Numerical Analysis and Computation

Calculus Review; Computer Artihmetic and Finite Precision

Lecture Notes #2

Joe Mahaffy
Department of Mathematics and Statistics
San Diego State University
San Diego, CA 92182-7720
mahaffy@math.sdsu.edu
http://www-rohan.sdsu.edu/~jmahaffy

$Id: lecture.tex,v 1.25 2007/08/31 19:49:27 blomgren Exp $

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!

Key concepts from Calculus

- Limits
- Continuity
- Convergence
- Differentiability
- Rolle's Theorem
- Mean Value Theorem
- Extreme Value Theorem
- Intermediate Value Theorem
- Taylor's Theorem

***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

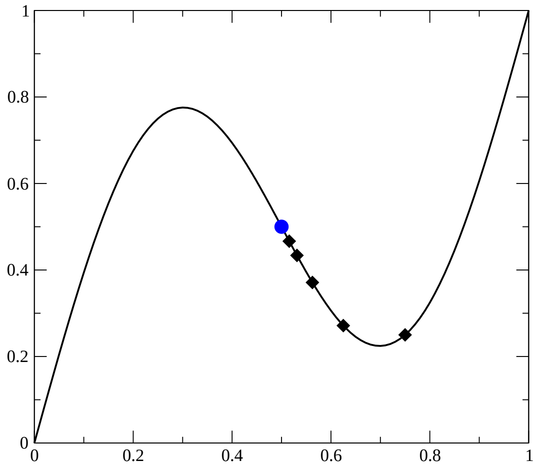
$$\lim_{x \to x_0} f(x) = L$$

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

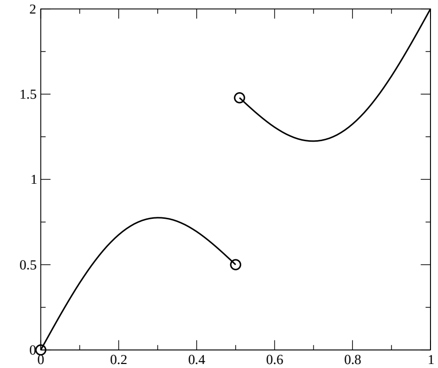
$$\lim_{x \to x_0} f(x) = f(x_0).$$

Example: Continuity at x_0



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

Examples: Jump Discontinuity

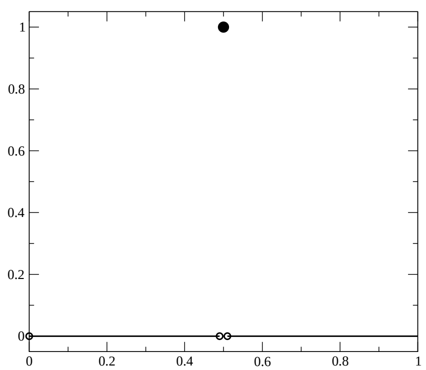


The function

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

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

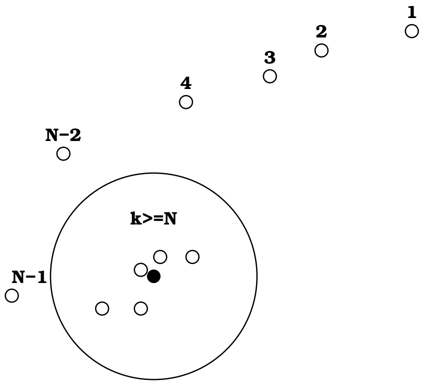
***Definition:* Continuity (in an interval) —**
The function f is continuous on the set X (denoted $f \in C(X)$) if it is continuous at each point x in X.

***Definition:* Convergence of a sequence —**
Let $\underline{\mathbf{x}} = \{x_n\}_{n=1}^{\infty}$ be an infinite sequence of real (or complex numbers). The sequence $\underline{\mathbf{x}}$ 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 $\{x_n\}_{n=1}^{\infty}$ converges to x.

Illustration: Convergence of a Complex Sequence



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

Differentiability

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

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

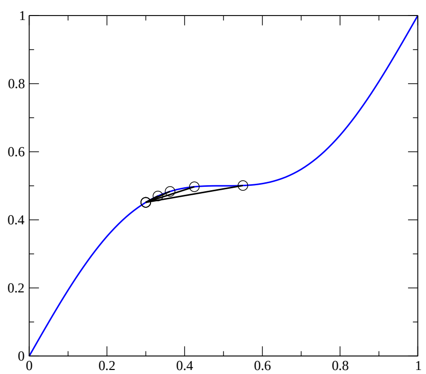
Definition: ***Differentiability (at a point)*** — Let f be a function defined on an open interval containing x_0 ($a < x_0 < b$). f is differentiable at x_0 if

$$f'(x_0) = \lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad \text{exists.}$$

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

Illustration: Differentiability



Here we see that the limit

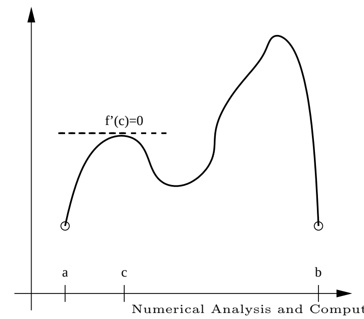
$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists — and approaches the slope / derivative at x_0, $f'(x_0)$.

Continuity / Rolle's Theorem

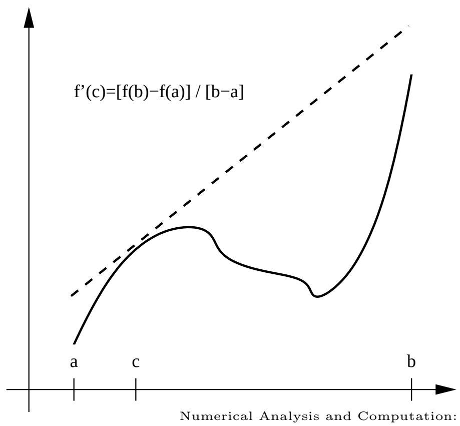
Theorem: ***Differentiability $\Rightarrow$ Continuity*** —
If f is differentiable at x_0, then f is continuous at x_0.

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



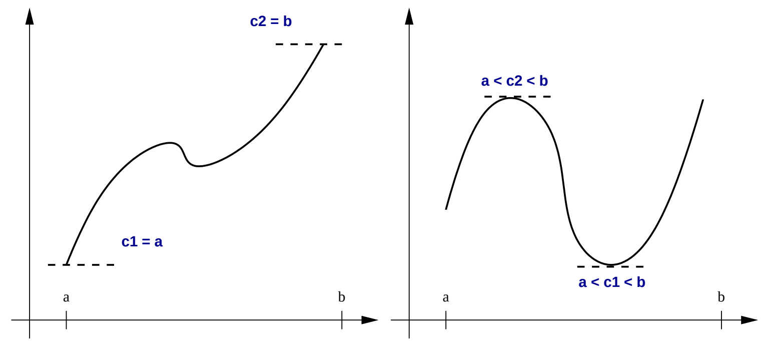
Mean Value Theorem

***Theorem:* Mean Value Theorem—**
If $f \in C[a,b]$ and f is differentiable on (a,b), then $\exists c \in (a,b)$: $f'(c) = \dfrac{f(b)-f(a)}{b-a}$.



Extreme Value Theorem

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



Taylor's Theorem

***Theorem:* Taylor's Theorem —**
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

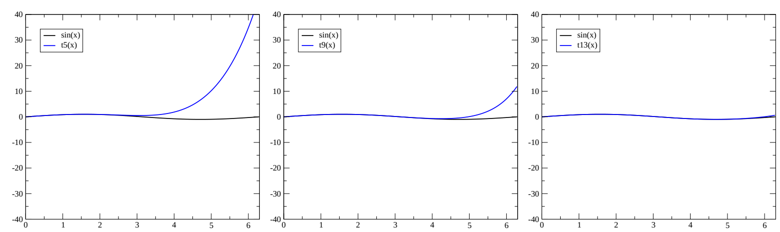
$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k,$$

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

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

Note: $f^{(n+1)}\exists$ on $[a,b]$, but is not necessarily continuous.

Illustration: Taylor's Theorem $\quad f(x) = \sin(x)$



$P_5(x)$ $\quad$ $P_9(x)$ $\quad$ $P_{13}(x)$

$$P_{13}(x) = \underbrace{\underbrace{x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5}_{P_5(x)} - \frac{1}{7!}x^7 + \frac{1}{9!}x^9}_{P_9(x)} - \frac{1}{11!}x^{11} + \frac{1}{13!}x^{13}$$

Taylor's Theorem: Computer Programming – Maple

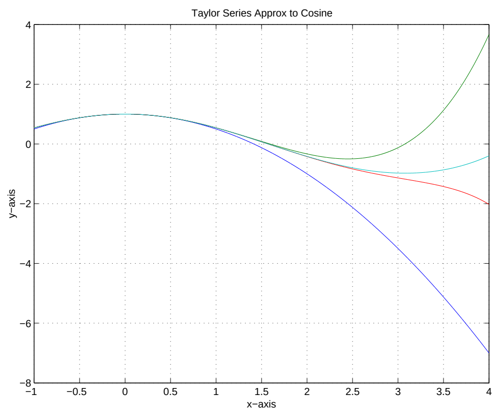
- A ***Taylor polynomial of degree*** n requires all derivatives up to order n and degree $n+1$ for the ***Remainder***.
- In general, derivatives may be complicated expressions.
- Maple computes derivatives accurately and efficiently – differentiation uses the command **diff(f(x), x);**
- Maple has a routine for Taylor series expansions – finding the Taylor's series uses the command **taylor(f(x), x=x0, n);**, meaning the Taylor series expansion about $x = x_0$ using n terms in the expansion.
- A Maple worksheet is available with many of these basic commands through my webpage for this class.

Taylor's Theorem: Computer Programming – MatLab

- Most versions of MatLab have a symbolic package that includes Maple, so this symbolic package can help with derivatives.
- Often easier to play to the strengths of each language and let Maple find the Taylor coefficients to employ in the MatLab code.
- MatLab provides relatively efficient numerical programs that are similar and based on C Programming.
- A MatLab code is provided to show the convergence of the Taylor series to the cosine function with increasing numbers of terms. This is shown on the Maple worksheet also, and the code is accessible through my webpage.

Taylor's Approximation for Cosine Function

- A series of Taylor polynomials approximating $\cos(x)$ with $n = 2$, 4, 6, and 8 are shown below.



Computer Arithmetic and Finite Precision

Computer Arithmetic and Finite Precision

Finite Precision — A single char

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:

2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
0	1	0	0	1	1	0	1

Here, $2^6 + 2^3 + 2^2 + 2^0 = 64 + 8 + 4 + 1 = 77$, which represents the upper-case character "M" (US-ASCII).

Finite Precision — A 64-bit real number, `double`

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:

$$\mathbf{s}\,\mathbf{c}_{10}\,\mathbf{c}_9\;\ldots\;\mathbf{c}_1\,\mathbf{c}_0\,\mathbf{m}_{51}\,\mathbf{m}_{50}\;\ldots\;\mathbf{m}_1\,\mathbf{m}_0$$

Where

Symbol	Bits	Description
s	1	The sign bit — 0=positive, 1=negative
c	11	The characteristic (exponent)
m	52	The mantissa

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

Burden-Faires' Description is not complete...

As described in previous slide, we cannot represent zero!

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

Type	S (1 bit)	C (11 bits)	M (52 bits)
signaling NaN	u	2047 (max)	.0uuuuu—u (with at least one 1 bit)
quiet NaN	u	2047 (max)	.1uuuuu—u
negative infinity	1	2047 (max)	.000000—0
positive infinity	0	2047 (max)	.000000—0
negative zero	1	0	.000000—0
positive zero	0	0	.000000—0

From: **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

0 10000000000 1000000000000000000000000000000000000000000000000000

$$r_1 = (-1)^0 \cdot 2^{2^{10}-1023} \cdot \left(1 + \frac{1}{2}\right) = 1 \cdot 2^1 \cdot \frac{3}{2} = 3.0$$

Example #2: The Smallest Positive Real Number

0 00000000000 0000000000000000000000000000000000000000000000000001

$$r_2 = (-1)^0 \cdot 2^{0-1023} \cdot (1 + 2^{-52}) = (1 + 2^{-52}) \cdot 2^{-1023} \cdot 1 \approx 10^{-308}$$

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 1111111111111111111111111111111111111111111111111111

$$\begin{aligned} r_3 &= (-1)^0 \cdot 2^{1023} \cdot \left(1 + \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{51}} + \frac{1}{2^{52}}\right) \\ &= 2^{1024} \cdot \left(2 - \frac{1}{2^{52}}\right) \approx 10^{308} \end{aligned}$$

Something is Missing — Gaps in the Representation

There are gaps in the floating-point representation!

Given the representation

0 00000000000 0000000000000000000000000000000000000000000000000001

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

The next larger floating-point value is

0 00000000000 0000000000000000000000000000000000000000000000000010

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!**

Something is Missing — Gaps in the Representation

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 1000000000000000000000000000000000000000000000000000

and the next value

0 10000000000 1000000000000000000000000000000000000000000000000001

The difference is $\frac{2}{2^{52}}$

Something is Missing — Gaps in the Representation

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:

Exponent	Gap	Relative Gap (Gap/Exponent)
2^{-1023}	2^{-1075}	2^{-52}
2^{1}	2^{-51}	2^{-52}
2^{1023}	2^{971}	2^{-52}

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

$$\pm 0.d_1 d_2 \cdots d_{k-1} d_k \cdot 10^n$$

where

$$1 \le d_1 \le 9, \quad 0 \le d_i \le 9,\ i \ge 2, \quad n \in \mathbb{Z}$$

k-Digit Decimal Machine Numbers

Any real number can be written in the form

$$\pm 0.d_1 d_2 \cdots d_\infty \cdot 10^n$$

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 $d_{k+1} \ge 5$ then add 1 to d_k. Truncate.

Examples

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

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

Quantifying the Error

Let p^* be and 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

$$\frac{|p - p^*|}{|p|} < 5 \cdot 10^{-t}$$

1) Representation — Roundoff.

2) Cancellation — Consider:

$$\begin{array}{rl} & 0.12345678012345 \cdot 10^{1} \\ - & 0.12345678012344 \cdot 10^{1} \\ \hline = & 0.10000000000000 \cdot 10^{-13} \end{array}$$

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!!!

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)

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 $\mathbf{0}$ (in 2 slides).

Subtractive cancellation produces an error which is approximately equal to the machine precision times $n!$.

The MatLab code for this example is provided on the webpage.

Maple has a routine **rsolve** that solves this recursive relation exactly, using the Gamma function.

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

Proof of Convergence to 0

The recursive relation is

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

with

$$x_0 = 1 - \frac{1}{e} = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \ldots$$

From the recursive relation

$$\begin{aligned}
x_1 &= 1 - x_0 = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \ldots \\
x_2 &= 1 - 2x_1 = \frac{1}{3} - \frac{2}{4!} + \frac{2}{5!} - \ldots \\
x_3 &= 1 - 3x_2 = \frac{3!}{4!} - \frac{3!}{5!} + \frac{3!}{6!} - \ldots \\
&\vdots \\
x_n &= 1 - nx_{n-1} = \frac{n!}{(n+1)!} - \frac{n!}{(n+2)!} + \frac{n!}{(n+3)!} - \ldots
\end{aligned}$$

This shows that

$$x_n = \frac{1}{n+1} - \frac{1}{(n+1)(n+2)} + \ldots \to 0 \quad \text{as} \quad n \to \infty.$$