

Wstęp do metod numerycznych

9. Rozwiązywanie równań algebraicznych

P. F. Góra

<http://th-www.if.uj.edu.pl/zfs/gora/>

2015

Co to znaczy rozwiązać równanie?

Przypuśmy, że postawiono przed nami problem rozwiązania równania

$$f(x) = 0. \quad (1)$$

Przed 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 jakiegolwiek rozwiązanie (1) istnieje, a jeśli tak, to ile ich jest. **Dlatego w przypadku ogólnym zadowolamy 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 $k-1$: $f(x_0) = f'(x_0) = f''(x_0) = \dots = f^{(k-1)}(x_0) = 0$. 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 $x = 0$.

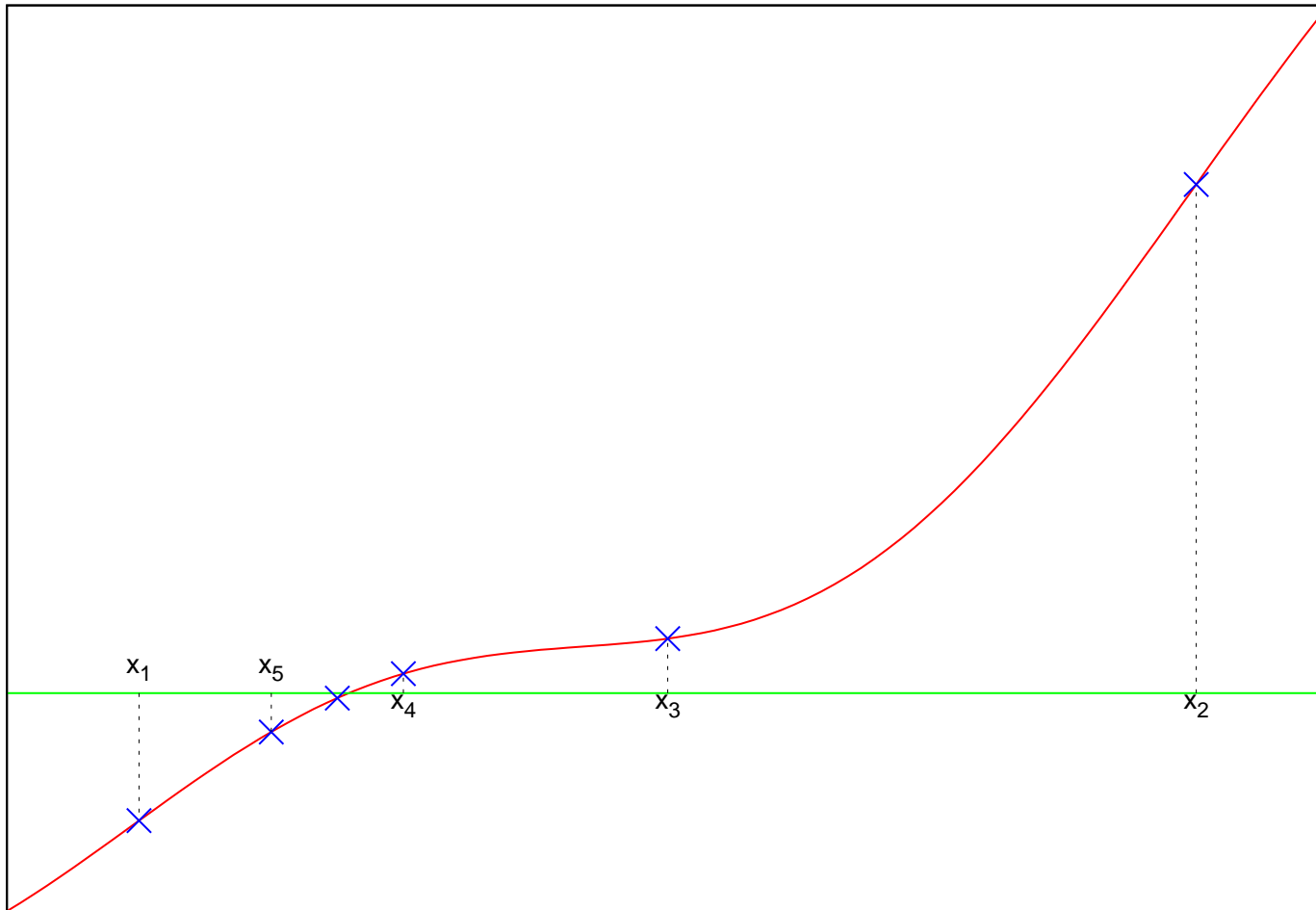
Funkcja zmienia znak w otoczeniu miejsca zerowego o krotności nieparzystej i *nie zmienia znaku* w otoczeniu miejsca zerowego o krotności parzystej.

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 znajdziemy x_n takie, że $|f(x_n)| \leq \varepsilon$, gdzie ε jest zadaną dokładnością poszukiwania rozwiązania równania (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.

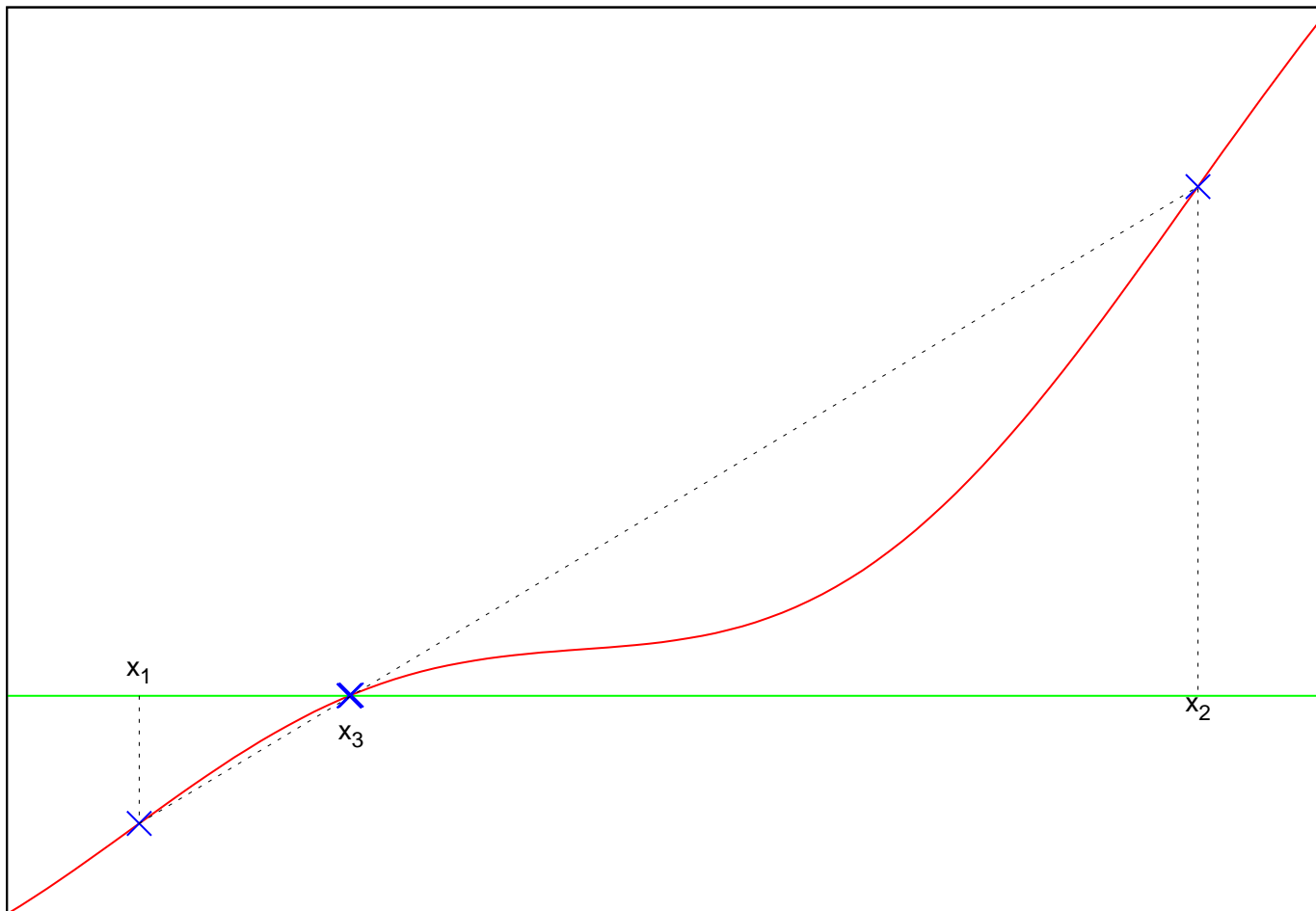


Metoda *regula falsi*

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 :

$$x_3 = \frac{f(x_1)x_2 - f(x_2)x_1}{f(x_1) - f(x_2)}. \quad (2)$$

Jeżeli $|f(x_3)| \leq \varepsilon$ (ε jak poprzednio), kończymy procedurę. Jeżeli nie, wybieramy ten z przedziałów $[x_1, x_3]$, $[x_3, x_2]$, **w którym funkcja zmienia znak** i postępujemy analogicznie.



Metoda siecznych

Metoda siecznych jest nągminnie 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 (2). 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 $x_1 = 0.8$, $x_2 = 1.2$, metody zbiegały się do $|f(0.879312)| \leq 10^{-6}$, przy czym liczba kroków wyniosła odpowiednio

metoda	kroków
bisekcji	17
<i>regula falsi</i>	8
siecznych	4

Interpolacja odwrotna

Przypuśćmy, że mamy stabelaryzowane wartości funkcji w węzłach:

$$\begin{array}{c|c|c|c|c|c} x_i & x_1 & x_2 & x_3 & \dots & x_n \\ \hline f_i = f(x_i) & f_1 & f_2 & f_3 & \dots & f_n \end{array} \quad (3)$$

przy czym — ważne! — stabelaryzowane wartości są **ściśle monotoniczne**, $f_1 > f_2 > \dots > f_n$ (lub $f_1 < f_2 < \dots < f_n$). Skoro funkcja jest monotoniczna, jest odwracalna, przy czym “węzły” i “wartości” zamieniają się miejscami:

$$\begin{array}{c|c|c|c|c|c} f_i & f_1 & f_2 & f_3 & \dots & f_n \\ \hline x_i = f^{-1}(f_i) & x_1 & x_2 & x_3 & \dots & x_n \end{array} \quad (4)$$

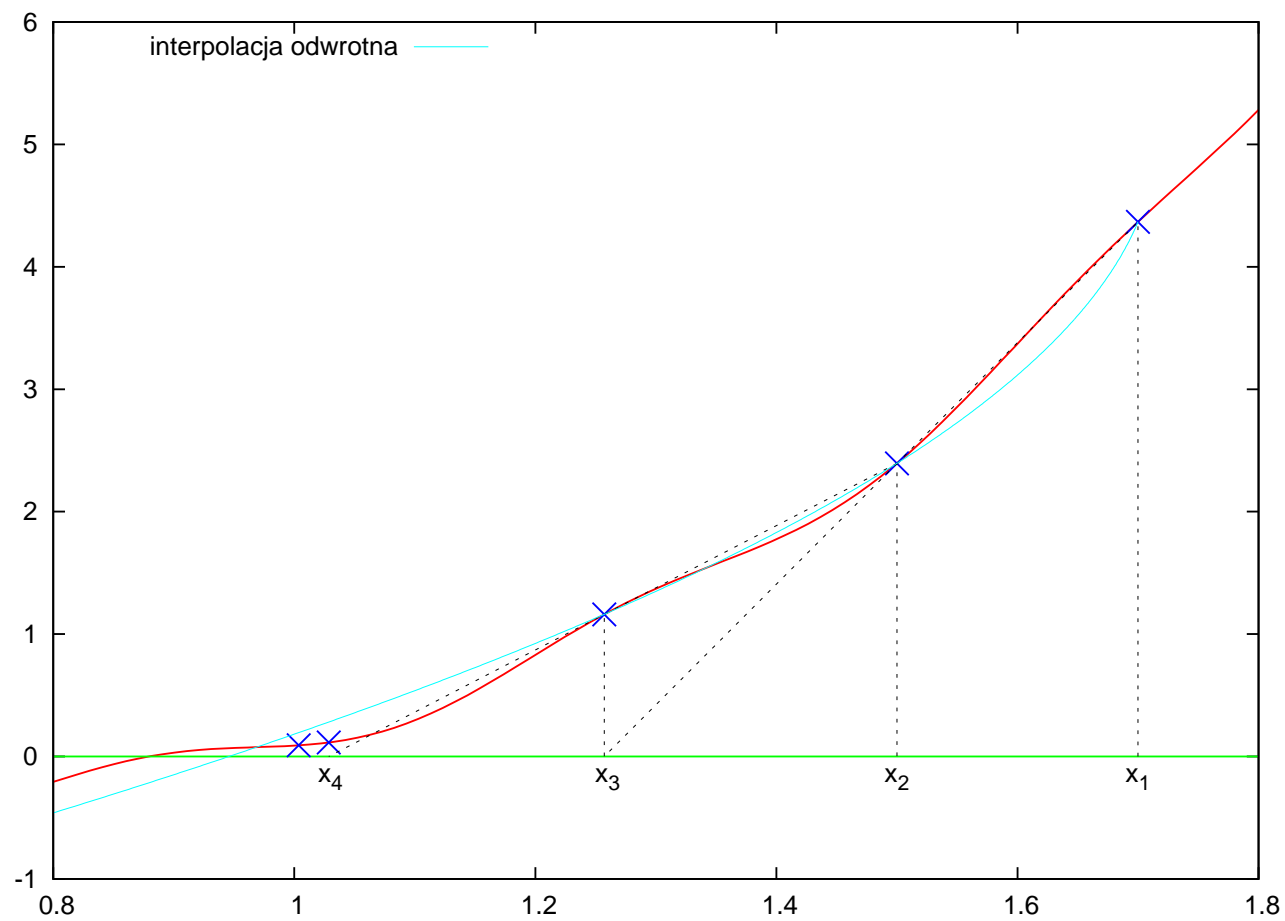
Wartość funkcji odwrotnej w zerze oznacza punkt, w którym funkcja ma **miejsce zerowe!** Aby znaleźć przybliżone miejsce zerowe funkcji $f(x)$,

tworzymy wielomian interpolacyjny według tabeli (4) i obliczamy wartość tego wielomianu, czyli przybliżenia funkcji odwrotnej, w zerze.

Ze względów praktycznych interpolację odwrotną stosuje się dla niewielkiej liczby węzłów. Wynik interpolacji odwrotnej może służyć jako punkt startowy innych, bardziej dokładnych metod.

Jeżeli prowadzimy interpolację odwrotną opartą o dwa punkty, jest to równoważne jednemu krokowi metody siecznych.

Metoda siecznych i interpolacja odwrotna



Metoda Newtona

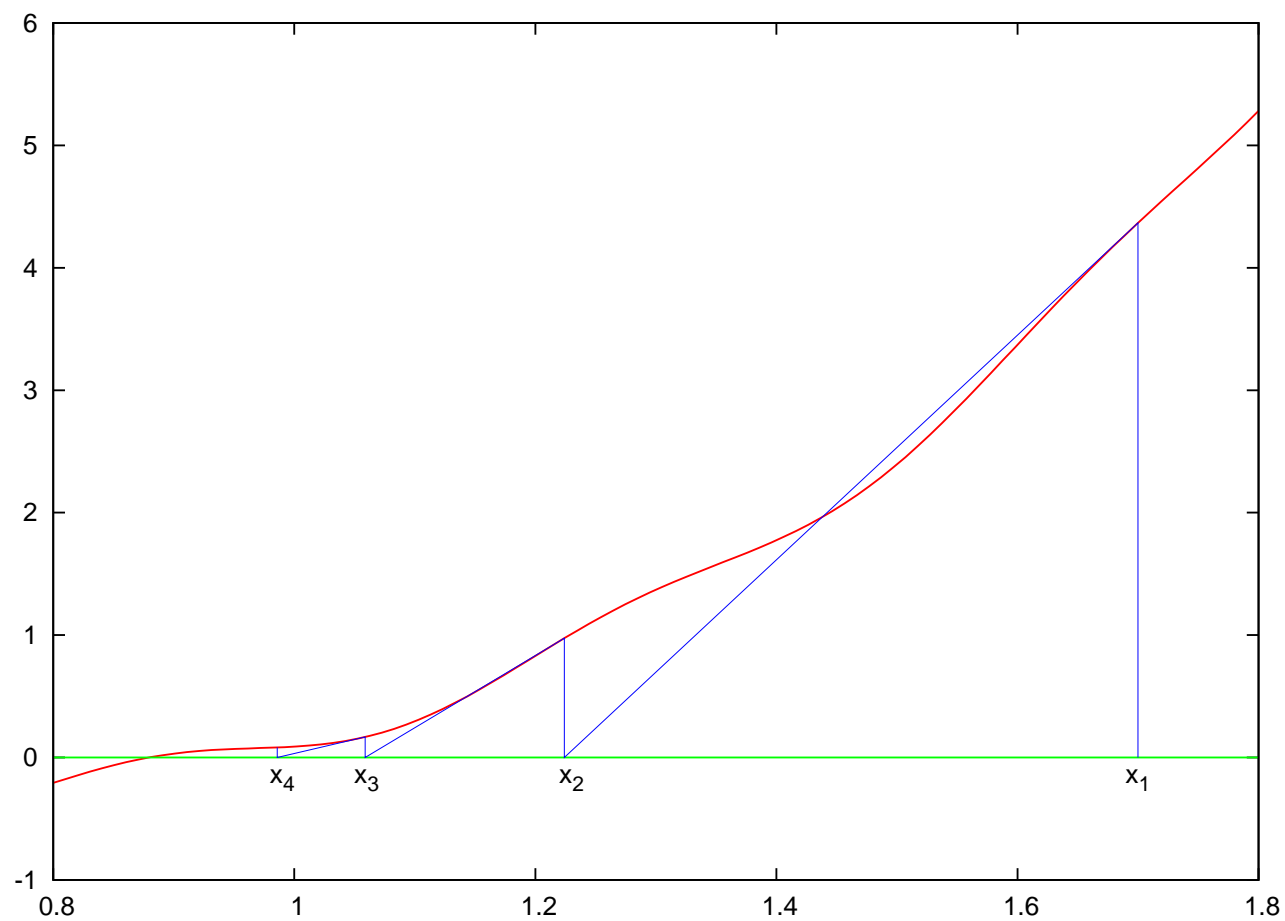
Przypuśćmy, że prawa strona równania (1) jest różniczkowalna. Rozwijamy tę funkcję w szereg Taylora wokół pewnego punktu

$$f(x_0 + \delta) \simeq f(x_0) + \delta \cdot f'(x_0) \quad (5)$$

a następnie **żądamy, aby lewa strona rozwinięcia (5) zniknęła**. 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

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (6)$$

Interpretacja geometryczna metody Newtona — metoda stycznych



Kryterium zbieżności

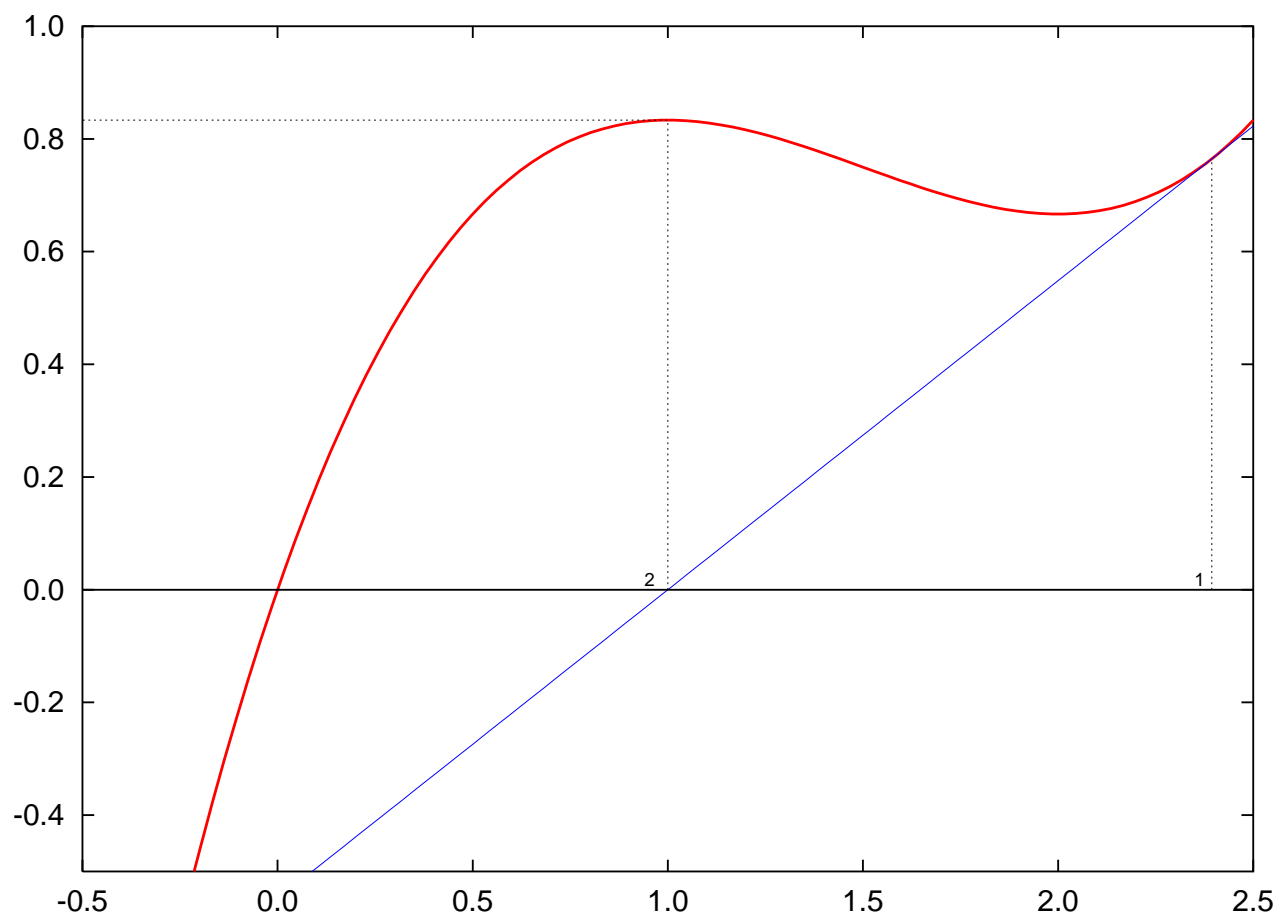
Zauważmy, że punkt stały iteracji (6) jest rozwiązaniem równania $f(x) = 0$. Formalnie, jeżeli funkcja

$$g(x) = x - \frac{f(x)}{f'(x)} \quad (7)$$

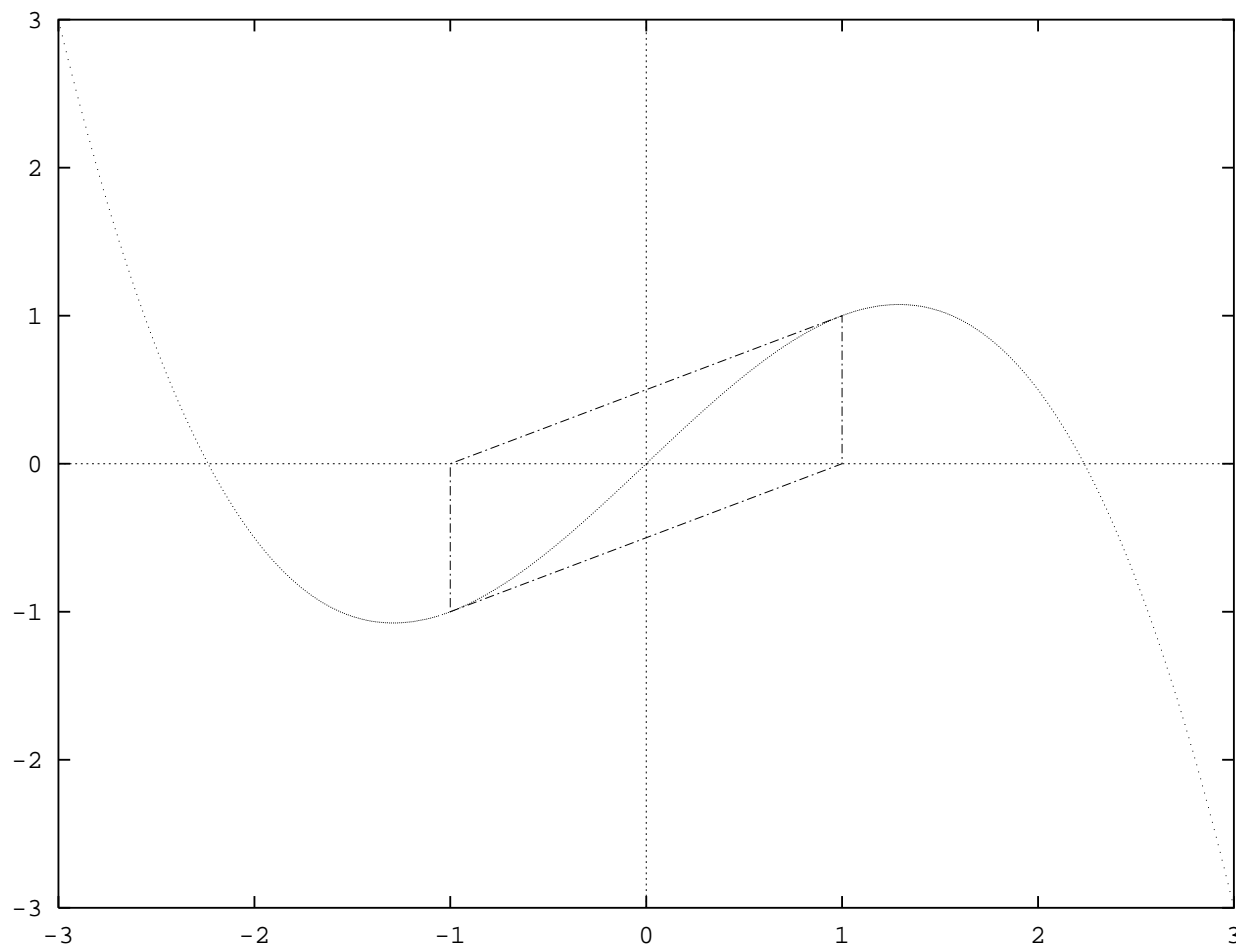
(i) jest ciągła oraz (ii) przeprowadza pewien przedział domknięty $[a, b]$ w ten sam przedział domknięty $[a, b]$, to na mocy twierdzenia Brouwera iteracja (6) ma w tym przedziale punkt stały, będący rozwiązaniem równania (1).

Problem leży w spełnieniu warunku (ii)

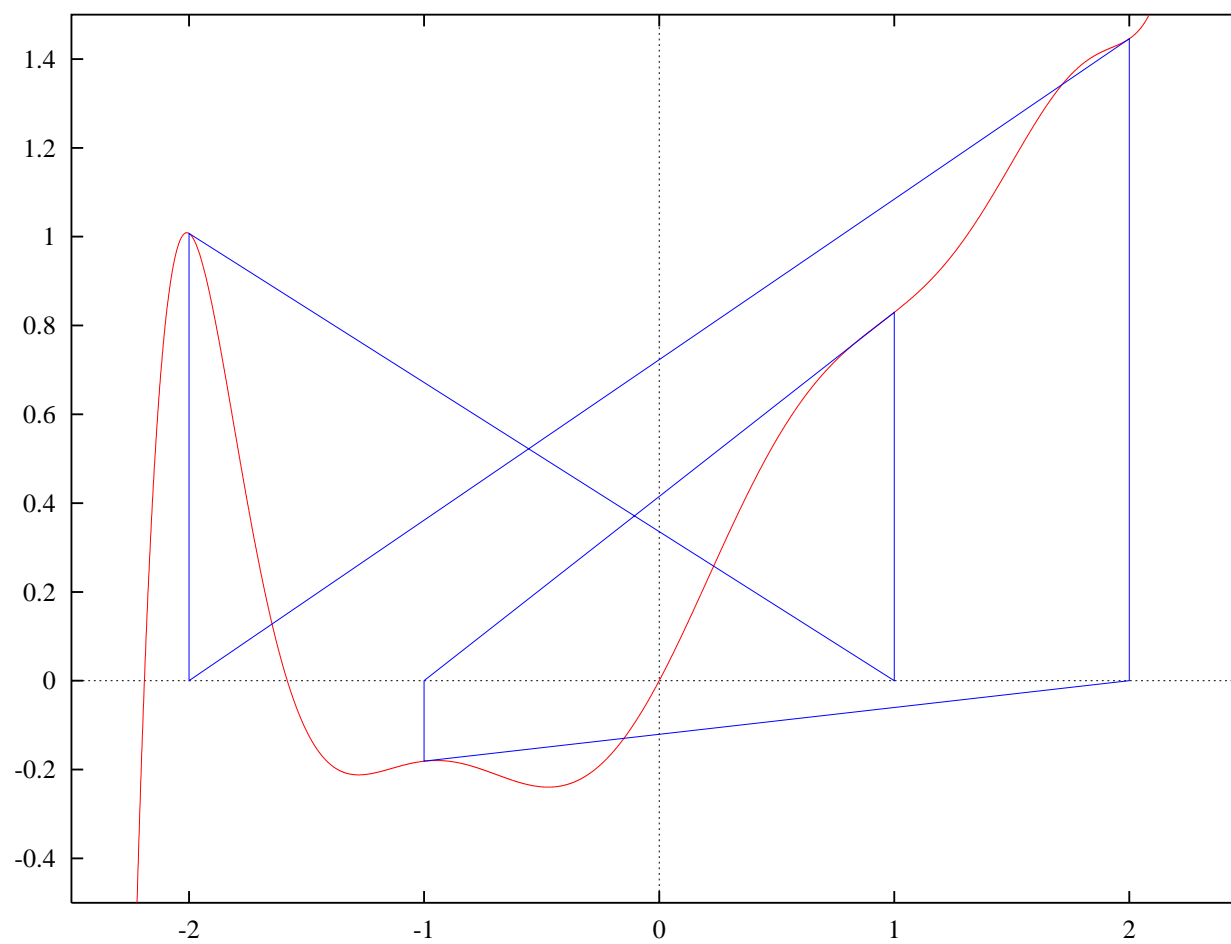
Metoda Newtona może być rozbieżna!



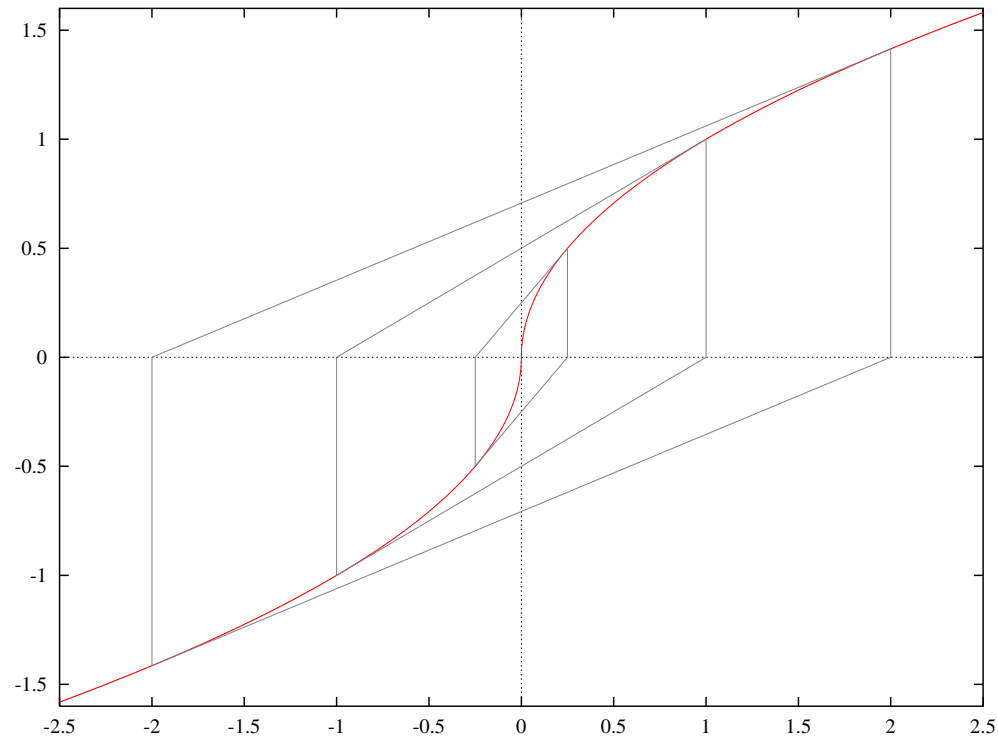
Metoda Newtona może prowadzić do cykli



Przykład czterocyklu



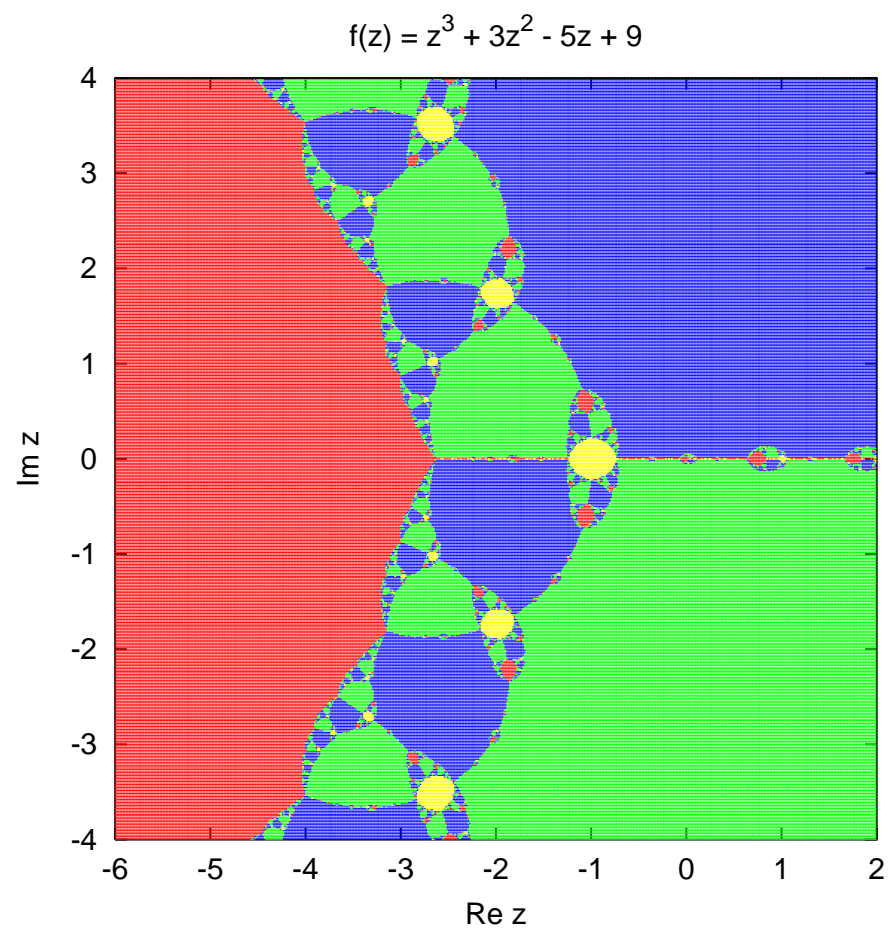
Inny przykład cyklu



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

- Metoda Newtona jest zbieżna *kwadratowo* do jednokrotnych miejsc zerowych (liczba iteracji potrzebnych do ustalenia każdego kolejnego miejsca dziesiętnego zmniejsza się o połowę).
- *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 (6) zainicjowana z rzeczywistego punktu początkowego dla rzeczywistej funkcji $f(x)$ pozostaje rzeczywista.
- Granice basenów atrakcji poszczególnych miejsc zerowych w metodzie Newtona na płaszczyźnie zespolonej bardzo często są *fraktalne*.

Wielomian ze stabilnym dwucyklem



Tłumiona metoda Newtona

W pewnych przypadkach — na przykład aby uciec ze stabilnego wielocyklu — zamiast metody Newtona (6) stosuje się *tłumioną metodę Newtona* (ang. *damped Newton method*)

$$x_{n_1} = x_n - \alpha \frac{f(x_n)}{f'(x_n)} \quad (8)$$

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 (5) do pierwszego rzędu. Możemy to uogólnić na rozwinięcie do rzędu drugiego:

$$f(x_0 + \delta) \simeq f(x_0) + \delta \cdot f'(x_0) + \frac{1}{2}\delta^2 \cdot f''(x_0). \quad (9)$$

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

$$\delta = \frac{-f'(x_0) \pm \sqrt{[f'(x_0)]^2 - 2f(x_0)f''(x_0)}}{f''(x_0)}, \quad (10)$$

a dalej, po prostych przekształceniach, do iteracji

$$x_{n+1} = x_n - \frac{2f(x_n)}{f'(x_n) \pm \sqrt{[f'(x_n)]^2 - 2f(x_n)f''(x_n)}}. \quad (11)$$

Znak w mianowniku (11) wbieramy tak, aby moduł mianownika był **więk-**
szy. W odróżnieniu od metody Newtona, metoda (11) może prowadzić do zespolonych iteratów także dla rzeczywistych wartości początkowych.

Metoda Halleya

Inną metodę daje zastosowanie metody Newtona do równania

$$g(x) = \frac{f(x)}{\sqrt{|f'(x)|}} = 0. \quad (12)$$

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

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2[f'(x_n)]^2 - f(x_n)f''(x_n)} \quad (13a)$$

lub w postaci alternatywnej

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \left[1 - \frac{f(x_n)}{f'(x_n)} \cdot \frac{f''(x_n)}{2f'(x_n)} \right]^{-1}. \quad (13b)$$

Układy równań algebraicznych

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

$$g(\mathbf{x}) = 0, \quad (14)$$

formalnie równoważne układowi równań

$$g_1(x_1, x_2, \dots, x_N) = 0, \quad (15a)$$

$$g_2(x_1, x_2, \dots, x_N) = 0, \quad (15b)$$

...

$$g_N(x_1, x_2, \dots, x_N) = 0. \quad (15c)$$

Rozwiązywanie układów równań algebraicznych jest trudne, gdyż geometrycznie oznacza znalezienie punktu (bądź punktów) przecięcia krzywych (15). O tych funkcjach na ogół nic nie wiemy, zmiana jednej nie wpływa na zmianę innej itd.

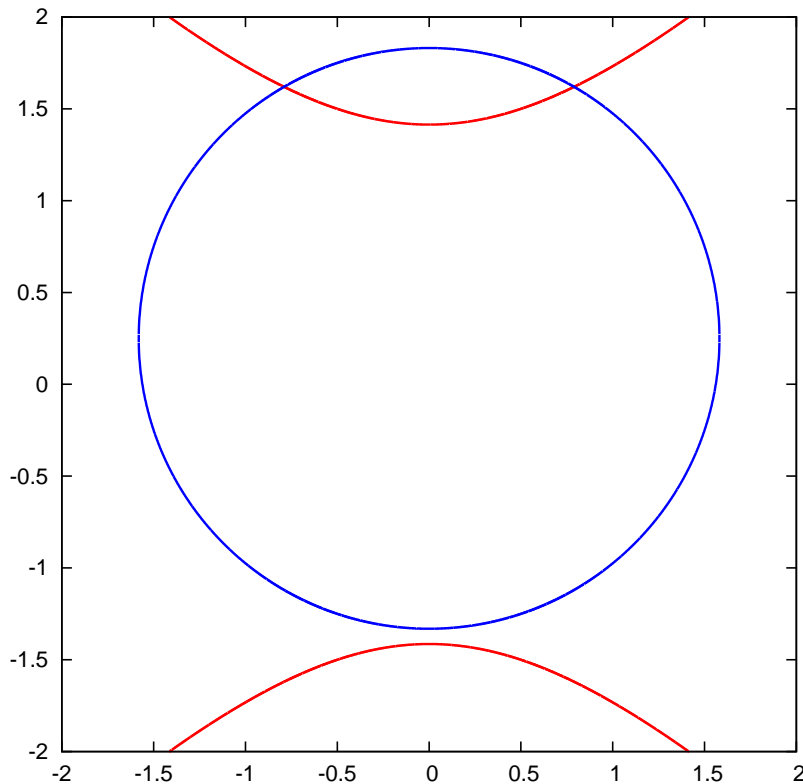
Przykład

W zależności od parametrów, układ równań

$$(x - x_0)^2 + (y - y_0)^2 - r^2 = 0 \quad (16a)$$

$$x^2 - y^2 - b^2 = 0 \quad (16b)$$

może mieć 0, 1, 2, 3 lub 4 rozwiązania, co wiemy z “zainwestowania” do analizy układu (16) 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).

Metoda Newtona

Rozwijając funkcję g w szereg Taylora do pierwszego rzędu otrzymamy

$$g(\mathbf{x} + \delta\mathbf{x}) \simeq g(\mathbf{x}) + \mathbf{J}\delta\mathbf{x}, \quad (17)$$

gdzie \mathbf{J} jest jakobianem funkcji g :

$$\mathbf{J}(\mathbf{x})_{ij} = \left. \frac{\partial g_i}{\partial x_j} \right|_{\mathbf{x}}. \quad (18)$$

Jaki krok $\delta\mathbf{x}$ musimy wykonać, aby znaleźć się w punkcie spełniającym równanie (14)? **Żądamy aby $g(\mathbf{x} + \delta\mathbf{x}) = 0$** , skąd otrzymujemy

$$\delta\mathbf{x} = -\mathbf{J}^{-1}\mathbf{g}(\mathbf{x}). \quad (19)$$

Prowadzi to do następującej iteracji:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{J}^{-1}(\mathbf{x}_k)\mathbf{g}(\mathbf{x}_k). \quad (20)$$

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 jacobianu.*

Uwaga: W metodzie (20) jacobian 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 $G: \mathbb{R}^N \rightarrow \mathbb{R}$

$$G(\mathbf{x}) = \frac{1}{2} \|\mathbf{g}(\mathbf{x})\|^2 = \frac{1}{2} (\mathbf{g}(\mathbf{x}))^T \mathbf{g}(\mathbf{x}) \quad (21)$$

i szukać jej minimum zamiast rozwiązywać (14). *Globalne* minimum $G = 0$ odpowiada co prawda rozwiązaniu (14), jednak G może mieć wiele minimumów 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 (21) i metody Newtona. Przypuśćmy, iż chcemy rozwiązywać równanie (14) metodą Newtona. Krok iteracji wynosi

$$\delta \mathbf{x} = -\mathbf{J}^{-1} \mathbf{g}. \quad (22)$$

Z drugiej strony mamy

$$\frac{\partial G}{\partial x_i} = \frac{1}{2} \sum_j \left(\frac{\partial g_j}{\partial x_i} g_j + g_j \frac{\partial g_j}{\partial x_i} \right) = \sum_j J_{ji} g_j \quad (23)$$

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

Jak zmienia się funkcja G (21) po wykonaniu kroku Newtona (22)?

$$(\nabla G)^T \delta \mathbf{x} = \mathbf{g}^T \mathbf{J} \left(-\mathbf{J}^{-1} \right) \mathbf{g} = -\mathbf{g}^T \mathbf{g} < 0, \quad (24)$$

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) $w \rightarrow w/2$
 - (b) *goto* 2

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

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

Jeśli wartość w spadnie poniżej pewnego akceptowalnego progu, obliczenia należy przerwać, jednak (24) 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 (14).

Jeżeli znajdziemy minimum lokalne $G_{\min} > \varepsilon$, 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 (14) jest tym większa, **im lepszy jest warunek początkowy**. Należy wobec tego zainvestować całą naszą wiedzę o funkcji g w znalezienie warunku początkowego; analogiczna uwaga obowiązuje w wypadku stosowania wielowymiarowej metody Newtona (20).

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 jacobianu, 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

$$g'(x_{i+1}) \simeq \frac{g(x_{i+1}) - g(x_i)}{x_{i+1} - x_i}, \quad (25)$$

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 jacobianu, poprawianym w każdym kroku iteracji. Otrzymujemy zatem

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \mathbf{B}_i^{-1} \mathbf{g}(\mathbf{x}_i), \quad (26)$$

natomiast poprawki \mathbf{B} obliczamy jako

$$\mathbf{B}_{i+1} = \mathbf{B}_i + \frac{(\Delta \mathbf{g}_i - \mathbf{B}_i \Delta \mathbf{x}_i) (\Delta \mathbf{x}_i)^T}{(\Delta \mathbf{x}_i)^T \Delta \mathbf{x}_i}, \quad (27)$$

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!**