Metody numeryczne

9. Rozwiązywanie równań algebraicznych

P. F. Góra

https://zfs.fais.uj.edu.pl/pawel_gora

12 grudnia 2023

Co to znaczy rozwiązać równanie?

Przypuśćmy, że postawiono przed nami problem rozwiązania równania

$$f(x) = 0\,. \tag{1}$$

Przede wszystkim musimy ustalić co oznacza słowo “rozwiązać”. Można bowiem mieć na myśli dwie rzeczy

I. Znaleźć *wszystkie* rozwiązania (1).

II. Znaleźć *jakieś* rozwiązanie (1).

Pierwszy przypadek zachodzi wtedy, gdy o równaniu możemy dużo powiedzieć od strony analitycznej — na przykład gdy jest to równanie trygonometryczne lub wielomianowe. W przypadku ogólnym na ogół nie wiemy nawet czy jakiekolwiek rozwiązanie (1) istnieje, a jeśli tak, to ile ich jest. Dlatego w przypadku ogólnym zadowalamy się znalezieniem *jakiegoś*, pojedynczego rozwiązania (o ile warunki zadania nie stanowią inaczej).

O funkcji $f(x)$ zakładamy, że jest ciągła i — na ogół — różniczkowalna odpowiednią ilość razy.

Krotność miejsca zerowego

Mówimy, że x_0 jest miejscem zerowym funkcji $f(x)$ o krotności k, jeżeli w tym punkcie zeruje się funkcja wraz ze swoimi pochodnymi do rzędu $k-1$: $f(x_0) = f'(x_0) = f''(x_0) = \cdots = f^{(k-1)}(x_0) = 0$. Na przykład wielomian $P(x) = x^4 - x^3 - x^2 + x$ ma jednokrotne miejsce zerowe w $x = -1$, jednokrotne miejsce zerowe w $x = 0$ i dwukrotne miejsce zerowe w $x = 1$. Natomiast funkcja $f(x) = (x^2 - 1)\sinh^3 x$ ma jednokrotne miejsca zerowe w $x = \pm 1$ i trzykrotne miejsce zerowe w $x = 0$.

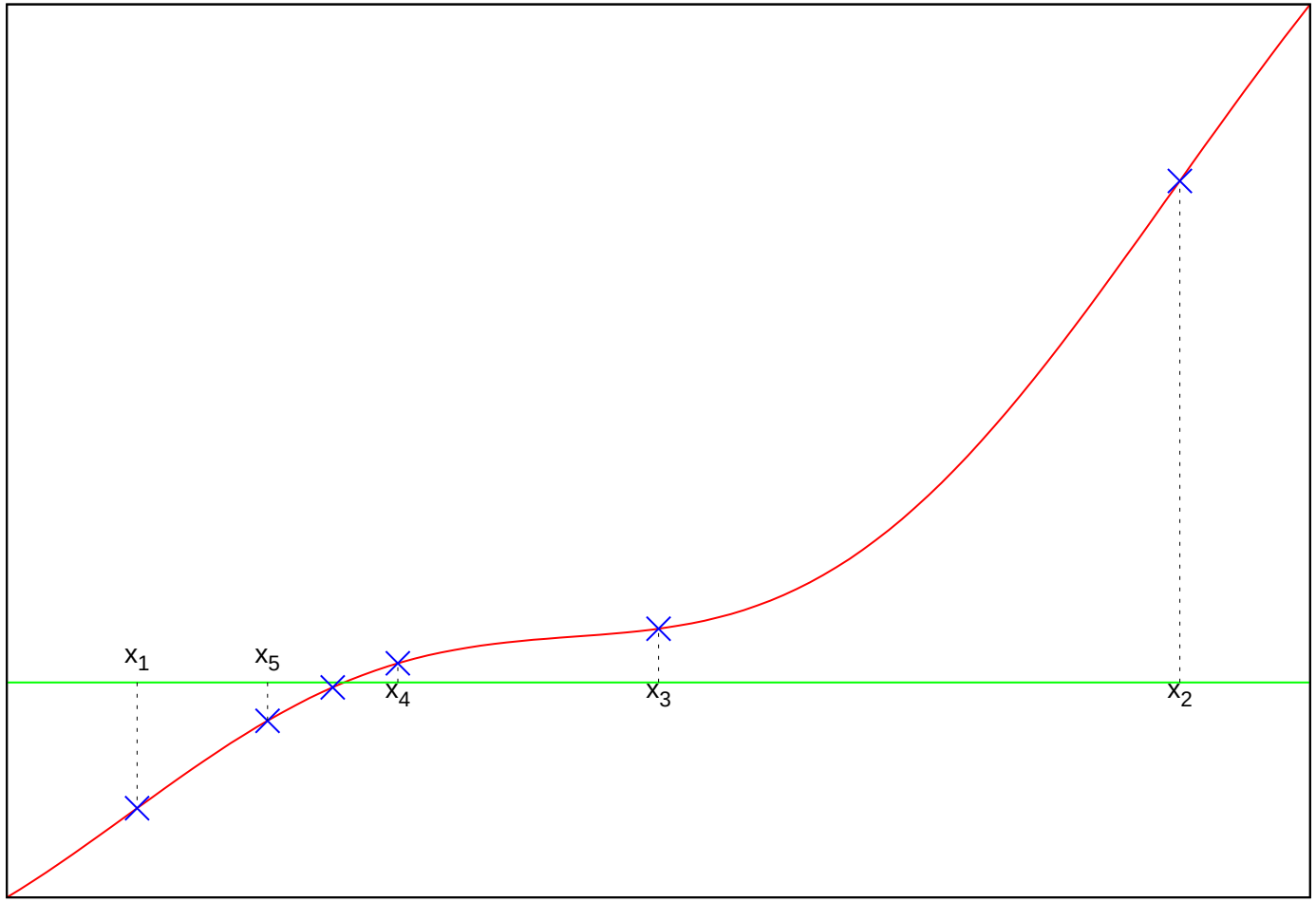
Funkcja zmienia znak w otoczeniu miejsca zerowego o krotności nieparzystej i *nie zmienia znaku* w otoczeniu miejsca zerowego o krotności parzystej.

Metoda bisekcji

Jeżeli funkcja $f(x)$ jest ciągła i jeżeli znajdziemy dwa punkty, w których znak funkcji jest przeciwny, $f(x_1) \cdot f(x_2) < 0$, jako przybliżenie miejsca zerowego bierzemy środkowy punkt przedziału $[x_1, x_2]$, $x_3 = (x_1 + x_2)/2$. Ustalamy, w którym z przedziałów $[x_1, x_3]$, $[x_3, x_2]$ funkcja zmienia znak, po czym powtarzamy całą procedurę dla tego przedziału. Procedurę kończymy, gdy odległość pomiędzy dwoma kolejnymi przybliżeniami stanie się dostatecznie mała, $|x_{n+1} - x_n| \leqslant \varepsilon \ll 1$.

Zbieżność metody bisekcji jest liniowa, to znaczy, że na ustalenie każdego kolejnego miejsca dziesiętnego w rozwinięciu miejsca zerowego potrzeba takiej samej liczby iteracji.

Metoda bisekcji działa dla miejsc zerowych o nieparzystej krotności i nie działa dla miejsc zerowych o krotności parzystej.



Uwaga o kryterium stopu

Jak powiedziano, kryterium stopu jest osiągnięcie dostatecznie małej odległości pomiędzy dwoma kolejnymi przybliżeniami:

$$|x_{n+1} - x_n| \leqslant \varepsilon \ll 1 \,. \tag{2a}$$

Dlaczego nie stosować "naiwnego" kryterium

$$|f(x_n)| \leqslant \varepsilon \ll 1 \,? \tag{2b}$$

Po pierwsze, poza bardzo rzadkimi, szczególnie dobranymi przypadkami, numerycznie prawie nigdy nie jesteśmy w stanie ściśle zlokalizować miejsca zerowego funkcji. Możemy natomiast wyznaczyć przedział, w którym leży poszukiwane miejsce zerowe. Chcemy, aby przedział ten był dostatecznie mały i w ten sposób kryterium (2a) pojawia się w sposób naturalny.

Po drugie, przypuśćmy, że rozwiązujemy problem

$$f(x) = 0\,. \tag{3a}$$

Przeskalujmy nasz problem:

$$10^{A} \cdot f(x) = 0 \tag{3b}$$

gdzie $A \gg 1$ jest pewną liczbą, lub też

$$10^{-A} \cdot f(x) = 0\,. \tag{3c}$$

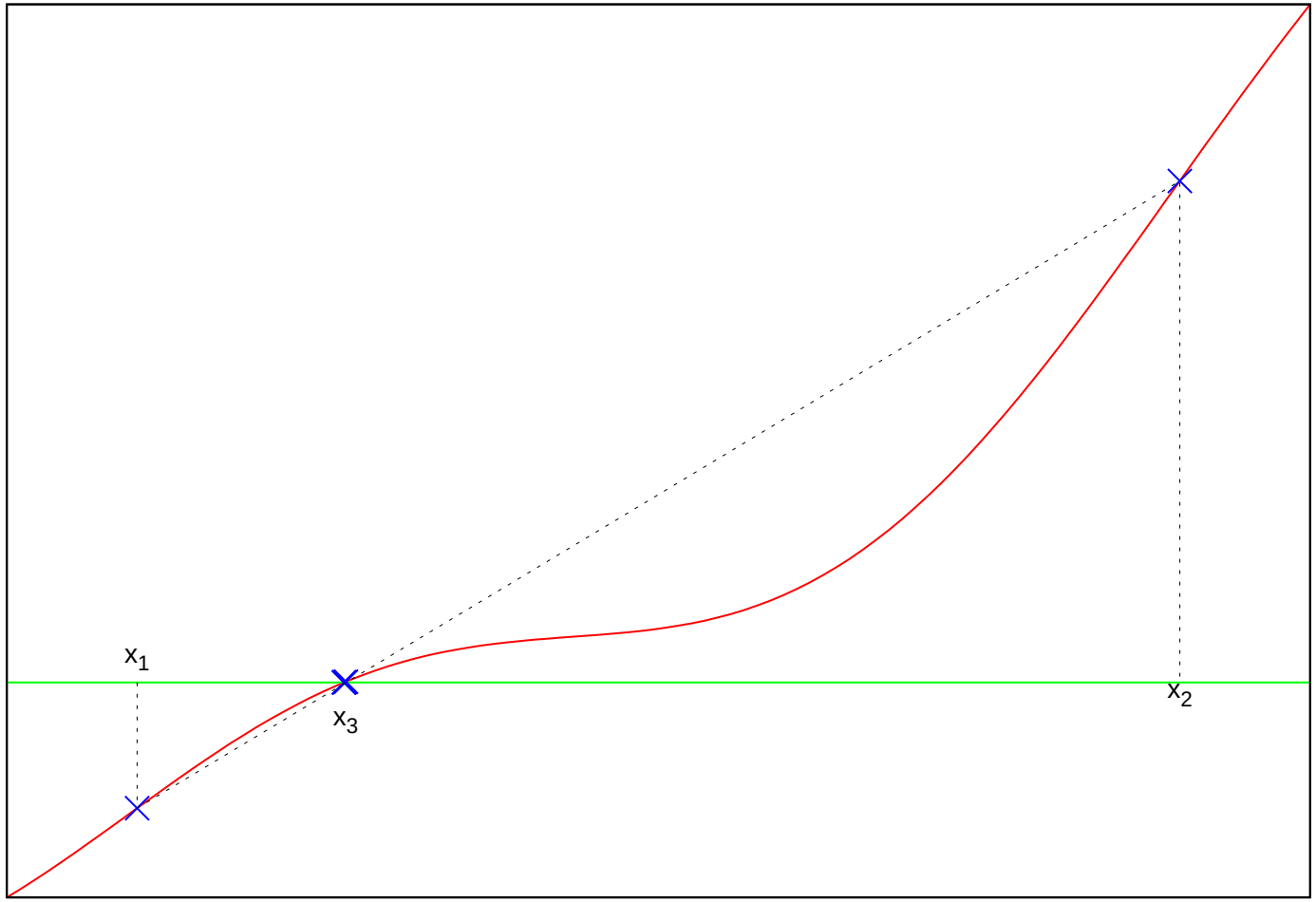
Analitycznie wszystkie równania (3) mają takie same rozwiązania. Tymczasem przy rozwiązywaniu numerycznym zastosowanie (2b) do drugiego z równań (3), dla dostatecznie dużego A, może *nigdy* nie doprowadzić do zbieżności, natomiast w zastosowaniu do trzeciego może "wykazać" zbieżność nawet dla bardzo kiepskiego przybliżenia. Kryterium (2a) jest odporne na problemy wynikające ze skalowania.

Metoda *regula falsi*

Metoda *regula falsi*, czyli "metoda fałszywego położenia", jest jedną z najczęściej stosowanych metod poszukiwania rozwiązań równania (1). Punkt wyjścia jest podobny do metody bisekcji: Jeżeli funkcja $f(x)$ jest ciągła i jeżeli znajdziemy dwa punkty, w których znak funkcji jest przeciwny, $f(x_1) \cdot f(x_2) < 0$, jako przybliżenie miejsca zerowego bierzemy punkt przecięcia siecznej przechodzącej przez punkty $(x_1, f(x_1))$, $(x_2, f(x_2))$ z osią OX:

$$x_3 = \frac{f(x_1)x_2 - f(x_2)x_1}{f(x_1) - f(x_2)} \,. \tag{4}$$

Jeżeli $|x_3 - x_2| \leqslant \varepsilon \ll 1$, kończymy procedurę. Jeżeli nie, wybieramy ten z przedziałów $[x_1, x_3]$, $[x_3, x_2]$, w którym funkcja zmienia znak i postępujemy analogicznie.



Interpolacja odwrotna

Przypuśćmy, że mamy stabelaryzowane wartości funkcji w węzłach:

x_i	x_1	x_2	x_3	$\ldots$	x_n
$f_i = f(x_i)$	f_1	f_2	f_3	$\ldots$	f_n

(5)

przy czym — ważne! — stabelaryzowane wartości są ściśle monotoniczne, $f_1 > f_2 > \cdots > f_n$ (lub $f_1 < f_2 < \cdots < f_n$). Moglibyśmy wówczas, za pomocą interpolacji, wyrysować przybliżony wykres funkcji, znaleźć punkt, w którym przecina on oś OX i wziąć ten punkt jako przybliżenie miejsca zerowego. Procedurę kontytuujemy, odrzucając najbardziej odległy "stary" punkt.

Jak to zrobić? Skoro funkcja jest monotoniczna, jest odwracalna, przy czym "węzły" i "wartości" zamieniają się miejscami:

f_i	f_1	f_2	f_3	$\ldots$	f_n
$x_i = f^{-1}(f_i)$	x_1	x_2	x_3	$\ldots$	x_n

(6)

Wartość funkcji odwrotnej w zerze oznacza punkt, w którym funkcja ma miejsce zerowe!

Aby znaleźć przybliżone miejsce zerowe funkcji $f(x)$, tworzymy wielomian interpolacyjny według tabeli (6) i obliczamy wartość tego wielomianu, czyli przybliżenia funkcji odwrotnej, w zerze. Ze względów praktycznych interpolację odwrotną stosuje się dla niewielkiej liczby węzłów.

Metoda siecznych

Metoda siecznych, będąca dwupunktową interpolacją odwrotną ☺, jest nagminnie mylona z metodą *regula falsi*. Punktem wyjścia są dowolne dwa punkty, dla których $f(x_1) \neq f(x_2)$. Prowadzimy sieczną przez te punkty (bez względu na znak $f(x_1) \cdot f(x_2)$), i jako x_3 bierzemy miejsce zerowe tej siecznej, dane *także* wzorem (4). W kolejnych krokach bierzemy zawsze dwa ostatnie punkty, bez względu na to, czy funkcja zmienia znak.

Metoda siecznych i metoda *regula falsi* to są inne metody! Metoda siecznych może być zbieżna *szybciej* niż metoda *regula falsi*, ale — w odróżnieniu od *regula falsi* i metody bisekcji — w niektórych przypadkach zawodzi (nie jest zbieżna do miejsca zerowego).

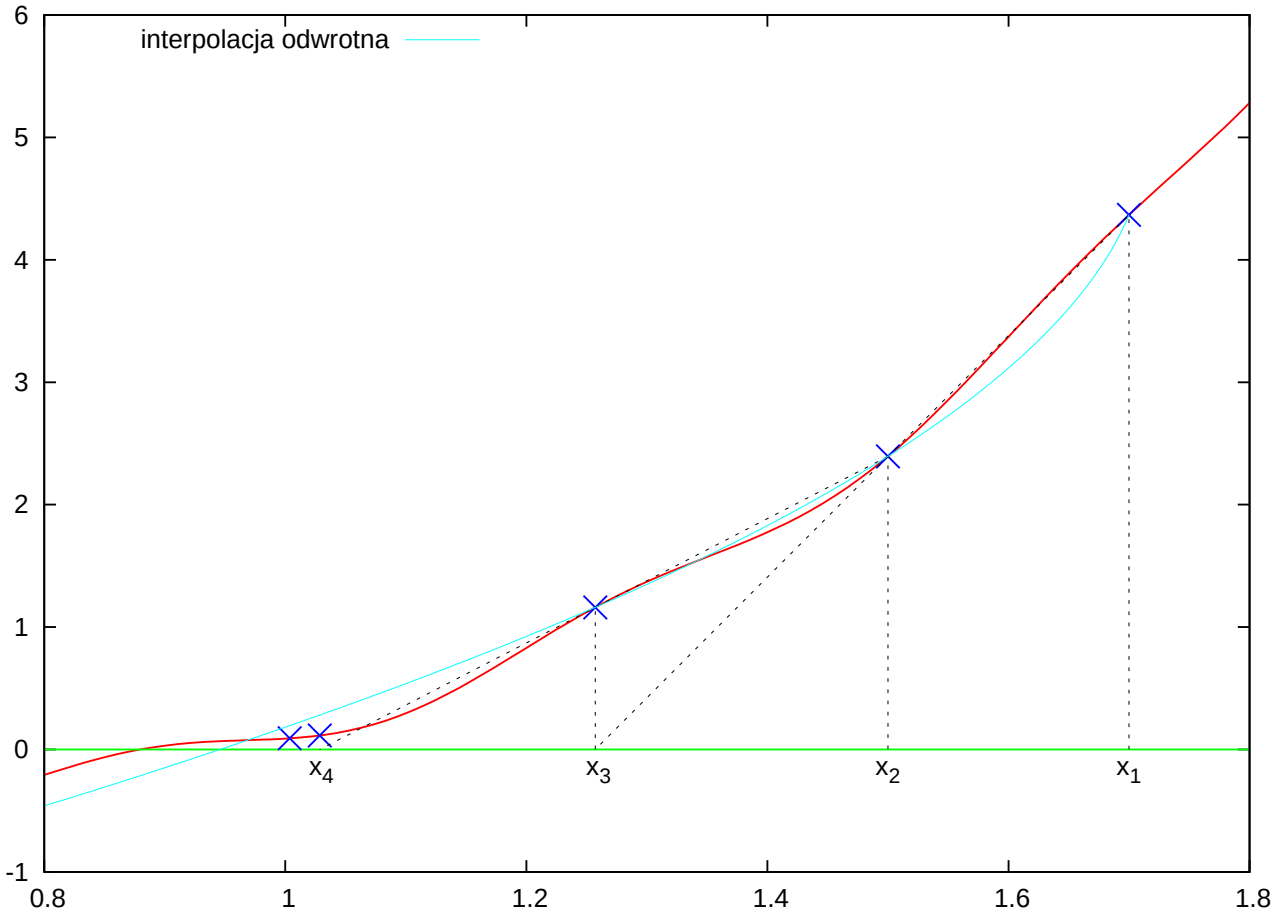
Porównanie

Dla funkcji $f(x) = \frac{1}{8}x^4 + x^3 - x + \frac{1}{8}\sin(16x)$ z punktami startowymi $x_1 = 0.8$, $x_2 = 1.2$, i przyjętej długości końcowego przedziału 10^{-8}, metody bisekcji, *regula falsi* i siecznych zbiegały się następująco:

metoda	liczba kroków	$\bar{x}$	$f(\bar{x})$
bisekcji	26	0.879311846	0.0000000029
regula falsi	12	0.879311849	0.0000000090
siecznych	5	0.879311845	0.0000000022

$\bar{x}$ oznacza środek osiągniętego przedziału. Miejsce zerowe leży w przedziale $[\bar{x} - 0.5 \cdot 10^{-8}, \bar{x} + 0.5 \cdot 10^{-8}]$.

Metoda siecznych i interpolacja odwrotna oparta na trzech punktach



Więcej o kryterium stopu

Wszystkie cztery przedstawione wyżej metody — metoda bisekcji, *regula falsi*, metoda siecznych i interpolacja odwrotna — są metodami iteracyjnymi, Kiedy więc należy zatrzymać iterację uznając, że osiągnęliśmy już wystarczającą dokładność w lokalizacji miejsca zerowego?

- Wszystkie te metody sprowadzają się do iteracyjnego pomniejszania przedziału, w którym znajduje się poszukiwane miejsce zerowe, różniąc się jedynie *sposobem* pomniejszania tego przedziału. Wobec tego iterację uznajemy za zakończoną, gdy przedział zawierający miejsce zerowe stanie się dostatecznie mały, $|x_n - x_{n-1}| < \varepsilon \ll 1$, gdzie x_{n-1}, x_n są kolejnymi iteratami.
- Dodatkowo, dla metody siecznych i interpolacji odwrotnej, które nie muszą być zbieżne, zakładamy największą dopuszczalną liczbę iteracji. Po przekroczeniu tej liczby przyjmujemy, że metoda nie osiągnęła

zbieżności. Nie należy się natomiast martwić, gdy długości przedziałów chwilowo nam wzrosną: $|x_n - x_{n-1}| > |x_{n-1} - x_{n-2}|$. Tak się w tej metodzie może zdarzyć i *nie musi* to świadczyć o rozbieżności metody.

- Dodatkowo, dla interpolacji odwrotnej, jeśli uzyskany ciąg wartości funkcji przestaje być monotoniczny, przerywamy iteracje, gdyż niespełnione są założenia analityczne leżące u podstaw metody.

Metoda Newtona

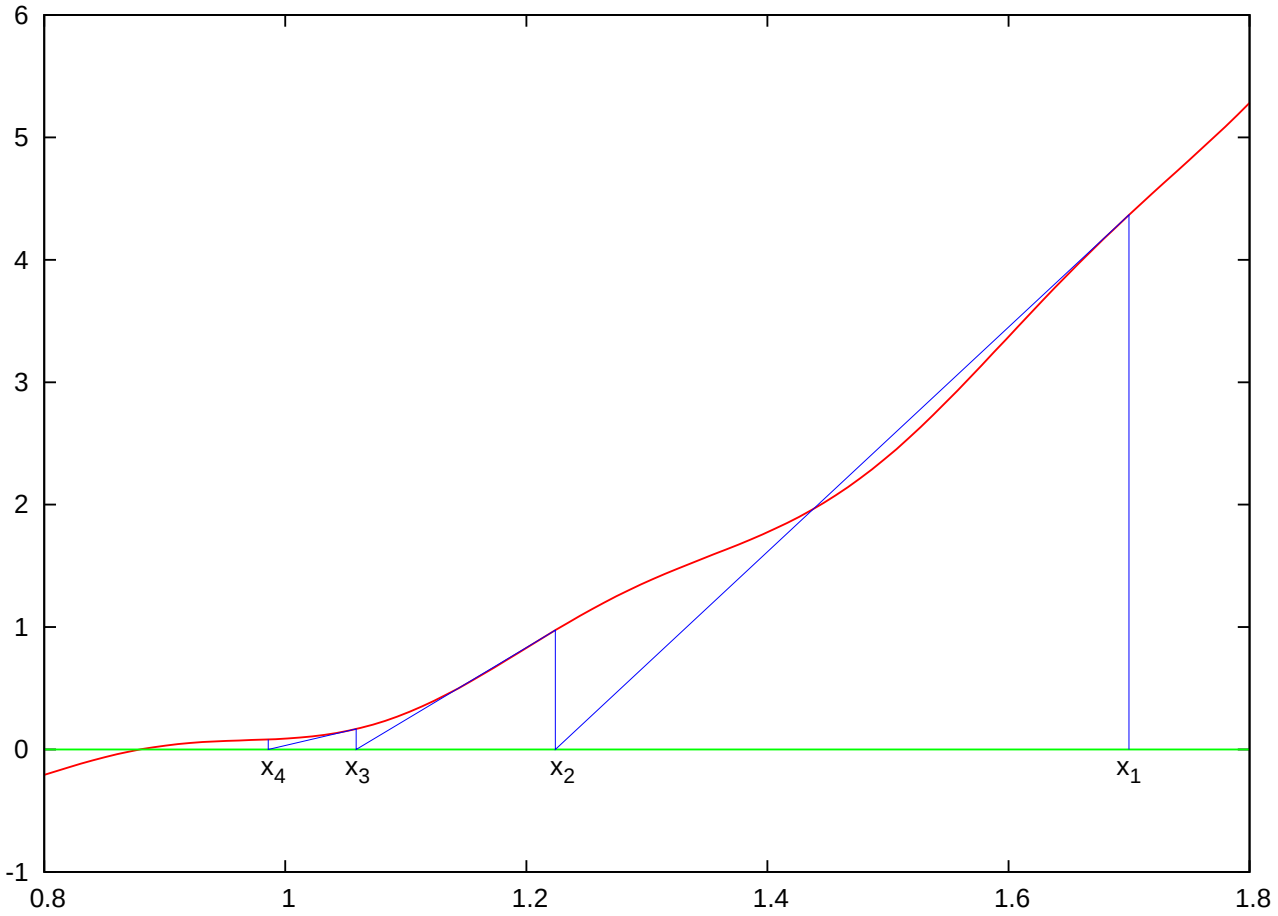
Przypuśćmy, że prawa strona rownania (1) jest różniczkowalna. Rozwijamy tę funkcję w szereg Taylora wokół pewnego punktu

$$f(x_0 + \delta) \simeq f(x_0) + \delta \cdot f'(x_0) \tag{7}$$

a następnie żądamy, aby lewa strona rozwinięcia (7) znikała. Jak duży krok δ powinniśmy wykonać? $\delta = -f(x_0)/f'(x_0)$. Przyjmujemy, że przesuwamy się do punktu $x_1 = x_0 + \delta$ i powtarzamy całą procedurę. Przesuwamy się do kolejnego punktu — i tak dalej. Prowadzi to do iteracji

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \,. \tag{8}$$

Interpretacja geometryczna metody Newtona — metoda stycznych



Kryterium zbieżności

Zauważmy, że punkt stały iteracji (8) jest rozwiązaniem równania $f(x) = 0$. Formalnie, jeżeli funkcja

$$g(x) = x - \frac{f(x)}{f'(x)} \tag{9}$$

(i) jest ciągła oraz (ii) przeprowadza pewien przedział domknięty $[a, b]$ w ten sam przedział domknięty $[a, b]$, to na mocy twierdzenia Brouwera iteracja (8) ma w tym przedziale punkt stały, będący rozwiązaniem równania (1).

Problem leży w spełnieniu warunku (ii)

Zbieżność kwadratowa

Niech funkcja $f(x)$ ma *jednokrotne* miejsce zerowe w $x = a$, to znaczy

$$f(x) = (x-a)g(x)\,,\ g(a) \neq 0 \tag{10}$$

i $g(x)$ jest różniczkowalna w otoczeniu $x = a$. Przyjmijmy, że $x_n = a + \varepsilon$, $|\varepsilon| \ll 1$. Wówczas metoda Newtona (8) prowadzi do

$$\begin{aligned} x_{n+1} &= x_n - \frac{(x_n - a)g(x_n)}{g(x_n) + (x_n - a)g'(x_n)} \\ &= a + \varepsilon - \frac{\varepsilon}{1 + \frac{g'(a+\varepsilon)}{g(a+\varepsilon)}\varepsilon} \\ &\simeq a + \varepsilon - \varepsilon\left(1 - \frac{g'(a+\varepsilon)}{g(a+\varepsilon)}\varepsilon\right) \simeq a + \frac{g'(a)}{g(a)}\varepsilon^2 \end{aligned} \tag{11}$$

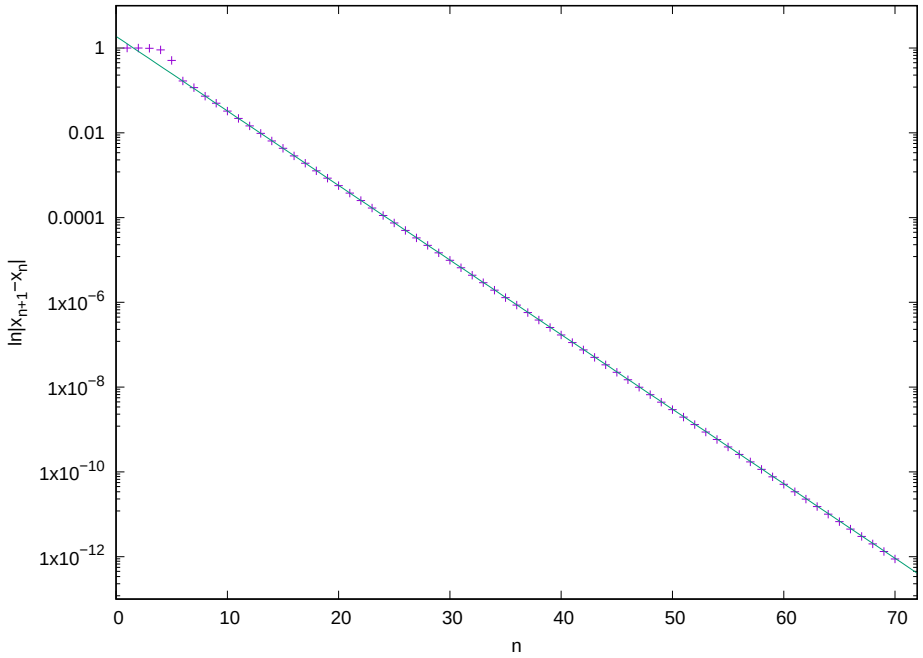
Widzimy, że dla *jednokrotnego* miejsca zerowego, jeśli w metodzie Newtona $|x_n - a| = |\varepsilon| \ll 1$, gdzie a jest miejscem zerowym, to $|x_{n+1} - a| \sim \varepsilon^2$, przy czym $\varepsilon^2 \ll |\varepsilon| \ll 1$. Zbieżność tego typu nazywamy zbieżnością kwadratową. Podkreślam, że zbieżność kwadratowa (11) zachodzi tylko dla punktów, które są dostatecznie blisko miejsca zerowego. Początkowe kroki mogą się wręcz oddalać od poszukiwanego miejsca zerowego, co *na ogół* nie szkodzi, gdyż największy wysiłek numeryczny wkłada się w precyzyjną lokalizację rozwiązania, nie w początkowe kroki.

Przykład

Za pomocą metody Newtona rozwiąż z dokładnością do 10^{-12} równanie

$$e^{2x} + e^x - 6 = 0 \tag{12}$$

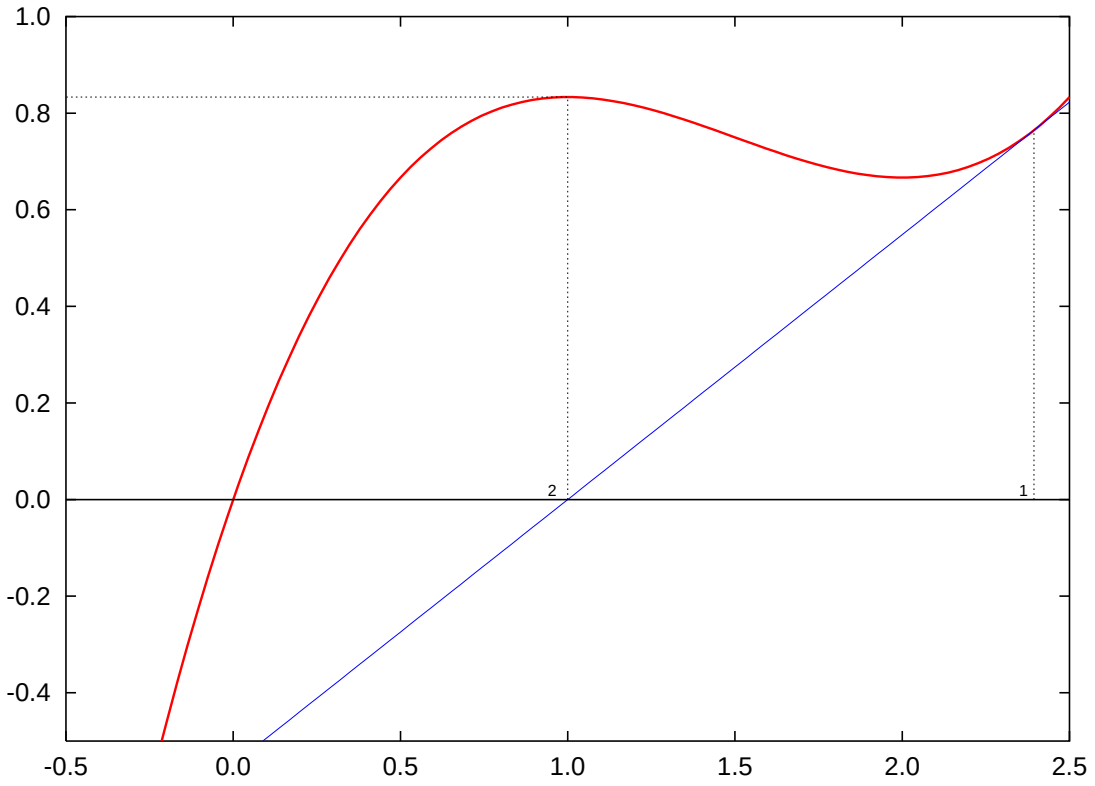
Po 70 iteracjach iteracja zatrzymuje się na $x = 0.693147180560$, gdyż $|x_{70} - x_{69}| < 10^{-12}$. Oznacza to, że numeryczne rozwiązanie leży gdzieś w przedziale $(0.693147180595, 0.693147180605)$. Analitycznym rozwiązaniem równania (12) jest $\ln 2 \simeq 0.693147180559945$.



Zwróćmy uwagę, że poza kilkoma początkowymi punktami, definiująca kryterium zbieżności wielkość $\ln |x_{n+1} - x_n|$ układa się niemalże idealnie liniowo wraz z n.

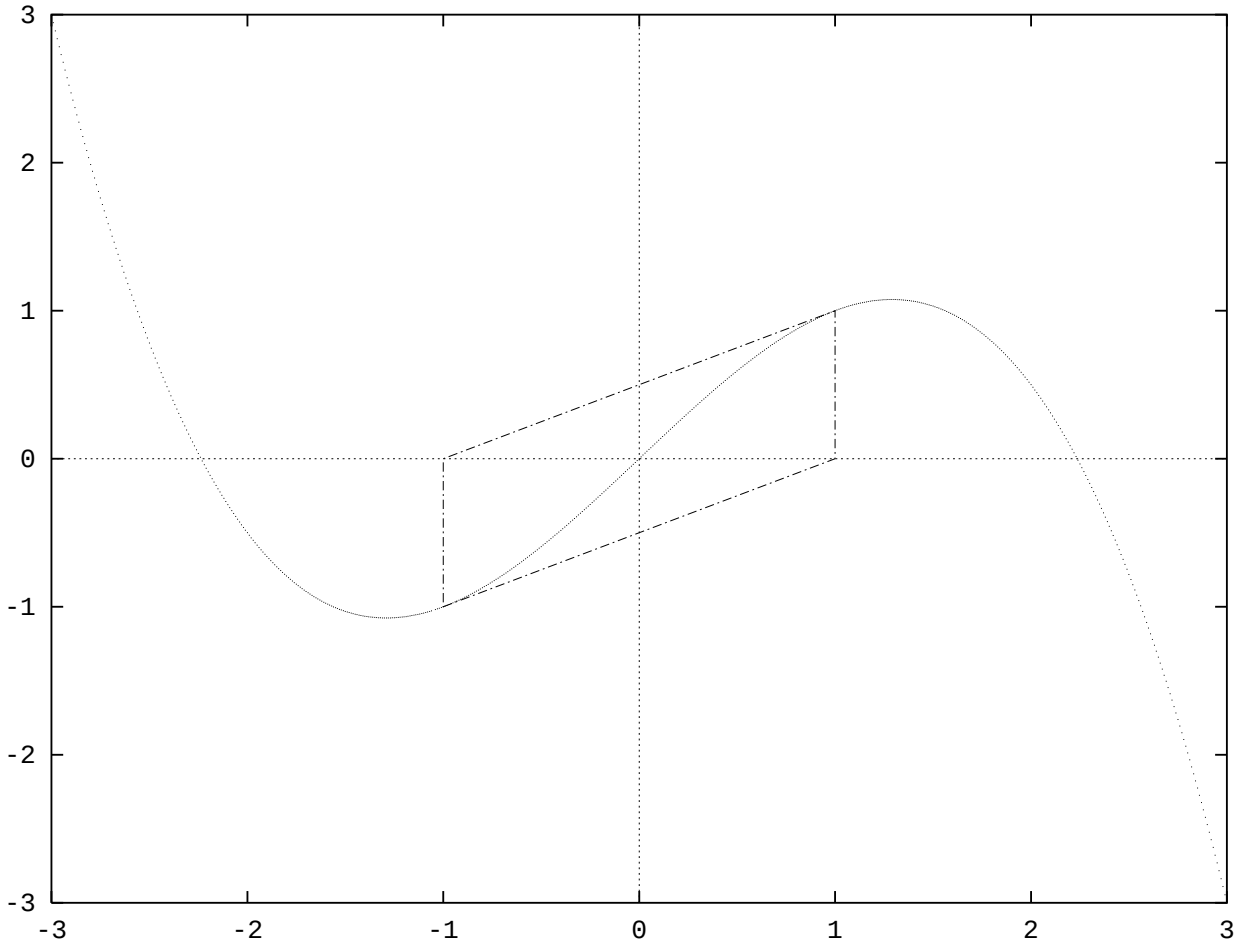
Niestety. . .

Metoda Newtona może być rozbieżna!

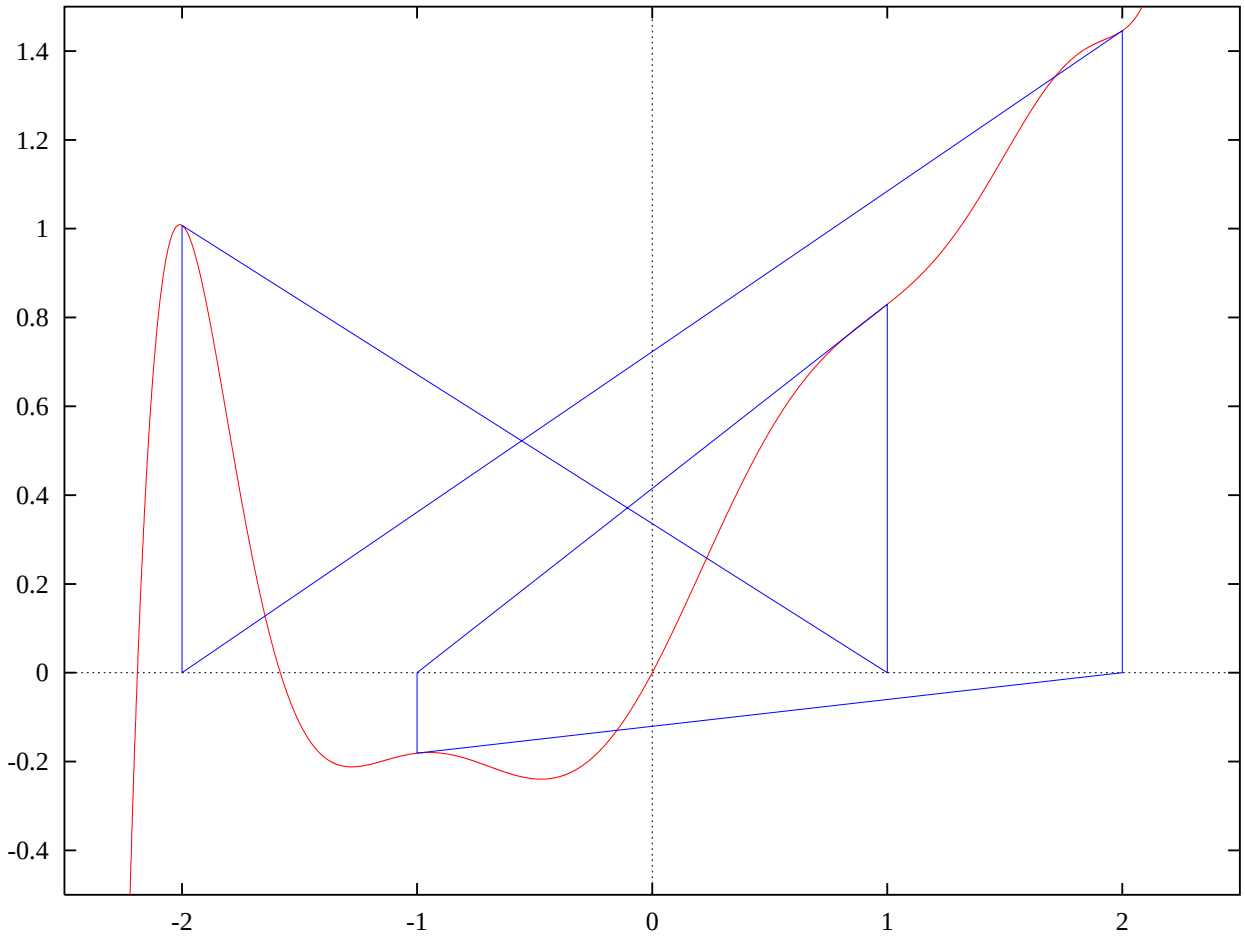


Metoda Newtona zawodzi w punktach, w których pochodna znika!

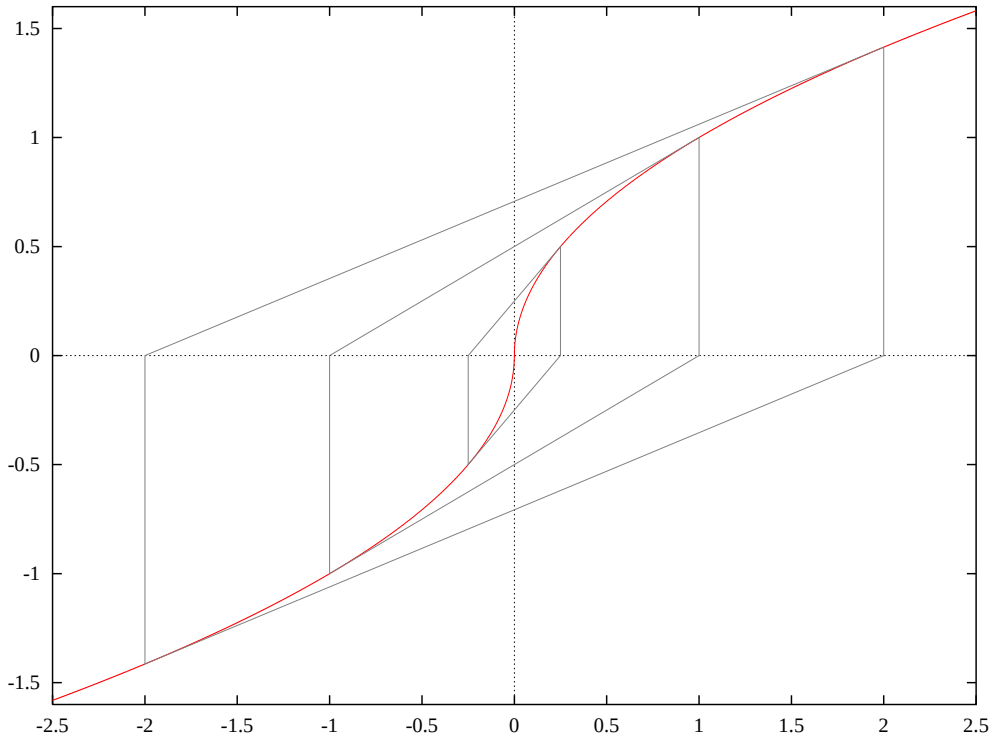
Metoda Newtona może prowadzić do cykli



Przykład czterocyklu



Inny przykład cyklu



Metoda Newtona zastosowana do funkcji $f(x) = \begin{cases} \sqrt{x} & x \geqslant 0 \\ -\sqrt{-x} & x < 0 \end{cases}$.

- Metoda Newtona jest zbieżna *kwadratowo* do jednokrotnych miejsc zerowych (liczba iteracji potrzebnych do ustalenia każdego kolejnego miejsca dziesiętnego zmniejsza się o połowę).
- Jako kryterium zbieżności, jak poprzednio, możemy *naiwnie* przyjąć warunek $|f(x_n)| < \varepsilon \ll 1$ lub — lepiej — że przedział, w którym znajduje się miejsce zerowe, jest już dostatecznie mały: $|x_n - x_{n-1}| < \varepsilon \ll 1$. Dodatkowo, ponieważ metoda może być rozbieżna, należy przyjąć maksymalną dopuszczalną liczbę iteracji.
- Metoda Newtona jest tym szybciej zbieżna, im bliżej poszukiwanego miejsca zerowego leży początkowe przybliżenie. Jeżeli początkowe przybliżenie jest "niedobre", metoda Newtona może zawieść.
- Metoda Newtona jest zbieżna liniowo do wielokrotnych miejsc zerowych.

- Metodę Newtona można łatwo uogólnić na przypadek zespolony, aczkolwiek iteracja (8) zastartowana z rzeczywistego punktu początkowego dla rzeczywistej funkcji $f(x)$ pozostaje rzeczywista.
- Istnienie wielocykli jest bardzo ciekawe, ale w praktyce nie stanowią one "numerycznego niebezpieczeństwa": zdecydowana większość wielocykli jest niestabilna i szanse na trafienie na taki wielocykl są równe zeru. Z drugiej strony można udowodnić*, że jeśli badaną funkcją jest wielomian o n_d różnych miejscach zerowych, a k jest liczbą pierwszą, metoda Newtona prowadzi do

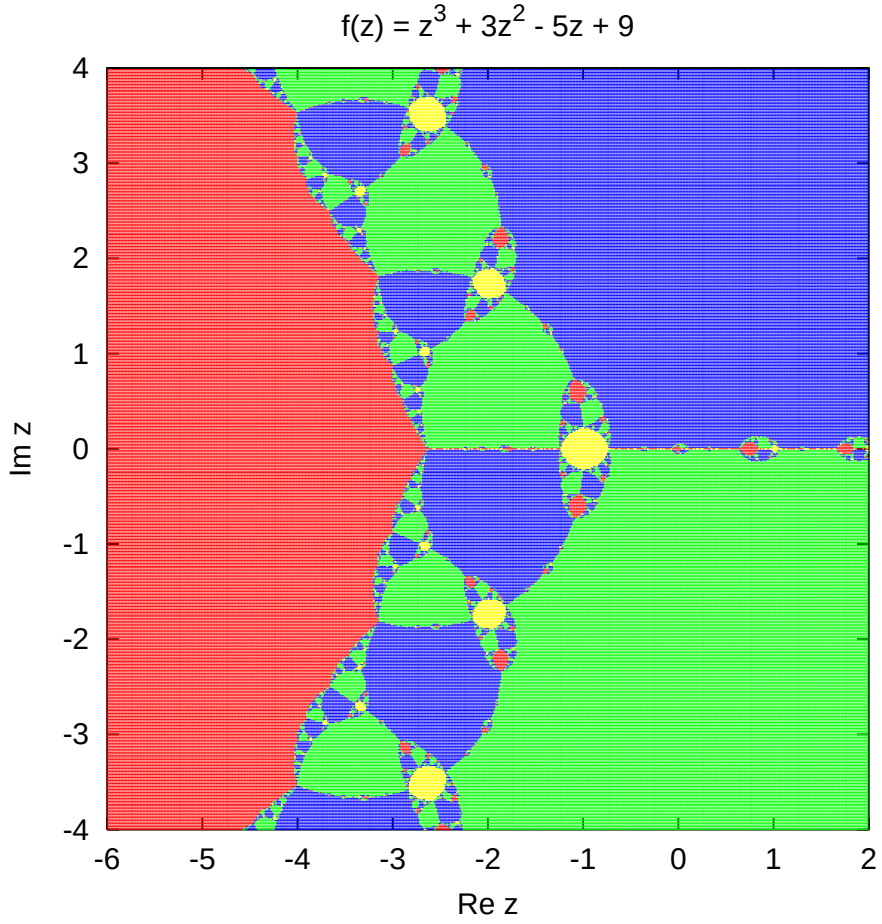
$$N_k = \frac{1}{k}\left(n_\mathrm{d}^k - n_\mathrm{d}\right) \tag{13}$$

różnych k-cykli.

*Ł. Skowronek, P. F. Góra, *Acta Phys. Pol.* B**38**, 1909 (2007)

- Basenem atrakcji jakiegoś miejsca zerowego nazywam zbiór punktów o tej własności, że metoda Newtona zastartowana z takiego punktu, prowadzi do wskazanego miejsca zerowego. Granice basenów atrakcji poszczególnych miejsc zerowych w metodzie Newtona na płaszczyźnie zespolonej bardzo często są *fraktalne*. Na tych właśnie granicach leżą te wszystkie niestabilne wielocykle, o których była mowa wyżej. Jeżeli natomiast jakiś wielocykl jest stabilny, numeryczne odkrycie go może być równie ważne, co znalezienie miejsca zerowego.

Wielomian ze stabilnym dwucyklem



Tłumiona metoda Newtona

W pewnych przypadkach — na przykład aby uciec ze stabilnego wielocyklu — zamiast metody Newtona (8) stosuje się *tłumioną metodę Newtona* (ang. *damped Newton method*)

$$x_{n_1} = x_n - \alpha \frac{f(x_n)}{f'(x_n)} \tag{14}$$

gdzie $\alpha \in (0, 1]$. Aby uciec z wielocyklu, na 2-3 kroki metodę Newtona zastępuje się metodą tłumioną.

Metody wykorzystujące drugą pochodną

Metoda Newtona opiera się na rozwinięciu Taylora (7) do pierwszego rzędu. Możemy to uogólnić na rozwinięcie do rzędu drugiego:

$$f(x_0+\delta) \simeq f(x_0) + \delta \cdot f'(x_0) + \frac{1}{2}\delta^2 \cdot f''(x_0)\,. \tag{15}$$

Jak poprzednio, żądamy, aby lewa strona znikała, co prowadzi do kroku

$$\delta = \frac{-f'(x_0) \pm \sqrt{\left[f'(x_0)\right]^2 - 2f(x_0)f''(x_0)}}{f''(x_0)}\,, \tag{16}$$

a dalej, po prostych przekształceniach, do iteracji

$$x_{n+1} = x_n - \frac{2f(x_n)}{f'(x_n) \pm \sqrt{\left[f'(x_n)\right]^2 - 2f(x_n)f''(x_n)}}\,. \tag{17}$$

Znak w mianowniku (17) wybieramy tak, aby moduł mianownika był większy. W odróżnieniu od metody Newtona, metoda (17) może prowadzić do zespolonych iteratów także dla rzeczywistych wartości początkowych.

Metoda Halleya

Inną metodę daje zastosowanie metody Newtona do równania

$$g(x) = \frac{f(x)}{\sqrt{|f'(x)|}} = 0\,. \tag{18}$$

Każdy pierwiastek $f(x)$, który *nie* jest miejscem zerowym pochodnej, jest pierwiastkiem $g(x)$; każdy pierwiastek $g(x)$ jest pierwiastkiem $f(x)$ (rozwiązaniem równania (1)). Po przekształceniach algebraicznych otrzymujemy iterację

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2\left[f'(x_n)\right]^2 - f(x_n)f''(x_n)} \tag{19a}$$

lub w postaci alternatywnej

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\left[1 - \frac{f(x_n)}{f'(x_n)}\cdot\frac{f''(x_n)}{2f'(x_n)}\right]^{-1}\,. \tag{19b}$$

Przykład

Skonstruujmy algorytm obliczania $\sqrt{z}\,, z > 0$. Poszukiwana liczba jest pierwiastkiem równania

$$x^2 - z = 0\,. \tag{20a}$$

Jeśli do równania (20a) zastosujemy metodę Newtona, otrzymamy iterację (zwaną nikiedy wzorem Herona)

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{z}{x_n}\right). \tag{20b}$$

Łatwo sprawdzić, że liczby $\pm\sqrt{z}$ są punktami stałymi iteracji (20b), a jeśli $x_1 > 0$, to $\forall n \geqslant 1 : x_n > 0$, a zatem starując z dodatniego warunku początkowego, zbiegniemy się do $+\sqrt{z}$.

Jak szybka będzie zbieżność? Jeśli $x_n = \sqrt{z} + \varepsilon$, $|\varepsilon| \ll 1$, rozwijając prawą stronę (20b) w szereg Taylora do najniższego nieznikającego rzędu otrzymamy

$$x_{n+1} \simeq \sqrt{z} + \frac{1}{4\sqrt{z}}\varepsilon^2 \,. \tag{20c}$$

Jest to, rzecz jasna, szczególny przypadek zależności (11), potwierdzający kwadratową zbieżność metody Newtona do jednokrotnych miejsc zerowych.

Spróbujmy teraz do obliczania pierwiastka zastosować metodę Halleya (19), gdzie $f(x) = x^2 - z$. Po przekształceniach algebraicznych otrzymujemy

$$x_{n+1} = x_n \cdot \frac{x_n^2 + 3z}{3x_n^2 + z} \,. \tag{20d}$$

Jak łatwo sprawdzić, jeśli $x_1 > 0$, to także wszystkie następne $x_n > 0$, a punkatmi stałymi iteracji (20d) są, jak poprzednio, liczby $\pm\sqrt{z}$.

Jak szybko zbieżna jest iteracja (19)? Zakładając, że $x_n = \sqrt{z} + \varepsilon$ i rozwijając (20d) w szerego Taylora do najniższego nieznikającego rzędu, otrzymujemy

$$x_{n+1} = \sqrt{z} + \frac{1}{4z}\varepsilon^3 \,. \tag{20e}$$

Otrzymujemy zbieżność sześcienną, a nie, jak w przypadku wzoru Herona (20b) kwadratową. Stosując metodę (20d) będziemy musieli wykonać *mniej* iteracji, niż stosując metodę (20b), aby obliczyć $\sqrt{z}$ z zadaną dokładnością.

Co to jednak oznacza w praktyce? Aby wykonać jedną iterację metodą (20b) potrzebujemy trzech operacji elementarnych (mnożenie, dzielenie, dodawanie). Żeby natiomiast wykonać jedną iterację (20d) potrzebujemy aż siedmiu operacji elementarnych. Zastosowanie metody (20d) będzie

oplacalne jeśli liczba iteracji spadnie ponad dwukrotnie w stosunku do metody (20b). Tak się jednak nie dzieje, gdy pożądana dokładność wyznaczenia pierwiastka jest $\sim 10^{-8}$: wykonujemy nieco mniej iteracji, ale ponieważ każda iteracja jest bardziej kosztowna, obliczenia będą trwały dłużej. Z drugiej strony przy bardzo dużych dokładnościach, $\sim 10^{-30}$ — takie dokładności niekiedy, choć raczej rzadko, mogą być wymagane w praktyce — skorzystanie z sześciennej zbieżności metody (20d) staje się bardzo opłacalne.

Układy równań algebraicznych

Niech $\mathrm{g}{:}\mathbb{R}^N \to \mathbb{R}^N$ będzie funkcją klasy co najmniej C^1. Rozważamy równanie

$$\mathrm{g}(\mathrm{x}) = 0\,, \tag{21}$$

formalnie równoważne układowi równań

$$\begin{aligned} g_1(x_1, x_2, \ldots, x_N) &= 0\,, && \text{(22a)} \\ g_2(x_1, x_2, \ldots, x_N) &= 0\,, && \text{(22b)} \\ &\ldots \\ g_N(x_1, x_2, \ldots, x_N) &= 0\,. && \text{(22c)} \end{aligned}$$

Rozwiązywanie układów równań algebraicznych jest trudne, gdyż geometrycznie oznacza znalezienie punktu (bądź punktów) przecięcia krzywych (22). O tych funkcjach na ogół nic nie wiemy, zmiana jednej nie wpływa na zmianę innej itd.

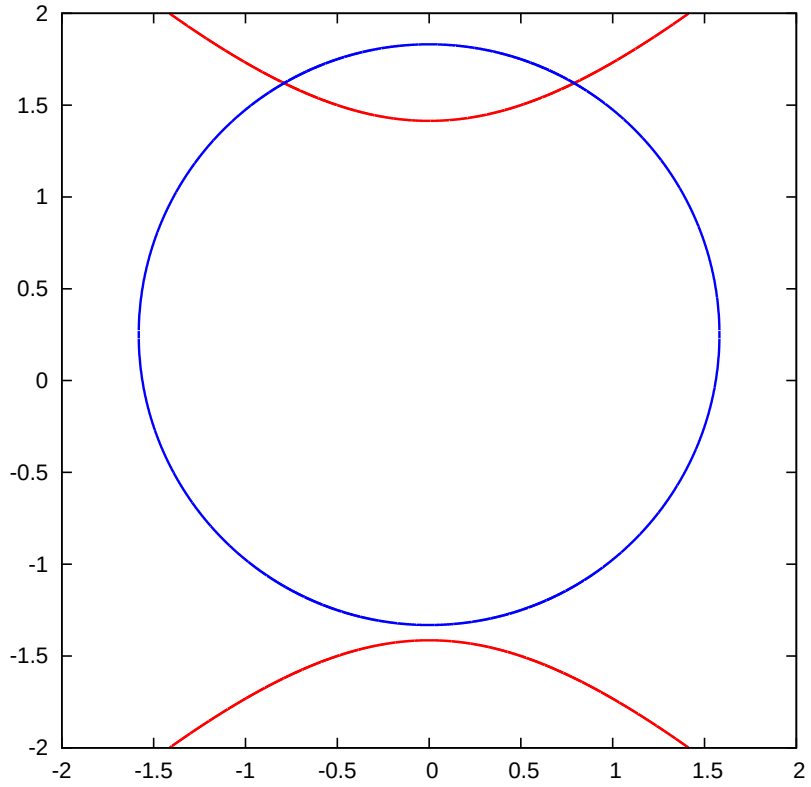
Przykład

W zależności od parametrów, układ równań

$$(x - x_0)^2 + (y - y_0)^2 - r^2 = 0 \quad (23a)$$

$$x^2 - y^2 - b^2 = 0 \quad (23b)$$

może mieć $0, 1, 2, 3$ lub 4 rozwiązania, co wiemy z "zainwestowania" do analizy układu (23) naszej wiedzy z zakresu krzywych stożkowych.



Interpretacja geometryczna układu równań

$$\begin{cases} x^2 + \left(y - \frac{1}{4}\right)^2 = \frac{5}{2} \\ y^2 - x^2 = 2 \end{cases}$$

Punkt leżący pomiędzy dolną gałęzią czerwonej hiperboli a niebieskim okręgiem odpowiada minimum *lokalnemu* funkcji G (patrz niżej).

Wielowymiarowa metoda Newtona

Rozwijając funkcję $\mathbf{g}$ w szereg Taylora do pierwszego rzędu otrzymamy

$$\mathbf{g}(\mathbf{x}+\delta\mathbf{x}) \simeq \mathbf{g}(\mathbf{x}) + \mathbf{J}\delta\mathbf{x}\,, \tag{24}$$

gdzie $\mathbf{J}$ jest jakobianem funkcji $\mathbf{g}$:

$$\mathbf{J}(\mathbf{x})_{ij} = \left.\frac{\partial g_i}{\partial x_j}\right|_{\mathbf{x}}\,. \tag{25}$$

Jaki krok $\delta\mathbf{x}$ musimy wykonać, aby znaleźć się w punkcie spełniającym równanie (21)? Żądamy aby $\mathbf{g}(\mathbf{x}+\delta\mathbf{x}) = 0$, skąd otrzymujemy

$$\delta\mathbf{x} = -\mathbf{J}^{-1}\mathbf{g}(\mathbf{x})\,. \tag{26}$$

Prowadzi to do następującej iteracji:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}^{-1}(\mathbf{x}_k)\mathbf{g}(\mathbf{x}_k)\,. \tag{27}$$

Oczywiście zapis $\mathbf{z} = \mathbf{J}^{-1}\mathbf{g}$ należy rozumieć w ten sposób, że $\mathbf{z}$ spełnia równanie $\mathbf{J}\mathbf{z} = \mathbf{g}$. *Nie należy konstruować jawnej odwrotności jakobianu.*

Uwaga: W metodzie (27) jakobian trzeba obliczać w każdym kroku. Oznacza to, że w każdym kroku trzeba rozwiązywać *inny* układ równań liniowych, co czyni metodę dość kosztowną, zwłaszcza jeśli N (wymiar problemu) jest znaczne. Często dla przyspieszenia obliczeń macierz $\mathbf{J}$ zmieniamy nie co krok, ale co kilka kroków — pozwala to użyć tej samej faktoryzacji $\mathbf{J}$ do rozwiązania kilku kolejnych równań $\mathbf{J}\mathbf{z} = \mathbf{g}(\mathbf{x}_k)$. Jest to dodatkowe uproszczenie, ale jest ono bardzo wydajne przy $N \gg 1$.

Rozwiązywanie równań nieliniowych a minimalizacja

Metoda Newtona czasami zawodzi ☹. Ponieważ rozwiązywanie równań algebraicznych jest “trudne”, natomiast minimalizacja jest “łatwa”, niektórzy skłonni są rozważać funkcję $G{:}\mathbb{R}^N \to \mathbb{R}$

$$G(\mathbf{x}) = \frac{1}{2} \left\| \mathbf{g}(\mathbf{x}) \right\|^2 = \frac{1}{2} \left(\mathbf{g}(\mathbf{x}) \right)^T \mathbf{g}(\mathbf{x}) \tag{28}$$

i szukać jej minimum zamiast rozwiązywać (21). *Globalne* minimum $G = 0$ odpowiada co prawda rozwiązaniu (21), jednak G może mieć wiele minimów lokalnych, *nie mamy także gwarancji*, że globalne minimum $G = 0$ istnieje. Nie jest to więc dobry pomysł.

Metoda globalnie zbieżna

Rozwiązaniem jest połączenie idei minimalizacji funkcji (28) i metody Newtona. Przypuśćmy, iż chcemy rozwiązywać równanie (21) metodą Newtona. Krok iteracji wynosi

$$\delta \mathbf{x} = -\mathbf{J}^{-1}\mathbf{g}\,. \tag{29}$$

Z drugiej strony mamy

$$\frac{\partial G}{\partial x_i} = \frac{1}{2}\sum_j \left(\frac{\partial g_j}{\partial x_i} g_j + g_j \frac{\partial g_j}{\partial x_i} \right) = \sum_j J_{ji} g_j \tag{30}$$

a zatem $\nabla G = \mathbf{J}^T \mathbf{g}$.

Jak zmienia się funkcja G (28) po wykonaniu kroku Newtona (29)?

$$(\nabla G)^T \, \delta\mathbf{x} = \mathbf{g}^T \mathbf{J} \left(-\mathbf{J}^{-1}\right) \mathbf{g} = -\mathbf{g}^T \mathbf{g} < 0 \,, \tag{31}$$

a zatem *kierunek kroku Newtona jest* lokalnym *kierunkiem spadku* G. Jednak przesunięcie się o pełną długość kroku Newtona nie musi prowadzić do spadku G. Postępujemy wobec tego jak następuje:

1. $w = 1$. Oblicz $\delta\mathbf{x}$.
2. $\mathbf{x}_{\text{test}} = \mathbf{x}_i + w\,\delta\mathbf{x}$.
3. Jeśli $G(\mathbf{x}_{\text{test}}) < G(\mathbf{x}_i)$, to
   (a) $\mathbf{x}_{i+1} = \mathbf{x}_{\text{test}}$
   (b) *goto* 1
4. Jeśli $G(\mathbf{x}_{\text{test}}) > G(\mathbf{x}_i)$, to
   (a) $w \to w/2$
   (b) *goto* 2

Jest to zatem forma *tłumionej* (damped) *metody Newtona*.

Zamiast połowienia kroku, można używać innych strategii poszukiwania w prowadzących do zmniejszenia się wartości G.

Jeśli wartość w spadnie poniżej pewnego akceptowalnego progu, obliczenia należy przerwać, jednak (31) gwarantuje, że istnieje takie w, iż $w\,\delta\mathbf{x}$ prowadzi do zmniejszenia się G.

Powyższa metoda jest zawsze zbieżna do *jakiegoś* minimum funkcji G, ale niekoniecznie do jej minimum globalnego, czyli do rozwiązania równania (21).

Jeżeli znajdziemy minimum lokalne $G_{\mathsf{min}} > \varepsilon$, gdzie $\varepsilon > 0$ jest pożądaną tolerancją, należy spróbować rozpocząć z innym warunkiem początkowym. Jeżeli kilka różnych warunków początkowych nie daje rezultatu, należy się poddać.

Szansa na znalezienie numerycznego rozwiązania układu równań (21) jest tym większa, im lepszy jest warunek początkowy. Należy wobec tego zainwestować całą naszą wiedzę o funkcji g w znalezienie warunku początkowego; analogiczna uwaga obowiązuje w wypadku stosowania wielowymiarowej metody Newtona (27).

Bardzo ważna uwaga

Wszystkie przedstawione tu metody wymagają znajomości analitycznych wzorów na pochodne odpowiednich funkcji.

Używanie powyższych metod w sytuacji, w których pochodne należy aproksymować numerycznie, *na ogół nie ma sensu*.

Wielowymiarowa metoda siecznych — metoda Broydena

Niekiedy analityczne wzory na pochodne są nieznane, niekiedy samo obliczanie jakobianu, wymagające obliczenia N^2 pochodnych cząstkowych, jest numerycznie zbyt kosztowne. W takich sytuacjach *czasami* używa się metody zwanej niezbyt ściśle "wielowymiarową metodą siecznych". Podobnie jak w przypadku jednowymiarowym, gdzie pochodną zastępuje się ilorazem różnicowym

$$g'(x_{i+1}) \simeq \frac{g(x_{i+1}) - g(x_i)}{x_{i+1} - x_i}\,, \tag{32}$$

jakobian w kroku Newtona zastępujemy wyrażeniem przybliżonym: Zamiast $\mathbf{J}\,\delta\mathbf{x} = -\mathbf{g}(\mathbf{x})$ bierzemy $\mathbf{B}\,\Delta\mathbf{x} = -\Delta\mathbf{g}$. Macierz $\mathbf{B}$ jest przybliżeniem jakobianu, poprawianym w każdym kroku iteracji. Otrzymujemy zatem

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{B}_i^{-1}\mathbf{g}(\mathbf{x}_i)\,, \tag{33}$$

natomiast poprawki $\mathbf{B}$ obliczamy jako

$$\mathbf{B}_{i+1} = \mathbf{B}_i + \frac{(\Delta\mathbf{g}_i - \mathbf{B}_i\Delta\mathbf{x}_i)\,(\Delta\mathbf{x}_i)^T}{(\Delta\mathbf{x}_i)^T\,\Delta\mathbf{x}_i}\,, \tag{34}$$

gdzie $\Delta\mathbf{x}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$, $\Delta\mathbf{g}_i = \mathbf{g}(\mathbf{x}_{i+1}) - \mathbf{g}(\mathbf{x}_i)$. Ponieważ poprawka do $\mathbf{B}_i$ ma postać iloczynu diadycznego dwu wektorów, do obliczania† $\mathbf{B}_{i+1}^{-1}\mathbf{g}(\mathbf{x}_{i+1})$ można skorzystać ze wzoru Shermana-Morrisona.

Metoda ta wymaga inicjalizacji poprzez podanie $\mathbf{B}_1$ oraz wektora $\mathbf{x}_1$. To drugie nie jest niczym dziwnym; co do pierwszego, jeśli to tylko możliwe, można przyjąć $\mathbf{B}_1 = \mathbf{J}(\mathbf{x}_1)$.

†Czyli tak naprawdę do rozwiązywania pewnego układu równań liniowych!