Univariate Optimization

Antonio Frangioni

Department of Computer Science
University of Pisa
https://www.di.unipi.it/~frangio
mailto:frangio@di.unipi.it

Computational Mathematics for Learning and Data Analysis
Master in Computer Science – University of Pisa

A.Y. 2024/25

Outline

- X any set, $f : X \to \mathbb{R}$ any function: optimization problem

$$(P) \qquad f_* = \min\{\, f(x) \,:\, x \in X \,\}$$

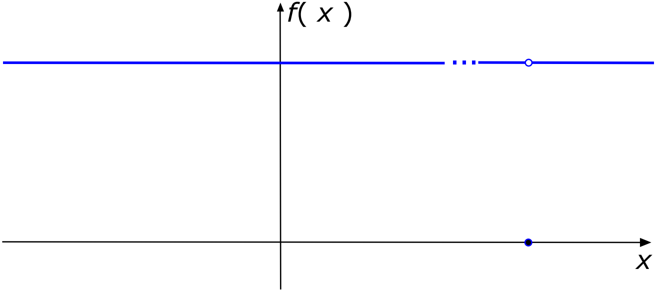
- Impossible (X inaccessible cardinal, f non computable function, ...)

- Let's start "easy":
  - X "very easy": $X = \mathbb{R}$ or (even better) bounded $X = [\, x_- \,,\, x_+ \,] \subset \mathbb{R}$
  - an (efficient, pointwise) oracle for f available:
    $\forall\, x \in X$, $f(x)$ is "easy to compute" (say, $O(1)$)

- Still not easy at all, in fact impossible in general [3, p. 408]

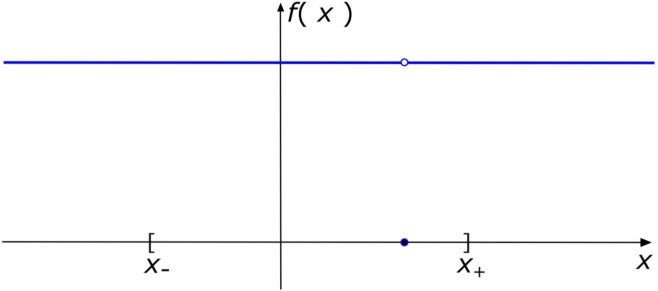
- Too trivial for $f(\cdot)$ linear or quadratic, $O(1)$ formulæ

- Need to find a middle ground (one must $\exists$)

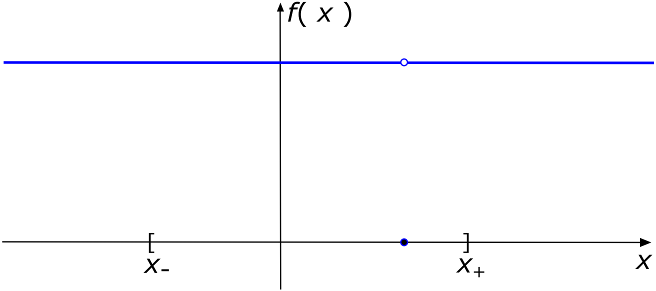


- Impossible because isolated minima can be anywhere [3, p. 408]

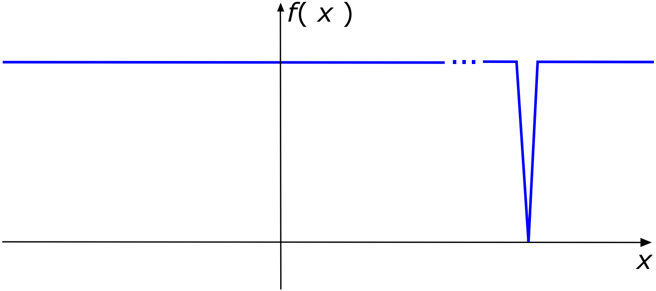
Recall: even approximate, optimization is hard / impossible



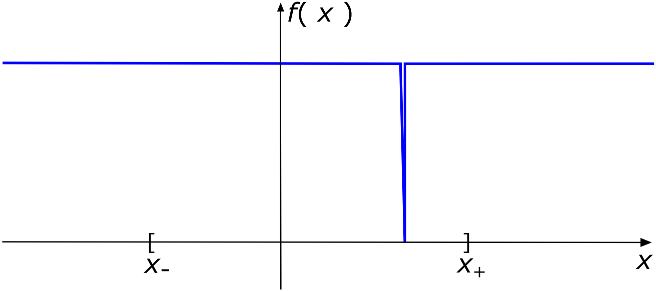
- Impossible because isolated minima can be anywhere [3, p. 408]
- Does it help restricting to $x \in X = [\, x_- \,, \, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?
- No: still uncountably many points to try



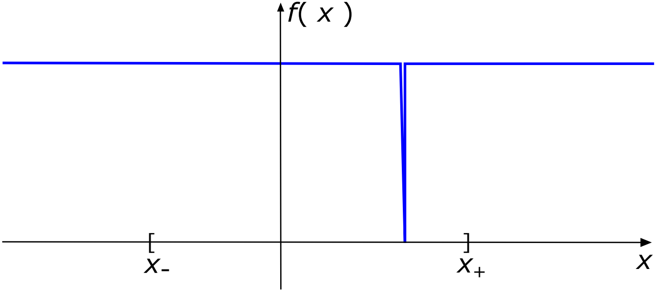
- Impossible because isolated minima can be anywhere [3, p. 408]
- Does it help restricting to $x \in X = [\, x_- \, , \, x_+ \,]$ ($-\infty < x_- < x_+ < +\infty$)?
- No: still uncountably many points to try
- Is it because f "jumps"?



- Impossible because isolated minima can be anywhere [3, p. 408]
- Does it help restricting to $x \in X = [\, x_- \,,\, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?
- No: still uncountably many points to try
- Is it because f "jumps"? No, f can have isolated $\downarrow$ spikes anywhere



- Impossible because isolated minima can be anywhere [3, p. 408]
- Does it help restricting to $x \in X = [\, x_- \,,\, x_+ \,]$ ($-\infty < x_- < x_+ < +\infty$)?
- No: still uncountably many points to try
- Is it because f "jumps"? No, f can have isolated $\downarrow$ spikes anywhere
  ...even on $X = [\, x_- \,,\, x_+ \,]$ as spikes can be aribtrarily narrow



- Impossible because isolated minima can be anywhere [3, p. 408]
- Does it help restricting to $x \in X = [\, x_- \,,\, x_+ \,]$ $(-\infty < x_- < x_+ < +\infty)$?
- No: still uncountably many points to try
- Is it because f "jumps"? No, f can have isolated $\downarrow$ spikes anywhere
  ...even on $X = [\, x_- \,,\, x_+ \,]$ as spikes can be aribtrarily narrow
- Making it possible $\equiv$ impose speed limits on the rate of change

- Impose spikes can't be arbitrarily narrow $\equiv f$ cannot change too fast $\equiv$ f Lipschitz continuous (L-c) on X [4, p. 624]: $\exists L > 0$ s.t.
  $$| f(x) - f(z) | \leq L | x - z | \qquad \forall x, z \in X$$
- f globally L-c $\equiv X = \mathbb{R}$, locally L-c at $x \equiv \exists \varepsilon > 0$ s.t. $X = [x - \varepsilon, x + \varepsilon]$
- Note: L depends on X (locally L-c $\not\Longrightarrow$ globally L-c)
- $f : \mathbb{R} \to \mathbb{R}$ continuous at $x \equiv \forall \{ x_i \} \to x \Longrightarrow \{ f(x_i) \} \to f(x) \equiv$ $\forall \varepsilon > 0 \, \exists \delta > 0$ s.t. $z \in [x - \delta, x + \delta] \Longrightarrow | f(z) - f(x) | \leq \varepsilon$
- continuous on $X \equiv \forall x \in X$, just "continuous" $\equiv X = \mathbb{R} \equiv f \in C^0$
- Many "simple" functions C^0 + continuity easily preserved:
  $f, g \in C^0 \Longrightarrow f + g$, $f \cdot g$, $\max\{ f, g \}$, $\min\{ f, g \}$, $f(g(\cdot)) \in C^0$
- f locally L-c at $x \Longrightarrow f$ continuous at x (**check**)

Exercise: Come up with f locally L-c everywhere but not globally L-c

Exercise: Come up with f continuous but not L-c on some finite $X = [x_- , x_+]$

- Still need to impose $X = [\, x_- \,,\, x_+ \,]$ with $D = x_+ - x_- < \infty$ (finite diameter), otherwise isolated $\downarrow$ spikes need not even be "very narrow"
- f L-c $\Longrightarrow$ one ε-optimum can be found with $O(\, LD \,/\, \varepsilon\,)$ evaluations: uniformly sample X with step $2\varepsilon \,/\, L$ [3, p. 411]

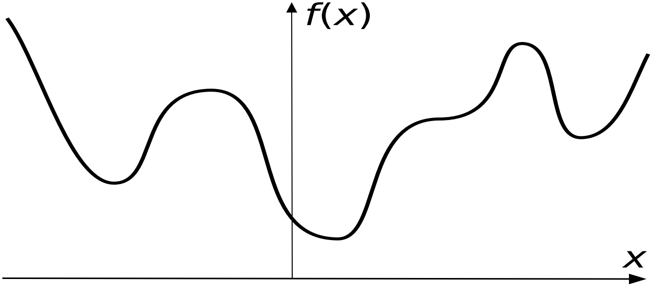
Exercise: Prove the above

- Bad news: no algorithm can work in less than $\Omega(\, LD \,/\, \varepsilon\,)$ [3, p. 413] (proof uses adversarial function, not typical in learning applications)
- $\#$ steps inversely proportional to accuracy, just not doable for "small" ε
- Even very dramatically worse if $X \subset \mathbb{R}^n$ (will see)
- No free lunch theorem says "all algorithms equally bad" [7], i.e., "if an algorithm is very good in some cases it has to be very bad in others"
- Also, L generally unknown and not easy to estimate (will see) but algorithms actually require/use it

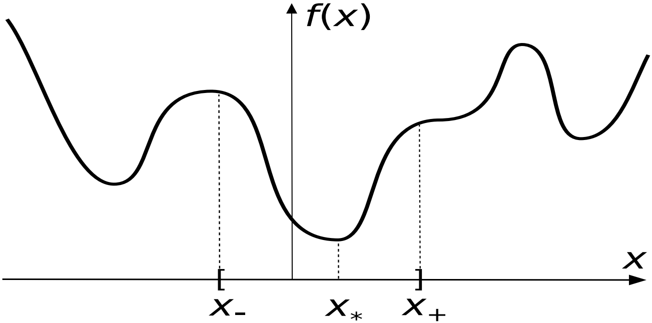
Outline

- Even if I should stumble in x_*, how do I recognize it?
- Turns out this is "the really difficult thing" (cf. knowing f_*)
- Simpler to start with a weaker condition: x_* is local minimum if
  $x_* = \operatorname{argmin}\{\, f(x) \,:\, x \in X(x_*, \varepsilon) = [\, x_* - \varepsilon \,,\, x_* + \varepsilon \,] \,\}$ for some $\varepsilon > 0$
- Stronger notion: strict local minimum if $f(x_*) < f(z) \;\; \forall z \in X(x_*, \varepsilon) \setminus \{ x_* \}$
- Why useful? Because "near x_*, f typically has a predictable shape"
- f (strictly) unimodal on $X = [\, x_- \,,\, x_+ \,]$:
  - has minimum $x_* \in X$
  - is (strictly) decreasing in $[\, x_- \,,\, x_* \,]$ and increasing in $[\, x_* \,,\, x_+ \,]$
- x_* local minimum $\Longrightarrow$ typically $\exists\, \varepsilon > 0$ s.t. f (strictly) unimodal on $X(x_*, \varepsilon)$

- Most functions are not unimodal (although some are, will see)

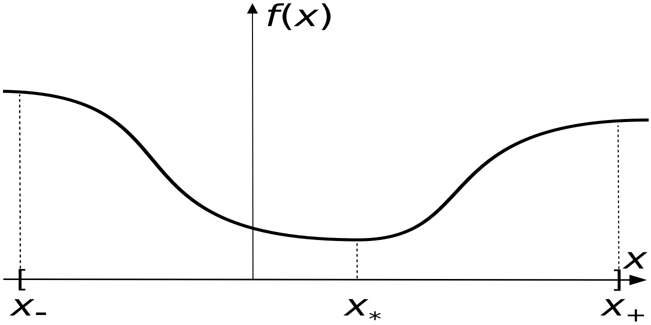


- Most functions are not unimodal (although some are, will see)



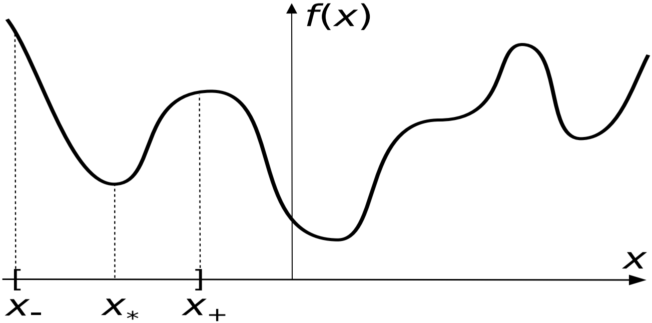
- But they are if you focus on the attraction basin of x_* and

- Most functions are not unimodal (although some are, will see)



- But they are if you focus on the attraction basin of x_* and restrict there

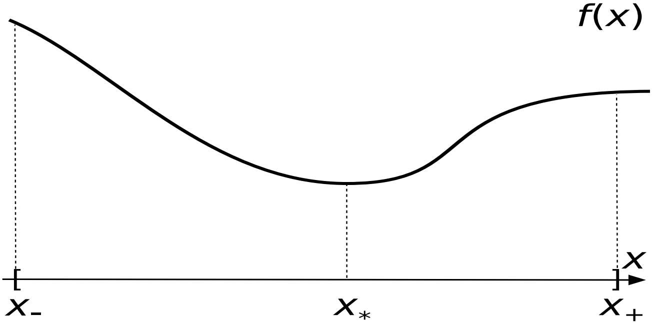
- Most functions are not unimodal (although some are, will see)



- But they are if you focus on the attraction basin of x_* and restrict there

- Unfortunately, this is true for every local optimum

- Most functions are not unimodal (although some are, will see)

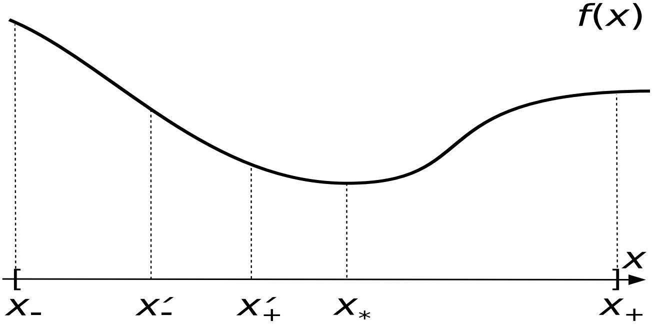


- But they are if you focus on the attraction basin of x_* and restrict there
- Unfortunately, this is true for every local optimum
- All local optima "look the same", comprised the global one
- Yet, this makes it finding some local optimum a lot easier
- Finding the right (global) one another matter entirely

Restricting Attraction Basins

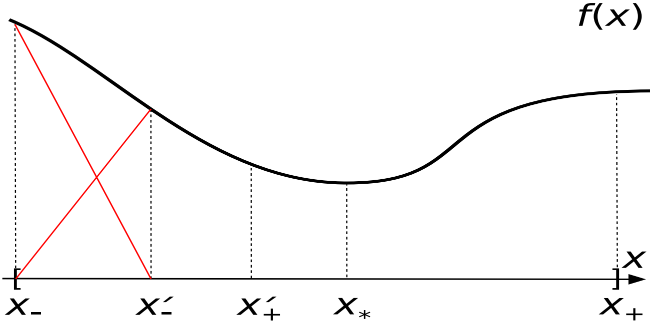
- Once in an attraction basin, we can restrict it by

- Once in an attraction basin, we can restrict it by evaluating f in two points



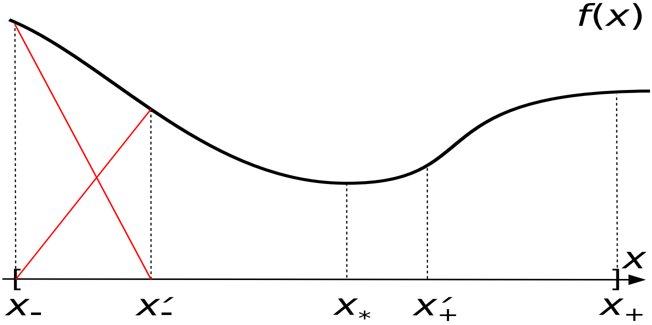
- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[\, x_- \, , \, x_+ \,]$ (minimum x_*), $x_- \leq x'_- \leq x'_+ \leq x_+$:

- Once in an attraction basin, we can restrict it by evaluating f in two points



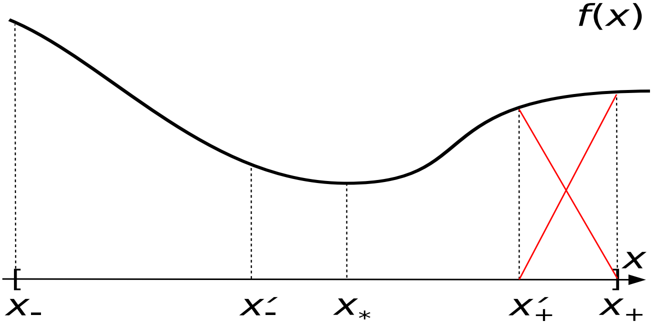
- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[\, x_- \, , \, x_+ \,]$ (minimum x_*), $x_- \leq x'_- \leq x'_+ \leq x_+$:
  $f(\, x'_- \,) \geq f(\, x'_+ \,) \implies x_* \in [\, x'_- \, , \, x_+ \,]$

- Once in an attraction basin, we can restrict it by evaluating f in two points



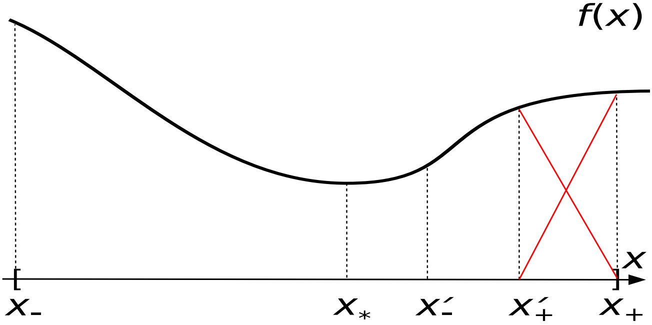
- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[\,x_-\,,\,x_+\,]$ (minimum x_*), $x_- \leq x_-' \leq x_+' \leq x_+$:
  $f(\,x_-'\,) \geq f(\,x_+'\,) \implies x_* \in [\,x_-'\,,\,x_+\,]$

- Once in an attraction basin, we can restrict it by evaluating f in two points



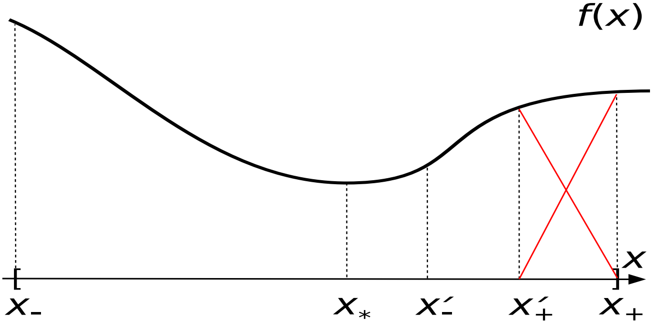
- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[x_-, x_+]$ (minimum x_*), $x_- \leq x'_- \leq x'_+ \leq x_+$:
  $f(x'_-) \geq f(x'_+) \implies x_* \in [x'_-, x_+]$ $\quad$ $f(x'_-) \leq f(x'_+) \implies x_* \in [x_-, x'_+]$

- Once in an attraction basin, we can restrict it by evaluating f in two points



- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[\, x_- \,,\, x_+ \,]$ (minimum x_*), $x_- \leq x'_- \leq x'_+ \leq x_+$:
  $f(\, x'_- \,) \geq f(\, x'_+ \,) \implies x_* \in [\, x'_- \,,\, x_+ \,]$ $\quad f(\, x'_- \,) \leq f(\, x'_+ \,) \implies x_* \in [\, x_- \,,\, x'_+ \,]$

- Once in an attraction basin, we can restrict it by evaluating f in two points



- [1, Th. 8.11 + Ex. 3.60 + Ex. 8.10]
  f (strictly) unimodal in $[\, x_- \,,\, x_+ \,]$ (minimum x_*), $x_- \leq x'_- \leq x'_+ \leq x_+$:
  $f(\, x'_- \,) \geq f(\, x'_+ \,) \implies x_* \in [\, x'_- \,,\, x_+ \,]$ $\quad f(\, x'_- \,) \leq f(\, x'_+ \,) \implies x_* \in [\, x_- \,,\, x'_+ \,]$

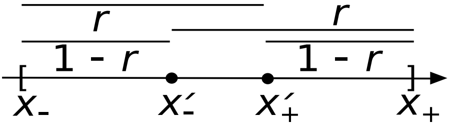
- By iterating this we can restrict the interval $\implies$ get close to x_* at will

- How should we choose x'_- and x'_+?

- General powerful concept: optimize worst-case behaviour $\Longrightarrow$ shrink the interval as quickly as possible
- Each iteration dumps either $[\,x_-\,,\,x'_-\,]$ or $[\,x'_+\,,\,x_+\,]$, don't know which $\Longrightarrow$ should be equal $\Longrightarrow$ select $r \in (\,1\,/\,2\,,\,1\,)$, $x'_- = x_- + (1-r)D$, $x'_+ = x_- + rD$
- Whatever the choice, new interval size $= Dr < D$
- Faster $\Longleftarrow$ r smaller (but $> 1\,/\,2$) $\equiv$ $r = 1\,/\,2 + \varepsilon$ $\equiv$ $x'_\pm = x_- + D\,/\,2 \pm \varepsilon$
- But next iteration will have two entirely different x'_- , x'_+ to evaluate f on

Optimally choosing the iterates

- General powerful concept: optimize worst-case behaviour $\Longrightarrow$ shrink the interval as quickly as possible
- Each iteration dumps either $[\,x_- \,,\, x'_-\,]$ or $[\,x'_+\,,\, x_+\,]$, don't know which $\Longrightarrow$ should be equal $\Longrightarrow$ select $r \in (\,1\,/\,2\,,\,1\,)$, $x'_- = x_- + (1-r)D$, $x'_+ = x_- + rD$
- Whatever the choice, new interval size $= Dr < D$
- Faster $\Longleftarrow$ r smaller (but $> 1\,/\,2$) $\equiv$ $r = 1\,/\,2 + \varepsilon$ $\equiv$ $x'_\pm = x_- + D\,/\,2 \pm \varepsilon$
- But next iteration will have two entirely different x'_- , x'_+ to evaluate f on
- Minimize function evaluations $\Longrightarrow$ re-use the surviving point
- $r : 1 = (1-r) : r$ $\equiv$ $r \cdot r = 1 - r$ $\equiv r = (\sqrt{5} - 1)\,/\,2$ (≈ 0.618)



- $r = 1\,/\,g$, g = golden ratio $= (\sqrt{5} + 1)\,/\,2 \approx 1.618$, $g = 1 + r = 1 + 1\,/\,g$

- Theorems breed algorithms: golden ratio search

$$
\begin{array}{l}
\textbf{procedure } [\, x_- \,,\, x_+ \,] = GRS\,(\, f \,,\, x_- \,,\, x_+ \,,\, \delta \,) \\
\quad x'_- \leftarrow x_- + (1-r)(x_+ - x_-);\ x'_+ = x_- + r(x_+ - x_-);\ \text{compute } f(\, x'_- \,),\ f(\, x'_+ \,); \\
\quad \textbf{while}(\ x_+ - x_- > \delta\ )\ \textbf{do} \\
\qquad \textbf{if}(\ f(\, x'_- \,) > f(\, x'_+ \,)\ ) \\
\qquad\quad \textbf{then } \{\ x_- \leftarrow x'_-;\ x'_- \leftarrow x'_+;\ x'_+ \leftarrow x_- + r(x_+ - x_-);\ \text{compute } f(\, x'_+ \,);\ \} \\
\qquad\quad \textbf{else } \{\ x_+ \leftarrow x'_+;\ x'_+ \leftarrow x'_-;\ x'_- \leftarrow x_- + (1-r)(x_+ - x_-);\ \text{compute } f(\, x'_- \,); \}
\end{array}
$$

- After k iterations, $x_+^k - x_-^k = Dr^k +$ stops when $Dr^k \leq \delta \implies$ stops when $k \approx 4.78 \log(\, D \,/\, \delta \,)$ (**check**): exponentially faster = can work with "small" δ

- With $r = 0.5$ but two $f(\cdot)$-evals it would be $k \approx 6.64 \log(\, D \,/\, \delta \,)$ (**check**)

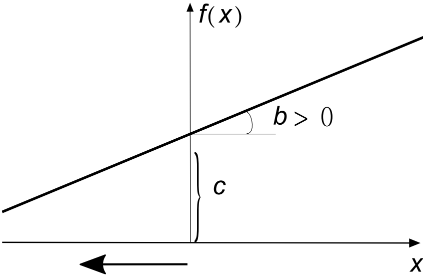
- Asymptotically optimal if no other information available [1, p. 355] ($r^k = F_{n-k}/F_{n-k+1}$, F_i = Fibonacci, slightly better if n fixed beforehand)

- $\delta \neq \varepsilon$, but f L-c $\implies A(\, x^k \,) \leq \varepsilon$ when $k \approx 4.78 \log(\, LD \,/\, \varepsilon \,)$ (**check**)

Outline

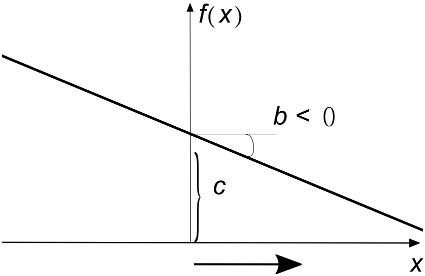
- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster

- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



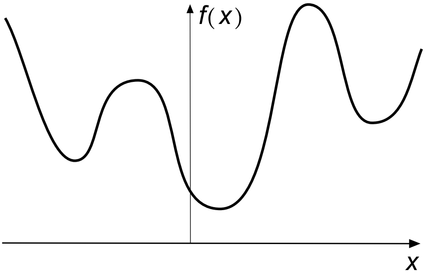
- easy for linear $f(x) = bx\ [+c]$: always left if $b > 0$,

- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\implies$ faster



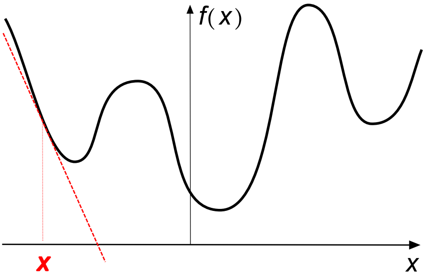
- easy for linear $f(x) = bx\ [+c]$: always left if $b > 0$, right if $b < 0$

- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



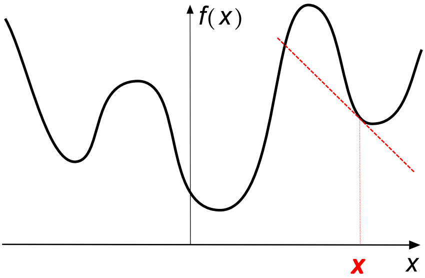
- easy for linear $f(x) = bx\ [+c]$:
  always left if $b > 0$, right if $b < 0$
- f nonlinear $\Longrightarrow$

- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



- easy for linear $f(x) = bx\ [+c]$: always left if $b > 0$, right if $b < 0$
- f nonlinear $\Longrightarrow$ first-order model of f at x: $L_x(z) = f'(x)(z - x) + f(x)$
- best linear approximation of f at x: $L_x(z) \approx f(z)\ \ \forall\, z \in [x - \varepsilon\,,\, x + \varepsilon]$ for some (small) $\varepsilon > 0$

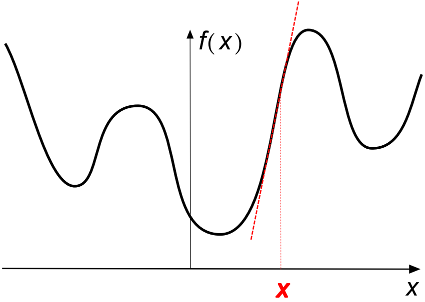
- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



- easy for linear $f(x) = bx\,[+c]$: always left if $b > 0$, right if $b < 0$
- f nonlinear $\Longrightarrow$ first-order model of f at x: $L_x(z) = f'(x)(z - x) + f(x)$
- best linear approximation of f at x: $L_x(z) \approx f(z)\ \forall z \in [x - \varepsilon, x + \varepsilon]$ for some (small) $\varepsilon > 0$

- Trusty old (first) derivative $f'(x)$ [6, §2.3]
- $f'(x)$ = slope of the tangent line to the graph of f in x: $f'(x) < 0 \Longrightarrow f$ decreasing at x,

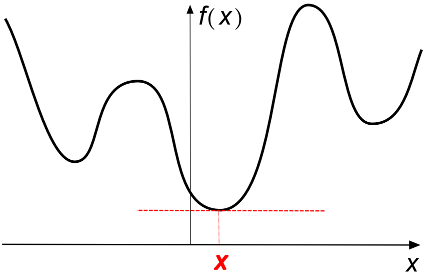
- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



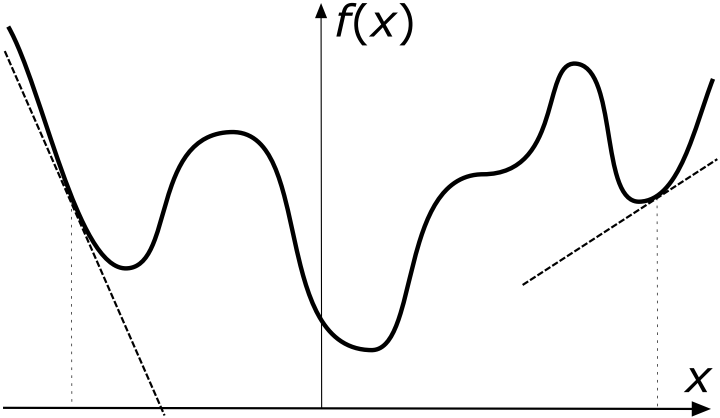
- easy for linear $f(x) = bx\ [+c]$:
  always left if $b > 0$, right if $b < 0$
- f nonlinear $\Longrightarrow$ first-order model of f at x: $L_x(z) = f'(x)(z - x) + f(x)$
- best linear approximation of f at x:
  $L_x(z) \approx f(z)\ \forall z \in [x - \varepsilon, x + \varepsilon]$
  for some (small) $\varepsilon > 0$

- Trusty old (first) derivative $f'(x)$ [6, §2.3]
- $f'(x) =$ slope of the tangent line to the graph of f in x:
  $f'(x) < 0 \Longrightarrow f$ decreasing at x, $f'(x) > 0 \Longrightarrow f$ increasing at x

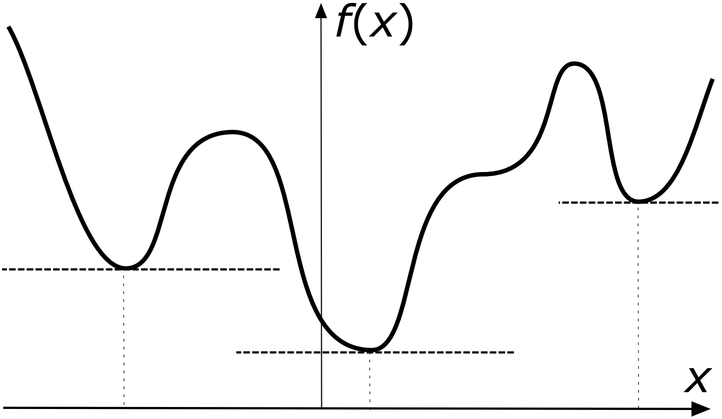
- Why do we need two points? To see in which direction f is decreasing
- If we could see this directly we could make it with one point $\Longrightarrow$ faster



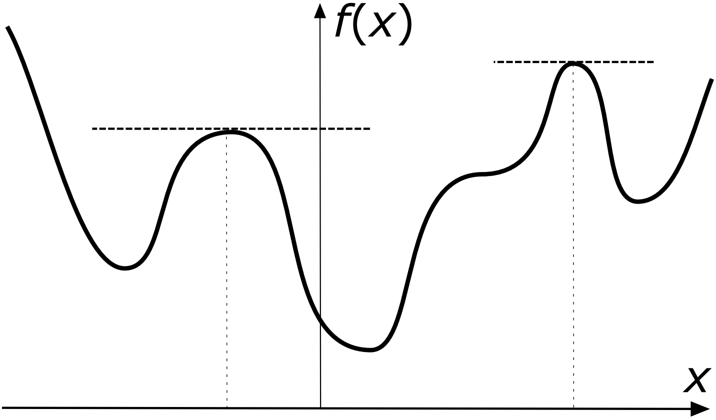
- easy for linear $f(x) = bx\,[+c]$: always left if $b > 0$, right if $b < 0$
- f nonlinear $\Longrightarrow$ first-order model of f at x: $L_x(z) = f'(x)(z - x) + f(x)$
- best linear approximation of f at x: $L_x(z) \approx f(z)\ \forall z \in [x - \varepsilon, x + \varepsilon]$ for some (small) $\varepsilon > 0$
- Trusty old (first) derivative $f'(x)$ [6, §2.3]
- $f'(x)$ = slope of the tangent line to the graph of f in x: $f'(x) < 0 \Longrightarrow f$ decreasing at x, $f'(x) > 0 \Longrightarrow f$ increasing at x
- x_* local minimum $\simeq f'(x_*) = 0 \equiv$ root of $f' \equiv$ stationary point



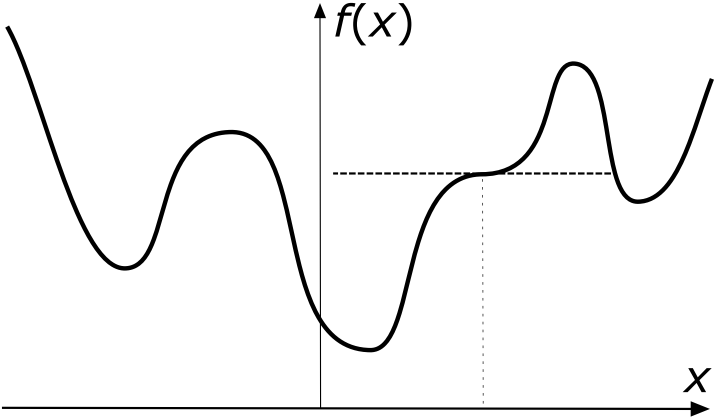
- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum



- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)

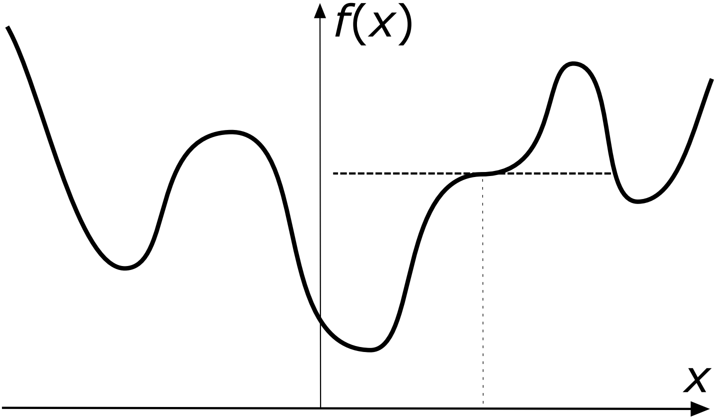


- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima



- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima

  . . . as well as in saddle points

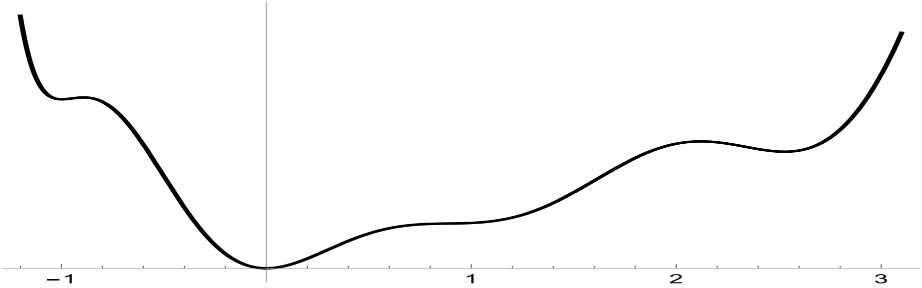


- If $f'(x) < 0$ or $f'(x) > 0$, x clearly cannot be a local minimum
- Hence, $f'(x) = 0$ in all local minima (hence in the global one as well)
- However, $f'(x) = 0$ also in local (hence global) maxima

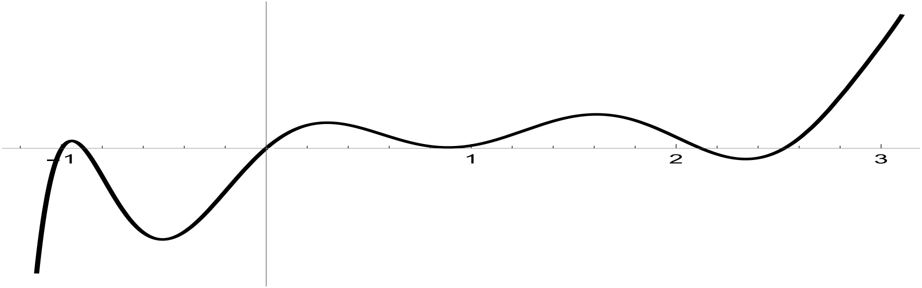
  ... as well as in saddle points
- How do I tell them apart? Look at $f'' = [f']' =$ second derivative

A polynominal example: roots of f' are the "interesting" points

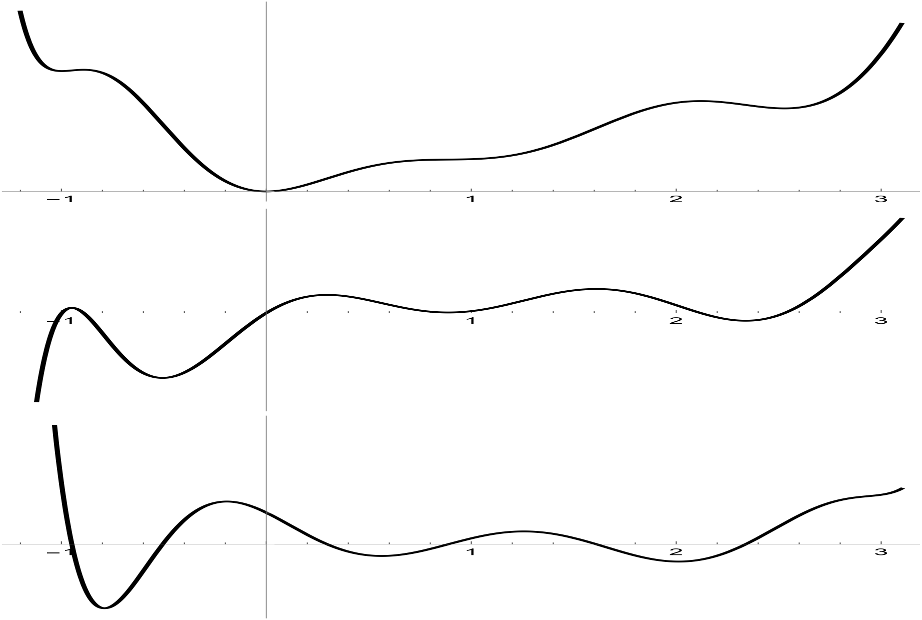
$$f(x) = \frac{91}{30}x^2 - \frac{19}{6}x^3 - \frac{54}{25}x^4 + \frac{93}{23}x^5 - \frac{23}{36}x^6 - \frac{121}{93}x^7 + \frac{72}{91}x^8 - \frac{13}{74}x^9 + \frac{9}{640}x^{10}$$



$$f'(x) = \frac{91}{15}x - \frac{19}{2}x^2 - \frac{216}{25}x^3 + \frac{465}{23}x^4 - \frac{23}{6}x^5 - \frac{847}{93}x^6 + \frac{576}{91}x^7 - \frac{117}{74}x^8 + \frac{9}{64}x^9$$



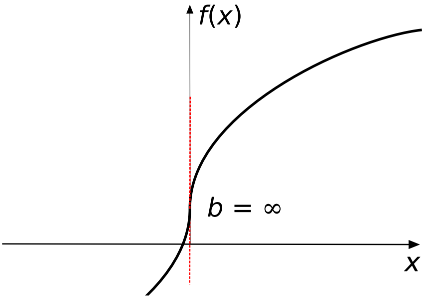
$$[\, f'(\,x\,)\,]' = \frac{91}{15} - 19x - \frac{648}{25}x^2 + \frac{1860}{23}x^3 - \frac{115}{6}x^4 - \frac{1694}{31}x^5 + \frac{576}{13}x^6 - \frac{468}{37}x^7 - \frac{81}{64}x^8$$



Mathematically speaking: Derivatives [6, §2.3]

- Derivative: $f'(x) = \lim_{t \to 0} [\, f(x+t) - f(x) \,] \,/\, t$
- Easy closed-forms for most reasonable functions

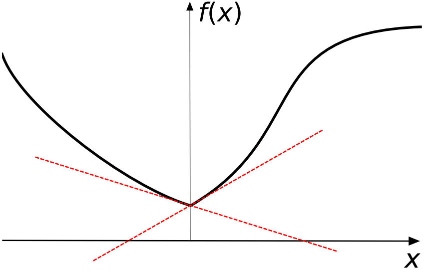
- Derivative: $f'(x) = \lim_{t \to 0}[f(x + t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



- . . . provided the limit is finite

Mathematically speaking: Derivatives [6, §2.3]

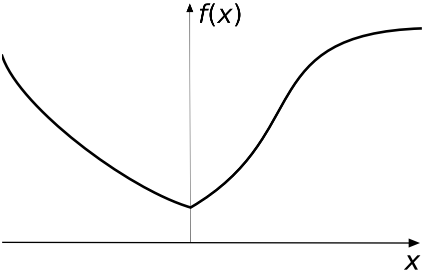
- Derivative: $f'(x) = \lim_{t\to 0}[f(x+t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



- . . . provided the limit is finite
- . . . and it exists at all

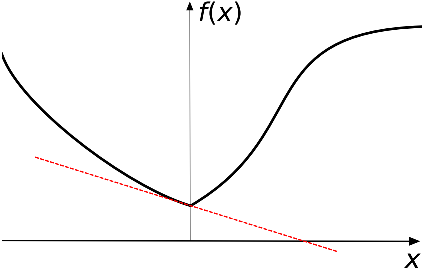
Mathematically speaking: Derivatives [6, §2.3]

- Derivative: $f'(x) = \lim_{t \to 0}[\, f(x+t) - f(x)\,] \,/\, t$
- Easy closed-forms for most reasonable functions



- ... provided the limit is finite
- ... and it exists at all
- Left and right derivatives:

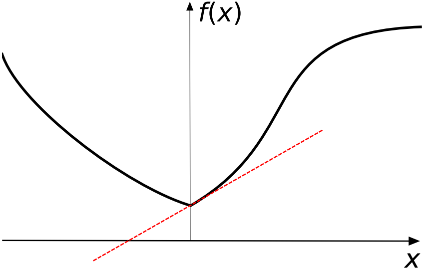
- Derivative: $f'(x) = \lim_{t \to 0}[f(x+t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



- ...provided the limit is finite
- ...and it exists at all
- Left and right derivatives:

  $f'_-(x) = \lim_{t \to 0_-}[f(x+t) - f(x)]/t$

- Derivative: $f'(x) = \lim_{t \to 0}[f(x+t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



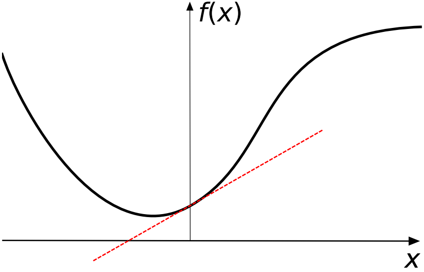
- ... provided the limit is finite
- ... and it exists at all
- Left and right derivatives:

$$f'_{-}(x) = \lim_{t \to 0_{-}}[f(x+t) - f(x)]/t$$

$$f'_{+}(x) = \lim_{t \to 0_{+}}[f(x+t) - f(x)]/t$$

Mathematically speaking: Derivatives [6, §2.3]

- Derivative: $f'(x) = \lim_{t \to 0} [f(x+t) - f(x)] / t$
- Easy closed-forms for most reasonable functions



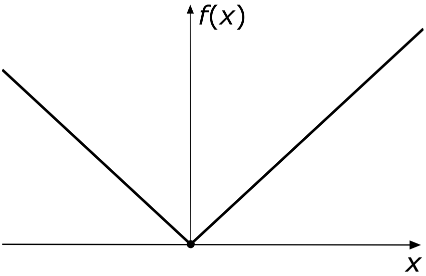
- ... provided the limit is finite
- ... and it exists at all
- Left and right derivatives:

  $f'_{-}(x) = \lim_{t \to 0_{-}} [f(x+t) - f(x)] / t$

  $f'_{+}(x) = \lim_{t \to 0_{+}} [f(x+t) - f(x)] / t$

- f differentiable at x if $f'(x)$ $\exists$ finite $\equiv$ $f'_{-}(x) = f'_{+}(x)$ ($\Longleftarrow$ $\exists$ finite)

- Derivative: $f'(x) = \lim_{t\to 0}[f(x+t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



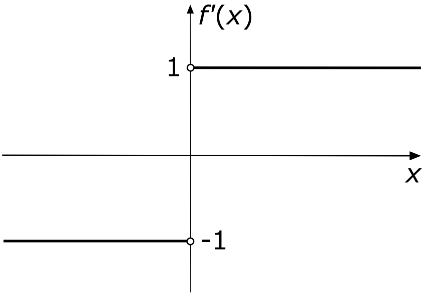
- ... provided the limit is finite
- ... and it exists at all
- Left and right derivatives:

  $f'_-(x) = \lim_{t\to 0_-}[f(x+t) - f(x)]/t$

  $f'_+(x) = \lim_{t\to 0_+}[f(x+t) - f(x)]/t$

- f differentiable at x if $f'(x)$ $\exists$ finite $\equiv$ $f'_-(x) = f'_+(x)$ ($\Longleftarrow$ $\exists$ finite)
- Nondifferentiable functions happen in practice: $f(x) = |x| = \max\{x, -x\}$

- Derivative: $f'(x) = \lim_{t\to 0}[f(x+t) - f(x)]/t$
- Easy closed-forms for most reasonable functions



- ...provided the limit is finite
- ...and it exists at all
- Left and right derivatives:

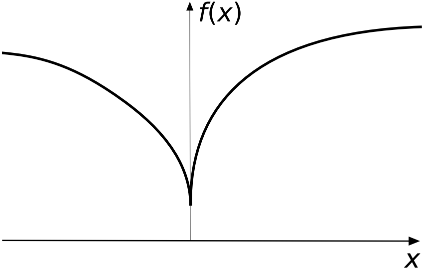
  $f'_-(x) = \lim_{t\to 0_-}[f(x+t) - f(x)]/t$

  $f'_+(x) = \lim_{t\to 0_+}[f(x+t) - f(x)]/t$

- f differentiable at x if $f'(x)$ $\exists$ finite $\equiv$ $f'_-(x) = f'_+(x)$ ($\Longleftarrow$ $\exists$ finite)
- $f'(x) = -1$ if $x < 0$, $f'(x) = +1$ if $x > 0$, $f'(x) =$??? if $x = 0$

Mathematically speaking: Derivatives [6, §2.3]

- Derivative: $f'(x) = \lim_{t\to 0}[\,f(x+t) - f(x)\,]\,/\,t$
- Easy closed-forms for most reasonable functions



- ...provided the limit is finite
- ...and it exists at all
- Left and right derivatives:

  $f'_-(x) = \lim_{t\to 0_-}[\,f(x+t) - f(x)\,]\,/\,t$

  $f'_+(x) = \lim_{t\to 0_+}[\,f(x+t) - f(x)\,]\,/\,t$

- f differentiable at x if $f'(x)$ $\exists$ finite $\equiv$ $f'_-(x) = f'_+(x)$ ($\Longleftarrow$ $\exists$ finite)
- Can be as different as $-\infty$ and $+\infty$
- f differentiable at x $\Longrightarrow$ f continuous at x, but $\Longleftarrow$ does not hold

Exercise: Prove it

- Derivatives of many simple functions are known, (almost always) continuous
  - $[\, x^k \,]' = k x^{k\,-1}$
  - $[\, e^x \,]' = e^x \quad , \quad [\, \ln(\, x \,) \,]' = 1 \,/\, x$
  - $[\, \sin(\, x \,) \,]' = \cos(\, x \,) \quad , \quad [\, \cos(\, x \,) \,]' = -\sin(\, x \,)$
- Many functional operations (almost always) preserve differentiability
  - $[\, \alpha f(\, x \,) + \beta g(\, x \,) \,]' = \alpha f'(\, x \,) + \beta g'(\, x \,)$
  - $[\, f(x) \cdot g(\, x \,) \,]' = f'(\, x \,) \cdot g(\, x \,) + f(\, x \,) \cdot g'(\, x \,)$
  - $[\, f(x) / g(\, x \,) \,]' = [\, f'(\, x \,) \cdot g(\, x \,) - f(\, x \,) \cdot g'(\, x \,) \,] \,/\, g(\, x \,)^2$
  - $[\, f(\, g(\, x \,) \,) \,]' = f'(\, g(\, x \,) \,) \cdot g'(\, x \,)$ (chain rule)
- A few common functional operations don't:

  $\max\{\, f(\, x \,) \,,\, g(\, x \,) \,\} \quad , \quad \min\{\, f(\, x \,) \,,\, g(\, x \,) \,\}$
- In general automatic differentiation well-developed, available, fast [8]

  $\Longrightarrow$ actually (writing code to) compute derivatives not our business

- $f' \in C^0 \;\equiv\; f \in C^1 \;\equiv\; f$ continuously differentiable $\Longrightarrow\; f \in C^0$

- $f'' \in C^0 \;\equiv\; f \in C^2 \;\equiv\; f' \in C^1 \;\Longrightarrow\; f' \in C^0 \;\Longrightarrow\; f \in C^1 \;\Longrightarrow\; f \in C^0$

- $f \in C^1$ globally L-c on (open) $X \;\Longrightarrow\; |\, f'(\, x \,)\,| \leq L \;\; \forall\, x \in X$

Exercise: Prove it, is $\Longleftarrow$ true?

Exercise: Formally prove $\exists\, f \in C^0$ but not L-c on some finite $X = [\, x_- \,,\, x_+ \,]$

- Extreme value theorem [6, Th. 2.2.9]: $f \in C^0$ on $X = [\, x_- \,,\, x_+ \,]$ (closed) finite
  $\Longrightarrow \quad \max\{\, f(\, x \,) \,:\, x \in X \,\} < \infty \quad , \quad \min\{\, f(\, x \,) \,:\, x \in X \,\} > -\infty$

- $f \in C^1$ on X finite (closed) $\Longrightarrow\; f$ globally L-c on X

- Best possible case ever: $f \in C^2$ (actually, C^3) on finite X
  $\Longrightarrow\;$ both f and f' globally L-c on X

- In simple cases, you get the answer by a closed formula (surprised?)
- $f(x) = bx\ [+c]$ (linear), $f'(x) = b = 0 \implies \nexists\, x$ if $b \neq 0$, $\forall\, x$ if $b = 0$
- $f(x) = ax^2 + bx\ [+c]$ (quadratic, $a \neq 0$), $f'(x) = 2ax + b = 0 \implies$ $x = -b \,/\, 2a$ unique minimum if $a > 0$, maximum if $a < 0$
- Generalise almost only to polynomials whose root have a closed formula (degree 3, some degree 4)
- Little hope for most trascendental / trigonometric / mixed unless you are very lucky
- Need an algorithm for solving nonlinear equations

- f' continuous + intermediate value theorem [6, Th. 2.2.10] $\Longrightarrow$
  $f'(x_-) < 0 \;\wedge\; f'(x_+) > 0 \Longrightarrow \exists\, x \in [\, x_- \,,\, x_+ \,]$ s.t. $f'(x) = 0$
- Theorems breed algorithms: dichotomic search

```
procedure x = DS ( f , x− , x+ , ε )
  do forever                 // invariant: f′( x− ) < −ε, f′( x+ ) > ε
      x ← in_middle_of( x− , x+ ); compute f′( x );
      if( | f′( x ) | ≤ ε )  then break;
      if( f′( x ) < 0 )      then x− ← x;
                             else x+ ← x;
```

- Trivial choice: in_middle_of(x_- , x_+){ **return**($(x_+ + x_-)/2$) }
- Linear convergence with $r = 0.5 < 0.618 \Longrightarrow$
  $k \approx 3.32 \log(D/\delta) < 4.78 \log(D/\delta)$ (err, who is δ?)
- f' L-c with constant L $\equiv$ L-smooth $\Longrightarrow$ $k \approx 3.32 \log(LD/2\varepsilon)$ (**check**)
- Does it show in practice?

- What if the assumption is not satisfied?

- Obvious solution:

$$
\boxed{
\begin{array}{ll}
\Delta x \leftarrow 1; & // \text{ or whatever value} > 0 \\
\textbf{while}(\ f'(\,x_+\,) \leq -\varepsilon\ )\ \textbf{do} & \\
\qquad x_+ \leftarrow x_+ + \Delta x;\ \Delta x \leftarrow 2\Delta x; & // \text{ or whatever factor} > 1
\end{array}
}
$$

- Of course, the same "in reverse" for x_- ($\Delta x = -1$)

- Will work in practice for all "reasonable" function

- Works if f coercive: $\lim_{|\,x\,|\to\infty} f(\,x\,) = \infty$

Exercise: construct an example where $x_+ \,/\, x_-$ exist but are not found

- If $f_* = -\infty$, $x_\pm$ may $\to \pm\infty$ "proving" unboundedness ($f(\,x_\pm\,) \to -\infty$) but how do you stop? (need a "finite $-\infty$")

Outline

- Choosing x "right in the middle" just the simplest approach:
  better if x is close to x_* (ideally, $x = x_*$ would stop in one iteration)
- One knows a lot about f: $f(x_-)$, $f(x_+)$, $f'(x_+)$, $f'(x_-)$, let's use that
- Powerful general idea: construct a model of f based on known information
- Quadratic interpolation: $ax^2 + bx + c$ that "agrees" with f at x_+, x_-
- Three parameters, four conditions, something's gotta give (three cases)
- One way: $2ax_+ + b = f'(x_+)$, $2ax_- + b = f'(x_-) \implies$
  $$a = \frac{f'(x_+) - f'(x_-)}{2(x_+ - x_-)} \quad , \quad b = \frac{x_+ f'(x_-) - x_- f'(x_+)}{x_+ - x_-}$$
- Minimum solves $2ax + b = 0$ (c irrelevant) $\equiv$
  $$x = \frac{x_- f'(x_+) - x_+ f'(x_-)}{f'(x_+) - f'(x_-)}$$
  "method of false position"
  a.k.a. "secant formula"

  always in the middle between x_+ and x_- (**check**)

Exercise: develop the other cases of quadratic interpolation and discuss them

- Very general issue: the model is an estimate $\Longrightarrow$ wrong $\Longrightarrow$ bad choices
- In this case, the model can be "very skewed":
  $f'(\,x_+\,) \gg -f'(\,x_-\,) \implies x \approx x_- \quad , \quad f'(\,x_+\,) \ll -f'(\,x_-\,) \implies x \approx x_+$
- Can lead to very short steps $\Longrightarrow$ slow convergence
- General remedy: never completely trust the model $\equiv$ regularise, stabilise, ...
- In this case: minimum guaranteed decrease $\sigma \leq 0.5$ (safeguard)
  $$x \leftarrow \max\{\,x_- + \sigma(x_+ - x_-)\,,\ \min\{\,x_+ - \sigma(x_+ - x_-)\,,\,x\,\}\,\}$$
- Worst case: linear convergence with $r = 1 - \sigma$
- Hopefully (much) faster than that when the model is "right"
- Does it really show in practice? And how much faster?

- Quadratic interpolation has superlinear convergence if started "close enough":
  [5, Th. 2.4.1] $f \in C^3$, $f'(x_*) = 0$ and $f''(x_*) \neq 0 \implies$
  $\exists\, \delta > 0$ s.t. $x^0 \in [\, x_* - \delta \,,\, x_* + \delta \,] \implies \{\, x^i \,\} \to x_*$ with $p = (1 + \sqrt{5}) \,/\, 2$
  ($1 < p = g \approx 1.618 < 2$, don't you just love maths?)

- This proves "very fast" already, but can we make it even faster?

- Quadratic interpolation has superlinear convergence if started "close enough":
  [5, Th. 2.4.1] $f \in C^3$, $f'(\, x_* \,) = 0$ and $f''(\, x_* \,) \neq 0 \implies$
  $\exists\, \delta > 0$ s.t. $x^0 \in [\, x_* - \delta \,,\, x_* + \delta \,] \implies \{\, x^i \,\} \to x_*$ with $p = (1 + \sqrt{5}) \,/\, 2$
  ($1 < p = g \approx 1.618 < 2$, don't you just love maths?)
- This proves "very fast" already, but can we make it even faster?
- Four conditions $\implies$ can fit a cubic polynomial and use its minima
- Rather tedious to write down, analyse and implement [5, § 2.4.2][4, p. 57]
- Theoretically pays: cubic interpolation has quadratic convergence ($p = 2$)
- Seems to work pretty well in practice

Exercise: (not for the faint of heart): develop cubic interpolation

- Better model of $f \equiv f'$ $\implies$ better guess of x_* $\implies$ faster
- Better model $\impliedby$ either more points or more (higher-order) derivatives
- Newton's method (tangent method): first-order model of f' at x^i

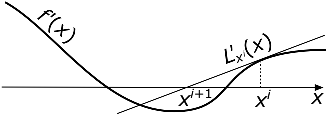
  $$L'_i(x) = L'_{x^i}(x) = f'(x^i) + f''(x^i)(x - x^i) \approx f'(x)$$

- Better model of $f \equiv f'$ $\Longrightarrow$ better guess of x_* $\Longrightarrow$ faster
- Better model $\Longleftarrow$ either more points or more (higher-order) derivatives
- Newton's method (tangent method): first-order model of f' at x^i

  $$L_i'(x) = L_{x^i}'(x) = f'(x^i) + f''(x^i)(x - x^i) \approx f'(x)$$

- Solve $L_i'(x) = 0 \approx f'(x) = 0 \quad \equiv$

  $x = x^i - f'(x^i) \,/\, f''(x^i)$

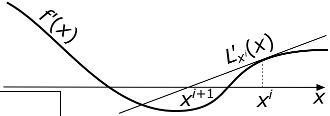


- Better model of $f \equiv f'$ $\Longrightarrow$ better guess of x_* $\Longrightarrow$ faster
- Better model $\Longleftarrow$ either more points or more (higher-order) derivatives
- Newton's method (tangent method): first-order model of f' at x^i

  $$L'_i(x) = L'_{x^i}(x) = f'(x^i) + f''(x^i)(x - x^i) \approx f'(x)$$

- Solve $L'_i(x) = 0 \approx f'(x) = 0 \quad \equiv$

  $x = x^i - f'(x^i) / f''(x^i)$



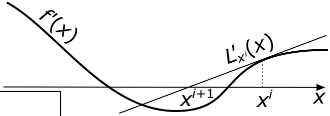

```
procedure x = NM ( f , x , ε )
  while( | f'( x ) | > ε ) do
    x ← x − f'( x ) / f''( x );    // what if f''( x ) = 0?
```

- Better model of $f \equiv f'$ $\Longrightarrow$ better guess of x_* $\Longrightarrow$ faster
- Better model $\Longleftarrow$ either more points or more (higher-order) derivatives
- Newton's method (tangent method): first-order model of f' at x^i

  $$L'_i(x) = L'_{x^i}(x) = f'(x^i) + f''(x^i)(x - x^i) \approx f'(x)$$

- Solve $L'_i(x) = 0 \approx f'(x) = 0 \quad \equiv$

  $x = x^i - f'(x^i) / f''(x^i)$



procedure $x = NM(f, x, \varepsilon)$
  while($|f'(x)| > \varepsilon$) **do**
    $x \leftarrow x - f'(x) / f''(x)$;   // what if $f''(x) = 0$?

- Alternative view (**check**): minimize second-order model of f at x^i

  $$Q_i(x) = Q_{x^i}(x) = f(x^i) + f'(x^i)(x - x^i) + f''(x^i)(x - x^i)^2 / 2$$

  (but Newton's actually a method to solve nonlinear equations)
- Converges fast (at all!) only if started "close enough" to x_* [1, Th. 8.2.3]
- Would require globalization (possible), will see in $\neq$ context

- Second-order Taylor's formula: $\forall\, z\ \exists\, w \in [\, x\,,\, z]$ s.t.

  $$f(\, z\,) - L_x(\, z\,) \;=\; f''(\, w\,)(\, z - x\,)^2/2 \qquad [6,\ \text{Th. } 2.5.4]$$

  "the error of L_x in z is $(\, z - x\,)^2\times$ the value of f'' somewhere in the middle"

- Hypotheses: $f \in C^3$, $f'(\, x_*\,) = 0$ and $f''(\, x_*\,) \neq 0$

- Thesis: $\exists\, \delta > 0$ s.t. $x^0 \in [\, x_* - \delta\,,\, x_* + \delta\,] \implies \{\, x^k\,\} \to x_*$ with $p = 2$

- Proof: $x^{i+1} - x_* = x^i - x_* + (\, f'(\, x_*\,) - f'(\, x^i\,)\,)\,/\,f''(\, x^i\,)$

  $$= [\, f'(\, x_*\,) - f'(\, x^i\,) - f''(\, x^i\,)(\, x_* - x^i\,)\,]\,/\,f''(\, x^i\,)$$

  Taylor's formula for f': $\exists\, w \in [\, x^i\,,\, x^*\,]$ s.t.

  $$f'(\, x_*\,) - f'(\, x^i\,) + f''(\, x^i\,)(\, x_* - x^i\,) = f'''(\, w\,)(\, x_* - x^i\,)^2/2$$

  $$\implies \qquad x^{i+1} - x_* = [\, f'''(\, w\,)\,/\,2f''(\, x^i\,)\,](\, x^i - x_*\,)^2$$

  $\exists\, \delta > 0$ s.t. $|\, f''(\, x\,)\,| \geq k_2 > 0$ and $|\, f'''(\, w\,)\,| \leq k_1 < \infty$ (**check**)

  $\forall\, x\,,\, w \in [\, x_* - \delta\,,\, x_* + \delta\,] \implies |\, x^{i+1} - x_*\,| \leq [\, k_1\,/\,2k_2\,](\, x^i - x_*\,)^2$

  $k_1(\, x^i - x_*\,)\,/\,2k_2 \leq 1 \implies |\, x^{i+1} - x_*\,| < |\, x^i - x_*\,| \implies$

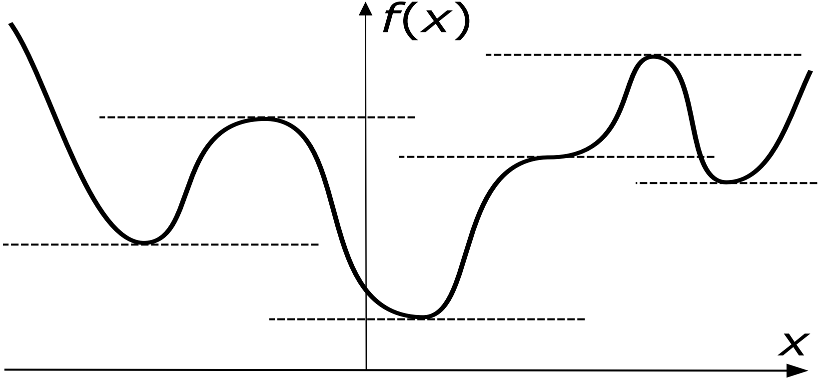
  $\{\, x^i\,\} \to x_*$ and the convergence is quadratic

Outline

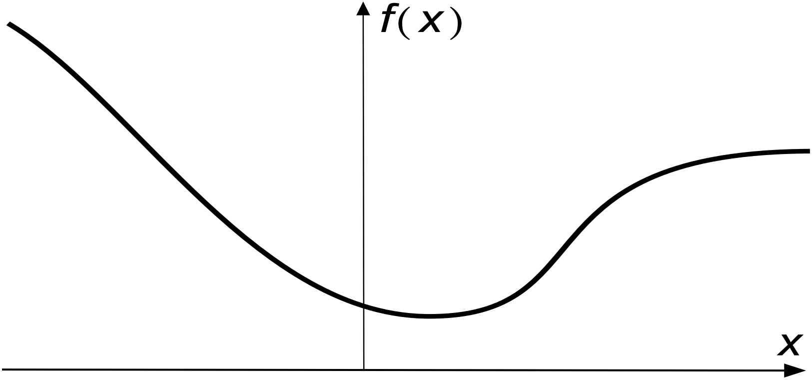
How about global optimization?

- What does this all tells about global optimization?

- What does this all tells about global optimization?
  Sadly, not much at all, unless strong assumptions are made

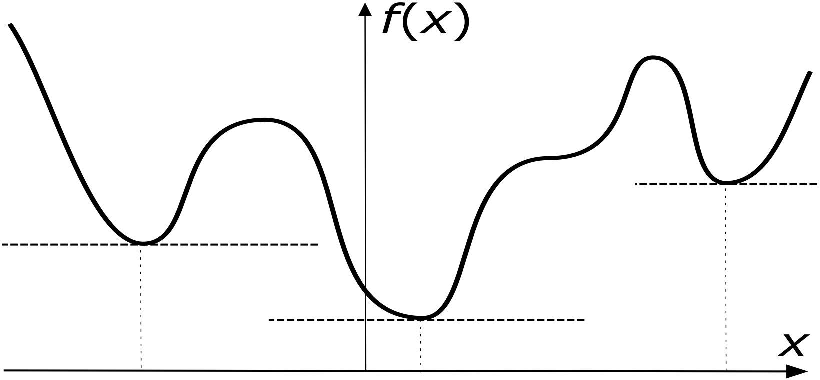


- What does this all tells about global optimization?
  Sadly, not much at all, unless strong assumptions are made



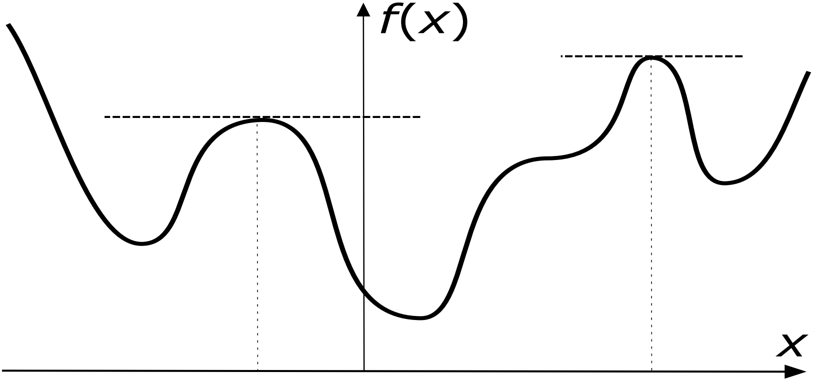
- The obvious one would be unimodal, but not easy to verify/construct

- What does this all tells about global optimization?
  Sadly, not much at all, unless strong assumptions are made



- Intuitively: f has local not global minima

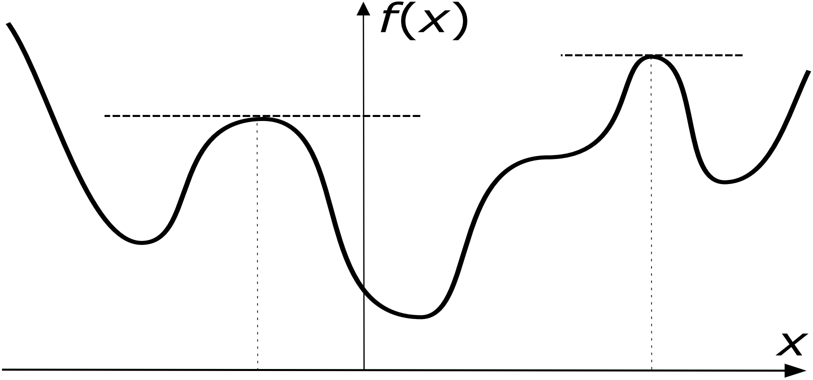
- What does this all tells about global optimization?
  Sadly, not much at all, unless strong assumptions are made



- Intuitively: f has local not global minima $\Longrightarrow$ has local maxima

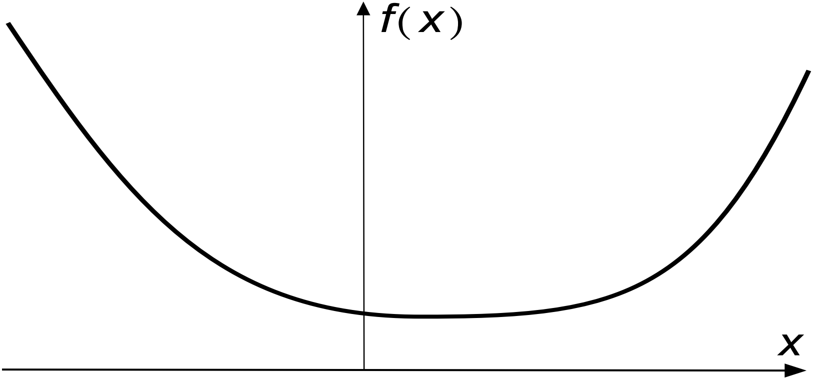
- What does this all tells about global optimization?

  Sadly, not much at all, unless strong assumptions are made



- Intuitively: f has local not global minima $\Longrightarrow$ has local maxima
- Avoid it: stationary point $\Longrightarrow$ local minima $\equiv$ $f'(x) = 0 \Longrightarrow f''(x) \geq 0$

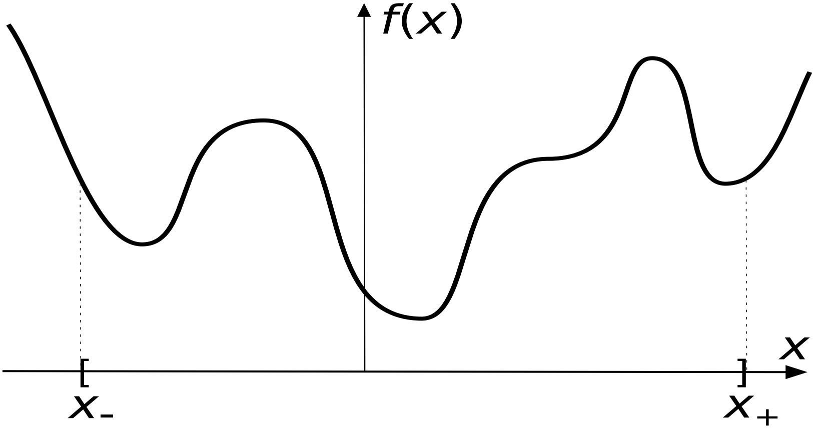
- What does this all tells about global optimization?
  Sadly, not much at all, unless strong assumptions are made



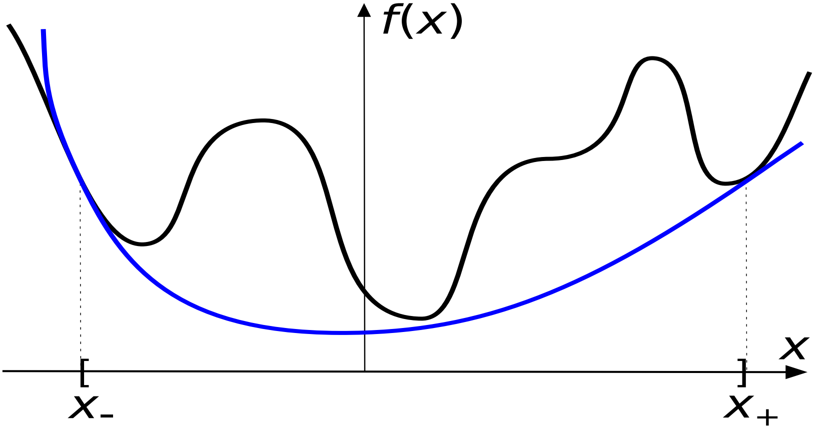
- Intuitively: f has local not global minima $\Longrightarrow$ has local maxima
- Avoid it: stationary point $\Longrightarrow$ local minima $\equiv$ $f'(x) = 0 \Longrightarrow f''(x) \geq 0$
- Sufficient condition: $f''(x) \geq 0 \ \forall x \in \mathbb{R} \Longrightarrow f$ convex

- Convex $\simeq$ f' is monotone nondecreasing $\simeq$ $f'' \geq 0$
- Not really because convex $\not\Longrightarrow$ C^1 (even less C^2), will see
- Some functions are convex + a few operations preserve convexity (will see)
  $\Longrightarrow$ the convex world is relatively large
  $\Longrightarrow$ can construct complicated (multivariate) convex functions/sets
- Plenty of theory [2] and software [10]
- Many models are purposely constructed convex (SVM) so that (global) optimization is easy
- "If you have the choice, choose convex"
- What if you don't and really need the global optimum?
- Will only say little here, but plenty of ways to satisfy your curiosity [9]

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, but using a clever guide

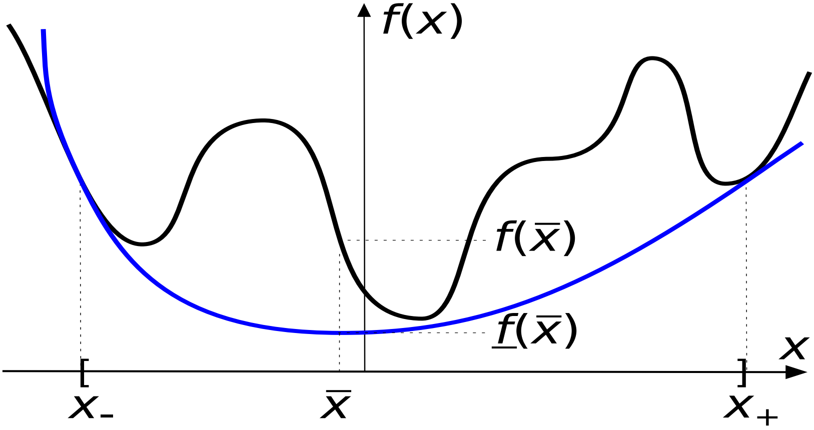


- Sift through all $X = [\, x_- \, , \, x_+ \,]$, but using a clever guide



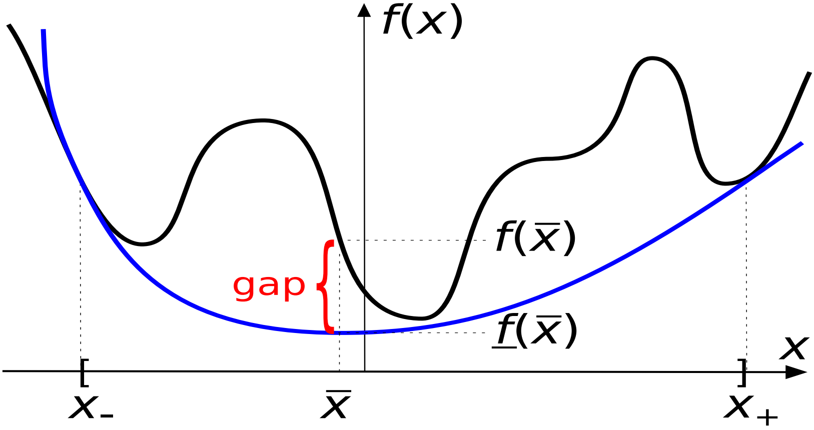
- Convex lower approximation $\underline{f}$ of nonconvex f on X

- Sift through all $X = [\, x_- \,, \, x_+ \,]$, but using a clever guide



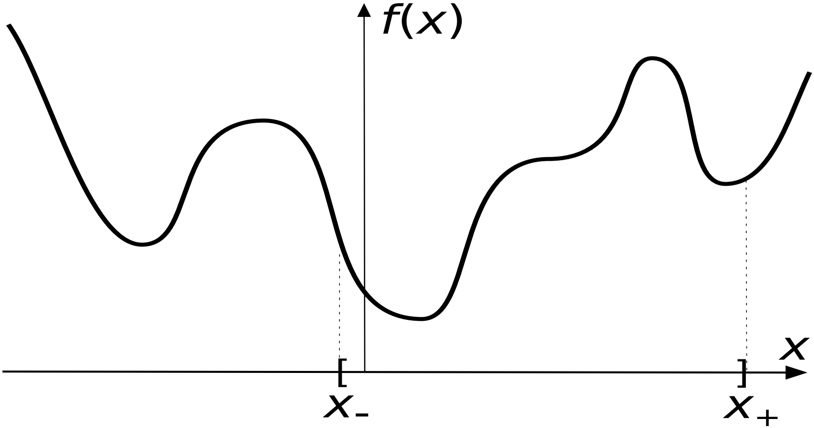
- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\, \bar{x}\,) \leq f_* \leq f(\, \bar{x}\,)$

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, but using a clever guide



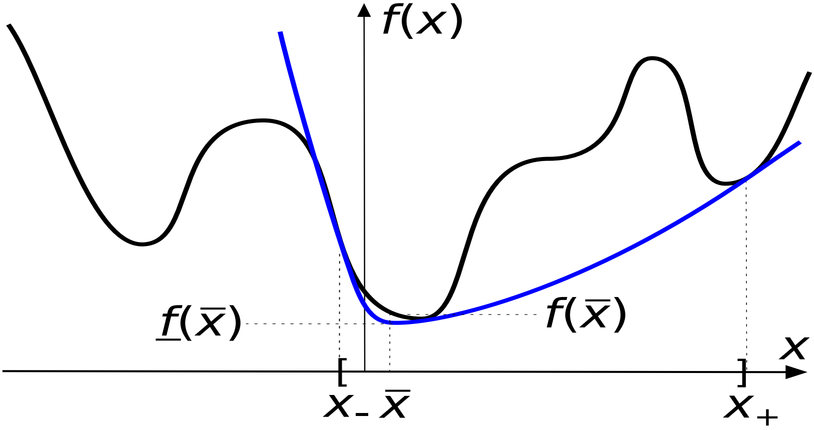
- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large,

- Sift through all $X = [\, x_- \,,\, x_+ \,]$, but using a clever guide



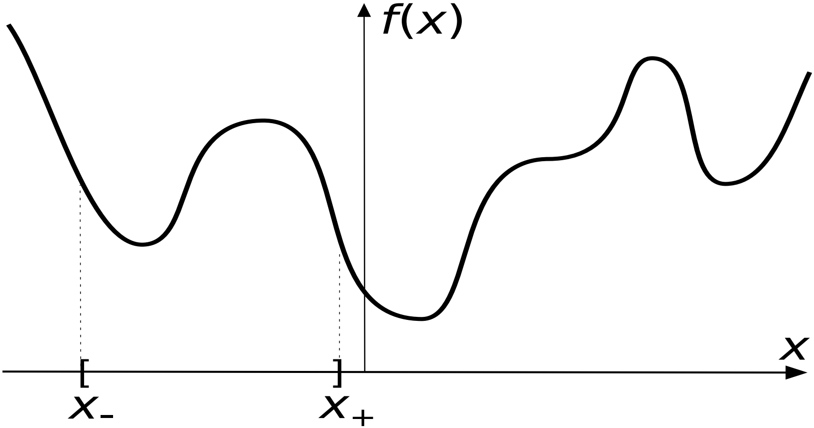
- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition X and iterate

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, but using a clever guide



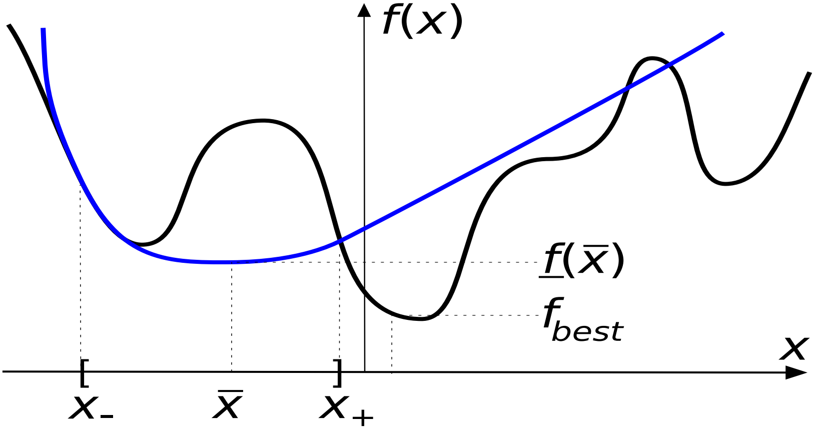
- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition X and iterate
- $\underline{f}$ depends on partition, smaller partition (hopefully) $\Longrightarrow$ better gap

- Sift through all $X = [\, x_- \,,\, x_+ \,]$, but using a clever guide



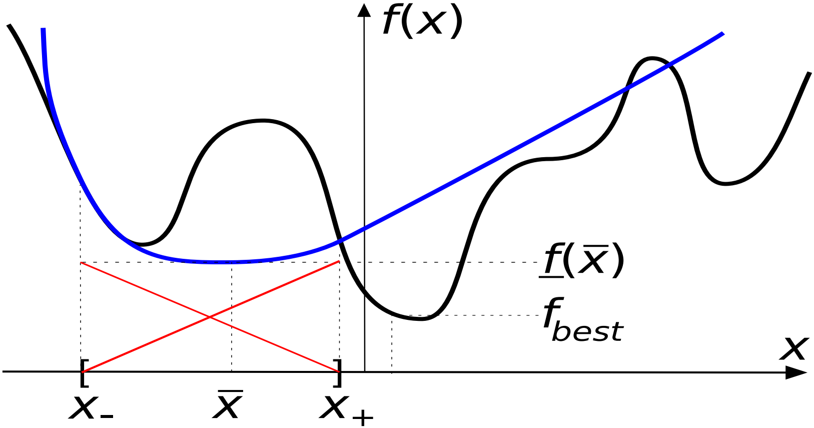
- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition X and iterate
- If on some partition

- Sift through all $X = [\, x_- \,,\, x_+ \,]$, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition X and iterate
- If on some partition $\underline{f}(\,\bar{x}\,) \geq$ best f-value so far,

- Sift through all $X = [\, x_- \, , \, x_+ \,]$, but using a clever guide



- Convex lower approximation $\underline{f}$ of nonconvex f on X
- "Easily" find local $\equiv$ global minimum $\bar{x}$, giving $\underline{f}(\,\bar{x}\,) \leq f_* \leq f(\,\bar{x}\,)$
- If gap $f(\,\bar{x}\,) - \underline{f}(\,\bar{x}\,)$ too large, partition X and iterate
- If on some partition $\underline{f}(\,\bar{x}\,) \geq$ best f-value so far, partition killed for good

Is Something Like This Efficient?

- In a word? Surely not in worst-case:
  keep dicing and slicing X until pieces "very small" $\Longrightarrow$ exponential

- However, in practice it depends on:
  - "how much nonconvex" f really is
  - how good $\underline{f}$ is as a lower approximation of f

- Clever approach: carefully choose your nonconvexities, e.g., integer variables

- In a word? Surely not in worst-case:
  keep dicing and slicing X until pieces "very small" $\Longrightarrow$ exponential
- However, in practice it depends on:
  - "how much nonconvex" f really is
  - how good $\underline{f}$ is as a lower approximation of f
- Clever approach: carefully choose your nonconvexities, e.g., integer variables
- Mixed-Integer Linear Programs: all is "trivial" when integer fixed/relaxed

Is Something Like This Efficient?

- In a word? Surely not in worst-case:
  keep dicing and slicing X until pieces "very small" $\Longrightarrow$ exponential

- However, in practice it depends on:
  - "how much nonconvex" f really is
  - how good $\underline{f}$ is as a lower approximation of f

- Clever approach: carefully choose your nonconvexities, e.g., integer variables

- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically)

Is Something Like This Efficient?

- In a word? Surely not in worst-case:
  keep dicing and slicing X until pieces "very small" $\Longrightarrow$ exponential
- However, in practice it depends on:
  - "how much nonconvex" f really is
  - how good $\underline{f}$ is as a lower approximation of f
- Clever approach: carefully choose your nonconvexities, e.g., integer variables
- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically)
  $\not\Longrightarrow$ always efficient, $\underline{f}$ often "bad" $\equiv$ bounds weak $\Longrightarrow$ exponential

- In a word? Surely not in worst-case:
  keep dicing and slicing X until pieces "very small" $\Longrightarrow$ exponential

- However, in practice it depends on:
  - "how much nonconvex" f really is
  - how good $\underline{f}$ is as a lower approximation of f

- Clever approach: carefully choose your nonconvexities, e.g., integer variables

- Mixed-Integer Nonlinear Convex Programs: still "easy" (less so numerically)
  $\not\Longrightarrow$ always efficient, $\underline{f}$ often "bad" $\equiv$ bounds weak $\Longrightarrow$ exponential

- (Mixed-Integer) Nonlinear Nonconvex Programs: finding any $\underline{f}$ complex
  - rewrite the expression of f in terms of unary/binary functions
  - apply specific convexification formulæ for each function

- Good news: implemented in available, well-engineered solvers and immensely less inefficient in practice than blind search

- Yet, immensely less efficient in practice than local optimization

Outline

- Global (constrained or not) optimization difficult (impossible) in general
- Local (unconstrained) optimization much easier, useful in general:
  once you know how to do unconstrained you can do constrained

Wrap up

- Global (constrained or not) optimization difficult (impossible) in general
- Local (unconstrained) optimization much easier, useful in general: once you know how to do local you can try global
- Algorithms are slow / medium / fast, "nicer" problems have faster algorithms
- The more continuous derivatives you have, the nicer the problem
- Derivatives $\Longrightarrow$ first- and second-order model
- f "complicated", model looks like f (close to x) and simple
- But the map is not the world, never blindly trust a model
- Fundamental concepts we will use all the time, let's move to $n > 1$

[1] M.S. Bazaraa, H.D. Sherali, C.M. Shetty *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, 2006

[2] S. Boyd, L. Vandenberghe *Convex Optimization*, https://web.stanford.edu/~boyd/cvxbook Cambridge University Press, 2008

[3] P. Hansen, B. Jaumard "Lipschitz Optimization" in *Handbook of Global Optimization – Nonconvex optimization and its applications*, R. Horst and P.M. Pardalos (Eds.), Chapter 8, 407–494, Springer, 1995

[4] J. Nocedal, S.J. Wright, *Numerical Optimization – second edition*, Springer Series in Operations Research and Financial Engineering, 2006

[5] W. Sun, Y.-X. Yuan, *Optimization Theory and Methods – Nonlinear Programming*, Springer Optimization and Its Applications, 2006

[6] W.F. Trench, *Introduction to Real Analysis* https://ramanujan.math.trinity.edu/wtrench/texts/TRENCH_REAL_ANALYSIS.PDF Free Hyperlinked Edition 2.04, December 2013

[7] L. Serafino *Optimizing Without Derivatives: What Does the No Free Lunch Theorem Actually Say?* Notices of the AMS 61(7):750–755, 2014 https://www.ams.org/notices/201407/rnoti-p750.pdf

[8] AutoDiff Org: https://www.autodiff.org

[9] CommaLab: https://commalab.di.unipi.it/courses

[10] CVX: https://cvxr.com

Outline

Optimization Problems

Local optimization

Faster local optimization

Fastest local optimization

A Fleeting Glimpse to Global Optimization

Wrap up & References

Solutions

- Take $\delta = \varepsilon \,/\, L$; then, $\forall\, y \in [\, x - \delta \,,\; x + \delta \,]$
  $|\, f(\, y\,) - f(\, x\,)\,| \leq L|\, x - y\,| \leq L\delta \leq L(\, \varepsilon \,/\, L\,) = \varepsilon$ [**back**]

- Note: we'll have a much simpler proof later, after we present the relationships between L-c and derivatives
  $f(\, x\,) = x^2$ is locally but not globally L-c; we prove this for $x \geq 0$, but the same arguments work for $x \leq 0$ (the function is symmetric)
  $\delta > 0 \implies 0 \leq f(\, x + \delta\,) - f(\, x\,) = (\, 2x + \delta\,)\delta \leq 3x\delta$ if $\delta \leq x$; hence, f is L-c at x with Lipschitz constant $3x$ in some (right) interval around x
  $\delta > 0 \implies f(\, x + \delta\,) - f(\, x\,) = (\, 2x + \delta\,)\delta \geq 2x\delta$; hence, f cannot be L-c at x with Lipschitz constant less than $2x$, and that value is not bounded as $x \to \infty$
  Symmetric arguments works for left intervals $(f(\, x\,) - f(\, x - \delta\,))$ [**back**]

- The standard example is $f(x) = \sqrt{|x|}$, which is easily verified to be continuous and to become "infinitely steep" as $x \to 0$, because it is the inverse function of $y = x^2$ (for $x \geq 0$), and x^2 becomes "infinitely flat" as $x \to 0$ Again, we'll have a much simpler proof later, after we present the relationships between L-c and derivatives; in fact, the proof is so much simpler that it is not worth proceeding now, we just wait until we have the right tools [**back**]

- Let x_* be any optimal solution in X; by definition it belongs to (at least) one interval $[x_i, x_{i+1}]$, with $x_{i+1} - x_i \leq 2\varepsilon / L$. Assume that $x_* - x_i \leq x_{i+1} - x_*$ (the other case is analogous); then $x_* - x_i \leq \varepsilon / L$. Hence, L-c gives $f(x_i) - f(x_*) \leq L| x_i - x_* | \leq \varepsilon$ [**back**]

- Basically done this already: $Dr^k < \varepsilon \equiv r^k < \varepsilon / D \equiv \log(r^k) < \log(\varepsilon / D) \equiv k \log(r) < \log(\varepsilon / D) \equiv k > \log(\varepsilon / D) / \log(r)$ as $r < 1 \implies \log(r) < 0$
  Hence, $k \geq \log(D / \varepsilon) / \log(1 / r) = \log(D / \varepsilon) / (-\log(r))$
  Now, $\log(1 / 0.618) \approx \log(1.618) \approx 0.21$, $1 / 0.21 \approx 4.78$ [**back**]

- Golden ratio search has $r = 0.5$: $1 / \log(1 / r) \approx 3.32$. But each iteration of the algorithm requires two function evaluations, so the factor is ≈ 6.64: less iterations, but more evaluations [**back**]

- Since $f(\cdot)$ is L-c, $d^i = |x^i - x_*| \leq \delta \implies r^i = f^i - f_* \leq Ld^i \leq L\delta$. Hence, to get $r^i \leq \varepsilon$ it is sufficient to ensure that $d^i \leq \varepsilon / L$, whence the bound [**back**]

- $\lim_{t\to 0}[f(x+t) - f(x)]/t = L$ finite $\implies \lim_{t\to 0} t([f(x+t) - f(x)]/t)$ $= \lim_{t\to 0} f(x+t) - f(x) = L \lim_{t\to 0} t = 0$ (limit of a product = product of the limits); that $\Longleftarrow$ does not hold is proven by $f(x) = |x|$ [**back**]

- $f(\cdot)$ L-c $\implies |f(x+t) - f(x)| \leq L|t| \equiv |[f(x+t) - f(x)]/t| \leq L$; now just take the $\lim_{t\to 0}$
  Yes, the other direction is also true: by the Mean Value Theorem [6, Theorem 2.3.9], $f(z) - f(x) = f'(w)(z - x)$ for some w in the interval of extremes x and z; take the $|\cdot|$ and use $|f'(w)| \leq L$ [**back**]

- Consider $f(x) = \sqrt[3]{x^2}$, whose derivative is $f'(x) = 2 / 3x^3$ (possibly written in the more complex but algebraic-proof form $(2x) / (3(x^2)^{2/3})$). Hence, $\lim_{x \to 0_-} f'(x) = -\infty$ and $\lim_{x \to 0_+} f'(x) = \infty$. In plain words, this is because the cubic root is the inverse function of x^3, which is "flat in 0"; inverse functions "exchange the axes", which means that if the graph of the function is "horizontal" as some x, then the graph of its inverse is "vertical" at the same x, which implies $f'(x) = \pm\infty$. Thus $f'(\cdot)$ is not bounded in any interval around 0, and therefore $f(\cdot)$ is not L-c there. Of course, $f'(x)$ is not continuous in 0 [**back**]

- $f'(x) = 0$ for some $x \in [\underline{x}^i, \overline{x}^i]$; $L - c$ of f' gives $|0 - f'(\underline{x}^i)| \leq L|x - \underline{x}^i|$ and $|f'(\overline{x}^i) - 0| \leq L|\overline{x}^i - x|$, whence $\min\{f'(\underline{x}^i), f'(\overline{x}^i)\} \leq L\min\{|x - \underline{x}^i|, |x - \underline{x}^i|\} \leq L\delta / 2$ (in the worst case, x is equidistant from the extremes); thus, the stopping criterion have to be satisfied when $\delta = 2\varepsilon / L$, i.e., within at most $3.32 \log(LD / 2\varepsilon)$ iterations [**back**]

- For $x_+ = \Delta x = 1$, the algorithm tries the iterates 1, 2, 4, 8, ..., i.e., 2^i. With $f(x) = \sin(\pi x + 3\pi/4) \implies f'(x) = \pi\cos(\pi x + 3\pi/4)$ we have $f'(2^i) = \pi\cos(\pi 2^i + 3\pi/4) = \pi\cos(3\pi/4) = -\pi\sqrt{2}/2 \approx -2.22$ (2^i is always even and $\cos(\cdot)$ has period 2π); that is, the algorithm always finds a "very negative" derivative and never stops, although $f(\cdot)$ has plenty of local minima. Clearly, by only very slightly changing the constants the counterexample would break down [**back**]

- $x_- f'(x_+) - x_+ f'(x_-) = x_- f'(x_+) - x_+ f'(x_-) + x_- f'(x_-) - x_- f'(x_-) = x_-(f'(x_+) - f'(x_-)) - f'(x_-)(x_+ - x_-)$. Divide by $f'(x_+) - f'(x_-)$ to get $x = x_+ + \alpha(x_+ - x_-)$ with $0 \leq \alpha = -f'(x_-)/(f'(x_+) - f'(x_-)) \leq 1$; it is then plain to see that $x_- \leq x \leq x_+$ [**back**]

- A full development would not be didactical. The four conditions are $ax_+^2 + bx_+ + c = f(x_+)$, $ax_-^2 + bx_- + c = f(x_-)$, $2ax_+ + b = f'(x_+)$, $2ax_- + b = f'(x_-)$; each three of them give a linear system with three equations in the three unknowns a, b, c that gives (not necessarily) different solutions (mind the special cases) and therefore quadratic models [**back**]

- No point to repeat [5, § 2.4.2][4, p. 57] here **[back]**

- $[\, Q_i\,]'(\, x\,) = L_i'(\, x\,) = f'(\, x^i\,) + f''(\, x^i\,)(\, x - x^i\,) = 0 \;\equiv$
  $x - x^i = -f'(\, x^i\,) \,/\, f''(\, x^i\,)$ **[back]**

- Since $f''(\, x_*\,) \neq 0$, $|\, f''(\, x_*\,)\,| > 0$; take e.g. $k_2 = |\, f''(\, x_*\,)\,| \,/\, 2$ [> 0], by continuity of $f''(\cdot)$ at x_*, $\exists\, \delta > 0$ s.t. $|\, 2k_2 - |\, f''(\, x\,)\,|\,| \leq k_2 \implies |\, f''(\, x\,)\,| \geq k_2 \;\forall\, x \in X$. Since $f'''(\cdot)$ is continuous, also $|\, f'''(\cdot)\,|$ is, hence $k_1 = \max\{\, |\, f'''(\, x\,)\,| \;:\; x \in X\,\} < \infty$ [6, Th. 2.2.9] **[back]**