MS-C2105 - Introduction to Optimization
Lecture 8

Fabricio Oliveira (with modifications by Harri Hakula)

Systems Analysis Laboratory
Department of Mathematics and Systems Analysis

Aalto University
School of Science

March 18, 2022

Outline of this lecture

Revision of calculus

General optimisation problems

Optimality conditions - unconstrained problems

Convexity of functions

One dimensional optimisation methods - line search

- Bisection method
- Newton's method

Reading: Taha: Chapter 20; Winston: Chapter 11

Tools from differential calculus

We focus on devising optimisation methods for general problems.

- No assumption of linearity.
- We consider first unconstrained problems.
- Later, we will include the consideration of constrains.

Let us first revise some important tools we will use.

Definition 1 (Limits and continuity)

Let $f : \mathbb{R} \to \mathbb{R}$ be a function. We say that $\lim_{x\to a} f(x) = c$ if, as x becomes closer to a, $f(x)$ becomes closer to c (asymptotically). Moreover, f is continuous at point a if $\lim_{x\to a} f(x) = f(a)$. If $\lim_{x\to a} f(x) = f(a)$ for all $a \in \mathbb{R}$, then function f is continuous.

Tools from differential calculus

Definition 2 (Differentiation)

The derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at $x = a$ (denoted $f'(a)$) is defined as

$$\lim_{\Delta x \to 0} \frac{f(a + \Delta x) - f(a)}{\Delta x}.$$

If $f'(x)$ exists for every $x \in \mathbb{R}$, then f is differentiable.

Differentiability is used to infer local behaviour of f

- It is direction dependent: $\lim_{\Delta x \to 0^+}$ and $\lim_{\Delta x \to 0^-}$. If they are the same for all x, $f(x)$ is continuous.
- $f'(a)$ can be thought as the slope of f at a.
- $f'(x) > 0$ means that the function is increasing at x; i.e., for a arbitrarily small $\epsilon > 0$, $f(x + \epsilon) > f(x)$.
- Likewise, $f'(x) < 0$ means that the function is decreasing at x.

Tools from differential calculus

Function	Derivative
a	0
x	1
$af(x)$	$af'(x)$
$f(x)+g(x)$	$f'(x)+g'(x)$
x^n	nx^{n-1}
e^x	e^x
a^x	$a^x\ln(a)$
$\ln(x)$	$\frac{1}{x}$
$[f(x)]^n$	$nf(x)^{n-1}f'(x)$
$e^{f(x)}$	$e^{f(x)}f'(x)$
$a^{f(x)}$	$a^{f(x)}f'(x)\ln a$
$\ln f(x)$	$\frac{f'(x)}{f(x)}$
$f(x)g(x)$	$f(x)g'(x)+f'(x)g(x)$
$\frac{f(x)}{g(x)}$	$\frac{g(x)f'(x)-f(x)g'(x)}{g(x)^2}$

Tools from differential calculus

Definition 3 (n-order derivative)

The n^{th}-order derivative $f^{(n)}(a)$ of f at a is the derivative of $f^{(n-1)}(a)$ at a.

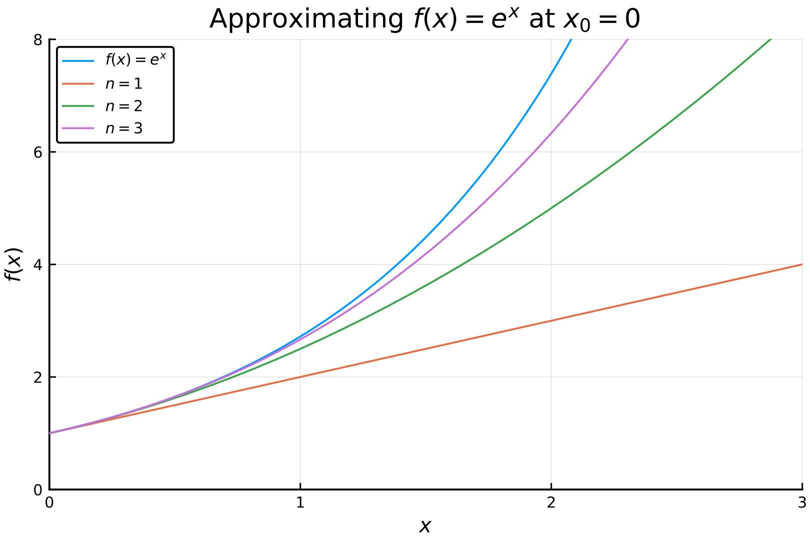
Higher derivatives are employed in Taylor series expansions, which are in turn used as general local approximations of function values.

Theorem 4 (Taylor's theorem)

Let f be n-times differentiable on an open interval containing x and x_0. Then, the Taylor series expansion of f is

$$\begin{aligned} f(x) &= f(x_0) + \frac{1}{1} f'(x_0)(x - x_0) + \frac{1}{1 \times 2} f''(x_0)(x - x_0)^2 + \dots \\ &+ \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n + R_{n+1}(x) \\ &= \sum_{i=0}^{n} \frac{1}{i!} f^{(i)}(x_0)(x - x_0)^i + R_{n+1}(x) \end{aligned}$$

Tools from differential calculus



Tools from differential calculus

Remarks:

1. The term $R_n(x)$ is called the residual. For some $c \in (x_0, x)$,

$$R_{n+1}(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}.$$

2. In particular, if $|f^{(n+1)}| \leq M$, then

$$R_{n+1}(x) \leq \frac{M|x - x_0|^{n+1}}{(n+1)!}.$$

3. Taylor's approximation is the Taylor's expansion without the residual term.
4. if $x_0 = 0$, Taylor's series reduce to the Maclaurin's series.

Nonlinear optimisation models

Nonlinear programming models are a more general class of optimisation problems.

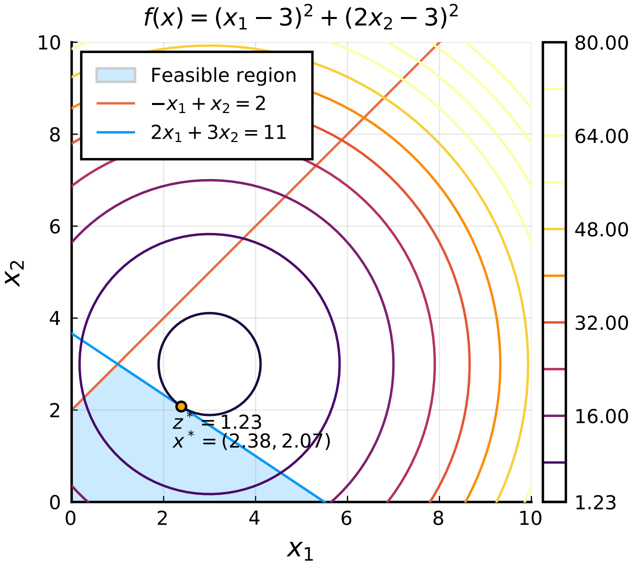
$$\begin{aligned} \text{min. } & f(x) \\ \text{s.t.: } & g_i(x) \leq 0, i = 1, \dots, m. \\ & x \in X \end{aligned}$$

Clearly, LP/ MIP models are a particular cases, to which dedicated efficient methods exist.

For more general problems

- optimal points might not be extreme points or be on the boundary of the feasible region.
- guarantees of global optimality might not exist.

Nonlinear optimisation models

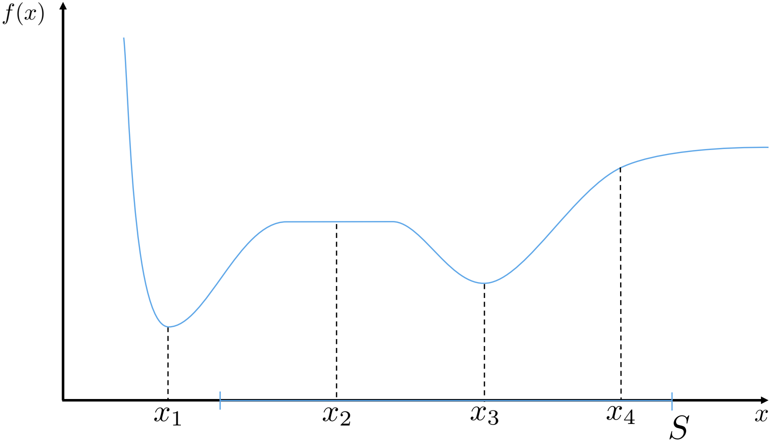


General optimality conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$. Consider the problem (P) : min. $\{f(x) : x \in S\}$. Some important terminology:

- feasible solution: $x \in S$;
- local optimal solution: $\overline{x} \in S$ that has a neighbourhood $N_\epsilon(\overline{x}) = \{x : ||x - \overline{x}|| \leq \epsilon\}$ for some $\epsilon > 0$ such that $f(\overline{x}) \leq f(x)$ for each $x \in S \cap N_\epsilon(\overline{x})$.
- global optimal solution: $\overline{x} \in S$ with $f(\overline{x}) \leq f(x)$ for all $x \in S$.

General optimality conditions



Necessary optimality conditions

Candidates to local optima must be critical points.

Theorem 5 (First-order optimality condition)

Let $\overline{x}$ be a local optimum for f in $N_\epsilon(\overline{x})$ and assume that f is differentiable. Then $f'(\overline{x}) = 0$.

Proof.

Suppose $\overline{x}$ is a local minimum. Then, there exists $\delta > 0$ for which $f(\overline{x}) \leq f(x)$ for all $x \in (\overline{x} - \delta, \overline{x} + \delta) \subset N_\epsilon(\overline{x})$.

1. for any $h \in (0, \delta)$, it holds that $\frac{f(\overline{x}+h)-f(\overline{x})}{h} \geq 0$.
   Thus, $\lim_{h \to 0^+} \frac{f(\overline{x}+h)-f(\overline{x})}{h} = f'(\overline{x}) \geq 0$.
2. for any $h \in (-\delta, 0)$, it holds that $\frac{f(\overline{x}+h)-f(\overline{x})}{h} \leq 0$.
   Thus, $\lim_{h \to 0^-} \frac{f(\overline{x}+h)-f(\overline{x})}{h} = f'(\overline{x}) \leq 0$.

From the above, we conclude that $f'(\overline{x}) = 0$. $\square$

Sufficient optimality conditions

The condition $f'(x) = 0$ does not imply local optimality.

- Points satisfying $f'(x) = 0$ are called stationary.
- An additional certificate is necessary to state optimality.

Theorem 6 (n^{th}-order optimality condition)

Suppose f has a stationary point at x_0 and that $f'(x_0) = \cdots = f^{(n-1)}(x_0) = 0$, while $f^{(n)}(x_0) \neq 0$. If $f^{(n)}$ is continuous, then

1. *if n is even and $f^{(n)}(x_0) > 0$, then x_0 is a local minimum.*
2. *if n is even and $f^{(n)}(x_0) < 0$, then x_0 is a local maximum.*
3. *if n is odd, then x_0 is an inflection point.*

Sufficient optimality conditions

Proof.

With the first $n-1$ derivatives vanishing, using Taylor's expansion, we have that

$$f(x) - f(x_0) = R_n(x_0) = \frac{f^{(n)}(c)}{n!}(x - x_0)^n$$

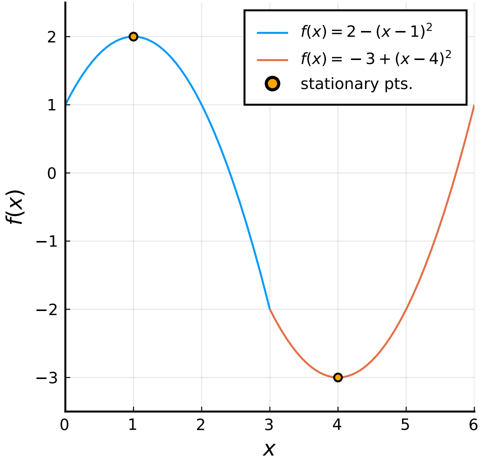
For n even, $(x - x_0)^n > 0$. $f^{(n)}(c)$ and $f^{(n)}(x^0)$ agree in sign, since they are arbitrarily close. Thus, $f(x) - f(x_0)$ will agree in sign of $f^{(n)}(x^0)$. If n is odd, $(x - x_0)^n$ and thus $f(x) - f(x_0)$ have opposite signs for $x < x_0$ and $x > x_0$. □

Sufficient conditions are posed considering $n = 2$, i.e.,

- if $f'(x_0) = 0$ and $f''(x_0) > 0$ then x_0 is a local minimum.
- if $f'(x_0) = 0$ and $f''(x_0) < 0$ then x_0 is a local maximum.
- if $f'(x_0) = 0$ and $f''(x_0) = 0$ then x_0 is a inflection point.

Necessary and sufficient optimality conditions

Example: $f(x) = \begin{cases} 2 - (x-1)^2, & \text{if } x < 3 \\ -3 + (x-4)^2, & \text{if } x \geq 3. \end{cases}$



Convexity of functions

Convexity is a key feature in optimisation. In convex optimisation problems, local optimality always implies global optimality.

Definition 7 (Convexity of a function)

Let $f : \mathbb{R}^n \to \mathbb{R}$. The function f is said to be convex if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

for each $x_1, x_2 \in \mathbb{R}^n$ and for each $\lambda \in (0, 1)$.

Remarks:

- f is concave if $-f$ is convex;
- if strict inequality holds, f is strictly convex.
- A nonconvex function can be convex within a specific set (e.g., $f(x) = x^3$ for $x \geq 0$)

Convexity of functions

Examples of convex functions:

1. $f(x) = a^\top x + b$;
2. $f(x) = e^x$;
3. $f(x) = x^p$ on $\mathbb{R}_+$ for $p \leq 0$ or $p \geq 1$; concave for $0 \leq p \leq 1$.
4. $f(x) = ||x||_p$ (p-norm);

Convexity of functions

Convexity preserving operations:

1. let $f_1, \dots, f_k : \mathbb{R}^n \to \mathbb{R}$ be convex. Then these are convex:
   - $f(x) = \sum_{j=1}^{k} \alpha_j f_j(x)$ where $\alpha_j > 0$ for $j = 1, \dots, k$;
   - $f(x) = \max\{f_1(x), \dots, f_k(x)\}$;
2. $f(x) = \frac{1}{g(x)}$ on S, where $g : \mathbb{R}^n \to \mathbb{R}$ is concave and $S = \{x : g(x) > 0\}$;
3. $f(x) = g(h(x))$, where $g : \mathbb{R} \to \mathbb{R}$ is a nondecreasing convex function and $h : \mathbb{R}^n \to \mathbb{R}$ is convex.
4. $f(x) = g(h(x))$, where $g : \mathbb{R}^m \to \mathbb{R}$ is convex and $h : \mathbb{R}^n \to \mathbb{R}^m$ is affine: $h(x) = Ax + b$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.

Example: $f(a) = (b - a^\top x)^2 + ||a||^2$. Is this function convex?

Convexity and optimality condition

The importance of convexity derives from this fundamental result:

Theorem 8 (Necessary and sufficient conditions)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. Then any local optimum $\overline{x}$ of f is also a global optimum.

Proof.

By contradiction. **Assume that $\overline{x}$ is a local minimum in $N_\epsilon(\overline{x})$**, but not a global minimum. Then, for some x we will have $f(x) < f(\overline{x})$. As f is convex, we have for every $\lambda \in [0, 1]$ that

$$\begin{aligned} f(\lambda\overline{x} + (1-\lambda)x) &\leq \lambda f(\overline{x}) + (1-\lambda)f(x) \\ &< \lambda f(\overline{x}) + (1-\lambda)f(\overline{x}) = f(\overline{x}). \end{aligned}$$

Now, $1-\lambda$ can be made arbitrarily small such that $\lambda\overline{x} + (1-\lambda)x$ belongs to $N_\epsilon(\overline{x})$, contradicting the **initial assumption**. $\square$

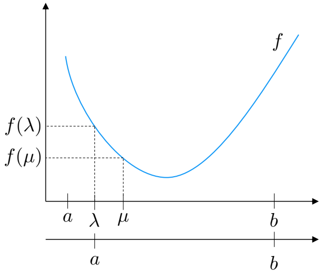
Line search methods

Most optimisation methods will iteratively search for points that satisfy first-order conditions.

One-dimensional (line) searches seek for $\overline{x}$ such that $f'(\overline{x}) = 0$.

Theorem 9 (Line search reduction)

Let $f : \mathbb{R} \to \mathbb{R}$ be convex over the interval $[a, b]$, and let $\lambda, \mu \in [a, b]$ such that $\lambda < \mu$. If $f(\lambda) > f(\mu)$, then $f(z) \geq f(\mu)$ for all $z \in [a, \lambda]$. If $f(\lambda) \leq f(\mu)$, then $f(z) \geq f(\lambda)$ for all $z \in [\mu, b]$.



Line search: bisection method

The bisection method uses gradient information to infer whether function is increasing or decreasing.

- Iteratively trim the search space (using Theorem 9).
- Relies on first-order conditions (presuming convexity/ sufficiency).

The main idea of the method is

1. if $f'(\lambda_k) = 0$, then λ_k is a minimiser.
2. if $f'(\lambda_k) > 0$, then, for $\lambda > \lambda_k$, we have $f(\lambda) \geq f(\lambda_k)$ since f is convex. Therefore, the new search interval becomes $[a_{k+1}, b_{k+1}] = [a_k, \lambda_k]$.
3. if $f'(\lambda_k) < 0$, the new search interval becomes $[a_{k+1}, b_{k+1}] = [\lambda_k, b_k]$.
4. To maximise interval reduction, we set $\lambda_k = \frac{1}{2}(b_k + a_k)$.

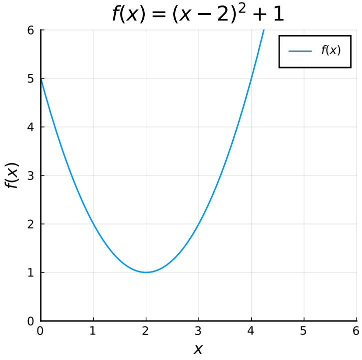
Line search: bisection method

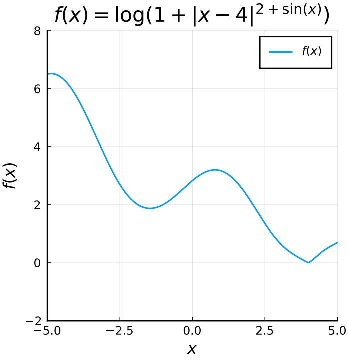
Algorithm Bisection method (minimisation)

```
1: initialise. tolerance l > 0, [a_0, b_0] = [a, b], k = 0.
2: while b_k − a_k > l do
3:     λ_k = (b_k + a_k)/2 and evaluate f′(λ_k).
4:     if f′(λ_k) = 0 then return λ_k.
5:     else if f′(λ_k) > 0 then
6:         a_{k+1} = a_k, b_{k+1} = λ_k.
7:     else
8:         a_{k+1} = λ_k, b_{k+1} = b_k.
9:     end if
10:    k = k + 1.
11: end while
12: return λ̄ = (a_k + b_k)/2.
```

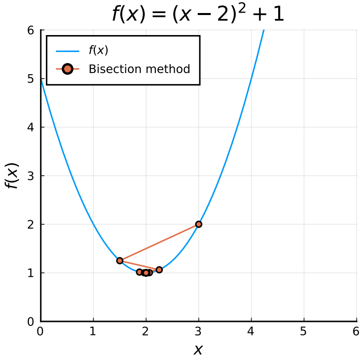
Remark: if maximising, the condition in Line 5 must be replaced with $f'(x) < 0$ and concavity is presumed.

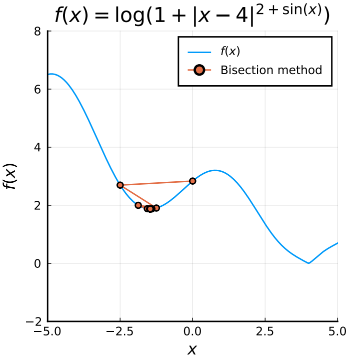
Bisection method: example



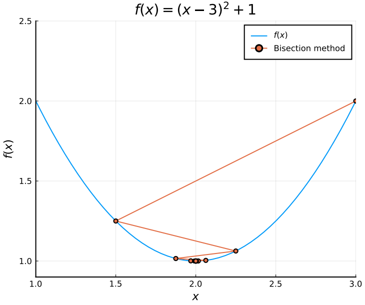


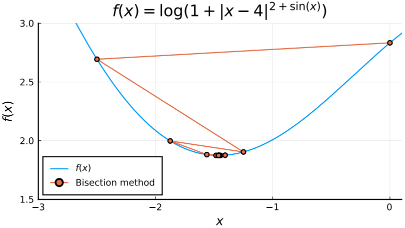
Bisection method: example





Bisection method: example (zoom in)





Line search: Newton's method

Explores the quadratic approximation q of f at a given point x_k:

$$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2$$

Letting x_{k+1} be the point at which $q'(x) = 0$, we have

$$q'(x_{k+1}) = f'(x_k) + f''(x_k)(x_{k+1} - x_k) = 0,$$

which implies

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

Remarks:

1. The search terminates when $|x_{k+1} - x_k| < \epsilon$ or $|f'(x_k)| < \epsilon$.
2. The same as applying Newton-Raphson's method (for finding roots of functions) to first-order optimality condition.

Line search: Newton's method

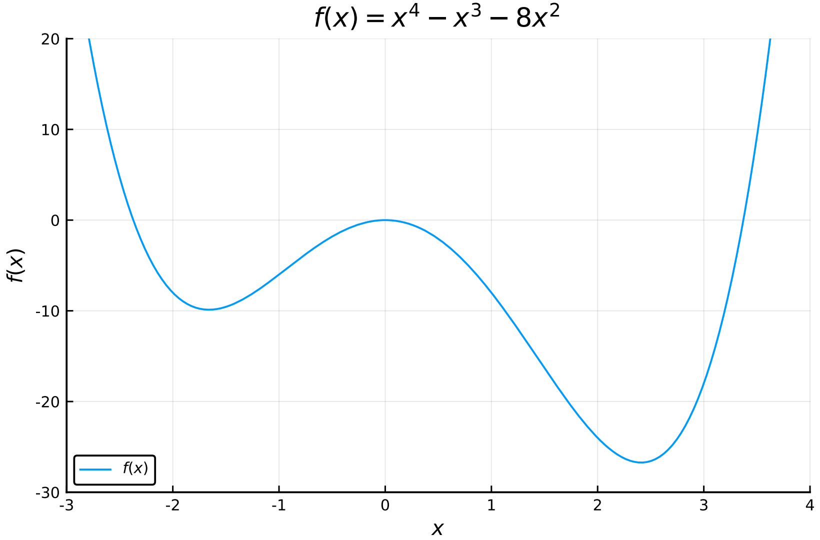
Algorithm Newton's method

1: **initialise.** tolerance $\epsilon > 0$, initial step size x_0, iteration count $k = 0$.
2: **while** $|f'(x_k)| > \epsilon$ **do**
3: $\quad x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$.
4: $\quad k = k + 1$.
5: **end while**
6: **return** $\overline{x} = x_k$.

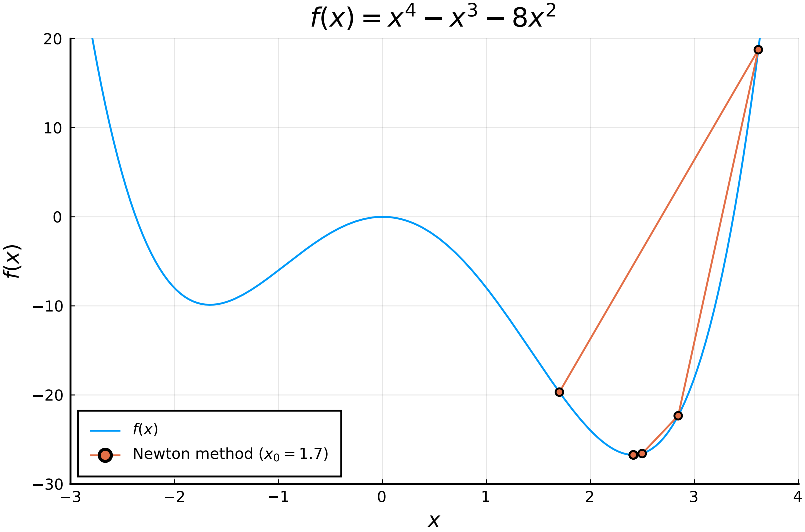
Remarks:

1. Newton's method is the backbone of several optimisation algorithms.
2. Has convergence issues if x_0 is too far away from optimal.
3. For quadratic problems, the approximation q is exact, meaning that only one iteration is needed.

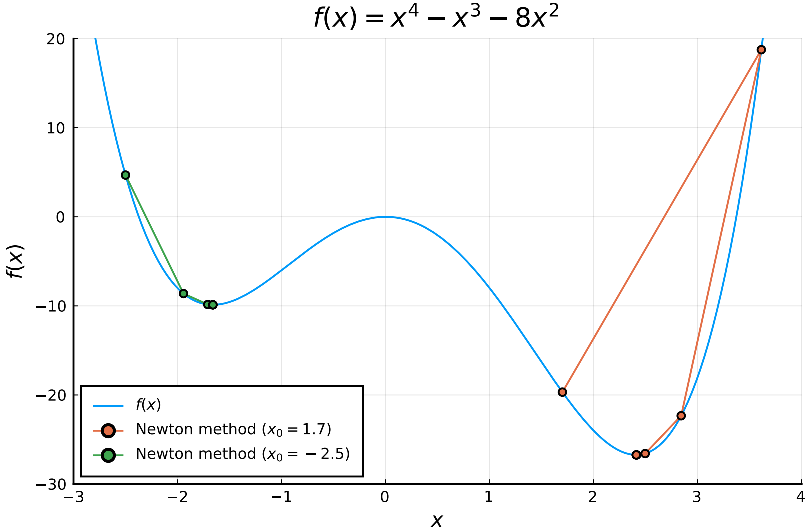
Newton's method: example



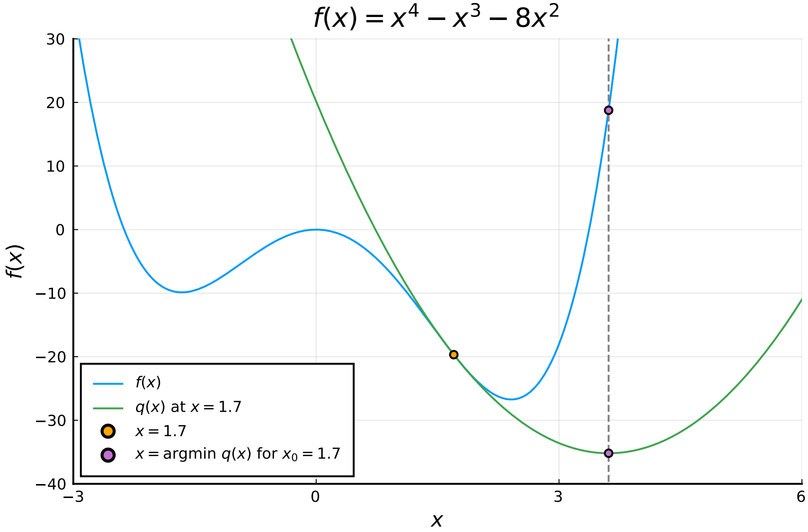
Newton's method: example



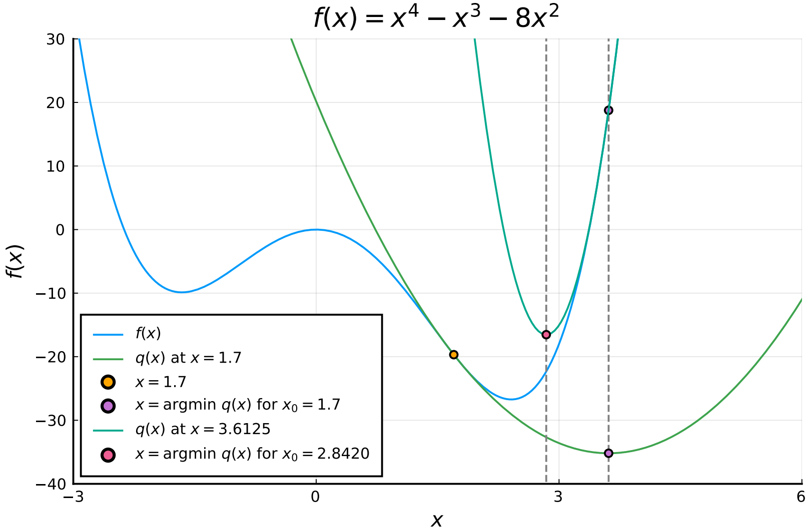
Newton's method: example



Newton's method: example



Newton's method: example



Newton's method: example