Numerical Analysis and Computing
Lecture Notes #2 — Calculus Review; Computer Arithmetic and Finite Precision; Algorithms and Convergence; Solutions of Equations of One Variable

Peter Blomgren,
⟨blomgren.peter@gmail.com⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://terminus.sdsu.edu/

Fall 2014
Outline

1 Calculus Review
   • Limits, Continuity, and Convergence
   • Differentiability, Rolle’s, and the Mean Value Theorem
   • Extreme Value, Intermediate Value, and Taylor’s Theorem

2 Computer Arithmetic & Finite Precision
   • Binary Representation, IEEE 754-1985
   • Something’s Missing...
   • Roundoff and Truncation, Errors, Digits
   • Cancellation

3 Algorithms
   • Algorithms, Pseudo-Code
   • Fundamental Concepts

4 Solutions of Equations of One Variable
   • \( f(x) = 0 \), “Root Finding”
   • The Bisection Method
   • When do we stop?!
   • *** Homework #1 ***
Why Review Calculus???

It’s a good warm-up for our brains!

When developing numerical schemes we will use theorems from calculus to guarantee that our algorithms make sense.

If the theory is sound, when our programs fail we look for bugs in the code!
Background Material — A Crash Course in Calculus

Key concepts from Calculus

- Limits
- Continuity
- Convergence
- Differentiability
- Rolle’s Theorem
- Mean Value Theorem
- Extreme Value Theorem
- Intermediate Value Theorem
- Taylor’s Theorem...
Limit / Continuity

Definition (Limit)
A function $f$ defined on a set $X$ of real numbers $X \subset \mathbb{R}$ has the limit $L$ at $x_0$, written

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

if given any real number $\epsilon > 0$ ($\forall \epsilon > 0$), there exists a real number $\delta > 0$ ($\exists \delta > 0$) such that $|f(x) - L| < \epsilon$, whenever $x \in X$ and $0 < |x - x_0| < \delta$.

Definition (Continuity (at a point))
Let $f$ be a function defined on a set $X$ of real numbers, and $x_0 \in X$. Then $f$ is continuous at $x_0$ if

$$\lim_{{x \to x_0}} f(x) = f(x_0).$$
Example: Continuity at $x_0$

Here we see how the limit $x \to x_0$ (where $x_0 = 0.5$) exists for the function $f(x) = x + \frac{1}{2} \sin(2\pi x)$.
Examples: Jump Discontinuity

The function

\[ f(x) = \begin{cases} 
  x + \frac{1}{2} \sin(2\pi x) & x < 0.5 \\
  x + \frac{1}{2} \sin(2\pi x) + 1 & x > 0.5 
\end{cases} \]

has a jump discontinuity at \( x_0 = 0.5 \).
Examples: “Spike” Discontinuity

The function

\[
f(x) = \begin{cases} 
1 & x = 0.5 \\
0 & x \neq 0.5
\end{cases}
\]

has a discontinuity at \( x_0 = 0.5 \).

The limit exists, but

\[
\lim_{x \to 0.5} f(x) = 0 \neq 1
\]
Continuity / Convergence

Definition (Continuity (in an interval))
The function \( f \) is continuous on the set \( X \) (\( f \in C(X) \)) if it is continuous at each point \( x \) in \( X \).

Definition (Convergence of a sequence)
Let \( x = \{x_n\}^\infty_{n=1} \) be an infinite sequence of real (or complex numbers). The sequence \( x \) converges to \( x \) (has the limit \( x \)) if \( \forall \epsilon > 0, \exists N(\epsilon) \in \mathbb{Z}^+ : |x_n - x| < \epsilon \ \forall n > N(\epsilon) \). The notation

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

means that the sequence \( \{x_n\}^\infty_{n=1} \) converges to \( x \).
Illustration: Convergence of a Complex Sequence

A sequence in $\mathbf{z} = \{z_k\}_{k=1}^{\infty}$ converges to $z_0 \in \mathbb{C}$ (the black dot) if for any $\epsilon$ (the radius of the circle), there is a value $N$ (which depends on $\epsilon$) so that the “tail” of the sequence $\mathbf{z}_t = \{z_k\}_{k=N}^{\infty}$ is inside the circle.
Differentiability

Theorem

If $f$ is a function defined on a set $X$ of real numbers and $x_0 \in X$, the following statements are equivalent:

(a) $f$ is continuous at $x_0$

(b) If $\{x_n\}_{n=1}^{\infty}$ is any sequence in $X$ converging to $x_0$, then $\lim_{n \to \infty} f(x_n) = f(x_0)$.

Definition (Differentiability (at a point))

Let $f$ be a function defined on an open interval containing $x_0$ ($a < x_0 < b$). $f$ is differentiable at $x_0$ if

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists.

If the limit exists, $f'(x_0)$ is the derivative at $x_0$.

Definition (Differentiability (in an interval))

If $f'(x_0)$ exists $\forall x_0 \in X$, then $f$ is differentiable on $X$. 

Peter Blomgren, ⟨blomgren.peter@gmail.com⟩

Calculus Review — (11/66)
Here we see that the limit

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists — and approaches the slope / derivative at $x_0$, $f'(x_0)$. 
Theorem (Differentiability $\Rightarrow$ Continuity)

*If $f$ is differentiable at $x_0$, then $f$ is continuous at $x_0$.*

Theorem (Rolle’s Theorem [Wiki-Link](#))

*Suppose $f \in C[a, b]$ and that $f$ is differentiable on $(a, b)$. If $f(a) = f(b)$, then $\exists c \in (a, b): f'(c) = 0$.  

\[ f'(c) = 0 \]  

[Diagram: A graph showing a function with a point $c$ where $f'(c) = 0$.]
Mean Value Theorem

Theorem (Mean Value Theorem [Wiki-Link])

If $f \in C[a, b]$ and $f$ is differentiable on $(a, b)$, then $\exists c \in (a, b)$:

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$
Extreme Value Theorem

Theorem (Extreme Value Theorem)  

If \( f \in C[a, b] \) then \( \exists c_1, c_2 \in [a, b] : f(c_1) \leq f(x) \leq f(c_2) \) \( \forall x \in [a, b] \). If \( f \) is differentiable on \((a, b)\) then the numbers \( c_1, c_2 \) occur either at the endpoints of \([a, b]\) or where \( f'(x) = 0 \).
Intermediate Value Theorem

Theorem (Intermediate Value Theorem) \([\text{Wiki-Link}]\)

If \( f \in C[a, b] \) and \( K \) is any number between \( f(a) \) and \( f(b) \), then there exists a number \( c \) in \((a, b)\) for which \( f(c) = K \).
Taylor’s Theorem

Theorem (Taylor’s Theorem \( \text{Wiki-Link} \) )

Suppose \( f \in C^n[a, b] \), \( f^{(n+1)} \exists \) on \( [a, b] \), and \( x_0 \in [a, b] \). Then \( \forall x \in (a, b), \ \exists \xi(x) \in (x_0, x) \) with \( f(x) = P_n(x) + R_n(x) \) where

\[
P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k, \quad R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!} (x-x_0)^{n+1}.
\]

\( P_n(x) \) is called the **Taylor polynomial of degree** \( n \), and \( R_n(x) \) is the **remainder term** (truncation error).

This theorem is **extremely important** for numerical analysis; Taylor expansion is a fundamental step in the derivation of many of the algorithms we see in this class (and in Math 542 & 693ab).
Illustration: Taylor's Theorem

\[ f(x) = \sin(x) \]

\[ P_{13}(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \frac{1}{9!}x^9 - \frac{1}{11!}x^{11} + \frac{1}{13!}x^{13} \]

\[ P_5(x) \]

\[ P_9(x) \]

\[ P_{13}(x) \]
Illustration: Taylor’s Theorem

$f(x) = \sin(x)$

$E_5(x)$

$E_9(x)$

$E_{13}(x)$

$P_{13}(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \frac{1}{9!}x^9 - \frac{1}{11!}x^{11} + \frac{1}{13!}x^{13}$

$\underbrace{P_5(x)}$

$\underbrace{P_9(x)}$
Taylor Expansions — Matlab

- A **Taylor polynomial of degree** \( n \) requires all derivatives up to order \( n \), and order \( n + 1 \) for the **remainder**.
- Derivatives may be [more] complicated expression [than the original function].
- **Matlab** can compute derivatives for you:

Matlab: Symbolic Computations

Try this!!!

```matlab
>> syms x
>> diff(sin(2*x))
>> diff(sin(2*x),3)
>> taylor(exp(x),x,0,'order',5)
>> taylor(exp(x),x,1,'order',5)
```
Computers use a finite number of bits (0’s and 1’s) to represent numbers.

For instance, an 8-bit unsigned integer (a.k.a a “char”) is stored:

\[
\begin{array}{cccccccc}
2^7 & 2^6 & 2^5 & 2^4 & 2^3 & 2^2 & 2^1 & 2^0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\
\end{array}
\]

Here, \(2^6 + 2^3 + 2^2 + 2^0 = 64 + 8 + 4 + 1 = 77\), which represents the upper-case character “M” (US-ASCII).
The *Binary Floating Point Arithmetic Standard* 754-1985 (IEEE — The Institute for Electrical and Electronics Engineers) standard specified the following layout for a 64-bit real number:

\[ s \, c_{10} \, c_{9} \, \ldots \, c_{1} \, c_{0} \, m_{51} \, m_{50} \, \ldots \, m_{1} \, m_{0} \]

Where

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s )</td>
<td>1</td>
<td>The sign bit — ( 0=)positive, ( 1=)negative</td>
</tr>
<tr>
<td>( c )</td>
<td>11</td>
<td>The characteristic (exponent)</td>
</tr>
<tr>
<td>( m )</td>
<td>52</td>
<td>The mantissa</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
    r &= (-1)^s \, 2^{c-1023} \, (1 + m) , \\
    c &= \sum_{k=0}^{10} c_k \, 2^k , \\
    m &= \sum_{k=0}^{51} \frac{m_k}{2^{52-k}}
\end{align*}
\]
As described in previous slide, we cannot represent zero!

There are some special signals in IEEE-754-1985:

<table>
<thead>
<tr>
<th>Type</th>
<th>S (1 bit)</th>
<th>C (11 bits)</th>
<th>M (52 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>signaling NaN</td>
<td>u</td>
<td>2047 (max)</td>
<td>.0uuuuuu—u (with at least one 1 bit)</td>
</tr>
<tr>
<td>quiet NaN</td>
<td>u</td>
<td>2047 (max)</td>
<td>.1uuuuu—u</td>
</tr>
<tr>
<td>negative infinity</td>
<td>1</td>
<td>2047 (max)</td>
<td>.000000—0</td>
</tr>
<tr>
<td>positive infinity</td>
<td>0</td>
<td>2047 (max)</td>
<td>.000000—0</td>
</tr>
<tr>
<td>negative zero</td>
<td>1</td>
<td>0</td>
<td>.000000—0</td>
</tr>
<tr>
<td>positive zero</td>
<td>0</td>
<td>0</td>
<td>.000000—0</td>
</tr>
</tbody>
</table>

From: [http://www.freesoft.org/CIE/RFC/1832/32.htm](http://www.freesoft.org/CIE/RFC/1832/32.htm)
Examples: Finite Precision

\[
 r = (-1)^s 2^{c-1023} (1 + f), \quad c = \sum_{k=0}^{10} c_k 2^k, \quad m = \sum_{k=0}^{51} \frac{m_k}{2^{52-k}}
\]

**Example #1: 3.0**

\[
\begin{align*}
 r_1 &= (-1)^0 \cdot 2^{2^{10} - 1023} \cdot \left(1 + \frac{1}{2}\right) = 1 \cdot 2^1 \cdot \frac{3}{2} = 3.0
\end{align*}
\]

**Example #2: The Smallest Positive Real Number**

\[
\begin{align*}
 r_2 &= (-1)^0 \cdot 2^{0-1023} \cdot (1 + 2^{-52}) = (1 + 2^{-52}) \cdot 2^{-1023} \cdot 1 \approx 10^{-308}
\end{align*}
\]
Examples: Finite Precision

\[ r = (-1)^s \, 2^{c-1023} \, (1 + f), \quad c = \sum_{k=0}^{10} c_k 2^k, \quad m = \sum_{k=0}^{51} \frac{m_k}{2^{52-k}} \]

Example #3: The Largest Positive Real Number

```
0 11111111110 111111111111111111111111111111111111111111111111111
```

\[
\begin{align*}
    r_3 &= (-1)^0 \cdot 2^{1023} \cdot \left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{51}} + \frac{1}{2^{52}}\right) \\
        &= 2^{1023} \cdot \left(2 - \frac{1}{2^{52}}\right) \approx 10^{308}
\end{align*}
\]
There are gaps in the floating-point representation!

Given the representation

\[ 0 \text{00000000000 00000000000000000000000000000000000000000000000001} \]

for the value \( \frac{2^{-1023}}{2^{52}} \).

The next larger floating-point value is

\[ 0 \text{00000000000 0000000000000000000000000000000000000000000000010} \]

i.e. the value \( \frac{2^{-1023}}{2^{51}} \).

The difference between these two values is \( \frac{2^{-1023}}{2^{52}} = 2^{-1075} \).

Any number in the interval \( \left( \frac{2^{-1023}}{2^{52}}, \frac{2^{-1023}}{2^{51}} \right) \) is not representable!
A gap of $2^{-1075}$ doesn’t seem too bad...

However, the size of the gap depend on the value itself...

Consider $r = 3.0$

$0\ 10000000000\ 100000000000000000000000000000000000000000000000000$

and the next value

$0\ 10000000000\ 100000000000000000000000000000000000000000000000001$

The difference is $\frac{2}{2^{52}} \approx 4.4 \cdot 10^{-16}$. 
At the other extreme, the difference between

\[ 0 \, 11111111110 \, 1111111111111111111111111111111111111111111111111111 \]

and the previous value

\[ 0 \, 11111111110 \, 1111111111111111111111111111111111111111111111111110 \]

is \[ \frac{2^{1023}}{2^{52}} = 2^{971} \approx 1.99 \cdot 10^{292} \].

That’s a “fairly significant” gap!!

The number of atoms in the observable universe can be estimated to be no more than \( \sim 10^{80} \).
The Relative Gap

It makes more sense to factor the exponent out of the discussion and talk about the relative gap:

<table>
<thead>
<tr>
<th>Exponent</th>
<th>Gap</th>
<th>Relative Gap (Gap/Exponent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2^{-1023}</td>
<td>2^{-1075}</td>
<td>2^{-52}</td>
</tr>
<tr>
<td>2^{1}</td>
<td>2^{-51}</td>
<td>2^{-52}</td>
</tr>
<tr>
<td>2^{1023}</td>
<td>2^{971}</td>
<td>2^{-52}</td>
</tr>
</tbody>
</table>

Any difference between numbers smaller than the local gap is not representable, e.g. any number in the interval

\[\left[3.0, 3.0 + \frac{1}{2^{51}}\right]\]

is represented by the value 3.0.
The Floating Point “Theorem”

“Theorem”

Floating point “numbers” represent intervals!

Since (most) humans find it hard to think in binary representation, from now on we will for simplicity and without loss of generality assume that floating point numbers are represented in the normalized floating point form as...

$k$-digit decimal machine numbers

\[ \pm 0.d_1d_2 \cdots d_{k-1}d_k \cdot 10^n, \]

where

\[ 1 \leq d_1 \leq 9, \quad 0 \leq d_i \leq 9, \quad i \geq 2, \quad n \in \mathbb{Z}. \]
The *Binary Floating Point Arithmetic Standard 754-1985* (IEEE — The Institute for Electrical and Electronics Engineers) standard specified the following layout for a 128-bit real number:

$$s c_{14} c_{13} \ldots c_1 c_0 m_{111} m_{110} \ldots m_1 m_0$$

Where

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s$</td>
<td>1</td>
<td>The sign bit — 0=positive, 1=negative</td>
</tr>
<tr>
<td>$c$</td>
<td>15</td>
<td>The exponent</td>
</tr>
<tr>
<td>$m$</td>
<td>112</td>
<td>The fraction</td>
</tr>
</tbody>
</table>

$$r = (-1)^s 2^{c-16,383} (1 + m), \quad c = \sum_{k=0}^{14} c_k 2^k, \quad m = \sum_{k=0}^{111} m_k 2^{52-k}$$
Finite Precision

Layout for a 256-bit real number:

\[
s c_{17} c_{16} \ldots c_1 c_0 m_{236} m_{235} \ldots m_1 m_0
\]

Where

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(s)</td>
<td>1</td>
<td>The sign bit — 0=positive, 1=negative</td>
</tr>
<tr>
<td>(c)</td>
<td>17</td>
<td>The exponent</td>
</tr>
<tr>
<td>(m)</td>
<td>237</td>
<td>The fraction</td>
</tr>
</tbody>
</table>

\[
r = (-1)^s 2^{c-131,071} (1 + m), \quad c = \sum_{k=0}^{17} c_k 2^k, \quad m = \sum_{k=0}^{236} \frac{m_k}{2^{52-k}}
\]
Any real number can be written in the form

$$\pm 0.d_1d_2 \cdots d_\infty \cdot 10^n$$

given infinite patience and storage space. We can obtain the floating-point representation $fl(r)$ in two ways:

1. Truncating (chopping) — just keep the first $k$ digits.

2. Rounding — if $d_{k+1} \geq 5$ then add 1 to $d_k$. Truncate.

**Examples**

$$fl_{t,5}(\pi) = 0.31415 \cdot 10^1, \quad fl_{r,5}(\pi) = 0.31416 \cdot 10^1$$

In both cases, the error introduced is called the **roundoff error**.
Quantifying the Error

Let \( p^* \) be an approximation to \( p \), then...

Definition (The Absolute Error)

\[
|p - p^*|
\]

Definition (The Relative Error)

\[
\frac{|p - p^*|}{|p|}, \quad p \neq 0
\]

Definition (Significant Digits)

The number of significant digits is the largest value of \( t \) for which

\[
\frac{|p - p^*|}{|p|} < 5 \cdot 10^{-t}
\]
Sources of Numerical Error

1) Representation — Roundoff.

2) Cancellation — Consider:

\[
\begin{align*}
0.12345678012345 \cdot 10^1 \\
- 0.12345678012344 \cdot 10^1 \\
= 0.100000000000000 \cdot 10^{-13}
\end{align*}
\]

this value has (at most) 1 significant digit!!!

If you assume a “canceled value” has more significant bits (the computer will happily give you some numbers) — I don’t want you programming the autopilot for any airlines!!!
Examples: 5-digit Arithmetic

Rounding 5-digit arithmetic

\[(96384 + 26.678) - 96410 =\]
\[(96384 + 00027) - 96410 =\]
\[96411 - 96410 = 1.0000\]

Truncating 5-digit arithmetic

\[(96384 + 26.678) - 96410 =\]
\[(96384 + 00026) - 96410 =\]
\[96410 - 96410 = 0.0000\]

Rearrangement changes the result:

\[(96384 - 96410) + 26.678 = -26.000 + 26.678 = 0.67800\]

Numerically, order of computation matters! (This is a HARD problem)
Examples: 5-digit Arithmetic

**Rounding 5-digit arithmetic**

\[
(0.96384 \cdot 10^5 + 0.26678 \cdot 10^2) - 0.96410 \cdot 10^5 = \\
(0.96384 \cdot 10^5 + 0.00027 \cdot 10^5) - 0.96410 \cdot 10^5 = \\
0.96411 \cdot 10^5 - 0.96410 \cdot 10^5 = 0.10000 \cdot 10^1
\]

**Truncating 5-digit arithmetic**

\[
(0.96384 \cdot 10^5 + 0.26678 \cdot 10^2) - 0.96410 \cdot 10^5 = \\
(0.96384 \cdot 10^5 + 0.00026 \cdot 10^5) - 0.96410 \cdot 10^5 = \\
0.96410 \cdot 10^5 - 0.96410 \cdot 10^5 = 0.00000 \cdot 10^0
\]

**Rearrangement changes the result:**

\[
(0.96384 \cdot 10^5 - 0.96410 \cdot 10^5) + 0.26678 \cdot 10^2 = \\
-0.26000 \cdot 10^2 + 0.26678 \cdot 10^2 = 0.67800 \cdot 10^0
\]
Example: Loss of Significant Digits due to Subtractive Cancellation

Consider the recursive relation

\[ x_{n+1} = 1 - (n + 1)x_n \quad \text{with} \quad x_0 = 1 - \frac{1}{e}. \]

This sequence can be shown to converge to 0 (in 2 slides).
Subtractive cancellation produces an error which is approximately equal to the machine precision times \( n! \).
Subtractive Cancellation Example: Output

<table>
<thead>
<tr>
<th>$n$</th>
<th>$x_n$</th>
<th>$n!$</th>
<th>$n$</th>
<th>$x_n$</th>
<th>$n!$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.63212056</td>
<td>1</td>
<td>11</td>
<td>0.07735223</td>
<td>3.99e+007</td>
</tr>
<tr>
<td>1</td>
<td>0.36787944</td>
<td>1</td>
<td>12</td>
<td>0.07177325</td>
<td>4.79e+008</td>
</tr>
<tr>
<td>2</td>
<td>0.26424112</td>
<td>2</td>
<td>13</td>
<td>0.06694778</td>
<td>6.23e+009</td>
</tr>
<tr>
<td>3</td>
<td>0.20727665</td>
<td>6</td>
<td>14</td>
<td>0.06273108</td>
<td>8.72e+010</td>
</tr>
<tr>
<td>4</td>
<td>0.17089341</td>
<td>24</td>
<td>15</td>
<td>0.05903379</td>
<td>1.31e+012</td>
</tr>
<tr>
<td>5</td>
<td>0.14553294</td>
<td>120</td>
<td>16</td>
<td>0.05545930</td>
<td>2.09e+013</td>
</tr>
<tr>
<td>6</td>
<td>0.12680236</td>
<td>720</td>
<td>17</td>
<td>0.05719187</td>
<td>3.56e+014</td>
</tr>
<tr>
<td>7</td>
<td>0.11238350</td>
<td>5.04e+003</td>
<td>18</td>
<td>−0.02945367</td>
<td>6.4e+015</td>
</tr>
<tr>
<td>8</td>
<td>0.10093197</td>
<td>4.03e+004</td>
<td>19</td>
<td>1.55961974</td>
<td>1.22e+017</td>
</tr>
<tr>
<td>9</td>
<td>0.09161229</td>
<td>3.63e+005</td>
<td>20</td>
<td>−30.19239489</td>
<td>2.43e+018</td>
</tr>
<tr>
<td>10</td>
<td>0.08387707</td>
<td>3.63e+006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Proof of Convergence to 0

The recursive relation is

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

with

\[ x_0 = 1 - \frac{1}{e} = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + ... \]

From the recursive relation

\[ x_1 = 1 - x_0 = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - ... \]
\[ x_2 = 1 - 2x_1 = \frac{1}{3} - \frac{2}{4!} + \frac{2}{5!} - ... \]
\[ x_3 = 1 - 3x_2 = \frac{3!}{4!} - \frac{3!}{5!} + \frac{3!}{6!} - ... \]

\[ \vdots \]

\[ x_n = 1 - nx_{n-1} = \frac{n!}{(n+1)!} - \frac{n!}{(n+2)!} + \frac{n!}{(n+3)!} - ... \]

This shows that

\[ x_n = \frac{1}{n+1} - \frac{1}{(n+1)(n+2)} + ... \to 0 \quad \text{as} \quad n \to \infty. \]
Matlab code: Loss of Significant Digits

clear
x(1) = 1 - 1/exp(1);
s(1) = 1;
f(1) = 1;
for i = 2:21
    x(i) = 1 - (i-1)*x(i-1);
s(i) = 1/i;
f(i) = (i-1)*f(i-1);
end
n = 0:20;
z = [n; x; s; f];
fprintf(1, '
\n n x(n) 1/(n+1) n!
\n', z)
Definition (Algorithm)

An **algorithm** is a procedure that describes, in an **unambiguous manner**, a finite sequence of steps to be performed in a specific order.

In this class, the objective of an algorithm is to implement a procedure to solve a problem or approximate a solution to a problem.

Most homes have a collection of algorithms in printed form — we tend to call them “recipes.”

There is a collection of algorithms “out there” called **Numerical Recipes**, Google for it!
Pseudo-code

Definition (Pseudo-code)

**Pseudo-code** is an algorithm description which specifies the input/output formats.

Note that pseudo-code is **not** computer language specific, but should be easily translatable to any procedural computer language.

**Examples of Pseudo-code statements:**

```plaintext
for i = 1,2,...,n
    Set \( x_i = a_i + i \times h \)

While \( i < N \) do Steps 17 - 21
    If ... then ... else
```
Definition (Stability)
An algorithm is said to be stable if small changes in the input, generates small changes in the output.

At some point we need to quantify what “small” means!

If an algorithm is stable for a certain range of initial data, then it is said to be conditionally stable.

Stability issues are discussed in great detail in Math 543.
Suppose $E_0 > 0$ denotes the initial error, and $E_n$ represents the error after $n$ operations.

If $E_n \approx C E_0 \cdot n$ (for a constant $C$ which is independent of $n$), then the growth is **linear**.

If $E_n \approx C^n E_0$, $C > 1$, then the growth is **exponential** — in this case the error will dominate very fast (undesirable scenario).

**Linear error growth** is usually unavoidable, and in the case where $C$ and $E_0$ are small the results are generally acceptable. — **Stable algorithm**.

**Exponential error growth** is unacceptable. Regardless of the size of $E_0$ the error grows rapidly. — **Unstable algorithm**.
Example BF-1.3.3

The recursive equation

\[ p_n = \frac{10}{3} p_{n-1} - p_{n-2}, \quad n = 2, 3, \ldots, \infty \]

has the exact solution

\[ p_n = c_1 \left( \frac{1}{3} \right)^n + c_2 3^n \]

for any constants \( c_1 \) and \( c_2 \). (Determined by starting values.)

In particular, if \( p_0 = 1 \) and \( p_1 = \frac{1}{3} \), we get \( c_1 = 1 \) and \( c_2 = 0 \), so

\[ p_n = \left( \frac{1}{3} \right)^n \text{ for all } n. \]

Now, consider what happens in 5-digit rounding arithmetic...
Now, consider what happens in 5-digit rounding arithmetic...

\[ p_0^* = 1.0000, \quad p_1^* = 0.33333 \]

which modifies

\[ c_1^* = 1.0000, \quad c_2^* = -0.12500 \cdot 10^{-5} \]

The generated sequence is

\[ p_n^* = 1.0000 (0.33333)^n - 0.12500 \cdot 10^{-5} (3.0000)^n \]

\[ \text{Exponential Growth} \]

\[ p_n^* \] quickly becomes a very poor approximation to \( p_n \) due to the exponential growth of the initial roundoff error.
The effects of roundoff error can be reduced by using higher-order-digit arithmetic such as the double or multiple-precision arithmetic available on most computers.

Disadvantages in using double precision arithmetic are that it takes more computation time and the growth of the roundoff error is not eliminated but only postponed.

Sometimes, but not always, it is possible to reduce the growth of the roundoff error by restructuring the calculations.
Key Concepts

Definition (Rate of Convergence)

Suppose the sequence $\beta = \{\beta_n\}_{n=1}^{\infty}$ converges to zero, and $\alpha = \{\alpha_n\}_{n=1}^{\infty}$ converges to a number $\alpha$.

If $\exists K > 0$: $|\alpha_n - \alpha| < K\beta_n$, for $n$ large enough, then we say that $\{\alpha_n\}_{n=1}^{\infty}$ converges to $\alpha$ with a Rate of Convergence $O(\beta_n)$ ("Big Oh of $\beta_n$").

We write

$$\alpha_n = \alpha + O(\beta_n)$$

Note: The sequence $\beta = \{\beta_n\}_{n=1}^{\infty}$ is usually chosen to be

$$\beta_n = \frac{1}{n^p}$$

for some positive value of $p$. 
Examples: Rate of Convergence

**Example #1:** If

\[ \alpha_n = \alpha + \frac{1}{\sqrt{n}} \]

then for any \( \epsilon > 0 \)

\[ |\alpha_n - \alpha| = \frac{1}{\sqrt{n}} \leq (1 + \epsilon) \frac{1}{\sqrt{n}} \]

hence

\[ \alpha_n = \alpha + \mathcal{O} \left( \frac{1}{\sqrt{n}} \right) \]
Examples: Rate of Convergence

**Example #2**: Consider the sequence (as $n \to \infty$)

$$
\alpha_n = \sin \left( \frac{1}{n} \right) - \frac{1}{n}
$$

We **Taylor expand** $\sin(x)$ about $x_0 = 0$:

$$
\sin \left( \frac{1}{n} \right) \sim \frac{1}{n} - \frac{1}{6n^3} + O \left( \frac{1}{n^5} \right)
$$

Hence

$$
|\alpha_n| = \left| \frac{1}{6n^3} + O \left( \frac{1}{n^5} \right) \right|
$$

It follows that

$$
\alpha_n = 0 + O \left( \frac{1}{n^3} \right)
$$

Note:

$$
O \left( \frac{1}{n^3} \right) + O \left( \frac{1}{n^5} \right) = O \left( \frac{1}{n^3} \right), \quad \text{since} \quad \frac{1}{n^5} \ll \frac{1}{n^3}, \quad \text{as} \quad n \to \infty
$$
Generalizing to Continuous Limits

Definition (Rate of Convergence)
Suppose
\[ \lim_{h \to 0} G(h) = 0, \quad \text{and} \quad \lim_{h \to 0} F(h) = L \]
If \( \exists K > 0: \)
\[ |F(h) - L| \leq K |G(h)| \]
\( \forall h < H \) (for some \( H > 0 \)), then
\[ F(h) = L + O(G(h)) \]
we say that \( F(h) \) converges to \( L \) with a Rate of Convergence \( O(G(h)) \).

Usually \( G(h) = h^p, \; p > 0 \).
Examples: Rate of Convergence

Example #2-b: Consider the function \( \alpha(h) \) (as \( h \to 0 \))

\[
\alpha(h) = \sin(h) - h
\]

We **Taylor expand** \( \sin(x) \) about \( x_0 = 0 \):

\[
\sin(h) \sim h - \frac{h^3}{6} + \mathcal{O}(h^5)
\]

Hence

\[
|\alpha(h)| = \left| \frac{h^3}{6} + \mathcal{O}(h^5) \right|
\]

It follows that

\[
\lim_{h \to 0} \alpha(h) = 0 + \mathcal{O}(h^3)
\]

Note:

\[
\mathcal{O}(h^3) + \mathcal{O}(h^5) = \mathcal{O}(h^3), \quad \text{since} \quad h^5 \ll h^3, \quad \text{as} \quad h \to 0
\]
Our new favorite problem:

\[ f(x) = 0. \]
We are going to solve the equation $f(x) = 0$ (i.e. finding root to the equation), for functions $f$ that are complicated enough that there is no closed form solution (and/or we are too lazy to find it?)

In a lot of cases we will solve problems to which we can find the closed form solutions — we do this as a training ground and to evaluate how good our numerical methods are.
The Bisection Method

Suppose $f$ is continuous on the interval $(a_0, b_0)$ and $f(a_0) \cdot f(b_0) < 0$ — This means the function changes sign at least once in the interval.

The **intermediate value theorem** guarantees the existence of $c \in (a_0, b_0)$ such that $f(c) = 0$.

Without loss of generality (just consider the function $-f(x)$), we can assume (for now) that $f(a_0) < 0$.

We will construct a sequence of intervals containing the root $c$:

$$(a_0, b_0) \supset (a_1, b_1) \supset \cdots \supset (a_{n-1}, b_{n-1}) \supset (a_n, b_n) \ni c$$
The sub-intervals are determined recursively:

Given \((a_{k-1}, b_{k-1})\), let \(m_{k-1} = \frac{a_{k-1} + b_{k-1}}{2}\) be the mid-point.

If \(f(m_{k-1}) = 0\), we’re done, otherwise

\[
(a_k, b_k) = \begin{cases} 
(m_{k-1}, b_{k-1}) & \text{if } f(m_{k-1}) < 0 \\
(a_{k-1}, m_{k-1}) & \text{if } f(m_{k-1}) > 0 
\end{cases}
\]

This construction guarantees that \(f(a_k) \cdot f(b_k) < 0\) and \(c \in (a_k, b_k)\).
The Bisection Method

After $n$ steps, the interval $(a_n, b_n)$ has the length

$$|b_n - a_n| = \left(\frac{1}{2}\right)^n |b_0 - a_0|.$$

We can take

$$m_n = \frac{a_n + b_n}{2}$$

as the estimate for the root $c$ and we have

$$c = m_n \pm d_n, \quad d_n = \left(\frac{1}{2}\right)^{n+1} |b_0 - a_0|.$$
Convergence is slow:

At each step we gain one binary digit in accuracy. Since $10^{-1} \approx 2^{-3.3}$, it takes on average 3.3 iterations to gain one decimal digit of accuracy.

Note: The rate of convergence is completely independent of the function $f$.

We are only using the sign of $f$ at the endpoints of the interval(s) to make decisions on how to update. — By making more effective use of the values of $f$ we can attain significantly faster convergence.

First an example...
The bisection method applied to

\[ f(x) = \left(\frac{x}{2}\right)^2 - \sin(x) = 0 \]

with \((a_0, b_0) = (1.5, 2.0)\), and \((f(a_0), f(b_0)) = (-0.4350, 0.0907)\) gives:

<table>
<thead>
<tr>
<th>(k)</th>
<th>(a_k)</th>
<th>(b_k)</th>
<th>(m_k)</th>
<th>(f(m_k))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.5000</td>
<td>2.0000</td>
<td>1.7500</td>
<td>-0.2184</td>
</tr>
<tr>
<td>1</td>
<td>1.7500</td>
<td>2.0000</td>
<td>1.8750</td>
<td>-0.0752</td>
</tr>
<tr>
<td>2</td>
<td>1.8750</td>
<td>2.0000</td>
<td>1.9375</td>
<td>0.0050</td>
</tr>
<tr>
<td>3</td>
<td>1.8750</td>
<td>1.9375</td>
<td>1.9062</td>
<td>-0.0358</td>
</tr>
<tr>
<td>4</td>
<td>1.9062</td>
<td>1.9375</td>
<td>1.9219</td>
<td>-0.0156</td>
</tr>
<tr>
<td>5</td>
<td>1.9219</td>
<td>1.9375</td>
<td>1.9297</td>
<td>-0.0054</td>
</tr>
<tr>
<td>6</td>
<td>1.9297</td>
<td>1.9375</td>
<td>1.9336</td>
<td>-0.0002</td>
</tr>
<tr>
<td>7</td>
<td>1.9336</td>
<td>1.9375</td>
<td>1.9355</td>
<td>0.0024</td>
</tr>
<tr>
<td>8</td>
<td>1.9336</td>
<td>1.9355</td>
<td>1.9346</td>
<td>0.0011</td>
</tr>
<tr>
<td>9</td>
<td>1.9336</td>
<td>1.9346</td>
<td>1.9341</td>
<td>0.0004</td>
</tr>
</tbody>
</table>
The Bisection Method

Example, 2 of 2

The Bisection Method Example, 2 of 2

Peter Blomgren, ⟨blomgren.peter@gmail.com⟩
Matlab code: The Bisection Method

```matlab
% WARNING: This example ASSUMES that f(a)<0<f(b)...
x = 1.5:0.001:2;
f = inline('(x/2).^2-sin(x)','x');
a = 1.5;
b = 2.0;
for k = 0:9
    plot(x,f(x),'k-','linewidth',2)
    axis([1.45 2.05 -0.5 .15])
    grid on
    hold on
    plot([a b],f([a b]),'ko','linewidth',5)
    hold off
    m = (a+b)/2;
    if( f(m) < 0 )
        a = m;
    else
        b = m;
    end
    pause
    print('-depsc',['bisec' int2str(k) '.eps'],'-f1');
end
```

Peter Blomgren, ⟨blomgren.peter@gmail.com⟩
Stopping Criteria

When do we stop?

We can (1) keep going until successive iterates are close:

\[ |m_k - m_{k-1}| < \epsilon \]

or (2) close in relative terms

\[ \frac{|m_k - m_{k-1}|}{|m_k|} < \epsilon \]

or (3) the function value is small enough

\[ |f(m_k)| < \epsilon \]

No choice is perfect. In general, where no additional information about \( f \) is known, the second criterion is the preferred one (since it comes the closest to testing the relative error).
Matlab command(s) of the day: help, lookfor

help — Display help text in Command Window
matlab>> help, by itself, lists all primary help topics. [...] 
matlab>> help help, gives help for the help command.

lookfor — Find all functions with
matlab>> lookfor function, will return a (long) list of things related to functions.
Homework #1

- Will open on 08/29/2014 at 09:30am PDT.
- Will close no earlier than 09/09/2014 at 09:00pm PDT.

http://webwork.sdsu.edu