jakie (jakiego rodzaju) osobliwości ma dana funkcja.

Aproksymacja jednostajna przez funkcje sklejane

Zauważmy, że gdyby funkcja f miała być przybliżana w przedziale dwukrotnie krótszym (np. w połowie przedziału $[a, b]$), w którym przyjęlibyśmy węzły interpolacyjne rozmieszczone w dwukrotnie mniejszych odstępach, to czynnik $\|p_{n+1}\|_\infty$ we wzorze (*) dla tego krótszego przedziału byłby 2^{n+1} razy mniejszy. Zatem skrócenie przedziału $[a, b]$ jest radykalnym sposobem zmniejszenia błędu aproksymacji jednostajnej i czasami jest to jedyny skuteczny sposób. Mając długi przedział, możemy podzielić go na krótsze podprzedziały i aproksymować funkcję f w każdym z nich innym wielomianem niskiego stopnia. Odpowiednio dobrana reprezentacja aproksymacyjnej funkcji sklejanej może zapewnić dobre własności analityczne w danym zastosowaniu, tj. ciągłość pochodnych dostatecznie wysokiego rzędu.

Znanych jest wiele twierdzeń na temat aproksymacji funkcjami sklejanymi różnych stopni. Bez dowodu, który jest dosyć żmudny (choć pouczający), podam jedno z tych twierdzeń.

Twierdzenie. *Niech* $f \in C^2[a, b]$ *i niech* s *oznacza kubiczną funkcję sklejaną klasy* $C^2[a, b]$ *z węzłami* $u_0 = a < u_1 < \cdots < u_N = b$*, taką że* $s(u_i) = f(u_i)$ *dla* $i = 0, \ldots, N$*. Niech* M_2 *oznacza stałą taką że* $|f''(x)| \leqslant M_2$ *dla każdego* $x \in [a, b]$*, oraz* $|s''(a)| \leqslant 3M_2$ *i* $|s''(b)| \leqslant 3M_2$*. Wtedy dla każdego* $x \in [a, b]$ *zachodzą nierówności*

$$|f(x) - s(x)| \leqslant \frac{1}{2} M_2 h^2, \qquad |f'(x) - s'(x)| \leqslant 2M_2 h,$$

gdzie $h = \max_i (u_{i+1} - u_i)$.

Twierdzenie to stosuje się m.in. do naturalnych kubicznych funkcji sklejanych (dla których $s''(a) = s''(b) = 0$), ale nie tylko. Wynika z niego, że do osiągnięcia dowolnie małego błędu wystarczy wybranie dostatecznie gęstego ciągu węzłów w przedziale $[a, b]$, a ponadto można w ten sposób również dowolnie zmniejszyć błąd aproksymacji pochodnej funkcji f.

Aproksymacja średniokwadratowa

Niech ρ oznacza funkcję określoną w przedziale A (który może być otwarty lub domknięty, ograniczony lub nieograniczony). Zakładamy, że funkcja ρ jest nieujemna, zbiór jej miejsc zerowych w A jest miary zero (np. pusty) i całka z funkcji ρ w zbiorze A jest skończona. Funkcję ρ nazywamy funkcją wagową albo wagą. Dla takiej funkcji wzór

$$\langle f, g \rangle_\rho = \int_A f(x) g(x) \rho(x)\, dx$$

określa iloczyn skalarny, a funkcjonał

$$\|f\|_\rho = \sqrt{\langle f, f \rangle_\rho} = \sqrt{\int_A f(x)^2 \rho(x)\, dx}$$

jest normą. Zadanie aproksymacji średniokwadratowej często jest uogólniane w ten sposób, że dla danej funkcji f należy znaleźć w ustalonej przestrzeni V (której wymiar jest skończony) funkcję g, taką że wyrażenie $\|f - g\|_\rho$ jest najmniejsze.

Rozwiązaniem zadania jest wektor (funkcja) $g^\star$, która jest rzutem prostopadłym wektora (funkcji) f na przestrzeń V; zadanie aproksymacji średniokwadratowej jest w istocie uogólnieniem liniowego zadania najmniejszych kwadratów. Mając bazę $p_0, \ldots, p_n$ przestrzeni V, wystarczy znaleźć współczynniki $x_0, \ldots, x_n$ wektora $g^\star = \sum_{j=0}^{n} x_j p_j$ w tej bazie. Wektor $f - g^\star$ jest prostopadły do wszystkich elementów bazy przestrzeni V. Na podstawie tego warunku możemy wyprowadzić układ równań normalnych

$$\sum_{j=0}^{n} \langle p_i, p_j\rangle_\rho x_j = \langle p_i, f\rangle_\rho, \quad \text{dla } i = 0, \ldots, n.$$

Macierz $A = [\langle p_i, p_j\rangle_\rho]_{i,j}$ tego układu równań jest symetryczna i dodatnio określona.

Wielomiany ortogonalne

Zadanie aproksymacji średniokwadratowej jest znacznie łatwiejsze do rozwiązania, jeśli dysponujemy bazą ortogonalną przestrzeni V, tj. układem wektorów (funkcji) $p_0, \ldots, p_n$, takich że $\operatorname{lin}\{p_0, \ldots, p_n\} = V$ oraz $\langle p_i, p_j\rangle_\rho = 0$ dla $i \neq j$. Dla takiej bazy macierz układu równań normalnych jest diagonalna. Mając dowolną bazę przestrzeni V, możemy znaleźć bazę ortogonalną za pomocą ortogonalizacji Grama-Schmidta. Jeśli $V = \mathbb{R}[x]_n$, tj. elementami przestrzeni V są wszystkie wielomiany stopnia co najwyżej n, to za pomocą ortogonalizacji bazy potęgowej możemy znaleźć bazę ortogonalną $p_0, \ldots, p_n$, w której dla każdego k stopień wielomianu p_k jest równy k. Bazę taką możemy również znaleźć za pomocą odpowiedniej formuły trójczłonowej; wcześniej otrzymaliśmy tym sposobem wielomiany Czebyszewa.

Twierdzenie. *Dla ustalonego przedziału $A \subset \mathbb{R}$ i funkcji wagowej ρ wielomiany p_k, tworzące układ ortogonalny dla $k = 0, 1, \ldots$ i takie, że dla każdego k stopień wielomianu p_k jest równy k, wyrażają się wzorem*

$$p_k(x) = (\alpha_k x + \beta_k) p_{k-1}(x) + \gamma_k p_{k-2}(x) \quad \textit{dla } k = 1, 2, \ldots,$$

dla pewnych liczb $\alpha_k \neq 0$ (można je wybrać dowolnie), oraz

$$\beta_k = -\frac{\alpha_k \langle x p_{k-1}, p_{k-1}\rangle_\rho}{\|p_{k-1}\|_\rho^2}, \qquad \gamma_k = -\frac{\alpha_k \langle x p_{k-1}, p_{k-2}\rangle_\rho}{\|p_{k-2}\|_\rho^2},$$

przy czym $p_{-1}(x) \stackrel{\text{def}}{=} 0$, $p_0(x) \stackrel{\text{def}}{=} a_0 \neq 0$.

Dowód. Dla każdego k stopień wielomianu p_k jest równy k, zatem jego współczynnik wiodący $a_k \neq 0$. Niech $\alpha_k = a_k/a_{k-1}$. Niech $w_k = p_k - \alpha_k x p_{k-1}$. Wielomian w_k jest stopnia mniejszego niż k. Dla iloczynu skalarnego określonego za pomocą całki z wagą zachodzi równość $\langle xf, g\rangle_\rho = \langle f, xg\rangle_\rho$, zatem dla $j < k - 2$

$$\langle w_k, p_j\rangle_\rho = \langle p_k - \alpha_k x p_{k-1}, p_j\rangle_\rho = \langle p_k, p_j\rangle_\rho - \alpha_k \langle p_{k-1}, x p_j\rangle_\rho = 0.$$

Wyrażając wielomian w_k w bazie $p_0, \ldots, p_{k-1}$, otrzymamy

$$\langle w_k, p_j\rangle_\rho = \langle \sum_{i=0}^{k-1} b_{ki} p_i, p_j\rangle_\rho = \sum_{i=0}^{k-1} b_{ki} \langle p_i, p_j\rangle_\rho = b_{kj} \langle p_j, p_j\rangle_\rho.$$

Stąd $b_{kj} = 0$ dla $j < k - 2$, a zatem mamy

$$p_k = \alpha_k x p_{k-1} + \beta_k p_{k-1} + \gamma_k p_{k-2},$$

dla $\beta_k = b_{k,k-1}$, $\gamma_k = b_{k,k-2}$.

Możemy obliczyć

$$0 = \langle p_k, p_{k-1}\rangle_\rho = \alpha_k \langle x p_{k-1}, p_{k-1}\rangle_\rho + \beta_k \langle p_{k-1}, p_{k-1}\rangle_\rho + \gamma_k \underbrace{\langle p_{k-2}, p_{k-1}\rangle_\rho}_{=0},$$

$$0 = \langle p_k, p_{k-2}\rangle_\rho = \alpha_k \langle x p_{k-1}, p_{k-2}\rangle_\rho + \beta_k \underbrace{\langle p_{k-1}, p_{k-2}\rangle_\rho}_{=0} + \gamma_k \langle p_{k-2}, p_{k-2}\rangle_\rho$$

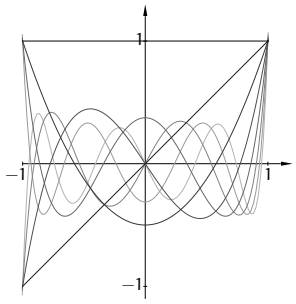
skąd otrzymujemy podane wyrażenia na β_k i γ_k. □

Oczywiście, można wybrać liczby a_0 i α_k tak, aby dostać bazę ortonormalną, ale nie zawsze się tak robi. Ważną własnością wielomianów ortogonalnych (dowód jest prostym ćwiczeniem) jest to, że wszystkie ich miejsca zerowe są rzeczywiste, jednokrotne i położone wewnątrz przedziału A.

Wielomiany ortogonalne są znane dla wielu różnych przedziałów i wag.

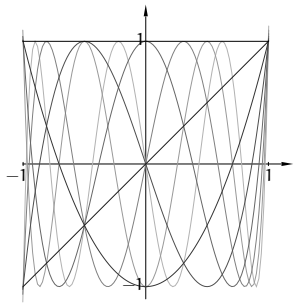
wielomiany Legendre'a: $A = (-1, 1)$, $\rho(x) = 1$,

$$P_0(x) = 1, \; P_1(x) = x, \; P_k(x) = \frac{2k-1}{k} x P_{k-1}(x) - \frac{k-1}{k} P_{k-2}(x),$$



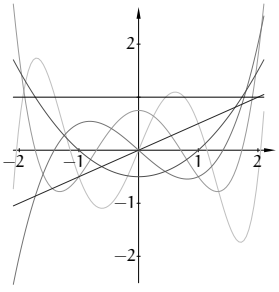
wielomiany Czebyszewa: $A = (-1, 1)$, $\rho(x) = (1 - x^2)^{-1/2}$,

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2x T_{k-1}(x) - T_{k-2}(x),$$



wielomiany Hermite'a: $A = \mathbb{R}$, $\rho(x) = e^{-x^2}$,

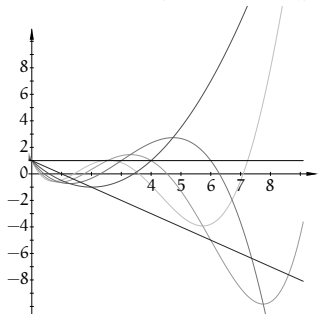
$$H_0(x) = 1, \; H_1(x) = x, \; H_k(x) = 2x H_{k-1}(x) - (2k-2) H_{k-2}(x),$$



(tu są wykresy funkcji $2^{-k} H_k$)

wielomiany Laguerre'a: $A = (0, +\infty)$, $\rho(x) = e^{-x}$, $L_0(x) = 1$,

$$L_1(x) = 1 - x, \quad L_k(x) = \frac{2k-1-x}{k} L_{k-1}(x) - \frac{k-1}{k} L_{k-2}(x).$$



Jeśli znamy bazę przestrzeni $\mathbb{R}[x]_n$ ortogonalną w sensie iloczynu skalarnego $\langle \cdot, \cdot \rangle_\rho$, to zadanie znalezienia wielomianu g_n^* stopnia co najwyżej n, najlepiej przybliżającego funkcję f, sprowadza się do obliczenia współczynników wielomianu g_n^* w tej bazie:

$$x_i = \frac{\langle f, p_i \rangle_\rho}{\|p_i\|_\rho^2}.$$

Mamy przy tym, na podstawie twierdzenia Pitagorasa,

$$\|f - g_n^*\|_\rho^2 = \|f\|_\rho^2 - \sum_{i=0}^{n} x_i^2 \|p_i\|_\rho^2.$$

Jeśli błąd jest za duży, możemy zwiększać n, obliczając tylko kolejne współczynniki x_i (ale uwaga: są takie funkcje f, dla których błąd nie maleje do zera, gdy $n \to \infty$ — trzeba uważać). Podstawą rozwiązywania zadań aproksymacji średniokwadratowej jest obliczanie całek, co można robić analitycznie (jeśli umiemy) lub numerycznie.

11. Numeryczne obliczanie całek

Def. Niech f oznacza pewną funkcję określoną w przedziale $[a, b]$. Kwadratura jest to kombinacja liniowa wartości funkcji f w pewnych punktach $x_i \in [a, b]$, zwanych węzłami kwadratury:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i).$$

Liczby A_i są nazywane współczynnikami kwadratury.
Ogólniejsza definicja określa kwadraturę jako kombinację liniową wartości funkcji f i jej pochodnych w węzłach kwadratury.

Kwadratura jest zatem funkcjonałem liniowym na przestrzeni funkcji określonych w przedziale $[a, b]$, podobnie jak całka oznaczona:

$$I(f) = \int_a^b f(x)\rho(x)\,dx.$$

W odróżnieniu od całki, mogąc obliczać wartości funkcji f w dowolnych punktach przedziału $[a, b]$, można obliczyć wartość kwadratury za pomocą skończenie wielu działań arytmetycznych. Numeryczne obliczanie całek polega na obliczaniu kwadratur. Ważne jest zapewnienie dostatecznej dokładności, tj. dostatecznie małego błędu aproksymacji całki przez kwadraturę. Temu celowi służy wybór węzłów i współczynników kwadratury. Jak zwykle, skuteczność wyboru zależy od własności funkcji, które mamy zamiar całkować.

Kwadratury interpolacyjne

Kwadratura interpolacyjna jest całką z wielomianu interpolacyjnego Lagrange'a lub Hermite'a funkcji f z węzłami w przedziale $[a, b]$. Jeśli jest to wielomian interpolacyjny Lagrange'a (tj. węzły są jednokrotne, obliczamy w nich tylko wartości funkcji f), to kwadratura ma współczynniki

$$A_i = \int_a^b \prod_{j \in \{0,\ldots,n-1\} \setminus \{i\}} \frac{x - x_j}{x_i - x_j} \rho(x)\,dx.$$

Wśród kwadratur interpolacyjnych wyróżniamy kwadratury Newtona-Cotesa, których węzły dzielą przedział $[a, b]$ na części o równych długościach (kwadratury te określa się z wagą $\rho(x) = 1$), kwadratury Gaussa, których węzły są miejscami zerowymi wielomianów ortogonalnych, a także inne kwadratury, dobierane specjalnie do zastosowań.

Błąd kwadratury jest to oczywiście różnica $I(f) - Q(f)$, która zależy od funkcji f. Błąd kwadratury interpolacyjnej opartej na n węzłach można oszacować, obliczając całkę z wyrażenia opisującego resztę interpolacji:

$$|I(f) - Q(f)| \leqslant \frac{M_n}{n!} \int_a^b |p_n(x)|\rho(x)\,dx,$$

ale to oszacowanie jest poprawne, jeśli funkcja f jest klasy $C^n[a, b]$, i możemy go użyć bezpośrednio, jeśli umiemy znaleźć stałą M_n, taką że $\|f^{(n)}\|_\infty \leqslant M_n$.

Def. Rząd kwadratury jest to liczba r, taka że kwadratura ma tę samą wartość co całka dla każdego wielomianu stopnia mniejszego niż r oraz inną wartość niż całka dla pewnego wielomianu stopnia r.

Z definicji kwadratury interpolacyjnej natychmiast wynika, że jej rząd jest nie mniejszy niż liczba węzłów. Rząd żadnej kwadratury opartej na n węzłach nie może być większy niż $2n$, ponieważ jeśli p_n jest wielomianem stopnia n, którego miejscami zerowymi są wszystkie węzły, to mamy $Q(p_n^2) = 0$ oraz $I(p_n^2) > 0$.

Możemy wybrać pewien ciąg kwadratur $Q_1, Q_2, \ldots$, np. kwadratur Newtona-Cotesa coraz wyższych rzędów, i zbadać zbieżność ciągu liczb $Q_1(f), Q_2(f), \ldots$ dla funkcji f spełniającej określone warunki (np. funkcji ciągłej). Chciałoby się, aby ten ciąg miał granicę, równą $I(f)$; jeśli ją ma, to istotna jest szybkość zbieżności do tej granicy.

Korzystając m.in. z twierdzenia Weierstrassa, można udowodnić

Twierdzenie. *Ciąg $Q_1(f), Q_2(f), \ldots$ jest zbieżny do granicy $I(f)$ dla dowolnej funkcji ciągłej f wtedy i tylko wtedy, gdy jest zbieżny dla każdego wielomianu i istnieje stała K, taka że suma wartości bezwzględnych współczynników każdej kwadratury w rozpatrywanym ciągu jest mniejsza niż K.*

Pierwszy warunek podany w twierdzeniu jest spełniony przez każdy ciąg kwadratur interpolacyjnych coraz wyższych rzędów, natomiast aby spełnić drugi warunek, wystarczy zapewnić, że współczynniki każdej kwadratury są nieujemne. Niestety, ciąg kwadratur Newtona-Cotesa tego warunku nie spełnia, co więcej, sumy wartości bezwzględnych współczynników tych kwadratur rosną nieograniczenie. Praktycznie użyteczne kwadratury Newtona-Cotesa mają tylko kilka (mniej niż 8) węzłów. Zbadamy dwie najprostsze z nich.

Kwadratura trapezów oparta jest na dwóch węzłach, będących końcami przedziału $[a, b]$:

$$T(f) = \frac{b-a}{2}\Big(f(a) + f(b)\Big).$$

Łatwo jest sprawdzić, że rząd tej kwadratury jest równy 2. Jeśli funkcja f jest klasy $C^2[a, b]$, to $p_2(x) = (x-a)(x-b)$ i mamy oszacowanie błędu

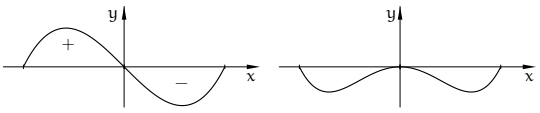
$$|I(f) - T(f)| \leqslant \frac{M_2}{2}\int_a^b |p_2(x)|\,dx = \frac{M_2}{12}(b-a)^3,$$

ze stałą M_2, taką że $|f''(x)| \leqslant M_2$ dla każdego $x \in [a, b]$.

Kwadratura Simpsona oparta jest na trzech węzłach: końcach i środku przedziału $[a, b]$; oznaczymy $c = (a+b)/2$:

$$S(f) = \frac{b-a}{6}\Big(f(a) + 4f(c) + f(b)\Big).$$

Okazuje się, że rząd kwadratury Simpsona jest równy 4. Jest tak dlatego, że to jest kwadratura interpolacyjna, której środkowy węzeł jest dwukrotny, ale współczynnik, przez który należałoby pomnożyć $f'(c)$, jest równy 0. Inne wyjaśnienie jest na rysunku.



Błąd kwadratury Simpsona możemy oszacować na dwa sposoby:

$$|I(f) - S(f)| \leqslant \frac{M_3}{6}\int_a^b |(x-a)(x-c)(x-b)|\,dx = \frac{M_3}{192}(b-a)^4,$$
$$|I(f) - S(f)| \leqslant \frac{M_4}{24}\int_a^b |(x-a)(x-c)^2(x-b)|\,dx = \frac{M_4}{2880}(b-a)^5,$$

gdzie M_3 i M_4 to oszacowania wartości bezwzględnych pochodnych trzeciego i czwartego rzędu funkcji f w przedziale $[a, b]$. Oczywiście, każdego z tych oszacowań możemy używać pod warunkiem, że odpowiednia pochodna funkcji f jest ciągła.

Zamiana zmiennych

Jeśli $f(u) = g(x)$ dla $x = su + t$, gdzie $s > 0$ i t są ustalonymi liczbami, oraz $c = sa + t$, $d = sb + t$, to

$$\int_a^b f(u)\rho(u)\,du = \frac{1}{s}\int_c^d g(x)\rho((x-t)/s)\,dx, \quad \text{oraz}$$
$$Q_1(f) = \sum_{i=0}^{n-1} A_i f(u_i) = \frac{1}{s}\sum_{i=0}^{n-1} sA_i g(su_i + t) = \frac{1}{s}Q_2(g).$$

W ten sposób, mając dowolną kwadraturę Q_1:

$$Q_1(f) = \sum_{i=0}^{n-1} A_i f(u_i) \approx \int_a^b f(u)\rho(u)\,du,$$

możemy otrzymać nową kwadraturę Q_2:

$$Q_2(g) = \sum_{i=0}^{n-1} B_i g(x_i) \approx \int_c^d g(x)\rho((x-t)/s)\,dx,$$

z węzłami $x_i = su_i + t$ i współczynnikami $B_i = sA_i$.

Kwadratury Q_1 i Q_2 mają ten sam rząd. Ponadto, mając oszacowanie błędu kwadratury Q_1, podobne do podanych wcześniej oszacowań dla kwadratur trapezów i Simpsona, można podać oszacowanie błędu kwadratury Q_2. Mianowicie, jeśli funkcje f i g są klasy C^k w swoich przedziałach całkowania i błąd kwadratury Q_1 ma górne oszacowanie o postaci

$$C(b-a)^{k+1}\max_{u\in[a,b]}|f^{(k)}(u)|,$$

to błąd kwadratury Q_2 jest nie większy niż

$$C(c-d)^{k+1}\max_{x\in[c,d]}|g^{(k)}(x)|,$$

z tą samą stałą C.

Kwadratury Gaussa

Niech $A \subset \mathbb{R}$ oznacza (ograniczony lub nieograniczony) przedział całkowania, niech ρ oznacza funkcję wagową i niech $p_0, p_1, \ldots$ będzie ciągiem wielomianów ortogonalnych w sensie iloczynu skalarnego

$$\langle f, g\rangle_\rho \stackrel{\text{def}}{=} \int_A f(x)g(x)\rho(x)\,dx.$$

Ustalmy liczbę n i określmy kwadraturę interpolacyjną Q z węzłami, które są miejscami zerowymi $x_0, \ldots, x_{n-1}$ wielomianu p_n; możemy to zrobić, bo miejsca zerowe tego wielomianu są jednokrotne i znajdują się w przedziale A.

Dowolny wielomian w stopnia mniejszego niż $2n$ możemy przedstawić w postaci

$$w(x) = p_n(x)a(x) + r(x),$$

gdzie a i r to iloraz i reszta z dzielenia wielomianu w przez p_n; stopnie wielomianów a i r są mniejsze niż n. Dzięki temu zachodzą równości

$$I(w) = \int_A w(x)\rho(x)\,dx = \int_A \Big(p_n(x)a(x) + r(x)\Big)\rho(x)\,dx =$$
$$\underbrace{\langle p_n, a\rangle_\rho}_{=0} + \int_A r(x)\rho(x)\,dx = Q(r) = Q(w),$$

ponieważ wartości wielomianów w i r we wszystkich węzłach kwadratury są jednakowe. Skonstruowana w ten sposób kwadratura jest zatem rzędu $2n$.

Kwadratury interpolacyjne, których węzły są miejscami zerowymi wielomianów ortogonalnych (odpowiadających danemu przedziałowi i funkcji wagowej) są nazywane kwadraturami Gaussa; nazwisko rodziny, do której odpowiedni wielomian należy, jest dołączane do nazwiska Gauss, i w ten sposób mówi się np.
o kwadraturach Gaussa-Legendre'a:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-1}^{1} f(x)\,dx,$$

kwadraturach Gaussa-Czebyszewa:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx,$$

kwadraturach Gaussa-Hermite'a:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-\infty}^{\infty} f(x) e^{-x^2}\,dx$$

i kwadraturach Gaussa-Laguerre'a:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{0}^{\infty} f(x) e^{-x}\,dx.$$

Konstruując kwadraturę Gaussa, na ogół trzeba jej węzły znaleźć, rozwiązując numerycznie równanie $p_n(x) = 0$. Współczynniki kwadratury Gaussa można obliczyć tak, jak współczynniki dowolnej kwadratury interpolacyjnej, lub na podstawie wzoru

$$A_i = \frac{1}{\sum_{k=0}^{n-1} \tilde{p}_k(x_i)^2},$$

w którym występują wielomiany ortonormalne $\tilde{p}_k(x) = p_k(x)/\|p_k\|_\rho$. Wygodnie jest użyć w tym obliczeniu formuły trójczłonowej. Zatem współczynniki każdej kwadratury Gaussa są dodatnie i z podanego wcześniej twierdzenia wynika, że dla dowolnej funkcji ciągłej ciąg kwadratur Gaussa coraz wyższych rzędów zbiega do całki z tej funkcji (z odpowiednią wagą).

Największe znaczenie praktyczne mają kwadratury Gaussa-Legendre'a, ponieważ najczęściej oblicza się całki w skończonym przedziale, z wagą $\rho(x) = 1$. Najprostsza kwadratura Gaussa-Legendre'a jest iloczynem długości przedziału całkowania i wartości funkcji w środku tego przedziału. Jest to więc kwadratura rzędu 2, oparta na jedym węźle.

Niech p_n oznacza wielomian ortogonalny Legendre'a stopnia n, wyskalowany tak, aby jego współczynnik wiodący był równy 1. Błąd aproksymacji jednostajnej funkcji f klasy $C^{2n}[-1,1]$ przez wielomian interpolacyjny Hermite'a h_{2n-1} stopnia $2n-1$, oparty na węzłach kwadratury Gaussa-Legendre'a (czyli miejscach zerowych wielomianu p_n), z których każdy liczymy dwukrotnie, ma oszacowanie

$$\max_{x\in[-1,1]} |f(x) - h_{2n-1}(x)| \leqslant \frac{M_{2n}}{(2n)!} p_n(x)^2,$$

gdzie $M_{2n} = \max_{x\in[-1,1]} |f^{(2n)}(x)|$. Niech

$$C_n = \int_{-1}^{1} p_n(x)^2\,dx.$$

Po dokonaniu zamiany zmiennych, możemy oszacować błąd kwadratury Gaussa-Legendre'a rzędu $2n$ dla przedziału $[a,b]$:

$$|I(f) - Q(f)| \leqslant C_n \frac{M_{2n}}{(2n)!}\left(\frac{b-a}{2}\right)^{2n+1},$$

przy czym teraz M_{2n} oznacza oszacowanie pochodnej rzędu $2n$ funkcji f w przedziale $[a,b]$.

Kwadratury złożone

Tak jak w aproksymacji jednostajnej funkcji, skutecznym sposobem zmniejszenia błędu aproksymacji całki przez kwadraturę jest podzielenie przedziału całkowania na krótsze podprzedziały i obliczenie sumy kwadratur interpolacyjnych dla tych podprzedziałów. W ten sposób otrzymuje się kwadratury złożone. Błąd takiej kwadratury jest sumą błędów kwadratur dla podprzedziałów, przy czym błędy te mogą mieć różne znaki, a zatem mogą się znosić. Oszacowania błędów kwadratur złożonych zwykle są sumami oszacowań błędów w podprzedziałach, przez co często bywają pesymistyczne.

Dodatkową korzyścią z zastosowania kwadratury złożonej jest możliwość podziału przedziału całkowania w punktach nieciągłości funkcji podcałkowej lub jej pochodnych (jeśli punktów tych jest skończenie wiele i je znamy). Wtedy w każdym podprzedziale funkcja podcałkowa ma wyższą klasę ciągłości, co umożliwia stosowanie kwadratur odpowiednio wyższego rzędu. Ponadto, po dokonaniu podziału można stosować w podprzedziałach różne kwadratury, dostosowane do zachowania funkcji podcałkowej w tych podprzedziałach. Kolejna możliwość to adaptacja — dla konkretnej funkcji można znaleźć oszacowania błędów w poszczególnych podprzedziałach, i na tej podstawie podejmować decyzję o dalszym (rekurencyjnym) podziale niektórych z nich.

Kwadratury w podprzedziałach konstruujemy za pomocą opisanej wcześniej zamiany zmiennych. Zobaczmy przykłady kwadratur z podziałem przedziału $[a,b]$ na N części o tej samej długości $h = (b-a)/N$.

Złożona kwadratura trapezów powstaje w ten sposób, że w każdym z podprzedziałów przedziału $[a,b]$ stosujemy kwadraturę trapezów. W ten sposób otrzymamy liczbę

$$T_h(f) = h\Big(\frac{1}{2}f(x_0) + \sum_{i=1}^{N-1} f(x_i) + \frac{1}{2}f(x_N)\Big),$$

gdzie $x_i = a + ih$. Jeśli funkcja f jest klasy $C^2[a,b]$ i $|f''(x)| \leqslant M_2$ dla każdego $x \in [a,b]$, to wartość bezwzględna lokalnego błędu kwadratury trapezów w przedziale $[x_i, x_{i+1}]$ nie przekracza $\frac{M_2}{12}h^3$, a zatem suma tych błędów ma oszacowanie

$$|I(f) - T_h(f)| \leqslant \frac{M_2}{12}(b-a)h^2.$$

Złożoną kwadraturę Simpsona otrzymujemy analogicznie. Oznaczmy $x_i = a + ih/2$ dla $i = 0, \ldots, 2N$. Suma kwadratur Simpsona w N podprzedziałach o długości h jest równa

$$S_h(f) = \frac{h}{6}\Big(f(x_0) + 4f(x_1) + \sum_{i=1}^{N-1}\big(2f(x_{2i}) + 4f(x_{2i+1})\big) + f(x_{2N})\Big),$$

zaś dla funkcji f odpowiednio klasy $C^3[a,b]$ i $C^4[a,b]$ błąd ma oszacowania

$$|I(f) - S_h(f)| \leqslant \frac{M_3}{192}(b-a)h^3,$$
$$|I(f) - S_h(f)| \leqslant \frac{M_4}{2880}(b-a)h^4.$$

Konstruowanie złożonych kwadratur Gaussa jest utrudnione, jeśli funkcja wagowa nie jest stała, dlatego powyższe podejście stosuje się tylko do kwadratur Gaussa-Legendre'a. Jeśli funkcja f jest klasy $C^{2n}[a,b]$, to możemy w każdym przedziale o długości h użyć kwadratury Gaussa-Legendre'a opartej na n węzłach i wtedy dostaniemy oszacowanie błędu o postaci

$$|I(f) - Q_h(f)| \leqslant C_n M_{2n}(b-a)h^{2n},$$

w którym stała C_n zależy tylko od rzędu kwadratury. Jak widać, dla $h \to 0$ błąd bardzo szybko dąży do zera. Jeśli funkcja f nie ma ciągłych pochodnych aż tak wysokiego rzędu, to błąd nadal dąży do zera, choć wolniej.

Ekstrapolacja Richardsona i metoda Romberga

Niech f oznacza funkcję klasy $C^{2n+2}[a,b]$. Dowodzi się, że błąd złożonej kwadratury trapezów, z przedziałem $[a,b]$ podzielonym na podprzedziały o jednakowej długości h, można wyrazić wzorem

$$I(f) - T_h(f) = c_1h^2 + c_2h^4 + \cdots + c_nh^{2n} + O(h^{2n+2}),$$

zwanym wzorem sumacyjnym Eulera-Maclaurina. Współczynniki $c_1, \ldots, c_n$ zależą od pochodnych funkcji f w przedziale $[a,b]$, ale nie zależą od długości podprzedziałów.

Możemy ten wzór przepisać dla złożonej kwadratury trapezów z dwukrotnie drobniejszym podziałem przedziału całkowania:

$$I(f) - T_{h/2}(f) = \frac{c_1}{4}h^2 + \frac{c_2}{16}h^4 + \cdots + \frac{c_n}{4^n}h^{2n} + O(h^{2n+2}),$$

Jeśli strony powyższego wzoru pomnożymy przez $4/3$ i odejmiemy od nich strony wzoru dla kwadratury z podprzedziałami o długości h pomnożone przez $1/3$, to otrzymamy równość

$$I(f) - \Big(\frac{4}{3}T_{h/2}(f) - \frac{1}{3}T_h(f)\Big) = d_2h^4 + \cdots + d_nh^{2n} + O(h^{2n+2}).$$

Kombinacja liniowa $T_h^{(1)}(f) = 4/3T_{h/2}(f) - 1/3T_h(f)$ jest kwadraturą, której dominujący składnik błędu jest rzędu h^4, zatem znacznie szybciej maleje podczas zmniejszania h. Opisany sposób wyeliminowania dominującego składnika błędu (który można stosować także w innych przypadkach, gdy błąd jest opisany za pomocą szeregu potęgowego) jest nazywany ekstrapolacją Richardsona.

Ekstrapolację Richardsona możemy iterować. Mając kwadratury $T_h^{(j)}$ i $T_{h/2}^{(j)}$, których dominujące składniki błędów są proporcjonalne do h^{2j+2}, określamy kwadraturę

$$T_h^{(j+1)}(f) = \frac{2^{2j+2}}{2^{2j+2}-1}T_{h/2}^{(j)}(f) - \frac{1}{2^{2j+2}-1}T_h^{(j)}(f),$$

której błąd ma dominujący składnik błędu h^{2j+4}. Oparta na tym pomyśle metoda numerycznego całkowania jest nazywana metodą Romberga. Podprogram obliczający całkę, dla ustalonego h, oblicza kwadratury $T_h(f)$ i $T_{h/2}(f)$ i oblicza kwadraturę $T_h^{(1)}(f)$. Wyrażenie $|T_h(f) - T_{h/2}(f)|$ może być przyjęte za oszacowanie błędu, co jest analogią do przyrostowego kryterium stopu w metodach numerycznych rozwiązywania równań nieliniowych. Jeśli to oszacowanie jest zbyt duże, to obliczana jest kwadratura $T_{h/4}(f)$, a następnie $T_{h/2}^{(1)}(f)$ i $T_h^{(2)}(f)$ itd.

Obliczenie przebiega zgodnie ze schematem

$$\begin{array}{ccccccc}
T_h(f) & & & & & & \\
 & \searrow & & & & & \\
T_{h/2}(f) & \to & T_h^{(1)}(f) & & & & \\
 & \searrow & & \searrow & & & \\
T_{h/4}(f) & \to & T_{h/2}^{(1)}(f) & \to & T_h^{(2)}(f) & & \\
\vdots & & \vdots & & \vdots & \ddots & \\
 & \searrow & & \searrow & & \searrow & \\
T_{h/2^k}(f) & \to & T_{h/2^{k-1}}^{(1)}(f) & \to & \ldots & \to & T_h^{(k)}(f)
\end{array}$$

Za oszacowanie błędu każdej kwadratury otrzymanej przez ekstrapolację możemy przyjąć różnicę kwadratur, na podstawie których została ona obliczona. Zauważmy, że po każdym zmniejszeniu długości podprzedziałów dla kwadratury trapezów wartości funkcji podcałkowej wystarczy tylko obliczyć tylko w nowych węzłach i nie ma potrzeby przechowywania wartości funkcji f w tablicy.

Całkowanie funkcji wielu zmiennych

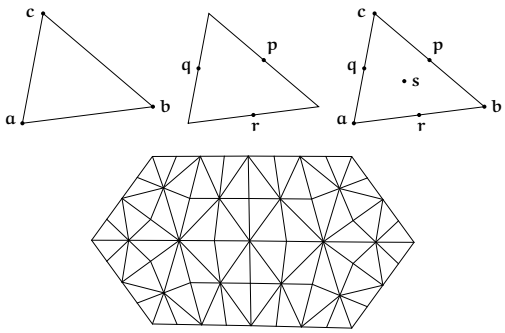
Znane z analizy twierdzenie Fubiniego umożliwia sprowadzenie zadania obliczenia całki z funkcji f określonej w wielowymiarowym obszarze A do obliczenia całek jednowymiarowych. Analogicznie można postępować z kwadraturami. Jest to szczególnie proste, gdy obszar A jest kostką. Powiedzmy, że jest to prostokąt: $A = [a,b] \times [c,d]$. Mając kwadratury przybliżające całki w przedziałach $[a,b]$ i $[c,d]$, odpowiednio z węzłami $x_0, \ldots, x_{n-1}$ i $y_0, \ldots, y_{m-1}$ oraz współczynnikami $A_0, \ldots, A_{n-1}$ i $B_0, \ldots, B_{m-1}$, możemy obliczyć

$$Q(f) = \sum_{i=0}^{n-1} A_i \sum_{j=0}^{m-1} B_j f(x_i, y_j) \approx \int_a^b \left(\int_c^d f(x,y)\,dy\right) dx.$$

Znane są również kwadratury odpowiednie dla obszarów o innym kształcie. Jeśli np. obszar całkowania A jest trójkątem, to kwadratury określone wzorami

$$Q_1(f) = \frac{T}{3}\big(f(\mathbf{a}) + f(\mathbf{b}) + f(\mathbf{c})\big),$$
$$Q_2(f) = \frac{T}{3}\big(f(\mathbf{p}) + f(\mathbf{q}) + f(\mathbf{r})\big),$$
$$Q_3(f) = \frac{T}{60}\Big(3\big(f(\mathbf{a}) + f(\mathbf{b}) + f(\mathbf{c})\big) + 8\big(f(\mathbf{p}) + f(\mathbf{q}) + f(\mathbf{r})\big) + 27f(\mathbf{s})\Big),$$

w których T oznacza pole trójkąta A o wierzchołkach $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, środkach boków $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$ i środku ciężkości $\mathbf{s}$, są dokładne, jeśli funkcja f jest odpowiednio wielomianem stopnia 1, 2 i 3. Dowolny obszar wielokątny możemy podzielić na trójkąty i stosować kwadratury złożone.



Numeryczne całkowanie jest kłopotliwe, jeśli wimiar obszaru A jest duży. Istnieją zadania praktyczne (biorące się m.in. z ekonomii), w których wymiar d obszaru całkowania jest rzędu kilkuset. Obliczenie całki w takiej kostce d-wymiarowej za pomocą kwadratury otrzymanej analogicznie, jak dla prostokąta, jest niewykonalne. Nawet gdyby w przedziale zmienności każdej zmiennej wybrać tylko dwa węzły, liczba punktów, w których trzeba by obliczyć wartości funkcji podcałkowej, byłaby równa 2^d. Zjawisko wykładniczego wzrostu złożoności obliczeniowej zadania ze wzrostem wymiaru dziedziny funkcji nosi nazwę przekleństwa wymiaru (ang. *dimensionality curse*).

Znanych jest kilka sposobów obliczania przybliżonych wartości całek wielowymiarowych za pomocą wartości funkcji obliczonych w znacznie mniejszej liczbie punktów. Sposób najprostszy i jednocześnie skuteczny dla najszerszej klasy takich zadań wynalazł Ulam w 1946 r. Sposób ten jest znany pod nazwą metody Monte Carlo. Obszar A uznajemy za przestrzeń zadrzeń elementarnych i określamy w nim jednostajny rozkład prawdopodobieństwa. Wtedy funkcja f jest zmienną losową. Iloczyn wartości oczekiwanej tej zmiennej losowej i miary $|A|$ obszaru A jest poszukiwaną całką, $\int_A f$. Dla n niezależnych losowań punktów $x_i \in A$ możemy określić nową zmienną losową wzorem

$$Q(f) = |A|\frac{1}{n}\sum_{i=0}^{n-1} f(x_i).$$

Jest to właśnie kwadratura Monte Carlo; jej wartość oczekiwana też jest równa poszukiwanej całce. Jeśli zmienna losowa f ma wariancję σ^2, to wariancja σ_n^2 kwadratury Monte Carlo jest równa $|A|\sigma^2/n$. Zatem odchylenie standardowe σ_n zmiennej losowej $Q(f)$ jest proporcjonalne do $n^{-1/2}$ i w szczególności nie zależy od wymiaru d obszaru A. Dla dostatecznie dużego n możemy oczekiwać, że błąd jest bardzo mały — z dużym prawdopodobieństwem, ale nie z całkowitą pewnością.