Numerical Analysis: Solving Nonlinear Equations

Computer Science, Ben-Gurion University

(slides based mostly on Prof. Ben-Shahar’s notes)

2019/2020, Fall Semester
1. Introduction
2. The Bisection Method
3. The Regula Falsi (False Position) Method
4. The Secant Method
5. Fixed Point
6. Convergence
7. The Newton (-Raphson) Method
   - The Method
   - Using Newton’s Method to Find a Square Root
   - A Root with Multiplicity
Our next topic for the several next lectures is solving nonlinear equations:

\[ f : \mathbb{R} \rightarrow \mathbb{R} \quad f(x) = 0 \quad x = ? \]

if \( x \) satisfies \( f(x) = 0 \), then it’s called a root of \( f \).
The seemingly-more general cases of solving

\[ f : \mathbb{R} \to \mathbb{R} \quad f(x) = b \quad x =? \]

can be rewritten in the original form above by defining a new function:

\[ f_{\text{new}} : \mathbb{R} \to \mathbb{R} \quad f_{\text{new}} : x \mapsto f(x) - b \quad f_{\text{new}}(x) = 0 \quad x =? \]

Note:

\[ f_{\text{new}}(x) = 0 \iff f(x) = b \]
Likewise,

\[
\begin{align*}
  f : \mathbb{R} &\to \mathbb{R} \\
  g : \mathbb{R} &\to \mathbb{R} \\
  f(x) &= g(x) \\
  x &= ?
\end{align*}
\]

can be handled via

\[
\begin{align*}
  f_{\text{new}} : \mathbb{R} &\to \mathbb{R} \\
  x &\mapsto f(x) - g(x) \\
  f_{\text{new}}(x) &= 0 \\
  x &= ?
\end{align*}
\]

Note:

\[
  f_{\text{new}}(x) = 0 \iff f(x) = g(x)
\]

Thus, WLOG, we will focus attention on solving \( f(x) = 0 \).
If $f$ has the form

$$f(x) = ax + b$$

(that is, $f$ is affine), where $a$ and $b$ are known real numbers, then we say the equation is linear, and the solution to

$$f(x) = 0$$

is given (assuming $a \neq 0$), of course, by

$$x = -\frac{b}{a}.$$
Some Nonlinear Equations are Easy

- Sometimes we know how to solve nonlinear equations.
- Examples:
  
  \[ x^3 - 8 = 0 \Rightarrow x = 2 \]  
  \( \text{(one solution)} \)

  \[ ax^2 + bx + c = 0 \Rightarrow x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]  
  \( \text{(at most two solutions)} \)

  \[ \sin(x) = 0 \Rightarrow x = k\pi, \ k \in \mathbb{Z} \]  
  \( \text{(\( \infty \)-many solutions)} \)
In general, however, if the equation is nonlinear we usually don’t know how to solve it analytically.
Example: The Ladder in the Mine
Based on Example 0.1 from Gerald and Wheatly

- $a$, $w_1$, and $w_2$ are fixed (determined by the mine’s geometry).
- Find the longest ladder (of 0 thickness) that can make the turn.
- The maximal ladder’s length, at a given angle $c$, is $L = L_1 + L_2$ where

\[
L_1 = \frac{w_1}{\sin(b)} \quad L_2 = \frac{w_2}{\sin(c)} \quad b = \pi - a - c
\]

\[
L = L_1 + L_2 = \frac{w_1}{\sin(\pi - a - c)} + \frac{w_2}{\sin(c)}
\]

- The angle $c$ determines $b$ (and vice versa).
- Thus, $L = \text{func}(c)$.
Example: The Ladder in the Mine

Based on Example 0.1 from Gerald and Wheatly

\[ L(c) = \frac{w_1}{\sin(\pi - a - c)} + \frac{w_2}{\sin(c)} \]

The longest ladder that can pass is

\[ \min_c L(c) . \]

Yes, minimum, not maximum. It's the bottleneck that matters.
Example: The Ladder in the Mine

Based on Example 0.1 from Gerald and Wheatly

\[ \min_c L(c) = \min_c \frac{w_1}{\sin(\pi - a - c)} + \frac{w_2}{\sin(c)} \]

- From calculus, we know that \( \hat{c} \triangleq \arg \min_c L(c) \) satisfies

\[ \frac{dL}{dc} \bigg|_{\hat{c}} = 0. \]

- Thus, we need to solve, for the unknown \( c \), the following equation:

\[ w_1 \frac{\cos(\pi - a - c)}{\sin^2(\pi - a - c)} - w_2 \frac{\cos(c)}{\sin^2(c)} = 0 \]

Equivalently:

\[ -w_1 \frac{\cos(a + c)}{\sin^2(a + c)} - w_2 \frac{\cos(c)}{\sin^2(c)} = 0 \]

- How can we solve this?
Nonlinear Equations

- Sometimes, in fact, even if a solution exists, an analytical form for it doesn’t exist.
- For example, the Abel-Ruffini theorem (also known as Abel’s impossibility theorem) states that this is the case for polynomials of degree higher than 4.

The theorem is named after Paolo Ruffini, who made an incomplete proof in 1799, and Niels Henrik Abel, who provided a proof in 1824.
Another example we mentioned earlier: how can we find the square root of a number using only the 4 arithmetic operations (addition, subtraction, multiplication, and division) and the operation of comparison?

Example

\[ x \in \mathbb{R}_{>0} \quad x^2 - 5 = 0 \quad x = ? \]
Our discussion so far motivates the need for non-analytical methods, based on simple operations, as means to solving

\[ f : \mathbb{R} \to \mathbb{R} \quad f(x) = 0 \quad x = ? \]
Trial and Error

- One possible approach is trial and error.
- Its disadvantage is that very little can be said, in general, about its expected performance in terms of how fast it will find an approximated solution.
So we want methods that not only approximate a root but also lend themselves to some useful analysis.

More generally (than the context of solving nonlinear equations), here are our requirements from numerical methods:

- Speed
- Reliability
- Ease to use
- Easy to analyze
The Bisection Method

- Sometimes, if a certain property holds for $f$ in a certain domain (e.g., some interval), it guarantees the existence of root in that domain.
- The Bisection method is an example for a method that exploits such a relation, together with iterations, to find the root of a function.
The Bisection Method

- The Bisection method (AKA “interval halving” or “binary search”), is a particular case of the so-called Bracketing methods.
- Bracketing methods determine successively smaller intervals (brackets) that contain a root.
The bisection method utilizes the Intermediate Value Theorem.

**Theorem (intermediate value)**

Let $f \in \mathcal{C}([a, b])$.

If $L \in \mathbb{R}$ is a number between $f(a)$ and $f(b)$, that is, $\min(f(a), f(b)) < L < \max(f(a), f(b))$, then there exists $c \in (a, b)$ such that $f(c) = L$.

$\mathcal{C}([a, b])$ is the set of all continuous functions from $[a, b]$ to $\mathbb{R}$. 
A Useful Variant in Our Context

- Particularly (taking $L = 0$):
  
  \[ f \in C([a, b]) \text{ and } f(a) \cdot f(b) < 0 \implies \exists c \in (a, b) \text{ such that } f(c) = 0. \]

- In other words if $f(a)$ and $f(b)$ have opposite signs, then $f$ must vanish at some $c \in (a, b)$. 
Definition (limit of a function at a point)

Let a function $f$ be defined in an open interval containing $x_0$, except possibly at $x_0$ itself. Then $f$ has limit $L$ at $x = x_0$, denoted

$$
\lim_{x \to x_0} f(x) = L,
$$

if $\forall \varepsilon > 0 \exists \delta$ such that

$$
|x - x_0| < \delta \Rightarrow |f(x) - L| < \varepsilon.
$$
**Definition (continuity of a function at a point)**

Let $f$ be defined over the open interval $(a, b)$ and let $x_0 \in (a, b)$. The function $f$ is said to be continuous at $x = x_0$ if

$$\lim_{x \to x_0} f(x) = f(x_0).$$

**Definition (continuity of a function over an open interval)**

A function $f$ is called called continuous over the open interval $(a, b)$ if it is continuous at every $x \in (a, b)$.

The definitions can be extended to the cases of $[a, b)$, $(a, b]$, or $[a, b]$ using one-sided limits at the relevant endpoints.
The Bisection Method

Notation

\[ C^n([a, b]) \triangleq \{ f \mid f : [a, b] \to \mathbb{R}, f \text{ and its first } n \text{ derivatives are continuous} \} \]

Thus, \( C^0([a, b]) \) is simply \( C([a, b]) \) (defined in a previous slide).

Example

\( f : x \mapsto x^{4/3} \) is in \( C^1([-1, 1]) \) but not in \( C^2([-1, 1]) \), since

\[
f' = \frac{4}{3}x^{1/3} \quad f'' = \frac{4}{9}x^{-2/3} = \frac{4}{9} \frac{1}{x^{2/3}}
\]

so \( f'' \) is discontinuous at \( x = 0 \).
The variant of the intermediate value theorem suggests the following algorithm.

**Input:** \([a, b]\) (a bracket) such that \(f(a) \cdot f(b) < 0\); \(\delta > 0\) (a tolerance value)

**Output:** \(\tilde{z}\) such that \(|\tilde{z} - z| < \delta\) where \(z\) is a root of \(f\) (i.e., \(f(z) = 0\))

```
repeat
   \(\tilde{z} \leftarrow \frac{1}{2}(a + b)\)
   if \(f(a) \cdot f(\tilde{z}) < 0\) then
      \(b \leftarrow \tilde{z}\)
   else
      \(a \leftarrow \tilde{z}\)
   until \(|b - a| < 2\delta;\)
   \(\tilde{z} \leftarrow \frac{1}{2}(a + b)\)
```

**Algorithm 1: The Bisection Method**
Example

Figure: Wikipedia
Example (Multiple Roots)
Properties of the Bisection Method

- Always works (assuming \( f \) is continuous): i.e., convergence to a true root is guaranteed:
  \[
  \lim_{n \to \infty} \tilde{z} = z
  \]
- Why? Since the approximation error after \( n \) iterations satisfies
  \[
  E(n) < \left| \frac{b - a}{2^n} \right|
  \]
  (so \( E(n) \to 0 \))
- The number of iterations required for a given approximation accuracy, \( \delta \), is known beforehand:
  \[
  \frac{b - a}{2^n} \leq \delta \Rightarrow n \geq \log_2 \left( \frac{b - a}{\delta} \right)
  \]
Properties of the Bisection Method (Continued)

- Convergence is relatively slow (i.e., a relative large number of iterations is needed) – we will discuss this topic later on.
- \( E(n) \) might not be a monotonic decreasing function of \( n \).
- A big advantage: it is fairly easy to find a good initial guess. – as we will see, in other methods convergence is guaranteed to happen only when the initial guess is close to the root. This requirement is unnecessary in the bisection method.
- Multiple roots for which the function doesn’t change sign will be missed.
One disadvantage of the bisection method is that, except the continuity of $f$ on $[a, b]$ and the opposite signs of $f(a)$ and $f(b)$, it does not exploit any additional knowledge we may have on $f$.

Particularly, the bisection does not exploit the fact that, in the vicinity of the root, we can approximate $f$ using functions that may be easier to handle in order to get a more educated guess about the root at each iteration.
Linear Approximation

For example, it is easy to approximate the function $f$ via a “linear” function\(^1\) for which it is easy to find a root.

\(^1\)This is the usual misnomer: this function is usually affine, not linear.
The Regula Falsi (False Position) Method

$c$ is easily found via similar triangles. Let $l : [a, b] \rightarrow \mathbb{R}$ denote the “linear” function. By construction, $l(a) = f(a)$, $l(b) = f(b)$, and $l(c) = 0$. Thus:

$$\frac{c - a}{0 - f(a)} \sim \frac{b - c}{f(b) - 0}$$

$$\Rightarrow \frac{c - a}{-f(a)} = \frac{b - c}{f(b)}$$

$$\Rightarrow (c - a)f(b) = (c - b)f(a)$$

$$\Rightarrow cf(b) - af(b) = cf(a) - bf(a)$$

$$\Rightarrow c[f(b) - f(a)] = af(b) - bf(a)$$

$$\Rightarrow c = \frac{af(b) - bf(a)}{f(b) - f(a)}$$
The Regula Falsi (False Position) Method

The following manipulation is useful:

\[
c = \frac{af(b) - bf(a)}{f(b) - f(a)} = \frac{af(b) - bf(a) + bf(b) - bf(b)}{f(b) - f(a)}
\]

\[
= \frac{f(b)(a - b) + b(f(b) - f(a))}{f(b) - f(a)}
\]

\[
= b - f(b) \frac{b - a}{f(b) - f(a)}
\]

\[
\Rightarrow \quad c = b - f(b) \frac{b - a}{f(b) - f(a)}
\]

(5 arithmetic operations)

Let \( m_f(a, b) \triangleq \frac{f(b) - f(a)}{b - a} \) denote the line’s slope.

\[
c = b - \frac{f(b)}{m_f(a, b)}
\]

We will later run again into expressions of this type.
The Regula Falsi (False Position) Method

The Process Can Be Repeated

Figure: Wikipedia
We can now formulate a new bracketing algorithm.

**Input:** $x_0, x_1$: initial guess s.t. $f(x_0) \cdot f(x_1) < 0$; $\delta > 0$ (a tolerance value)

**Output:** $\tilde{z}$ such that $|f(\tilde{z})| < \delta$ // note the difference from the bisection

```
\begin{algorithm}
  \textbf{repeat}
  \begin{algorithmic}
    \State $i \leftarrow i + 1$
    \State $x_i \leftarrow x_{i-1} - f(x_{i-1}) \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})}$
    \If {$f(x_i) \cdot f(x_{i-1}) < 0$}
      \State $x_{i-2} \leftarrow x_i$
    \Else
      \State $x_{i-1} \leftarrow x_i$
    \EndIf
  \Endalgorithmic
  \textbf{until} $|f(x_i)| < \delta$  // Should we check $|x_i - x_{i-1}| < \delta$? No. See # 4 next slide
\end{algorithm}
```

\textbf{Algorithm 2:} The Regula Falsi Method
The Regula Falsi (False Position) Method

Properties

1. Convergence to the root is guaranteed:

\[ \lim_{i \to \infty} x_i = z \quad (\text{where } f(z) = 0) \]

2. May converge faster or slower than the bisection method, depending on \( f \). Performance gets better the more the function, on \([a, b] \), can be well approximated by a line.

3. Can’t predict in advance the number of iterations.

4. The bracket might not decrease to zero (despite property 1)

5. The calculations, per iteration, are more expensive than bisection:
   - 3 addition/subtraction
   - 1 multiplication
   - 1 division

while a bisection iteration requires only one addition and one division.
The Secant Method

- Like the regula falsi method, it is based on a linear approximation.
- Unlike regula falsi and bisection, this is not a bracketing method. (particularly, this method enables to start from any pair of points, even if the root is not between them)
- Might fail to converge.
The Secant Method
The Secant Method

Again, by triangle similarity:

\[
\frac{x_1 - x_2}{f(x_1)} = \frac{x_0 - x_2}{f(x_0)}
\]

\[
\Rightarrow x_i = x_{i-1} - f(x_{i-1}) \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})}
\]

Note \( \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})} \) is the reciprocal of the line’s slope.
The Secant Method

Input: $x_0, x_1$: initial guess near the root (complicating an initial guess); $\delta > 0$ (a tolerance value)
Output: $\tilde{z}$ such that $|f(\tilde{z})| < \delta$ // as in regula falsi
/* The Repeat-Until loop below: similar to regula falsi, but w/o the if/else */
i = 1
repeat
   i ← i + 1
   $x_i \leftarrow x_{i-1} - f(x_{i-1}) \frac{x_{i-1} - x_{i-2}}{f(x_{i-1}) - f(x_{i-2})}$
until $|f(x_i)| < \delta$
$\tilde{z} \leftarrow x_i$

Algorithm 3: The Secant Method
Properties

- A natural extension of regula falsi without bracketing.
- If it converges, it tends to do it faster than the bisection and regaul falsi methods.
- The price: might fail to converge.
A Failure Case
A Failure Case

Figure: Wikipedia
$x_0 \in \mathbb{R}$ is called a fixed point of $g : \mathbb{R} \to \mathbb{R}$ if

$$g(x_0) = x_0.$$
In order to solve $f(x) = 0$, reorganize the equation to an equivalent form,

$$x = g(x)$$

for some suitable function $g$, and try to find a fixed point of $g$; i.e., a point $x_0$ such that $g(x_0) = x_0$. 

Idea
There is always at least one way to define such a $g$:

$$f(x_0) = 0 \Rightarrow x_0 + f(x_0) = x_0$$

so define

$$g(x) \triangleq x + f(x).$$

In which case $x = x_0$, which is a root of $f$, is a fixed point of $g$. 
Usually, however, there is more than one way to do it.

Example

\[ f(x) = x^2 - 2x - 3 \], so \( f(x) = 0 \) \( \iff \) \( x_{1,2} = -1, 3 \). Now:

\[
\begin{align*}
    x &= f(x) + x = x^2 - x - 3 \\
    2x &= f(x) + 2x = x^2 - 3 \\
    x^2 &= f(x) + x^2 = 2x^2 - 2x - 3 \\
    -x^2 + 2x &= f(x) - x^2 + 2x = -3
\end{align*}
\]

\[ \Rightarrow \quad g_1(x) = x^2 - x - 3 \]

\[ \frac{1}{2} \]

\[ g_2(x) = \frac{1}{2}(x^2 - 3) \]

\[ \frac{1}{x} \]

\[ g_3(x) = 2x - 2 - \frac{3}{x} \]

\[ \frac{1}{2-x} \]

\[ g_3(x) = \frac{3}{x - 2} \]
In any case, the function $g$ is called the **fixed-point iteration function** for solving $f(x) = 0$.

If $g$ is the fixed-point iteration function for solving $f(x) = 0$, then, if $p$ is a root of $f$, then (by constriction) $p$ is fixed point of the iteration

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

This suggest the following algorithm.
The Fixed-point Method

**Input:** $x_1$: initial guess close to the a root; $\delta > 0$ (a tolerance value)

**Output:** $\tilde{z}$, an approximation to the root

Rearrange the equation $f(x) = 0$ to the form $x = g(x)$.

1. $i = 1$
2. repeat
3.  
4.    $i \leftarrow i + 1$
5.    $x_i \leftarrow g(x_{i-1})$
6. until $|x_i - x_{i-1}| < \delta$ // Note it’s $x$ values, not $f(x)$ values
7. 
8. $\tilde{z} \leftarrow x_i$

**Algorithm 4:** The fixed-point method
The Fixed-point Method

- What does this procedure do?
- How can we interpret it graphically?
- What are its properties?

To answer the questions, we need some math first.
Theorem

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous function, and let \( \{x_i\}_{n=0}^{\infty} \) be a point sequence generated via the iteration

\[
x_{n+1} = g(x_n).
\]

If

\[
\lim_{n \to \infty} x_n = p
\]

then $p$ is a fixed point of $g$; i.e.,

\[
p = g(p).
\]

In other words: if (that’s a big “if” . . . ) the iteration converges then it converges to a fixed point of $g$, hence a root of $f$. 
Proof.

\[
\lim_{n \to \infty} x_n = p \Rightarrow \lim_{n \to \infty} x_{n+1} = p
\]

\[\Rightarrow g(p) = g\left(\lim_{n \to \infty} x_n\right) \quad \text{continuity} = \lim_{n \to \infty} g(x_n) \quad \text{by def. of } x_{n+1} \Rightarrow \lim_{n \to \infty} x_{n+1} = p\]
This is nice, but when can we expect convergence?
Fixed Point

Theorem

If \( g \in C([a, b]) \) and the range of \( y = g(x) \) satisfies

\[
\forall x \in [a, b] \quad y \in [a, b]
\]

(i.e., \( g \) is continuous and is “into” \([a, b]\)) then \( g \) has a fixed point in \([a, b]\):

\[
\exists x_0 \in [a, b] \quad g(x_0) = x_0.
\]
Example (existence of a fixed point)
Remark

Note well! The theorem, which holds when the domain of $g$ is a finite interval, $[a, b]$, says nothing about what happens if the domain is unbounded (e.g., $\mathbb{R}$, or $[a, \infty)$, etc.).
Proof.

If \( g(a) = a \) or \( g(b) = b \), we are done. Otherwise,

\[
a < g(a) \leq b \quad \text{and} \quad a \leq g(b) < b.
\]

Thus, \( f(x) \triangleq x - g(x) \) satisfies

\[
f(a) < 0 \quad \text{and} \quad f(b) > 0.
\]

(1)

By the intermediate value theorem,

\[
\exists c \in (a, b) \text{ s.t. } f(c) = 0.
\]

If follows that \( c \) is a fixed point of \( g \):

\[
g(c) \overset{\text{by def. of } f}{=} c - f(c) = c.
\]
Reminder: Lagrange’s Mean Value Theorem

Theorem (mean value)

Let $f \in C([a, b])$ and suppose $f'(x)$ exists for every $x \in (a, b)$. Then there exists $c \in (a, b)$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a} \quad \text{(the secant’s slope)}$$
Example (mean value theorem)

Figure: Wikipedia
Proof.

Define the function

\[ g(x) \triangleq f(x) - \left( f(a) + (x - a) \frac{f(b) - f(a)}{b - a} \right) \]

the secant between \((a, f(a))\) and \((b, f(b))\)

Thus, \(g(a) = g(b) = 0\). By Rolle’s theorem\(^a\), \(\exists c \in (a, b)\) such that \(g'(c) = 0\). Now:

\[ g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a} \]

\[ 0 = g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a} \Rightarrow f'(c) = \frac{f(b) - f(a)}{b - a} \]

\(^a\)Any real-valued differentiable function that attains equal values at two distinct points must have at least one point somewhere between them where the first derivative is zero.
Theorem (existence and uniqueness of a fixed point)

If \( g \in C([a, b]) \), \( g(x) \in [a, b] \), \( g'(x) \) exists on \( (a, b) \) and

\[
\forall x \in (a, b) \quad |g'(x)| \leq k < 1
\]

then \( g \) has a unique fixed point \( p \in [a, b] \).
Proof.

Proof by contradiction. Assume there exist two fixed points:

\[ p_1, p_2 \in [a, b] \quad p_1 < p_2 \quad g(p_1) = p_1 \quad g(p_2) = p_2 \]

By the mean value theorem, there exists \( d \in (p_1, p_2) \) such that

\[
g'(d) \quad \text{mean value thm} \quad \frac{g(p_2) - g(p_1)}{p_2 - p_1} = \frac{p_2 - p_1}{p_2 - p_1} = 1
\]

violating the assumption that \( k < 1 \).
Definition (contraction map)

A contraction map is a function that $f$ that satisfies

$$|f(y) - f(x)| \leq k|x - y| \quad k < 1$$

for every $x$ and $y$ in its domain.

- A contraction mapping “shrinks” distances between points.
- A contraction mapping need not be differentiable (but it is always Lipschitz continuous, hence uniformly continuous, hence continuous).
- A differentiable function need not be a contraction mapping.
A differentiable function $f$ is a contraction mapping \iff it satisfies $|f'(x)| \leq k < 1$.

In this class, we will often just say “contraction mapping” but will mean “differentiable contraction mapping”.

Thus, we can rephrase the last theorem as follows: If $g \in C([a, b])$, $g(x) \in [a, b]$, $g'(x)$ exists on $(a, b)$ and $g$ is a contraction map on $(a, b)$, then $g$ has a unique fixed point $p \in [a, b]$.

As we are starting to see, convergence of fixed-point iterations will be guaranteed for contraction mapping.
Theorem

Suppose:

- \( g \in C([a,b]) \)
- \( g \) has a fixed point \( p \in (a,b) \);
- \( g \) is defined in \( (a,b) \);
- \( \forall x \in [a,b], \ g(x) \in [a,b] \);
- \( x_0 \in (a,b) \).

If \( g \) is a contraction mapping then the iteration \( x_n = g(x_{n-1}) \) will converge to \( p \). In which case, \( p \) is called the **attractive fixed point**.

If \( g \) satisfies
\[
|g'(x)| > 1
\]
for every \( x \in [a,b] \) then the iteration \( x_n = g(x_{n-1}) \) will (locally) diverge; in which case \( p \) is called an **expelling fixed point**.
Proof.

We will prove for the convergent case. Note that \( \{x_n\}_{n=0}^{\infty} \subset (a, b) \). Now:

\[
|p - x_n| \quad \text{since } p \text{ is a fixed pt} \quad \Rightarrow \quad |g(p) - x_n| \quad \text{by def. of } x_n \quad \Rightarrow \quad |g(p) - g(x_{n-1})| \\
\text{mean value thm} \quad \Rightarrow \quad |g'(c_{n-1})(p - x_{n-1})| = |g'(c_{n-1})| \cdot |p - x_{n-1}| \\
\text{contraction} \quad \leq \quad k|p - x_{n-1}|
\]

for some \( k \in (0, 1) \); particularly, \( |p - x_n| < |p - x_{n-1}| \). WTS:

\[
limit_{n \to \infty} |p - x_n| = 0. \quad \text{But by induction,} \quad |p - x_n| \leq k^n |p - x_0|.
\]

Thus,

\[
0 \leq \limit_{n \to \infty} |p - x_n| \leq \limit_{n \to \infty} k^n |p - x_0| = |p - x_0| \limit_{n \to \infty} k^n = |p - x_0| \cdot 0 = 0.
\]
The divergent case is handled similarly, by showing that

\[ |g'(x)| > 1 \Rightarrow |p - x_n| > |p - x_{n-1}|, \]

in which case we are getting further and further from \( p \).

The divergence is not to \( \pm \infty \) as the sequence is contained in \((a,b)\), hence the term “locally”.

The theorem addresses the cases where \( |g'(x)| < 1 \) for every \((a,b)\) and where \( |g'(x)| > 1 \) for every \((a,b)\). It does not handle other cases.
Example: Convergence
Example: Divergence

\[ y - x \]
More Examples

(a) Convergence

(b) Divergence

Figure: Justin Solomon’s book
Example: $0 < K \ll 1 \Rightarrow \text{Fast Convergence}$
Example: $0 < K < 1 \ & \ K \approx 1 \Rightarrow \text{Slow Convergence}$
Order of Convergence

Recall that, proving the convergent case from the theorem, we saw

\[ |p - x_{n+1}| \leq k |p - x_n| \]

i.e.,

\[ |e_{n+1}| \leq k |e_n| . \]

Thus,

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|} \leq k < 1 . \]

So there is at least linear convergence.
(with additional knowledge, we may be able to show better rates)
This result can be generalized and extended. But first, we need another reminder to a result that generalizes the Mean Value Theorem.
Theorem (Taylor)

If $f \in C^{n+1}([a, b])$ then,

$$f(x + h) = \left( \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) \right) + E_n(h) \quad \forall \ x, x + h \in [a, b]$$

where

$$E_n(h) = \frac{h^{n+1}}{(n + 1)!} f^{(n+1)}(\xi)$$

for some $\xi \in [x, x + h]$. Moreover, $E_n(h) = O(h^{n+1})$. 
Taylor

\[ f(x + h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(x) + \ldots + \frac{h^n}{n!} f^n(x) + E_n(h) \]

\[ = P_n(h) + E_n(h), \text{ a polynomial of order } n \]

\[ P_n(h), \text{ an error} \]
Remark

Taking \( n = 0 \), we get there exists some \( \xi \in [x, x + h] \) such that

\[
f(x + h) = f(x) + h \cdot f'(\xi)
\]

Equivalently, there exists some \( \xi \in [x, x + h] \) such that

\[
\frac{f(x + h) - f(x)}{h} = f'(\xi).
\]

and this is exactly the mean value theorem.
Back to Convergence of Fixed Point Iteration

\[ e_n = x_n - p \Rightarrow x_n = p + e_n \]

\[ |e_{n+1}| = |p - x_{n+1}| = |g(p) - g(x_n)| \quad \text{reverse order} \]
\[ = |g(x_n) - g(p)| \]

\[ = \left| g(p) + \left[ \sum_{k=1}^{q-1} \frac{e_n^k}{k!} g^{(k)}(p) \right] + \frac{e_n^q}{q!} g^{(q)}(c_n) \right| - g(p) \]

\[ = \left| \sum_{k=1}^{q-1} \frac{e_n^k}{k!} g^{(k)}(p) \right| + \frac{e_n^q}{q!} g^{(q)}(c_n) \]

for some \( c_n \in (p, x_n) \).
We have just obtained that

\[ |e_{n+1}| = \left| \sum_{k=1}^{q-1} \frac{e_n^k}{k!} g^{(k)}(p) \right| + \frac{e_n^q}{q!} g^{(q)}(c_n) \]

for some \( c_n \in (p, x_n) \).

\( \Rightarrow \) the order of convergence of a fixed-point iteration is the smallest \( q \) such that \( g^{(q)} \neq 0 \) since then

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^q} = \lim_{n \to \infty} \frac{1}{q!} \left| g^{(q)}(c_n) \right| \]

or, since \( x_n \to p \),

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^q} = \frac{1}{q!} \left| g^{(q)}(p) \right| \]
Recall that, in order to solve \( f(x) = 0 \), we can construct multiple fixed-point iterations:

\[
\begin{align*}
  x_{i+1} &= g_1(x_1) \\
  x_{i+1} &= g_2(x_1) \\
  x_{i+1} &= g_3(x_1) \\
  \vdots
\end{align*}
\]

How should we choose from these?

Answer: the considerations are

- Order of convergence
- Number of calculations per iteration
- Ease of finding an initial guess (the size of the range around the root such that \(|g'(x)| < 1\) and \(g\) maps into it)

Usually we will prefer \( g \) with the smallest \(|g'(p)|\) as possible.
So far we saw 3 methods of computing (approximated) roots of equations of the form \( f(x) = 0 \).

One of the aspects in which these methods differ from each other is the speed in which they “lead” us to the answer, i.e., their convergence rate.
Example

In this example, the secant is clearly the fastest while the bisection is the slowest.

Table 1.3: Comparison of methods, $f(x) = 3x + \sin(x) - e^x = 0$, $x_0 = 0$, $x_1 = 1$

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Interval halving</th>
<th>False position</th>
<th>Secant method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x$</td>
<td>$f(x)$</td>
<td>$x$</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.330704</td>
<td>0.470990</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>-0.286621</td>
<td>0.372277</td>
</tr>
<tr>
<td>3</td>
<td>0.375</td>
<td>0.036281</td>
<td>0.361598</td>
</tr>
<tr>
<td>4</td>
<td>0.3125</td>
<td>-0.121899</td>
<td>0.360538</td>
</tr>
<tr>
<td>5</td>
<td>0.34375</td>
<td>-0.041956</td>
<td>0.360433</td>
</tr>
</tbody>
</table>

Error after 5 iterations:

- Interval halving: 0.01667
- False position: $-1.17 * 10^{-5}$
- Secant method: $<-1 * 10^{-7}$

(Exact value of root is 0.360421703.)

In this example, the secant is clearly the fastest while the bisection is the slowest.

Figure: Gerald and Wheatley
Part of the goals of the field of Numerical Analysis is to characterize such relative performances in a more principled way.

For this, we need some definitions...
Definition (Big O notation for sequences)

Let $f_n$ and $g_n$ be two sequences. We write

$$f_n = O(g_n)$$

if there exist constants, $N \in \mathbb{Z}$ and $C \in \mathbb{R}_{>0}$, such that

$$|f_n| \leq C|g_n| \quad \forall n > N$$
Definition (Big O notation for functions)

Let $f$ and $g$ be two functions from $\mathbb{R}$ to $\mathbb{R}$. We write

$$f(x) = O(g(x)) \ (\text{as} \ x \to \infty)$$

$\iff$ there exist constants, $x_0 \in \mathbb{R}$ and $C \in \mathbb{R}_{>0}$, such that

$$|f(x)| \leq C|g(x)| \ \forall x > x_0$$

(often the “$x \to \infty$” is omitted and is implicitly assumed).

This definition lets us describe the rate of decay (or growth) of functions/sequences in terms of well-known functions/sequences such as $n^p$, $n^{1/p}$, $a^n$, $\log_a n$, etc.
What is the connection between sequences and convergence of numerical calculations?

It turns out that in many numerical calculations we can recognize several types of sequences.
Of particular importance, is the sequence that represents the approximation error as a function of the iteration number.

**Definition (approximation-error sequence)**

Let \( \{\tilde{x}_n\} \) be a sequence approximating the true value \( x \). The approximation-error sequence is

\[
e_n \triangleq \tilde{x}_n - x.
\]
Definition (order of convergence)

\( \{ \tilde{x}_n \} \) be a sequence such that

\[
\lim_{n \to \infty} \tilde{x}_n = x.
\]

The order of convergence of \( \{ \tilde{x}_n \} \) is \( R > 0 \) if there exist \( A \),

\[
0 < A < \infty,
\]

such that

\[
\lim_{n \to \infty} \frac{\left| \frac{\tilde{x}_{n+1} - x}{\tilde{x}_n - x} \right|^R}{\left| \frac{e_{n+1}}{e_n} \right|^R} = A.
\]

In which case, the number \( A \) is called the asymptotic error constant.

- The order of convergence is defined w.r.t. the true value, \( x \), which is usually unknown in numerical calculations.
- Still, we will see that this definition will enable practical calculations.
From the definition of $O(\cdot)$, we can say that the order of convergence is $R$ if

$$e_{n+1} = O(e_n^R).$$

**Remark**

In principle, one could think of more general types of convergence such as

$$e_{n+1} = O(g(e_n)),$$

but the use of these in practice is limited.
Particular Cases

- \( R = 1 \Rightarrow \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|} = A \in (0, 1) \)

  \( \Rightarrow \) linear convergence.

- \( R = 2 \Rightarrow \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = A \in (0, \infty) \)

  \( \Rightarrow \) quadratic convergence.
Comments

- When $R = 1$ we must have $0 < A < 1$ for convergence, since otherwise the error does not decrease (if $A = 1$) or might even increase (if $A > 1$).
- When $R = 2$, $A$ doesn’t have to be smaller than 1.
- The order of convergence, $R$, does not have to be an integer.
- If
  \[
  \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|} = 0
  \]
  the convergence is called superlinear for $R = 1$. 

BGU CS     Solving Nonlinear Equations (ver. 1.00)     AY '19/'20, Fall Semester
The Newton (-Raphson) Method

Both the secant method and the Regula Falsi method utilize the advantage in approximating $f$ via a “linear” function near the current guess in order to get us closer to the root.

Reminder, here is the iteration in these methods:

$$x_{i+1} = x_i - f(x_i) \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})}$$

or, equivalently,

$$x_{i+1} = x_i - \frac{f(x_i)}{m_{i,i+1}}$$

where $m_{i,i+1}$ is the slope of the secant between $x_i$ and $x_{i-1}$. 
Newton’s method obviates the need of two points; rather, it approximates $f$ at the current via its tangent:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}.$$ 

(we will derive this expression soon)

Note the same result would have been obtained by taking the expression from the secant’s method for $x_i \rightarrow x_{i-1}$. 
Example (tangent line)
Deriving the Iteration

\[ y(x_{i+1}) = m \cdot x_{i+1} + b = 0 \]
\[ y(x_i) = m \cdot x_i + b = f(x_i) \]

Thus, by subtraction,

\[ mx_i - m \cdot x_{i+1} = f(x_i) \]

The slope, by the definition of the tangent line, is

\[ m = f'(x_i). \]

Thus,

\[ f'(x_i)(x_i - x_{i+1}) = f(x_i) \]

\[ \Rightarrow x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}. \]
Newton’s Method

Input: $x_0$: initial guess near the root (complicating an initial guess); $\delta > 0$ (a tolerance value)
Output: $\tilde{z}$ such that $|f(\tilde{z})| < \delta$

1. $i = 0$
2. repeat
   3. $i \leftarrow i + 1$
   4. $x_i \leftarrow x_{i-1} - \frac{f(x_{i-1})}{f'(x_{i-1})}$
   5. until $|f(x_i)| < \delta$
6. $\tilde{z} \leftarrow x_i$ // or until $|x_i - x_{i-1}| < \delta$

Algorithm 5: Newton’s Method
Example (Newton’s method iterations)

Figure: http://tutorial.math.lamar.edu/
Note it is important to start near the root for otherwise it might diverge.

But how close to the root do we need to be? Under what conditions will we achieve convergence?
Example (Newton’s method failure: an endless loop)

Figure: Numerical Recipes in C++
Example (Newton’s method failure: an endless loop)

Figure 2.15  (b) Newton-Raphson iteration for $f(x) = x^3 - x - 3$ can produce a cyclic sequence.
Example (Newton’s method failure: an endless loop)

Figure: Gerald and Wheatley
Example (Newton’s method failure: divergence)

Figure 2.15 (c) Newton-Raphson iteration for $f(x) = \arctan(x)$ can produce a divergent oscillating sequence.
Example (Newton’s method failure: zero derivative)
**Theorem (convergence of Newton’s method)**

Let \( f \in C^2([a, b]) \) and let \( p \in [a, b] \) such that \( f(p) = 0 \). If \( f'(p) \neq 0 \), then \( \exists \delta > 0 \) such that the sequence defined via

\[
x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}
\]

converges to \( p \) for every \( x_0 \in (p - \delta, p + \delta) \).

The proximity to the root and the continuity of \( f'' \) are not intuitive requirements. The reason behind these requirements is revealed through the proof.
Theorem (convergence of Newton’s method)

Let $f \in C^2([a, b])$ and let $p \in [a, b]$ such that $f(p) = 0$. If $f'(p) \neq 0$, then $\exists \delta > 0$ such that the sequence defined via

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

converges to $p$ for every $x_0 \in (p - \delta, p + \delta)$.

The proximity to the root and the continuity of $f''$ are not intuitive requirements. The reason behind these requirements is revealed through the proof.
The Newton (-Raphson) Method

The Method

Proof.

Newton’s method is a fixed point iteration of the form

\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)} \, . \]

By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires

\[ |g'(x)| = \left| 1 - \frac{[f'(x)]^2 - f(x) \cdot f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x) \cdot f''(x)}{[f'(x)]^2} \right| \quad \text{required} \quad < \quad 1 \]

in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \((p - \delta, p + \delta)\) on which \( |g'(x)| < 1 \) it suffices to require continuity of \( g \) on it.

As \( |g'(x)| = \left| \frac{f(x) \cdot f''(x)}{[f'(x)]^2} \right| \), we have continuity \( \iff \ f \in C^2([p - \delta, p + \delta]) \).

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \((p - \delta, p + \delta) \subseteq [a, b] \) and \( g(x) \triangleq f - f(x)/f'(x) \) is a contraction mapping on \((p - \delta, p + \delta)\), implying the iteration’s convergence.
Proof.

Newton’s method is a fixed point iteration of the form $x_{i+1} = g(x_i) \triangleq x_i - f(x_i)/f'(x_i)$. By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if $g$ is a contraction mapping on it. This requires

$$|g'(x)| = \left|1 - \frac{f'(x)^2 - f(x) \cdot f''(x)}{[f'(x)]^2}\right| = \left|\frac{f(x) \cdot f''(x)}{[f'(x)]^2}\right| \text{ required } < 1$$

in some interval about $p$. Since $f(p) = 0$ and $f'(p) \neq 0$ it follows that $g'(p) = 0$. So to ensure the existence of an interval $(p - \delta, p + \delta)$ on which $|g'(x)| < 1$ it suffices to require continuity of $g$ on it. As $|g'(x)| = \left|\frac{f(x) \cdot f''(x)}{[f'(x)]^2}\right|$, we have continuity $\iff f \in C^2([p - \delta, p + \delta])$.

In other words: If $p \in [a, b]$ and $f \in C^2[a, b]$ then $\exists \delta > 0$ such that $(p - \delta, p + \delta) \subseteq [a, b]$ and $g(x) \triangleq f - f(x)/f'(x)$ is a contraction mapping on $(p - \delta, p + \delta)$, implying the iteration’s convergence.
The Newton (-Raphson) Method

Proof.

Newton’s method is a fixed point iteration of the form

\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)} \]

By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires

\[
|g'(x)| = \left| 1 - \frac{f'(x)^2 - f(x) \cdot f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| \quad \text{required} < 1
\]

in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \( (p - \delta, p + \delta) \) on which \( |g'(x)| < 1 \) it suffices to require continuity of \( g \) on it.

As \( |g'(x)| = \left| \frac{f(x)f''(x)}{[f'(x)]^2} \right| \), we have continuity \( \iff f \in C^2([p - \delta, p + \delta]) \).

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \( (p - \delta, p + \delta) \subseteq [a, b] \) and \( g(x) \triangleq f - f(x)/f'(x) \) is a contraction mapping on \( (p - \delta, p + \delta) \), implying the iteration’s convergence.
Proof.

Newton’s method is a fixed point iteration of the form

\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)}. \]

By the fixed point iteration's convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires

\[ |g'(x)| = \left| 1 - \frac{[f'(x)]^2 - f(x) \cdot f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| \text{ required} < 1 \]

in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \((p - \delta, p + \delta)\) on which \( |g'(x)| < 1 \) it suffices to require continuity of \( g \) on it.

As \( |g'(x)| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| \), we have continuity \( \iff f \in C^2([p - \delta, p + \delta]) \).

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \((p - \delta, p + \delta) \subseteq [a, b]\) and \( g(x) \triangleq f - f(x)/f'(x) \) is a contraction mapping on \((p - \delta, p + \delta)\), implying the iteration’s convergence. \( \square \)
Proof.

Newton’s method is a fixed point iteration of the form
\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)}. \]
By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires
\[
|g'(x)| = \left|1 - \frac{[f'(x)]^2 - f(x) \cdot f''(x)}{[f'(x)]^2}\right| = \left|\frac{f(x) f''(x)}{[f'(x)]^2}\right| \quad \text{required} < 1
\]
in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \((p - \delta, p + \delta)\) on which \( |g'(x)| < 1 \) it suffices to require continuity of \( g \) on it.

As \(|g'(x)| = \left|\frac{f(x) f''(x)}{[f'(x)]^2}\right|\), we have continuity \( \iff f \in C^2([p - \delta, p + \delta]) \).

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \((p - \delta, p + \delta) \subseteq [a, b]\) and \( g(x) \triangleq f - \frac{f(x)}{f'(x)} \) is a contraction mapping on \((p - \delta, p + \delta)\), implying the iteration’s convergence.
Proof.

Newton’s method is a fixed point iteration of the form

\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)}. \]

By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires

\[
|g'(x)| = \left| 1 - \frac{[f'(x)]^2 - f(x) \cdot f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| < 1
\]

in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \((p - \delta, p + \delta)\) on which \(|g'(x)| < 1\) it suffices to require continuity of \( g \) on it.

As \(|g'(x)| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right|\), we have continuity \( \iff f \in C^2([p - \delta, p + \delta]). \)

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \((p - \delta, p + \delta) \subseteq [a, b] \) and \( g(x) \triangleq f - f(x)/f'(x) \) is a contraction mapping on \((p - \delta, p + \delta)\), implying the iteration’s convergence.
Proof.

Newton’s method is a fixed point iteration of the form

\[ x_{i+1} = g(x_i) \triangleq x_i - \frac{f(x_i)}{f'(x_i)} \]

By the fixed point iteration’s convergence theorem, convergence is guaranteed in an open interval containing the root if \( g \) is a contraction mapping on it. This requires

\[ |g'(x)| = \left| 1 - \frac{[f'(x)]^2 - f(x) \cdot f''(x)}{[f'(x)]^2} \right| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| \]  \text{required} \quad < \quad 1

in some interval about \( p \). Since \( f(p) = 0 \) and \( f'(p) \neq 0 \) it follows that \( g'(p) = 0 \). So to ensure the existence of an interval \((p - \delta, p + \delta)\) on which \( |g'(x)| < 1 \) it suffices to require continuity of \( g \) on it.

As \( |g'(x)| = \left| \frac{f(x) f''(x)}{[f'(x)]^2} \right| \), we have continuity \( \iff \) \( f \in C^2([p - \delta, p + \delta]) \).

In other words: If \( p \in [a, b] \) and \( f \in C^2[a, b] \) then \( \exists \delta > 0 \) such that \((p - \delta, p + \delta) \subseteq [a, b] \) and \( g(x) \triangleq \frac{f - f(x) / f'(x)}{f'(x)} \) is a contraction mapping on \((p - \delta, p + \delta)\), implying the iteration’s convergence.
Here is the theorem again:

**Theorem (convergence of Newton’s method)**

Let \( f \in C^2([a, b]) \) and let \( p \in [a, b] \) such that \( f(p) = 0 \).
If \( f'(p) \neq 0 \), then \( \exists \delta > 0 \) such that the sequence defined via
\[
x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}
\]
converges to \( p \) for every \( x_0 \in (p - \delta, p + \delta) \).

From the theorem’s assumption and what we saw in its proof:

- \( f(p) = 0 \)
- \( f'(p) \neq 0 \)
- \( g(x) = x - \frac{f(x)}{f'(x)} \)
- \( |g'(x)| = \left| \frac{f(x)f''(x)}{(f'(x))^2} \right| \)
- \( g(p) = p \)
- \( g'(p) = 0 \)
Newton’s Method: Order of convergence

- Identifying the order of convergence here is an immediate consequence of what we learned for a fixed point iteration.
- Reminder: The order of convergence of a fixed-point iteration is the smallest \( q \) such that \( g^{(q)} \neq 0 \).
- For Newton’s method, we just saw that (if it converges)

\[
g'(p) = 0,
\]

hence the convergence order of Newton’s method is 2.
Newton’s Method: Order of convergence

- Identifying the order of convergence here is an immediate consequence of what we learned for a fixed point iteration.

- Reminder: The order of convergence of a fixed-point iteration is the smallest \( q \) such that \( g^{(q)} \neq 0 \).

- For Newton’s method, we just saw that (if it converges)
  \[
g'(p) = 0,
\]
hence the convergence order of Newton’s method is 2.
Newton’s Method: Order of convergence

- Identifying the order of convergence here is an immediate consequence of what we learned for a fixed point iteration.
- Reminder: The order of convergence of a fixed-point iteration is the smallest $q$ such that $g^{(q)} \neq 0$.
- For Newton’s method, we just saw that (if it converges)
  \[ g'(p) = 0, \]
  hence the convergence order of Newton’s method is 2.
Deriving it Again, Specifically for Newton’s Case

Now:

\[ |e_{n+1}| = |x_{n+1} - p| = |g(x_n) - g(p)| = |g(p + e_n) - g(p)| \]

\[ = |g(p) + e_n \cdot g'(p) + \frac{e_n^2}{2} \cdot g''(\xi_n) - g(p)| \quad \text{(for some } \xi_n \text{ between } p \text{ & } p + e_n) \]

\[ = |e_n \cdot g'(p) + \frac{e_n^2}{2} \cdot g''(\xi_n)| = |e_n \cdot \frac{f(p)f''(p)}{(f'(p))^2} + \frac{e_n^2}{2} \cdot g''(\xi_n)| \]

\[ f(p) = 0 \quad \frac{e_n^2}{2} \cdot g''(\xi_n) \quad \Rightarrow \quad |e_{n+1}| = \left| \frac{e_n^2}{2} \cdot g''(\xi_n) \right| \]

Thus,

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = \lim_{n \to \infty} \left| \frac{1}{2} g''(\xi_n) \right| \quad \text{since } \xi_n \to p \text{ and } g'' \text{ is cont.} \quad \left| \frac{1}{2} g''(p) \right| \]
Deriving it Again, Specifically for Newton’s Case

Now:

\[ |e_{n+1}| = |x_{n+1} - p| = |g(x_n) - g(p)| = |g(p + e_n) - g(p)| \]

\[ = |g(p) + e_n \cdot g'(p) + \frac{e_n^2}{2} \cdot g''(\xi_n) - g(p)| \quad \text{(for some } \xi_n \text{ between } p \& p + e_n) \]

\[ = |e_n \cdot g'(p) + \frac{e_n^2}{2} \cdot g''(\xi_n)| = |e_n \cdot \frac{f(p)f''(p)}{(f'(p))^2} + \frac{e_n^2}{2} \cdot g''(\xi_n)| \]

\[ \text{if } f(p) = 0 \quad \text{or } f'(p) \neq 0 \quad \Rightarrow |e_{n+1}| = \left| \frac{e_n^2}{2} \cdot g''(\xi_n) \right| \]

Thus,

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = \lim_{n \to \infty} \left| \frac{1}{2} \cdot g''(\xi_n) \right| = \frac{1}{2} g''(p) \quad \text{since } \xi_n \to p \text{ and } g'' \text{ is cont.} \]
Just saw:

\[
\lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = \left| \frac{1}{2} g''(p) \right|
\]

Is the RHS smaller than \( \infty \)?

\[
g''(x) = \frac{d}{dx} \frac{f(x)f'''(x)}{[f''(x)]^2} = \frac{(ff''' + f'f'')(f')^2 - (ff'')(2f''f')}{(f')^4}
\]

\[
f(p)=0 \Rightarrow g''(p) = \frac{(f'(p)f''(p))(f'(p))^2}{(f'(p))^4} = \frac{f''(p)}{f'(p)}
\]

\[
\Rightarrow \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = A \triangleq \frac{1}{2} \left| \frac{f''(p)}{f'(p)} \right|
\]

and \( A < \infty \) if \( f'(p) \neq 0 \), which is exactly one of the requirements in the convergence theorem for Newton’s method.
The Newton (-Raphson) Method

The Method

Just saw:

\[ \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = \left| \frac{1}{2} g''(p) \right| \]

Is the RHS smaller than \( \infty \)?

\[ g''(x) = \frac{d}{dx} \frac{f(x)f'''(x)}{[f''(x)]^2} = \frac{(f f''') + f' f'') (f')^2 - (f f'') (2 f'' f')}{(f')^4} \]

\[ \Rightarrow \quad g''(p) = \frac{(f'(p)f''(p))(f'(p))^2}{(f'(p))^4} = \frac{f''(p)}{f'(p)} \]

\[ \Rightarrow \quad \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = A \triangleq \frac{1}{2} \left| \frac{f''(p)}{f'(p)} \right| \]

and \( A < \infty \) if \( f'(p) \neq 0 \), which is exactly one of the requirements in the convergence theorem for Newton’s method.
Finding a Square Root

- Goal: Given $M > 0$, compute $\sqrt{M}$ up to any given accuracy, using only arithmetic operations.
- Method: Newton’s iterations.
• \( \sqrt{M} \) is a root of

\[
f(x) = x^2 - M
\]

• In other words, we want to solve the equation

\[
x^2 - M = 0.
\]

• According to Newton’s method, the iteration is

\[
x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}
\]

\( \Rightarrow \)

\[
x_{i+1} = x_i - \frac{x_i^2 - M}{2x_i}
\]

\( \Rightarrow \)

\[
x_{i+1} = \frac{1}{2} \left( x_i + \frac{M}{x_i} \right)
\]

• This computation requires:
  • 1 division
  • 1 addition
  • 1 division by 2 (shifts bits to the right)
How to Pick an Initial Guess?

- Given $M$, we want an initial guess that will satisfy the conditions that guarantee convergence.

\[ g(x) = \frac{1}{2} \left( x + \frac{M}{x} \right) \quad \Rightarrow \quad \left| g'(x) \right| = \frac{1}{2} \left| 1 - \frac{M}{x^2} \right| \]

- Want:

\[ \frac{1}{2} \left| 1 - \frac{M}{x^2} \right| < 1 \]
Equivalently, want:

\[-2 < 1 - \frac{M}{x^2} < 2\]

Inspecting the right inequality:

\[1 - \frac{M}{x^2} < 2 \iff -\frac{M}{x^2} < 1 \iff \frac{M}{x^2} > -1 \iff -M < x^2\]

so this inequality always holds.
As for the left inequality:

\[-2 < 1 - \frac{M}{x^2} \iff -3 < -\frac{M}{x^2} \iff 3 > \frac{M}{x^2} \iff x > \sqrt{\frac{M}{3}}\]
So, it seems we have solved our problem: we “just” need to start in some some $x_0 > \sqrt{\frac{M}{3}}$.

The problem, of course, is that requires us to compute a square root – which is exactly the problem we are trying to solve.
Idea: transform the problem to a range where it is easy to pick a safe initial guess. Apply Newton’s method to the new problem, and then transform the result back to the original problem.
For every $M > 0$, there exist a real number $m$ such that $1/4 \leq m \leq 1$ and an integer number $e$ such that

$$M = m \cdot 4^e$$

Thus,

$$\sqrt{M} = \sqrt{m} \cdot 2^e$$
• Since $1/4 \leq m \leq 1$ we know that

$$1/2 \leq \sqrt{m} \leq 1$$

• We can get a good initial guess for finding $\sqrt{m}$ using a linear approximation w/o having to use $\sqrt{\cdot}$.

• The line between the two points

$$(m = 1/4, \sqrt{m} = 1/2)$$

and

$$(m = 1, \sqrt{m} = 1)$$

is

$$\frac{1}{3}(2m + 1)$$
Taking \( x_0 = \frac{1}{3}(2m + 1) \) ensures \( x_0 > \sqrt{\frac{m}{3}} \).

Indeed:

\[
\frac{1}{3}(2m + 1) > \sqrt{\frac{m}{3}}
\]

\[
\frac{1}{9}(2m + 1)^2 > \frac{m}{3}
\]

\[
(2m + 1)^2 > 3m
\]

\[
4m^2 + 4m + 1 > 3m
\]

\[
4m^2 + m + 1 > 0
\]

but this always holds since \( m > 0 \).

Thus, convergence of Newton’s iterations is guaranteed.
How Many Iterations Are Required?

\[ e_0 = \sqrt{m} - x_0 = m^{1/2} - \frac{1}{3}2m + 1 \]

\[ e'_0 = \frac{1}{2}m^{-1/2} - \frac{2}{3} \]

Set \( e'_0 \) to zero find \( m \) that will give the maximal error:

\[ m_{\text{max}}^{-1/2} = \frac{4}{3} \Rightarrow m_{\text{max}} = \frac{9}{16} \Rightarrow \sqrt{m_{\text{max}}} = \frac{3}{4} \]

In which case, \( e_{0\text{max}} = \frac{3}{4} - \frac{1}{3} \left(2 \cdot \frac{9}{16} + 1\right) \approx 0.042 \)
We know the convergence rate is quadratic.

\[ |e_{n+1}| = |x_{i+1} - \sqrt{m}| = \left| \frac{1}{2}(x_i + m/x_i) - \sqrt{m} \right| \]

\[ = \frac{1}{2x_i}(x_i - \sqrt{m})^2 = \frac{1}{2x_i}e_i^2 \]

Since \(1/2 \leq x_i \leq 1\) we get that \(|e_{n+1}| \leq e_i^2\).

Thus,

\[ e_3 \leq e_2^2 \leq e_1^4 \leq e_0^8 \leq 0.042^8 \approx 10^{-11} \]

Point: even just three iterations suffice for accuracy of at least 10 digits!
The Condition \( f'(p) \neq 0 \)

- The condition \( f'(p) \neq 0 \)

follows since the convergence theorem for Newton’s method.

- Without it, we could not have made key claims. For example:
  - \( g'(p) = 0 \) (was needed to prove convergence)
  - \[
  \lim_{n \to \infty} \frac{|e_{n+1}|}{|e_n|^2} = A \overset{\Delta}{=} \frac{1}{2} \left| \frac{f''(p)}{f'(p)} \right| < \infty
  \]
    (was needed to determine order of convergence)
This raises the question: when does this condition doesn’t hold? In effect, when do we have
\[ f'(p) = 0 \]?

And what can be done in such a case?
Example

Consider the family of functions of the form \( f(x) = x^2 - 1 + d \)

\( f(x) = 0 \) has two roots for \( d < 1 \) (since then \( x^2 = 1 - d > 0 \)), and a single root for \( d = 1 \) (since then \( x^2 = 0 \). As \( d \) approaches 1 from below, the inter-root distance decreases, becoming zero when \( d = 1 \); i.e., the roots coincide. Such a root is called a “multiple root” or “a root with multiplicity”. This is in contrast to a “simple root” (here, for \( d < 1 \), we have two simple roots.)
More Formally

Definition

If \( f \in C^M([a, b]) \), and \( \exists p \in (a, b) \) such that

\[
\begin{align*}
f^{(m)}(p) &= 0 \quad \forall m = 0, 1, \ldots, M - 1 \\
f^{(M)}(p) &\neq 0
\end{align*}
\]

then \( p \) is called a root of multiplicity \( M \). A root of multiplicity 1 is called simple.

- Clearly, some of our consequences so far about the convergence of Newton’s method are invalid for a root of multiplicity \( M > 1 \).
Theorem (convergence of Newton’s method for a multiple root)

Let $f \in C^M([a, b])$. Let $p \in [a, b]$ be a root of multiplicity $M > 1$. Then:

1. $\exists \delta > 0$ s.t. the sequence

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad n = 0, 1, 2, \ldots,$$

will converge to $p$ for every $x_0 \in (p - \delta, p + \delta)$ with a linear order of convergence, s.t.

$$|e_{n+1}| \approx \frac{M-1}{M} |e_n|$$

2. $\exists \delta > 0$ s.t. the sequence

$$x_{n+1} = x_n - M \frac{f(x_n)}{f'(x_n)} \quad n = 0, 1, 2, \ldots,$$

will converge to $p$ for every $x_0 \in (p - \delta, p + \delta)$ with a quadratic order of convergence.
Remark

- The constant $\frac{M-1}{M}$ in the first case, might seem disappointing: For $M > 1$ this is more than $1/2$. Even in the bisection method we had a $1/2$ factor. However, in the case of an even $M$, bisection is not even applicable here.

- We don’t know $M$ in advance. It is usually not worth it to do a trial-and-error search for the $M$ that will give us a quadratic convergence, and it is better to stick with the linear convergence here.
Before the Proof

- We already saw that the order of convergence for Newton’s method followed from the convergence theorem for fixed point iterations: $q$ is the smallest number for which $g^{(q)}(p) \neq 0$, where $p$ is a root of $f$. We also saw that $g(x) = x - f(x)/f'(x)$ and that

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

- Thus, at a simple root $f(p) = 0$ and $f'(p) \neq 0$, and it was easy to see that $g'(p) = 0$ and to claim the existence of an interval $(p - \delta, p + \delta)$ on which $|g'(x)| < 1$.

- But here, with $M > 1$, we can say something similar about $g'(p)$ since it gets the form $0/0$. 
Theorem

If \( f(x) = 0 \) has a root of multiplicity \( M \) at \( p \), then there exists a continuous function, \( h(x) \), such that \( f \) can be written as

\[
f(x) = (x - p)^M h(x)
\]

\( h(p) \neq 0 \)

Example

\( f(x) = x^2 \) can be written as

\[
f(x) = x^2 = (x - 0)^2 \cdot 1
\]

\( p = 0 \)
\( M = 2 \)
\( h(x) \equiv 1 \)
Proof of Part 1 in the Convergence Theorem

Proof.

By the last theorem, we can write $f$ as

$$f(x) = (x - p)^M h(x) \quad h(p) \neq 0$$

Recall that $g(x) = x - f(x)/f'(x)$ so $g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$. Now:

$$f'(x) = M (x - p)^{M-1} h(x) + (x - p)^M h'(x)$$
$$= (x - p)^{M-1} \left[ Mh(x) + (x - p)h'(x) \right]$$

$$f''(x) = (M - 1)(x - p)^{M-2} \left[ Mh(x) + (x - p)h'(x) \right]$$
$$+ (x - p)^{M-1} \left[ Mh'(x) + h'(x) + (x - p)h''(x) \right]$$

With some algebra, it follows that $g'(p) = \frac{M-1}{M}$ Thus, unless $M = 1$, we will have $|g'(p)| < 1$ and $g'(p) \neq 0$. Thus, there is convergence and it’s linear.
Proof of Part 2 in the Convergence Theorem

Proof.

The improved iteration is

\[ g(x) = x - \frac{M f(x)}{f'(x)}. \]

With some calculations, using expressions from earlier, we will get that

\[ g'(x) = (p - x) \frac{(1 - M)(p - x)h'(x) + Mh(x) [(p - x)h''(x) - 2h''(x)]}{[M h(x) + (x - p)h'(x)]^2} \]

Thus, \( g'(p) = 0 \). Thus, the convergence rate is at least quadratic. With some more work, we can show that \( g''(p) = \frac{2}{M} \frac{h'(p)}{h(p)} \). Thus, depending on \( h'(x) \), the convergence order may be even higher (i.e., better), e.g., if \( h'(p) = 0 \).
Version Log

- 4/11/2019, ver 1.00.