

## **WYKŁAD 4**

# **PODSTAWOWE METODY PRZYBLIŻONEGO ROZWIĄZYWANIA NIELINIOWYCH RÓWNAŃ ALGEBRAICZNYCH**

## MOTYWACJA

Wykład nr 4 jest poświęcony omówieniu elementarnych algorytmów wyznaczania przybliżonych rozwiązań (pierwiastków) nieliniowego równania algebraicznego postaci

$$f(x) = 0$$

gdzie symbol  $f$  oznacza zadaną funkcję. Zwykle rozwiązania takiego równania nie mogą być wyznaczone ściśle metodami analitycznymi. Oto kilka przykładów takich równań

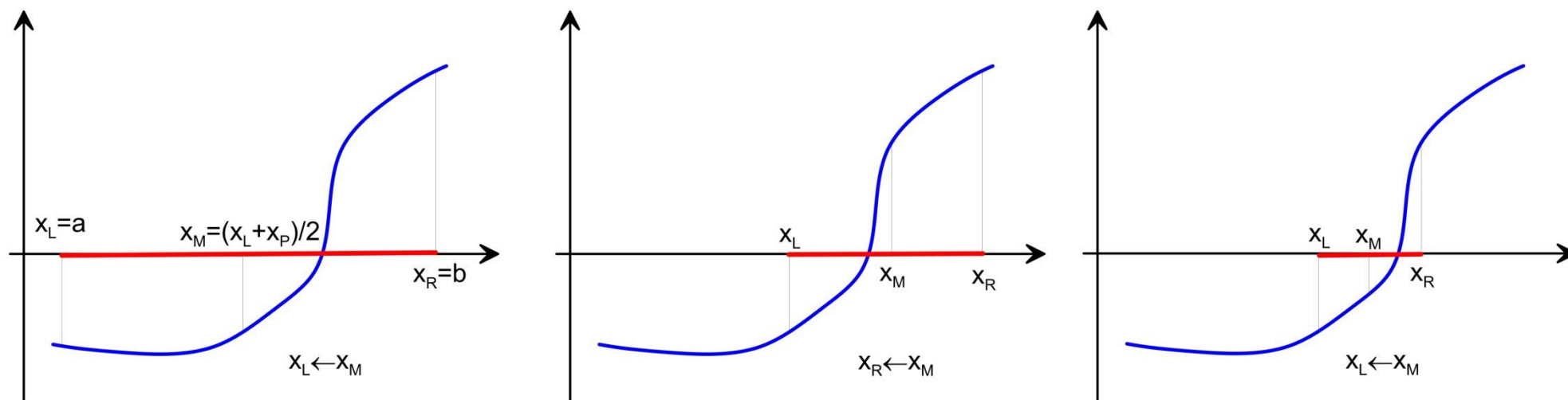
$$x^5 - x^2 + 4x - 1 = 0 \quad , \quad e^{-x} + \sin(x) = 0 \quad , \quad 4 - x - \tan^{-1}(x) = 0$$

Wspólną cechą wszystkich zaprezentowanych dalej algorytmów jest to, że są to metody iteracyjne. W metodach iteracyjnych, rozwiązanie wyznaczane jest na drodze kolejnych przybliżeń, a liczba powtórzeń (iteracji) niezbędnych do uzyskania przybliżenia o satysfakcjonującej precyzji nie jest – na ogół – a priori znana. Wynika stąd, że w **metodzie iteracyjnej potrzebne są: (a) przybliżenie początkowe (punkt startowy) i (b) kryterium stopu, tj. warunek przerywania obliczeń oparty na jakimś kryterium „jakości” przybliżonego rozwiązania.**

**UWAGA:** *zajmiemy się tu wyłącznie przypadkiem pojedynczego równania – metody dla układów algebraicznych równań nieliniowych są tematem kolejnych, bardziej zaawansowanych kursów metod numerycznych.*

## METODA BISEKCJI

Zacznijemy od bardzo elementarnego (i intuicyjnego) sposobu znanego pod nazwą Metody Bisekcji (dalej MB). „Mechanikę” tej metody wyjaśniają poniższe rysunki.



Wyznaczamy przedział zawierający pierwiastek, dzielimy go na połowy i określamy, w której z nich położony jest pierwiastek. Kryterium wyboru polega na określeniu na końcach której połowy funkcja  $f$  przyjmuje wartości o przeciwnym znaku. Ponieważ funkcja  $f$  jest z założenia ciągła to zmiana znaku w przedziale gwarantuje istnienie w nim miejsca zerowego.

Przebieg obliczeń wg MB pokazuje przedstawiony obok „pseudokod”.

**Zadanie domowe:** napisać w pełni funkcjonalny program w języku C/C++ realizujący algorytm MB

### Co z warunkiem stopu w MB?

Założmy, że chcemy wyznaczyć  $x_*$  z błędem nie większym niż  $\varepsilon$ , czyli godzimy się przerwać obliczenia jeśli zajdzie warunek  $|x_* - x_M| \leq \varepsilon$ .

Zauważmy, że po  $M$  krokach obliczeń MB mamy oszacowanie

$$|x_* - x_M| \leq \frac{b-a}{2^n}$$

Zatem, liczna iteracji niezbędna do uzyskania założonej dokładności to

$$n \geq \log_2 \frac{b-a}{\varepsilon}$$

```
START : fL := f(xL) ; fR := f(xR)  
if (fL · fR < 0) then  
  while (xR - xL > 2ε) do  
    xM := (xR + xL) / 2  
    fM := f(xM)  
    if (fL · fM < 0) then  
      xR := xM ; fR := fM  
    else  
      xL := xM ; fL := fM  
    endif  
  enddo  
  return(xM)  
else  
  change the interval [xL, xR];  
  return to START;  
endif
```

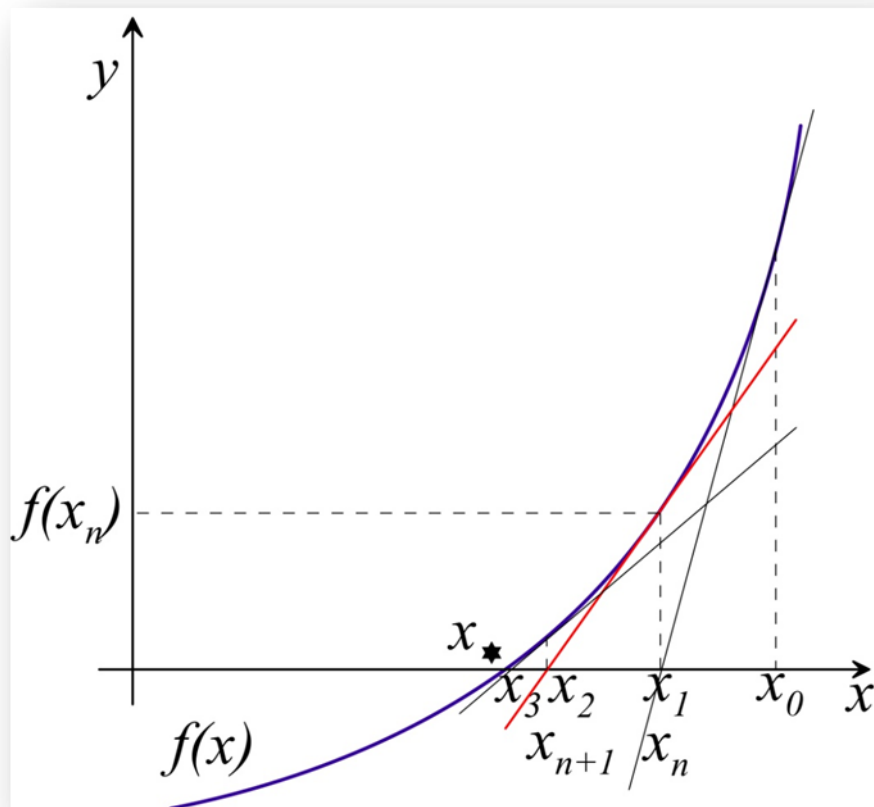
## Plusy i minusy Metody Bisekcji:

**Plus:** jeśli początkowy przedział poszukiwań jest wybrany poprawnie to MB zawsze zbiegnie do miejsca zerowego (jakiegoś miejsca zerowego) funkcji ciągłej  $f$

**Minus:** MB jest nieczuła na lokalne cechy (takie np. nachylenie wykresu) funkcji  $f$  w otoczeniu jej miejsca zerowego; uzyskanie rozwiązania dobrej jakości wymaga znaczącej liczby iteracji, które mogą być kosztowne obliczeniowo, jeśli obliczanie wartości funkcji  $f$  jest złożone.

## METODA STYCZNYCH (NEWTONA)

**Metoda Stycznych** (MST) to bardzo wydajna metoda wyznaczania miejsc zerowych funkcji ciągłych wraz z pochodnymi do rzędu równego dwa lub więcej. Ideę **MST** ilustruje poniższy rysunek.



Tak więc następne przybliżenie pierwiastka  $x_{n+1}$  jest określone jako odcięta punktu przecięcia stycznej wystawionej do wykresu funkcji  $f$  w poprzednim punkcie  $x_n$  i osi  $Ox$ .

Równanie linii stycznej ma postać

$$y = f'(x_n)(x - x_n) + f(x_n)$$

Wobec tego, równanie dla  $x_{n+1}$  ma postać

$$0 = f'(x_n)(x_{n+1} - x_n) + f(x_n), \text{ a stąd}$$

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

Pokażemy, że jeśli funkcja  $f$  spełnia odpowiednie warunki, to **MST wykazuje kwadratową zbieżność**.

Zacznijmy od formalnego wykorzystania Twierdzenia Taylora. Jeśli funkcja  $f$  jest klasy  $C^2$  w pewnym przedziale zawierającym  $x_*$  i  $x_n$  to istnieje taka liczba  $\xi \in [\min(x_n, x_*), \max(x_n, x_*)]$ , że ma miejsce równość

$$f(x_*) = f(x_n) + f'(x_n)(x_* - x_n) + \frac{1}{2} f''(\xi)(x_* - x_n)^2$$

Dalej, korzystamy z faktu, że  $f(x_*) \equiv 0$ . Rachunki przebiegają następująco:

$$\begin{aligned} 0 &= f(x_n) + f'(x_n)(x_* - x_n) + \frac{1}{2} f''(\xi)(x_* - x_n)^2 = \\ &= f(x_n) - f'(x_n)x_n + f'(x_n)x_* + \frac{1}{2} f''(\xi)(x_* - x_n)^2 = \\ &= -f'(x_n) \underbrace{\left( x_n - \frac{f(x_n)}{f'(x_n)} \right)}_{x_{n+1}} + f'(x_n)x_* + \frac{1}{2} f''(\xi)(x_* - x_n)^2 = \\ &= f'(x_n)(x_* - x_{n+1}) + \frac{1}{2} f''(\xi)(x_* - x_n)^2 \end{aligned}$$

Z otrzymanej równości wynika konkluzja:  $|x_* - x_{n+1}| = \left| \frac{f''(\xi)}{2f'(x_n)} \right| |x_* - x_n|^2$

Przeprowadzony wyżej rachunek uprawnia do sformułowania **twierdzenia o lokalnej zbieżności Metody Stycznych (Newtona)**:

Założmy, że w pewnym przedziale  $I_\alpha = (x_* - \alpha, x_* + \alpha)$  funkcja  $f$  spełnia warunki:

$$1) \exists_{M>0} \forall_{x \in I_\alpha} |f''(x)| \leq M \quad , \quad 2) \exists_{m>0} \forall_{x \in I_\alpha} |f'(x)| \geq m$$

Wówczas, dla każdego  $x_0 \in I_\alpha$  i takiego, że  $\frac{M}{2m}|x_* - x_0| \leq \gamma < 1$  zachodzi zastępujące oszacowanie

$$|x_* - x_1| \leq \frac{M}{2m}|x_* - x_0|^2 \leq \gamma|x_* - x_0|$$

z którego wynika, że  $x_1 \in I_\alpha$ , a także - na podstawie argumentu indukcyjnego - że

$$|x_* - x_{n+1}| \leq \gamma|x_* - x_n| \leq \gamma^{n+1}|x_* - x_n|$$

Wobec tego  $|x_* - x_n| \xrightarrow{n \rightarrow \infty} 0$  (ma miejsce zbieżność)

Ponadto  $|x_* - x_{n+1}| \leq \frac{M}{2m}|x_* - x_n|^2$



**UWAGA:** Zauważmy, że powyższe twierdzenie gwarantuje jedynie **zbieżność lokalną**. Innymi słowy, MST będzie “typowo” osiągać kwadratową zbieżność, o ile początkowe przybliżenie  $x_0$  jest już dostatecznie blisko rozwiązania  $x_*$ . Jeśli jednak tak nie jest, to nie mamy gwarancji nie tylko zbieżności kwadratowej, ale ciąg  $(x_0, x_1, \dots, x_n, x_{n+1}, \dots)$  **może być w ogóle rozbieżny!**

Co się dzieje jeśli niektóre założenia twierdzenia nie są spełnione? Rozważmy przykład:

$$f(x) = x^2 - 2x + 1 = 0 \Rightarrow x_* = 1$$

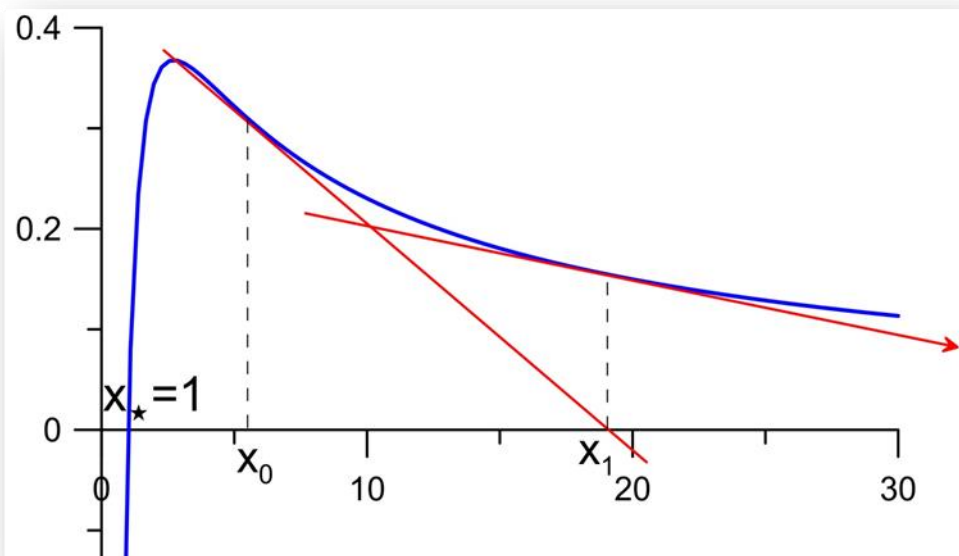
Mamy  $f'(x) = 2x - 2$ ,  $f'(1) = 0$ , a zatem

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{(x_n - 1)^2}{2(x_n - 1)} = x_n - \frac{1}{2}(x_n - 1) = \frac{1}{2}x_n + \frac{1}{2}$$

Wnioskujemy z powyższego, że  $x_{n+1} - 1 = \frac{1}{2}(x_n - 1)$ . Widzimy, że odległość kolejnych przybliżeń generowanych przez MST od ścisłego rozwiązania **maleje w tempie liniowym** (o czynnik  $\frac{1}{2}$  na każdą iterację), a nie kwadratowym. **Zbieżność kwadratowa nie jest osiągnięta ponieważ  $f'(x_*) = 0$  wbrew drugiemu z założeń twierdzenia.**

1. Niech  $f(x) = (x - a)^m$ . Pokazać, że MST prowadzi do formuły  $x_{n+1} - a = (1 - \frac{1}{m})(x_n - a)$ . Zauważmy, że dla dużych wartości  $m$  zbieżność będzie bardzo powolna!
2. Przeanalizować zbieżność (do rozwiązania zerowego) MST zastosowanej do równania  $x^{3/2} + 2x = 0$

Rozważmy teraz inne scenariusze „patologicznego zachowania” MST.

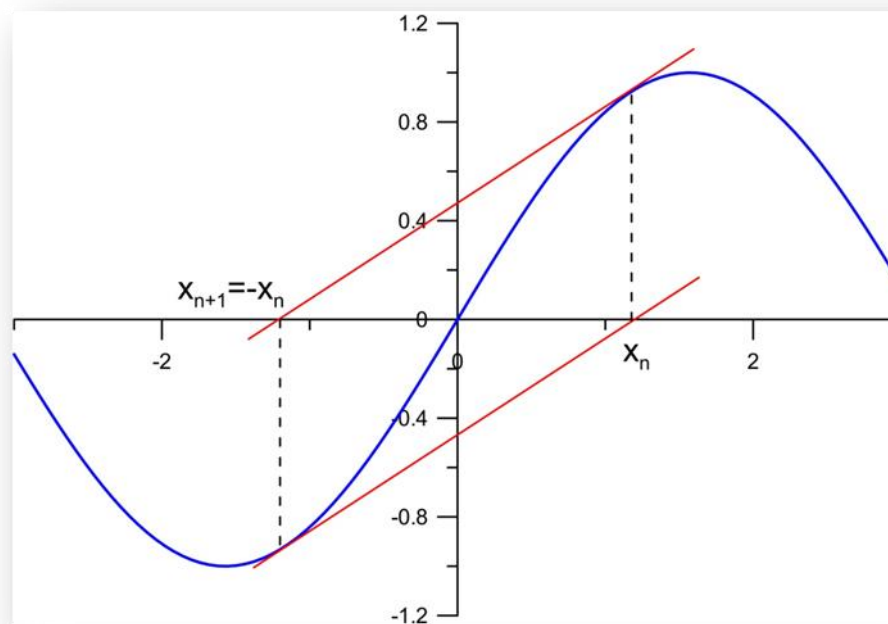


Przykład 1: Przedział lokalnej zbieżności może być bardzo wąski, tak jak to ma miejsce dla funkcji  $f(x) = \ln(x)/x$  z  $x_* = 1$ . Jak widać, na rysunku obok, wybranie jako punktu startowego MST wartości na prawo od ekstremum zawsze kończy się rozbieżnością iteracji do nieskończoności. Z drugiej strony, wybranie punktu zbyt blisko po lewej stronie ekstremum prowadzi do wyznaczenia liczby ujemnej, a zatem nie należącej do dziedziny funkcji ( $R_+$ )

Przykład 2: Iteracje MST mogą się „zapętlić”. Zachowanie takie demonstruje następujący przykład (rysunek po prawej).

Rozważmy  $f(x) = \sin(x)$  i założmy, że punkt początkowy  $x_0$  to

$$x_0 = \frac{1}{2} \pi g(x_0)$$



Wówczas

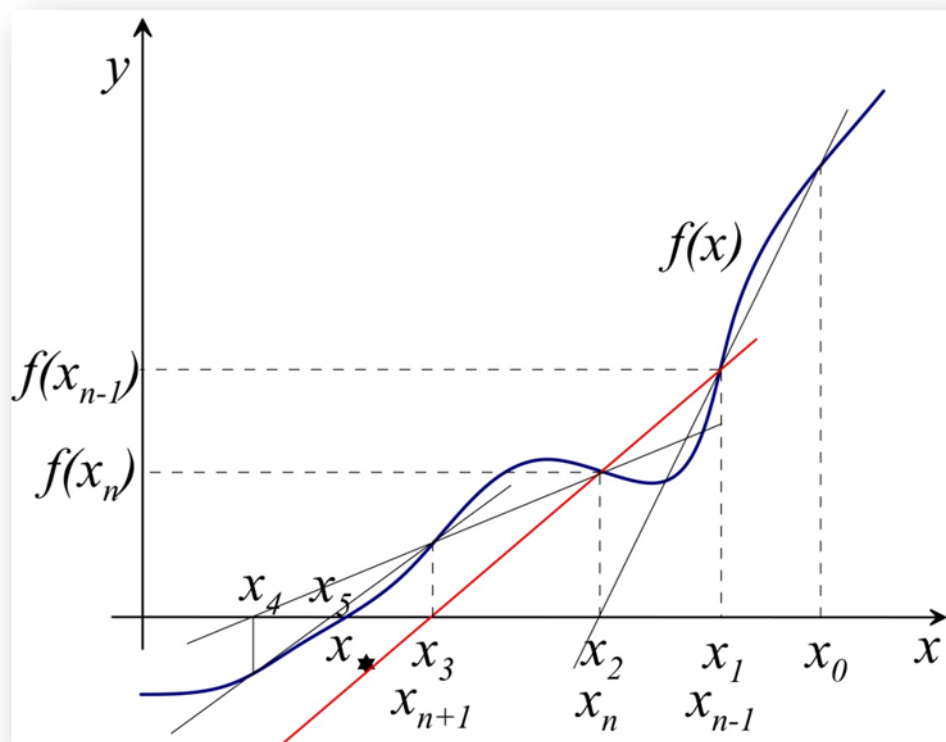
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = x_0 - \frac{\sin(x_0)}{\cos(x_0)} = x_0 - \tan(x_0) = -x_0$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = -x_0 - \frac{\sin(-x_0)}{\cos(-x_0)} = -x_0 + \tan(x_0) = x_0$$

Tak więc – przynajmniej w teorii – iteracje MST będą naprzemiennie zwracać tylko dwie wartości:  $x_0$  and  $-x_0$ .

**Zadanie:** Jaka jest wrażliwość opisanego wyżej cyklu na nieuniknione błędy zaokrągleń arytmetyki komputera? Przeprowadź taką analizę analitycznie (trudne!) lub za pomocą eksperymentu w komputerze.

## METODA SIECZNYCH (MSC)



Słabą stroną metody Newtona (MST) jest konieczność obliczania pochodnej. W pewnych zastosowaniach funkcja  $f$  może nie być zadana wzorem, ale raczej jako pewna procedura obliczeniowa. Trudność tę można obejść stosując **metodę siecznych** (MSC). Mechanikę tej metody wyjaśnia rysunek obok.

Tym razem, następne przybliżenie  $x_{n+1}$  pierwiastka  $x_*$  jest zdefiniowane jako odcięta punktu przecięcia osi  $Ox$  przez sieczną przechodzącą przez punkty  $(x_{n-1}, f(x_{n-1}))$  and  $(x_n, f(x_n))$ .

Nietrudno pokazać (ćwiczenie), że **równanie linii siecznej** ma postać

$$y = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x - x_n) + f(x_n)$$

Przyrównanie  $y$  do zera prowadzi do **wzoru ogólnego metody siecznych**, a mianowicie

$$0 = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} (x_{n+1} - x_n) + f(x_n)$$

⇓

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n)$$

Zajmiemy się teraz analizą zbieżności MSC, która jest nieco bardziej złożona niż dla metody stycznych (MST).

Stosując (formalnie) rozwinięcie z szereg potęgowy Taylora, możemy napisać

$$f(x_n) = f(x_*) + f'(x_*)\varepsilon_n + \frac{1}{2} f''(x_*)\varepsilon_n^2 + \dots$$

$$f(x_{n-1}) = f(x_*) + f'(x_*)\varepsilon_{n-1} + \frac{1}{2} f''(x_*)\varepsilon_{n-1}^2 + \dots$$

Trzy kropki po ostatnim plusie symbolizują pominięte człony wyższych rzędów.

Dalsze rachunki przebiegają następująco:

$$\begin{aligned}
 \varepsilon_{n+1} &= x_{n+1} - x_* = \underbrace{x_n - x_*}_{\varepsilon_n} - \frac{f(x_n)}{f(x_n) - f(x_{n-1})} \left[ \underbrace{x_n - x_*}_{\varepsilon_n} - \underbrace{(x_{n-1} - x_*)}_{\varepsilon_{n-1}} \right] = \\
 &= \varepsilon_n - \frac{[f'(x_*)\varepsilon_n + \frac{1}{2}f''(x_*)\varepsilon_n^2 + \dots](\varepsilon_n - \varepsilon_{n-1})}{f'(x_*)\varepsilon_n + \frac{1}{2}f''(x_*)\varepsilon_n^2 + \dots - [f'(x_*)\varepsilon_{n-1} + \frac{1}{2}f''(x_*)\varepsilon_{n-1}^2 + \dots]} = \\
 &= \varepsilon_n - \frac{f'(x_*)\varepsilon_n + \frac{1}{2}f''(x_*)\varepsilon_n^2 + \dots}{f'(x_*) + \frac{1}{2}f''(x_*)(\varepsilon_{n-1} + \varepsilon_n) + \dots} = \varepsilon_n - \frac{\varepsilon_n + \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}\varepsilon_n^2 + \dots}{1 + \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}(\varepsilon_{n-1} + \varepsilon_n) + \dots} = \\
 &= \varepsilon_n - \left( \varepsilon_n + \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}\varepsilon_n^2 + \dots \right) \cdot \left( 1 - \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}(\varepsilon_{n-1} + \varepsilon_n) + \dots \right) = \\
 &= \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}(\varepsilon_{n-1} + \varepsilon_n)\varepsilon_n - \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}\varepsilon_n^2 + \dots = \frac{1}{2}\frac{f''(x_*)}{f'(x_*)}\varepsilon_{n-1}\varepsilon_n + \dots
 \end{aligned}$$

W ten sposób otrzymaliśmy przybliżoną formułę na błąd aproksymacji miejsce zerowego:

$$\varepsilon_{n+1} \cong \frac{f''(x_*)}{2f'(x_*)} \varepsilon_n \varepsilon_{n-1} = K \varepsilon_n \varepsilon_{n-1}$$

Korzystniej byłoby wyrazić wielkość  $\varepsilon_{n+1}$  odwołując się wyłącznie do wielkości błędu z poprzedniej iteracji, czyli  $\varepsilon_n$ . W tym celu, założymy, że poszukiwany związek da się zapisać w postaci

$$\varepsilon_{n+1} = K^\beta \varepsilon_n^\alpha$$

gdzie nieznane na razie wielkości  $\alpha$  i  $\beta$  są stałe, tj. niezależne od numeru iteracji.

Wynika stąd, że  $\varepsilon_n = K^\beta \varepsilon_{n-1}^\alpha \Rightarrow \varepsilon_{n-1} = K^{-\beta/\alpha} \varepsilon_n^{1/\alpha}$ . Po podstawieniu do pierwotnej formuły przyjmie ona postać następującą

$$\varepsilon_{n+1} = K \varepsilon_n \varepsilon_{n-1} = K \varepsilon_n (K^{-\beta/\alpha} \varepsilon_n^{1/\alpha}) = K^{1-\frac{\beta}{\alpha}} \varepsilon_n^{\frac{1+\alpha}{\alpha}}$$

Ponieważ oba otrzymane związki pomiędzy  $\varepsilon_n$  i  $\varepsilon_{n+1}$  musi w szczególności zachodzić następująca równość dla liczby  $\alpha$

$$\alpha = \frac{1+\alpha}{\alpha} \Rightarrow \alpha^2 - \alpha - 1 = 0$$

Ponieważ jedynie dodatnia wartość ma sens, mamy ostatecznie  $\alpha = \frac{1+\sqrt{5}}{2} \approx 1.618$ .

Mamy też przy okazji wartość  $\beta$  (mniej istotną)

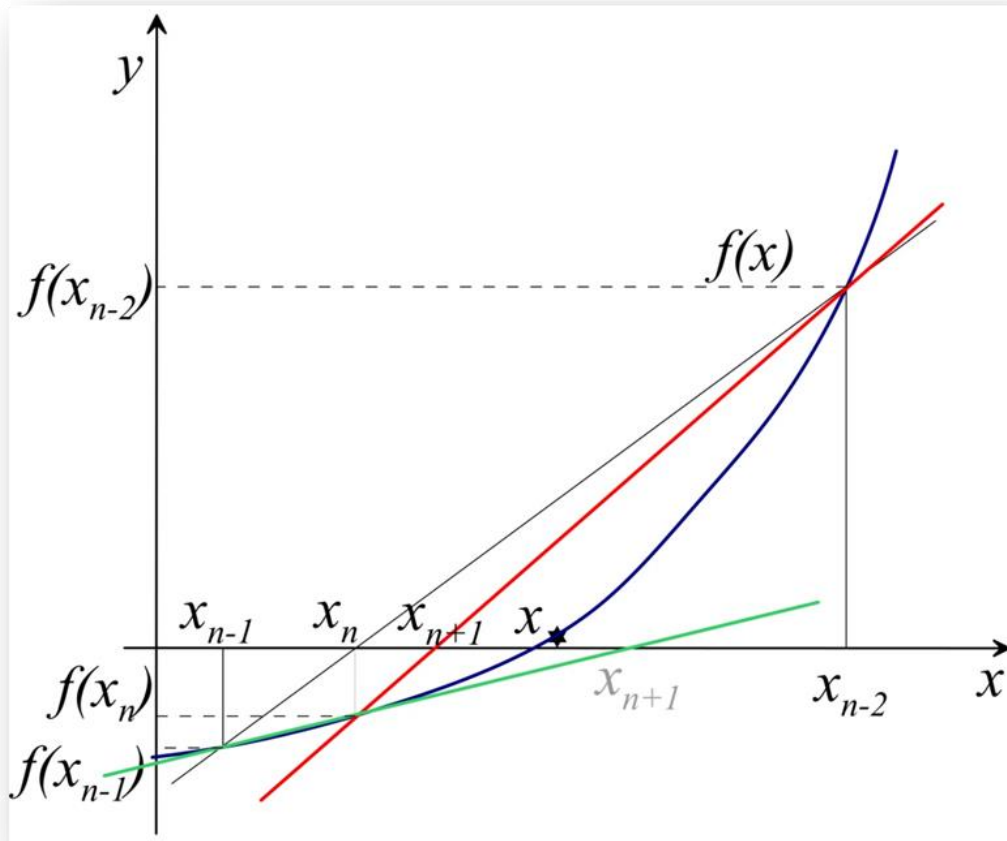
$$\beta = 1 - \frac{\beta}{\alpha} \Rightarrow \beta = \frac{\alpha}{1+\alpha} = \frac{1}{\alpha}$$

Ponieważ wykładnik  $\alpha$  spełnia warunek  $1 < \alpha < 2$ , mówimy, że **Metoda Siecznych jest superliniowo zbieżna** (lepiej niż liniowo, ale gorzej niż kwadratowo). Można również powiedzieć, że ceną za uniknięcie konieczności obliczania pochodnej w MST jest utrata kwadratowej zbieżności.



Metoda Siecznych jest wrażliwa na „patologie” podobne do tych, które opisaliśmy przy okazji Metody Stycznych. W praktyce **zbieżność MSC ma również charakter lokalny**, tj. początkowe punkty startowe (są tym razem dwa tj.  $x_0$  i  $x_1$ ) muszą być dostatecznie blisko pierwiastka  $x_*$ .

## METODA FAŁSZYWEGO PUNKTU (MFP) - REGUŁA FAŁSI



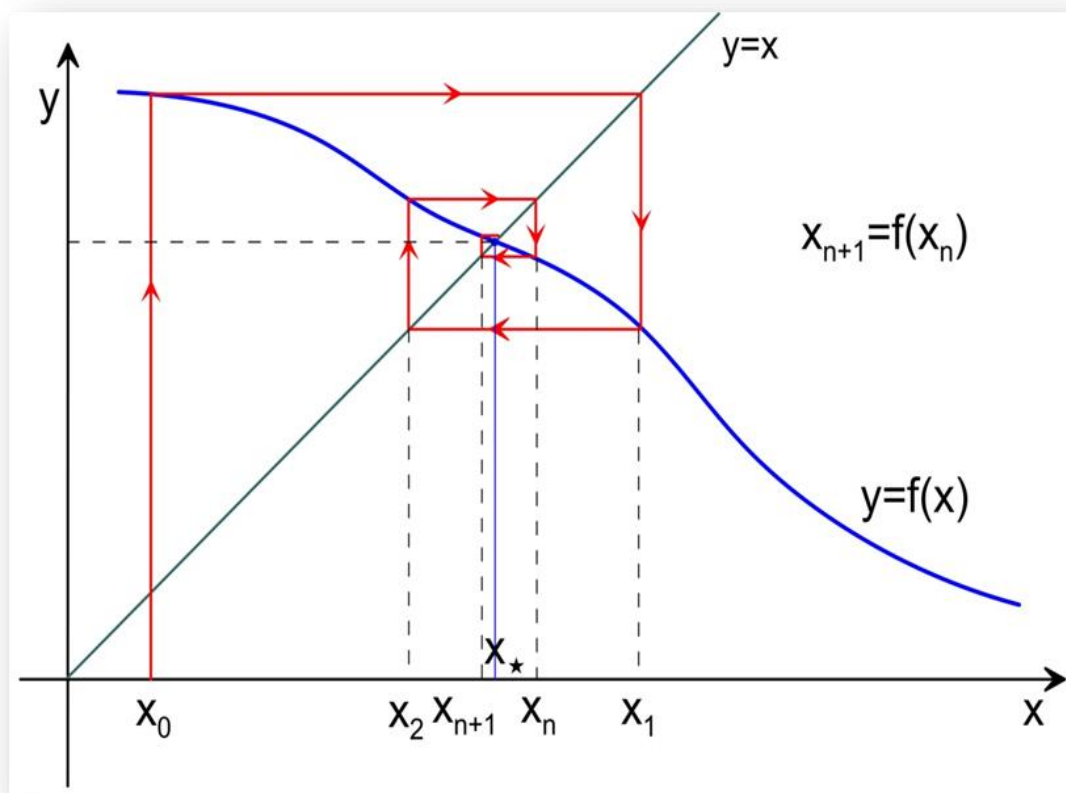
Podstawowa różnica pomiędzy standardową MSC a MFP polega na innym wyborze pary punktów. W sytuacji przedstawionej na rysunku po lewej, punkt  $x_{n+1}$  w metodzie siecznych byłby położony tam, gdzie sieczna w kolorze zielonym przecina oś  $Ox$ . W MFP, punkt  $x_{n+1}$  jest w tym przypadku miejscem przecięcia osi  $Ox$  przez sieczną w kolorze czerwonym (stąd “false position”).

**Innymi słowy: linia sieczna jest przeprowadzona przez najnowsze punkty położone po przeciwnej stronie osi  $Ox$ .**

**Jeśli dwa pierwsze punkty (startowe) obramowują rozwiązanie to zbieżność MFP jest zagwarantowana.** Z drugiej strony, MFP jest na ogół jedynie liniowo zbieżna. Praktykowany jest również wariant, w którym rzędna punktu „zamrożonego” (w tym przykładzie  $x_{n-2}$ ) jest sukcesywnie dzielona przez 2, dopóki nie pojawi się nowy punkt po tej samej stronie osi  $Ox$ . Wyrafinowanym wariantem MFP jest również **metoda Riddersa (liczne opisy w internecie).**

**Zadanie:** opracować sposób efektywnej implementacji metod MSC i MFP w języku C/C++ i porównać ich osiągi na kilku przykładach.

## METODA ITERACJI PROSTEJ



MIP to metoda dla równań nieliniowych zapisanych w postaci

$$x = g(x)$$

Naturalnym jest zdefiniować następujący proces iteracyjny

$$x_{n+1} = g(x_n)$$

Jeśli proces ten jest zbieżny do pewnego punktu  $x_*$  wówczas automatycznie

$$x_* \equiv g(x_*)$$

Innymi słowy: **rozwiązanie  $x_*$  jest punktem stałym tego procesu iteracyjnego.** Przypadek taki ilustruje powyższy rysunek.

Kiedy proces iteracji prostej jest zbieżny? Jakie jest tempo zbieżności? **Okazuje się, że odpowiedź na oba pytania odnosi się do jednej i tej samej wielkości, a mianowicie  $g'(x_*)$ .**

W celu oszacowania tempa zbieżności procesu iteracji prostej przeprowadzamy następujący rachunek:

$$x_{n+1} = g(x_n) \Rightarrow \underbrace{x_{n+1} - x_*}_{\varepsilon_{n+1}} = g(x_n) - g(x_*) = g'(x_*) \underbrace{(x_n - x_*)}_{\varepsilon_n} + \dots$$

Zatem, dla punktów położonych dostatecznie blisko pierwiastka  $x_*$  mamy

$$|\varepsilon_{n+1}| \cong |g'(x_*)| \cdot |\varepsilon_n|$$

Skąd wnioskujemy, że  $x_*$  **jest punktem przyciągającym (atraktorem) procesu iteracyjnego wtedy i tylko wtedy, gdy**

$$|g'(x_*)| < 1$$

**Ma też miejsce następujący fakt:** jeżeli w punkcie  $x_*$  funkcja  $g$  spełnia warunki

$$|g'(x_*)| = 0 \quad , \quad |g''(x_*)| < \infty$$

to proces iteracji prostej jest **kwadratowo zbieżny**. Istotnie, możemy napisać

$$x_{n+1} - x_* = g(x_n) - g(x_*) = g'(x_*)(x_n - x_*) + \frac{1}{2} g''(x_*)(x_n - x_*)^2 + \dots$$

$$\begin{array}{c} 0 \\ \Downarrow \end{array}$$

$$|\varepsilon_{n+1}| \cong \frac{1}{2} |g''(x_*)| |\varepsilon_n|^2$$

**Przykład:** Metoda stycznych (Newtona) może być interpretowana jako metoda iteracji prostej zapisana dla następującej funkcji  $g$

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

Obliczmy pochodną tej funkcji

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f(x)f''(x)}{[f'(x)]^2}$$

Jeśli teraz pierwsza pochodna funkcji  $f$  w punkcie  $x_*$  jest różna od zera, to mamy

$$g'(x_*) = \frac{f''(x_*)}{[f'(x_*)]^2} f(x_*) = 0$$

czyli **metoda stycznych jest zbieżna kwadratowo** (co już skądinąd wiemy).

**UWAGA:** jeżeli  $|g'(x_*)| \geq 1$  to metoda iteracji prostej będzie na ogół rozbieżna. Jeśli funkcja  $g$  ma inne miejsca zerowe to metoda iteracji prostej może doprowadzić czasami zbiec do innego pierwiastka niż „pożądany”. Możliwe są również inne zachowania np. cykle, a nawet przebiegi chaotyczne. Klasycznym przypadkiem procesu iteracyjnego o bardzo bogatej dynamice jest tzw. odwzorowanie logistyczne

$$x_{n+1} = \mu x_n (1 - x_n)$$

które ma postać iteracji prostej dla równania kwadratowego  $x = \mu x(1 - x)$ . Analiza zachowania tego procesu iteracyjnego (lub – jak mówią matematycy – układu dynamicznego z „dyskretnym czasem”) w funkcji parametru  $\mu$  należy do żelaznego repertuaru niemal każdego kursu tzw. **dynamiki chaotycznej** (np. trzeci rozdział monografii „Chaos deterministyczny. Wprowadzenie” H.G. Schustera, PWN 1993).