Computer Arithmetic and Errors

While numerical result is an approximation, this can usually be as <u>accurate as needed</u>.

The necessary accuracy is, of course, determined by the application. To achieve high accuracy, very many separate operations must be carried out, but <u>computers</u> do them so rapidly without ever making mistakes that is no significant problem.

Actually, evaluating an analytical result to get the numerical answer for a specific application is subject to the *<u>same errors</u>*.

How can we define 'error' in a computation? In its simplest form, it is the difference between the exact answer True, say, and the computed answer, Approximate.

Hence, we can write,

$$ERROR = True - Approximate$$

Since we are usually interested in the magnitude or absolute value of the error we can also define

$$ABSOLUTE\_ERROR = |True - Approximate|$$

In practical calculations, it is important to obtain an upper bound on the error i.e. a number, Upper, such that,

$$|True - Approximate| \langle Upper(TOLERANCE)$$

Clearly, we would like *Upper* to be small!

In practice we are often more interested in so-called 'relative error' rather than absolute error and we define,

$$RELATIVE\_ERROR = \frac{|True - Approximate|}{|True|}$$

The percentage error is 100% times the relative error.

Hence, an 'error' of 10^{-5} may be a good or bad 'relative error' depending on the answer.

For example,

Answer	**=**	**1000**	**error**	**=**	**10^{-5}**	**very good**
Answer	**=**	**1**	**error**	**=**	**10^{-5}**	**good**
Answer	**=**	**10^{-5}**	**error**	**=**	**10^{-5}**	**very bad**

Example: Find the absolute error and relative error in the following cases.

- Let y = 3.141592(True) and
  ŷ =3.14(Approximate);

Then the absolute error is

$$E_y = |3.141592 - 3.14| = 0.001592$$

and the relative error is

$$R_y = \frac{0.001592}{3.141592} = 0.0057.$$

- Let w = 1000000 and ŵ =999996; then the absolute error is

$$E_w = |1000000 - 999996| = 4$$

and the relative error is

$$R_w = \frac{4}{1000000} = 0.000004.$$

- Let s= 0.00012 and ŝ = 0.000009; then the absolute error is

$$E_s = |0.000012 - 0.000009| = 0.000003$$

and the relative error is

$$R_s = \frac{0.000003}{0.000012} = 0.25$$

Floating point decimal precision

Decimal32 supports 7 decimal digits of significand and an exponent range of −95 to +96, i.e. $\pm0.000000\times10^{-95}$ to $\pm9.999999\times10^{96}$. (Equivalently, $\pm0000000\times10^{-101}$ to $\pm9999999\times10^{90}$.) Because the significand is not normalized (there is no implicit leading "1"), most values with less than 7 significant digits have multiple possible representations; $1\times10^{2}=0.1\times10^{3}=0.01\times10^{4}$, etc. Zero has 192 possible representations (384 when both signed zeros are included).

Decimal64 supports 16 decimal digits of significand and an exponent range of −383 to +384, i.e. $\pm0.000000000000000\times10^{-383}$ to $\pm9.999999999999999\times10^{384}$. (Equivalently, $\pm0000000000000000\times10^{-398}$ to $\pm9999999999999999\times10^{369}$.) In contrast, the corresponding binary format, which is the most commonly used type, has an approximate range of $\pm0.000000000000001\times10^{-308}$ to $\pm1.797693134862315\times10^{308}$. Because the significand is not normalized, most values with less than 16 significant digits have multiple possible representations; $1\times10^{2}=0.1\times10^{3}=0.01\times10^{4}$, etc. Zero has 768 possible representations (1536 if you include both signed zeros).

Decimal128 supports 34 decimal digits of significand and an exponent range of −6143 to +6144, i.e. $\pm0.000000000000000000000000000000000\times10^{-6143}$ to $\pm9.999999999999999999999999999999999\times10^{6144}$. (Equivalently, $\pm0000000000000000000000000000000000\times10^{-6176}$ to $\pm9999999999999999999999999999999999\times10^{6111}$.) Therefore, decimal128 has the greatest range of values compared with other IEEE basic floating point formats. Because the significand is not normalized, most values with less than 34 significant digits have multiple possible representations; $1\times10^{2}=0.1\times10^{3}=0.01\times10^{4}$, etc. Zero has 12288possible representations (24576 if you include both signed zeros).

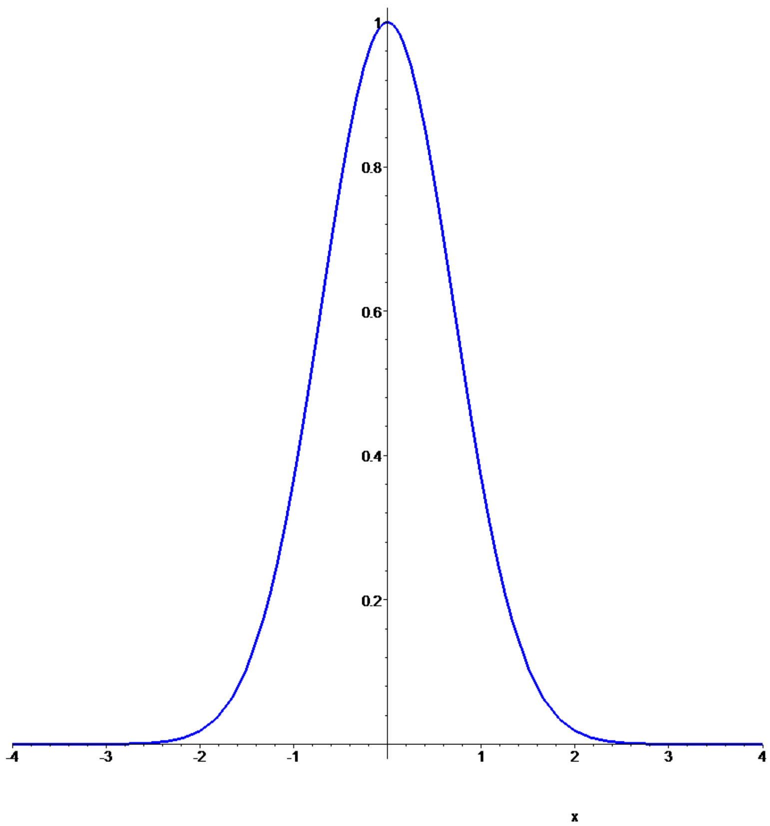
Numerical Integration

Simple Functions that do not have simple antiderivatives

The **normal distribution** is a very important functions in **statistics**. **Gaussian noise** is one of many ways in which this function is used in **engineering and science**. The normal distribution function is scaled form of the function $f(x)=e^{-x^2}$. The indefinite integral of this function **cannot be represented as a simple function.**

```
> int(exp(-x^2),x);
```

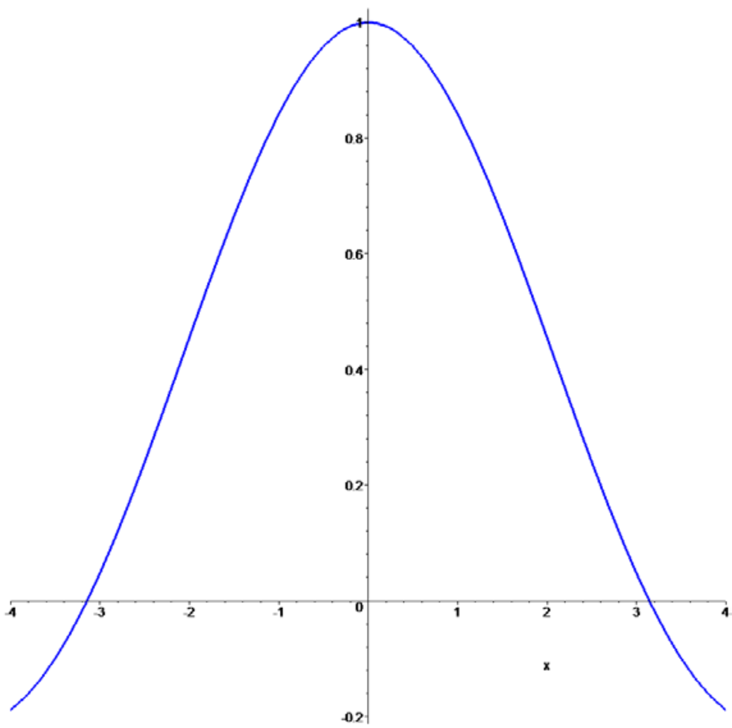
$$\frac{1}{2}\sqrt{\pi}\,\mathrm{erf}(x)$$

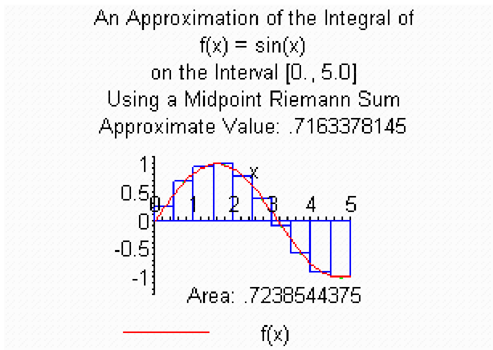


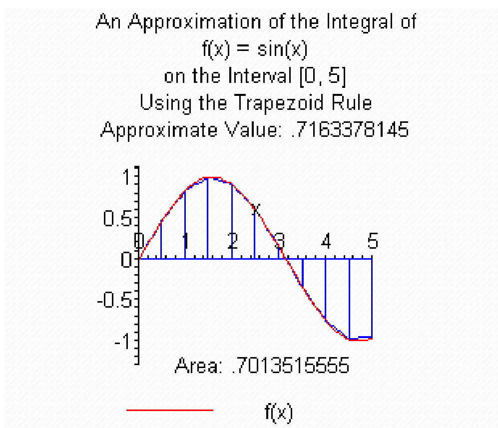
Another function that is important in optics and other applications, but does not have a simple antiderivative is, $f(x) = \sin(x)/x$.

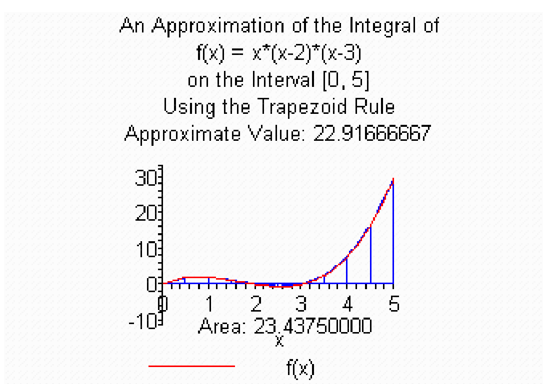
```
> int(sin(x)/x,x);
```

$$\mathrm{Si}(x)$$

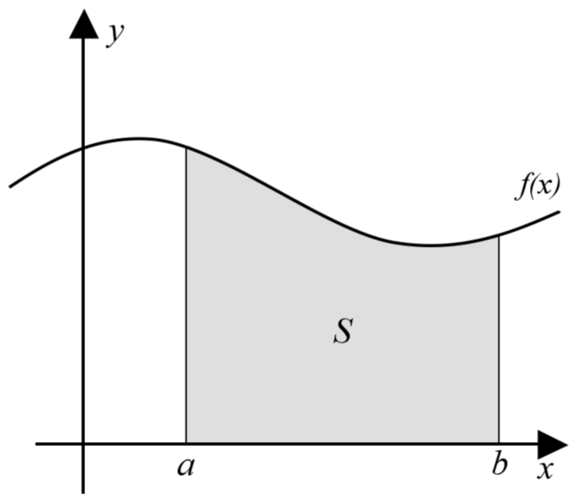



An Approximation of the Integral of
f(x) = sin(x)
on the Interval [0., 5.0]
Using a Midpoint Riemann Sum
Approximate Value: .7163378145
Area: .7238544375
f(x)


An Approximation of the Integral of
f(x) = sin(x)
on the Interval [0, 5]
Using the Trapezoid Rule
Approximate Value: .7163378145
Area: .7013515555
f(x)


An Approximation of the Integral of
f(x) = x*(x-2)*(x-3)
on the Interval [0, 5]
Using the Trapezoid Rule
Approximate Value: 22.91666667
Area: 23.43750000
f(x)

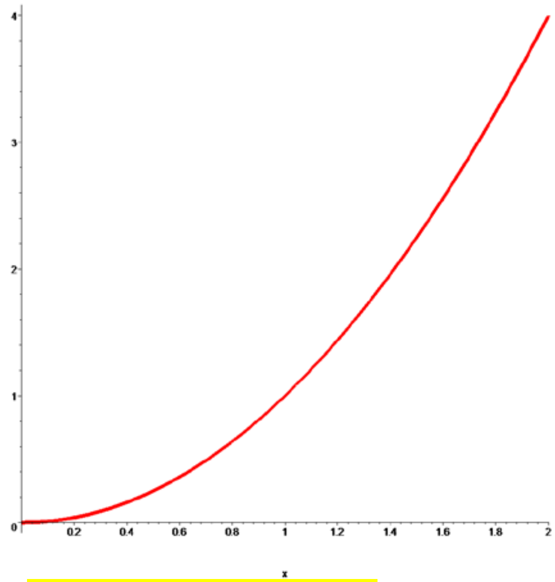
Theorem: ***If a function f(x) is continuous on a finite interval*** $a \le x \le b$ ***, then the definite integral of f(x) with respect to x from x=a to x=b exists.***



Theorem:(The Fundamental Theorem of Calculus) ***If f(x) is continuous on the interval*** $a \le x \le b$ ***, and F(x) is an antiderivative of f(x) then***

$$\int_a^b f(x)dx = F(b) - F(a)$$

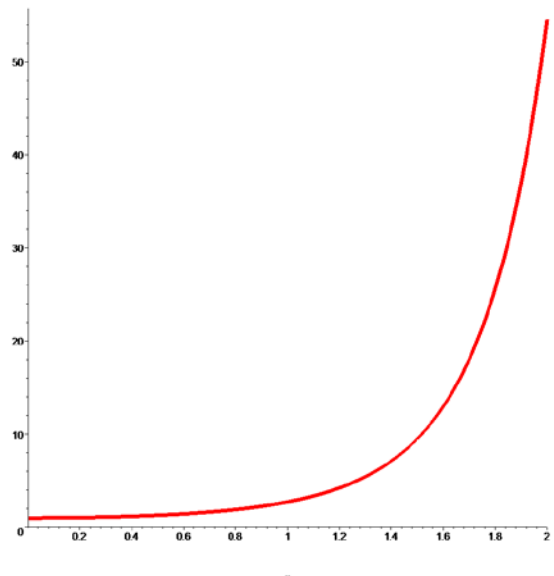
Evaluate $\int_0^2 x^2 dx$

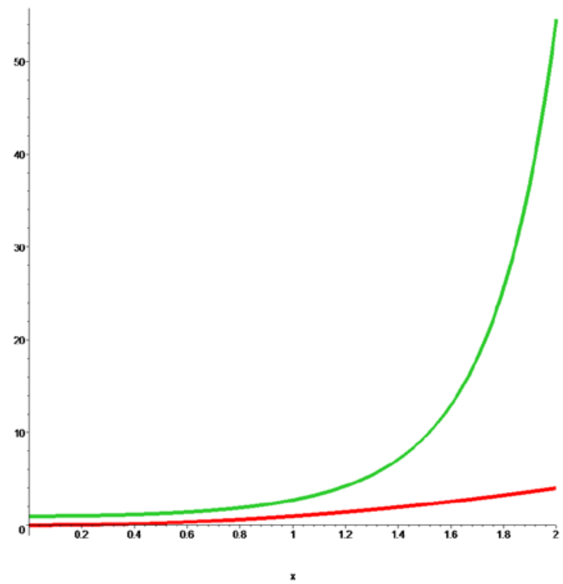


$$\int_0^2 x^2 dx = \left|\frac{x^3}{3}\right|_0^2 = \frac{8}{3}$$

Numerical integration rules are very important because even simple functions <u>may not have exact formulas for their</u> antiderivatives (indefinite integrals).Even when an exact formula for the <u>antiderivative does exist, it may be difficult to find.</u>

$$\int_0^2 e^{x^2}\,dx = ?$$





$$y = e^{x^2}, and \quad y = x^2$$

$$\int_0^2 e^{x^2}\,dx = -\frac{1}{2} I erf(2I)\sqrt{\pi}$$

```
> int(exp(x^2),x=0..2);
```

$$\frac{-1}{2}\, I\, \mathrm{erf}(\,2\,I\,)\,\sqrt{\pi}$$

In general, a numerical integration formula approximates a definite integral by a weighted sum of function values at points within the interval of integration has the form

$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} c_i f(x_i)$$

where the coefficients c_i depend on the particular method.

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x_i$$

$\Delta x_i = x_i - x_{i-1}$

We consider the most common numerical integration formulas that are based on equally spaced data points : these are known as *Newton-Cotes formulas.*

Subdivide the interval into n equal subintervals by n-1 points.

$$a = x_0 < x_1 < x_2 < .... < x_i .... < x_{n-1} < x_n = b$$

Evaluate $\int_0^2 x^2 dx$

$$\Delta x = \frac{b-a}{n} \qquad \Delta x = \frac{2-0}{n} = \frac{2}{n}$$

$$x_i = a + i\Delta x = 0 + i\frac{2}{n} = \frac{2i}{n}$$

$$\sum_{i=1}^{n} f(x_i)\Delta x = \sum_{i=1}^{n} f\left(\frac{2i}{n}\right)\left(\frac{2}{n}\right)$$

$$= \sum_{i=1}^{n}\left(\frac{2i}{n}\right)^2\left(\frac{2}{n}\right)$$

$$= \frac{8}{n^3}\sum_{i=1}^{n} i^2$$

$$= \left(\frac{8}{n^3}\right)\left(\frac{n(n+1)(2n+1)}{6}\right) = \frac{4(n+1)(2n+1)}{3n^2}$$

$$\int_0^2 x^2 dx = Lim_{n\to\infty}\frac{4(n+1)(2n+1)}{3n^2} = \frac{8}{3}$$

```
> int(x^2,x=0..2);
```

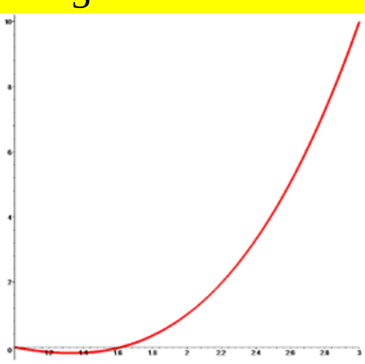
$$\frac{8}{3}$$

Evaluate $\int_1^3 (x^3 - 2x^2 + 1)dx$

$$\Delta x = \frac{3-1}{n} = \frac{2}{n} \qquad x_i = a + i\Delta x = 1 + i\frac{2}{n} = 1 + \frac{2i}{n}$$

$$\sum_{i=1}^{n} f(x_i)\Delta x = \sum_{i=1}^{n} f\left(1+\frac{2i}{n}\right)\left(\frac{2}{n}\right)$$
$$= \sum_{i=1}^{n} \left\{\left(1+\frac{2i}{n}\right)^3 - 2\left(1+\frac{2i}{n}\right)^2 + 1\right\}\left(\frac{2}{n}\right)$$
$$= \frac{16}{n^4}\sum_{i=1}^{n} i^3 + \frac{8}{n^3}\sum_{i=1}^{n} i^2 - \frac{4}{n^2}\sum_{i=1}^{n} i$$
$$= \left(\frac{16}{n^4}\right)\left(\frac{n^2(n+1)^2}{4}\right) + \left(\frac{8}{n^3}\right)\left(\frac{n(n+1)(2n+1)}{6}\right) - \left(\frac{4}{n^2}\right)\left(\frac{n(n+1)}{2}\right)$$

$$\int_1^3 (x^3 - 2x^2 + 1)dx = \lim_{n\to\infty}\left\{\frac{4(n+1)^2}{n^2} + \frac{4(n+1)(2n+1)}{3n^2} - \frac{2(n+1)}{n}\right\}$$
$$= 4 + \frac{8}{3} - 2 = \frac{14}{3}$$



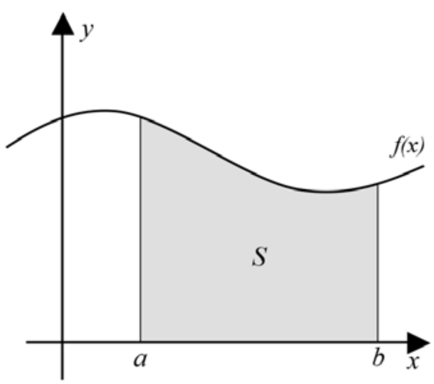
```
> int(x^3-2*x^2+1,x=1..3);
```

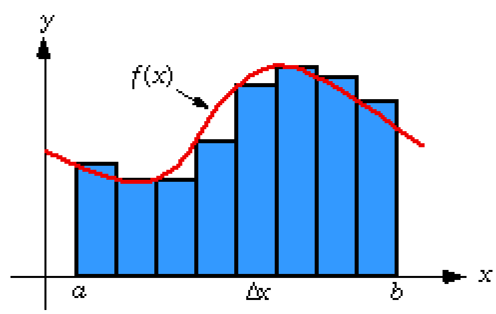
$$\frac{14}{3}$$

Riemann Sum

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i)\Delta x$$

where $\Delta x = \dfrac{b-a}{n}$





This method of integral approximation is known as a **Riemann sum**. There are **3 basic types** of Riemann sums:

1) The left or lower sum

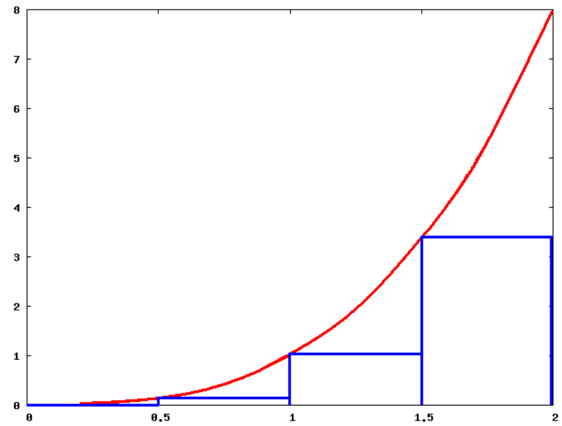
$$\int_a^b f(x)dx \approx \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$

2) The right or upper sum

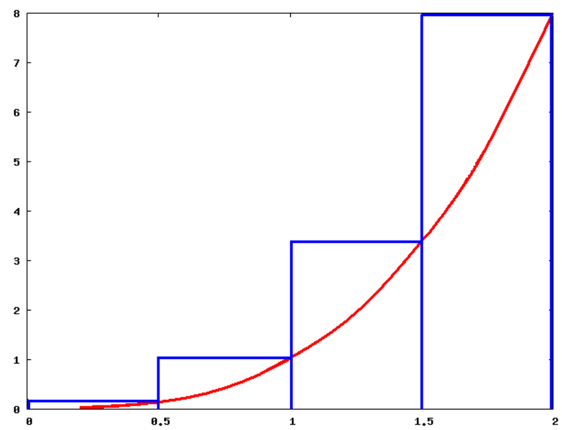
$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1})$$

3) The middle or midpoint sum

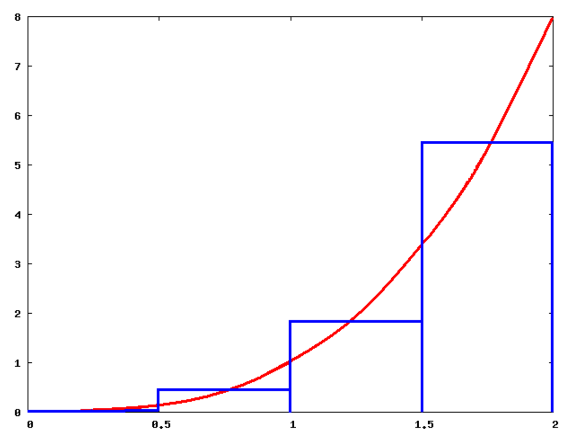
$$\int_a^b f(x)dx \approx \sum_{i=1}^{n} f\left(\frac{x_i + x_{i-1}}{2}\right)(x_i - x_{i-1})$$



LOWER

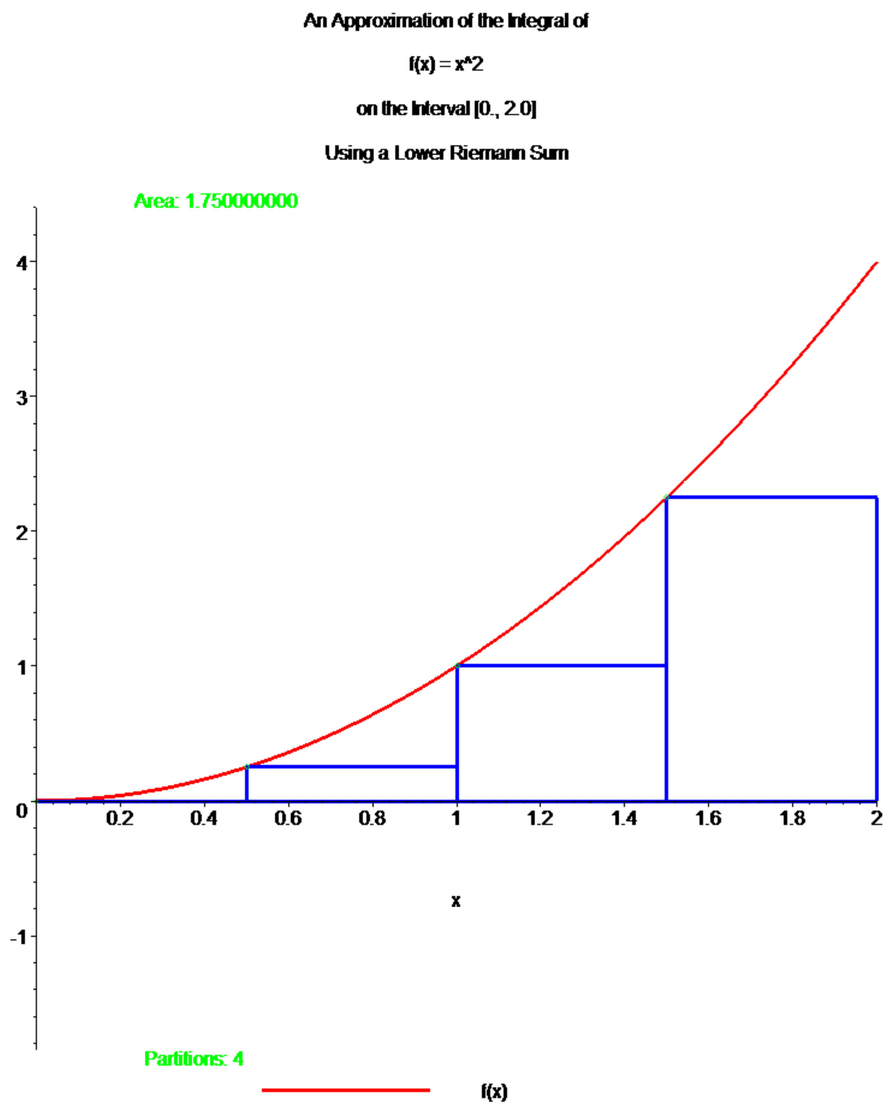


UPPER



MIDPOINT

```
> with(Student[Calculus1]):
> RiemannSum(x^2, x=0.0..2.0, method = lower,output=plot,partition=4);
```

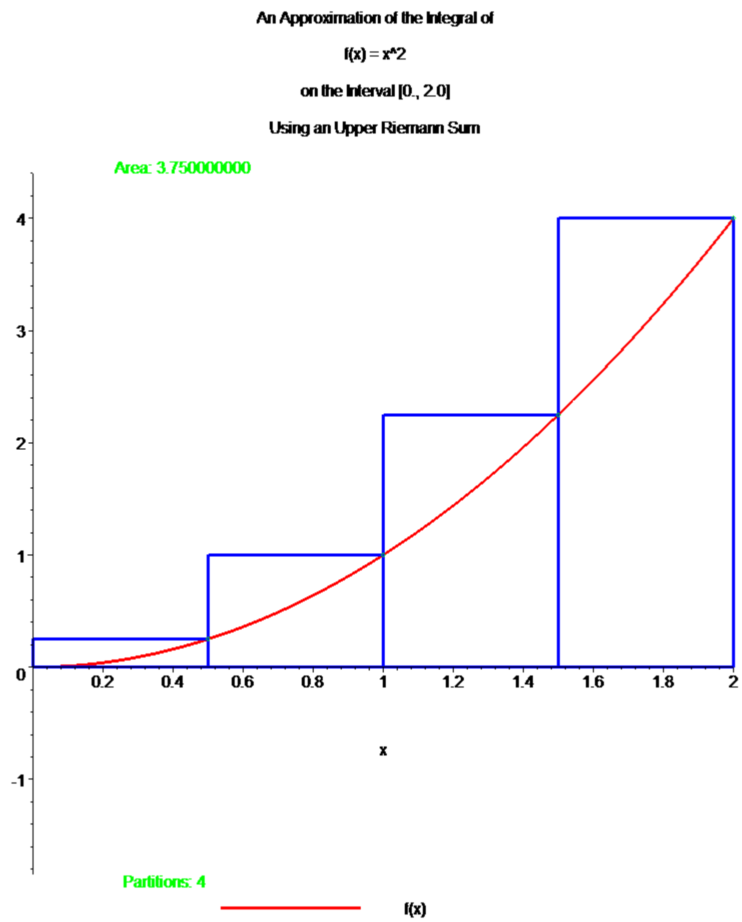


$$\int_0^2 x^2 dx \approx \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$

$$= 0^2(0.5-0) + 0.5^2(1-0.5) + 1^2(1.5-1) + 1.5^2(2-1.5)$$

$$= 1.75$$

```
> with(Student[Calculus1]):
> RiemannSum(x^2, x=0.0..2.0, method = upper,output=plot,partition=4);
```

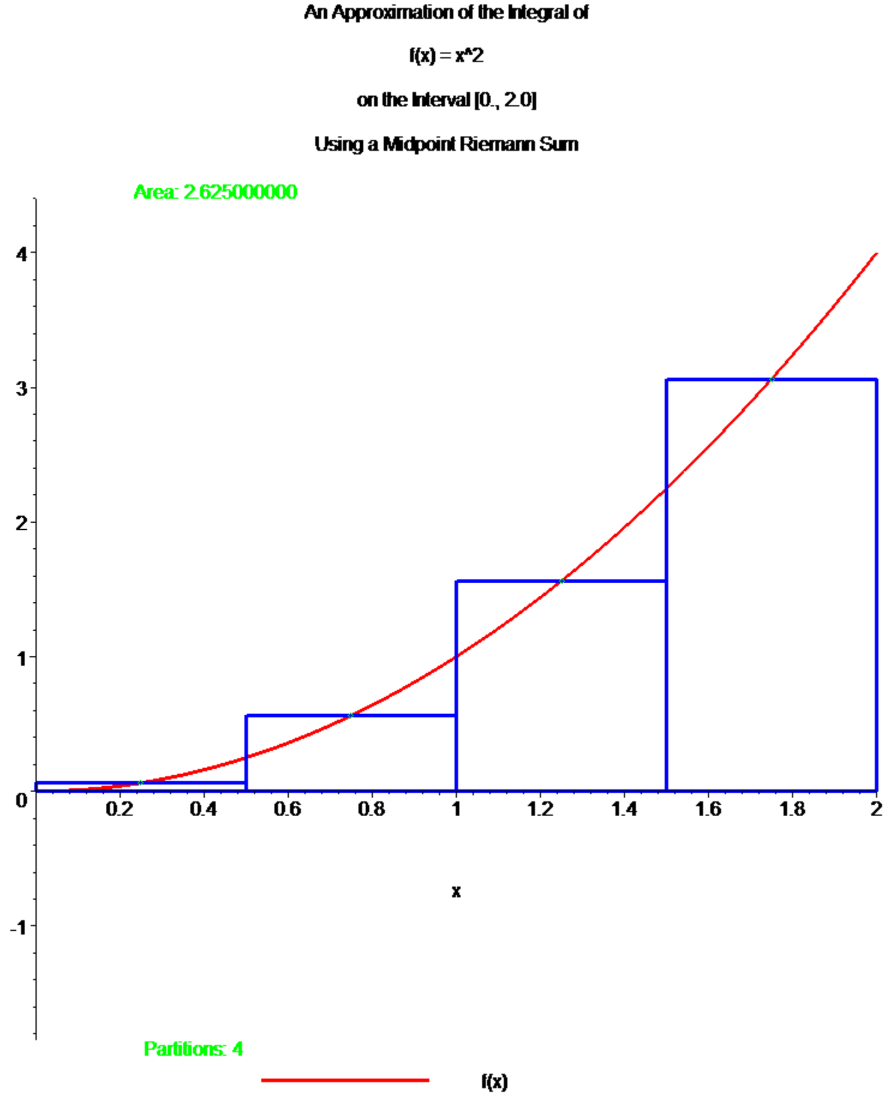


$$\int_0^2 x^2 dx \approx \sum_{i=1}^{n} f(x_i)(x_i - x_{i-1})$$

$$= 0.5^2(0.5-0) + 1^2(1-0.5) + 1.5^2(1.5-1) + 2^2(2-1.5)$$

$$= 3.75$$

```
> with(Student[Calculus1]):
> RiemannSum(x^2, x=0.0..2.0, method = midpoint,output=plot,partition=4);
```



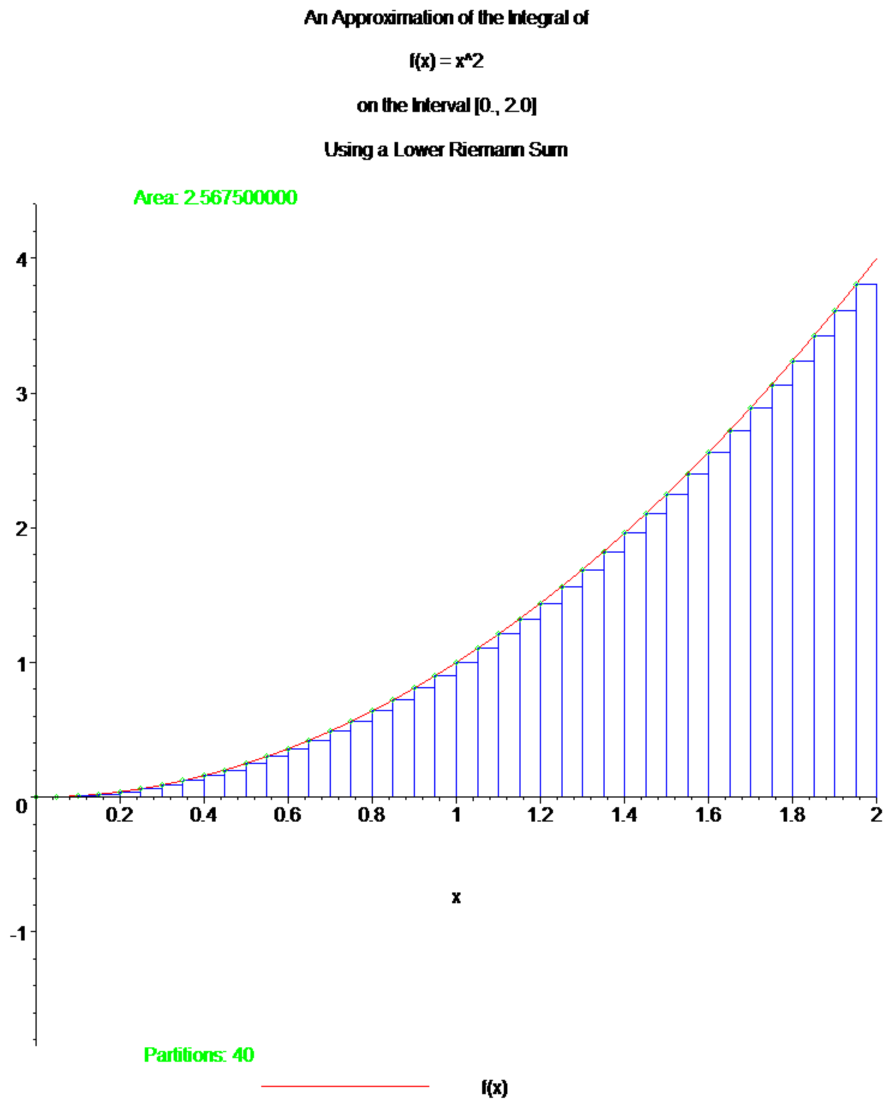
$$\int_0^2 x^2 dx \approx \sum_{i=1}^{n} f\left(\frac{x_i + x_{i-1}}{2}\right)(x_i - x_{i-1})$$

$$= 0.25^2(0.5-0) + 0.75^2(1-0.5)$$

$$+ 1.25^2(1.5-1) + 1.75^2(2-1.5)$$

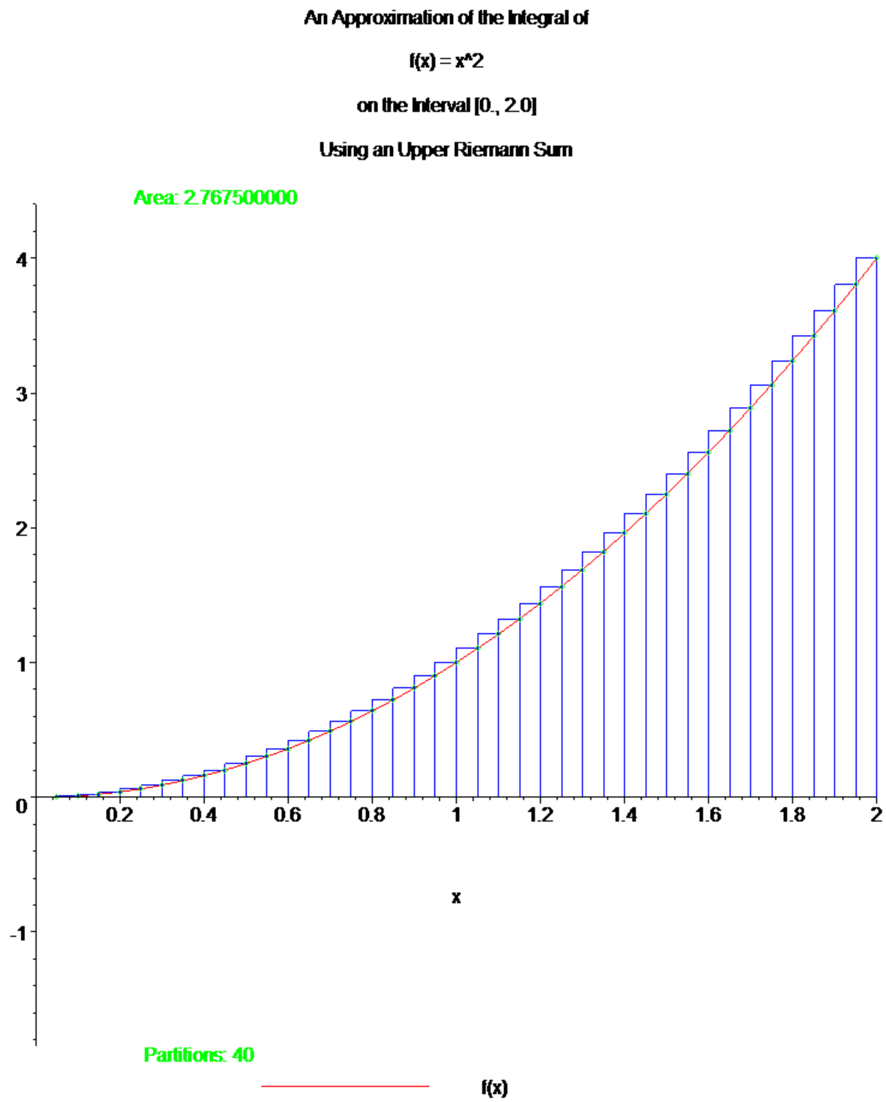
$$= 0.5(0.25^2 + 0.75^2 + 1.25^2 + 1.75^2) = 2.625$$

Lower Riemann Sum (Partitions=40)



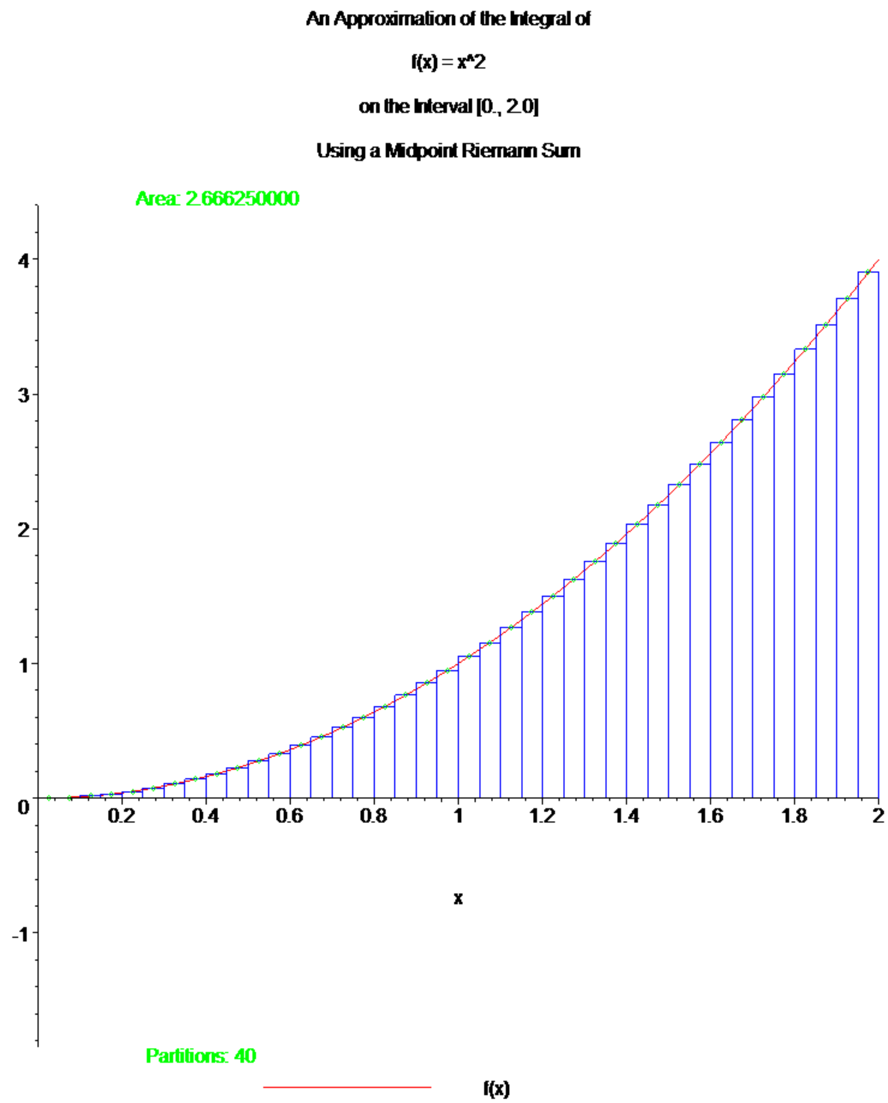
Area=2.56750000

Upper Riemann Sum (Partitions=40)



Area=2.76750000

Midpoint Riemann Sum (Partitions=40)



Area=2.66625000

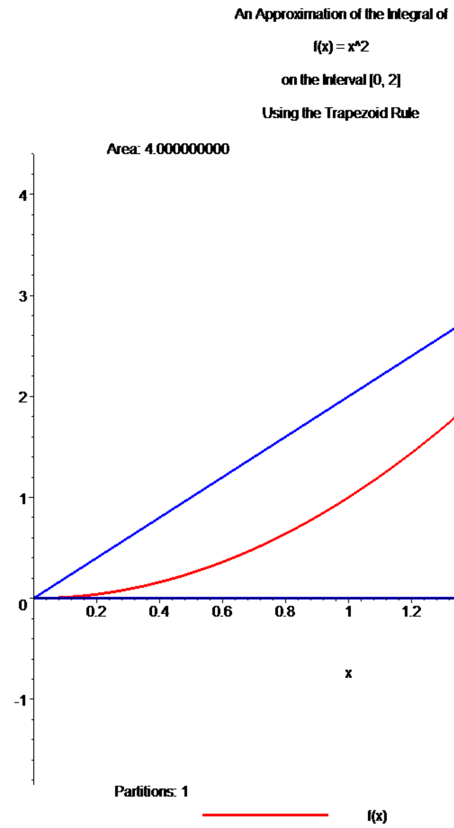
Trapezoidal Rule

One of the simplest ways to approximate the area under a curve is to approximate the curve by a straight line. The trapezoidal rule approximates the curve by the straight line that passes through the points $(a, f(a))$ **and** $(b, f(b))$**, the two ends of the interval of interest.**

We have $x_0 = a$, $x_1 = b$ and $h = b - a$

$$\int_a^b f(x)dx \approx \frac{h}{2}[f(x_0) + f(x_1)]$$

$$\int_a^b f(x)dx \approx \frac{b-a}{2n}[f(a) + 2f(x_1) + 2f(x_2) + \ldots + 2f(x_{n-1}) + f(b)]$$



The area of the i_{th} trapezoid is given by

$$\Delta x\left(\frac{f(x_i)+f(x_{i-1})}{2}\right)$$

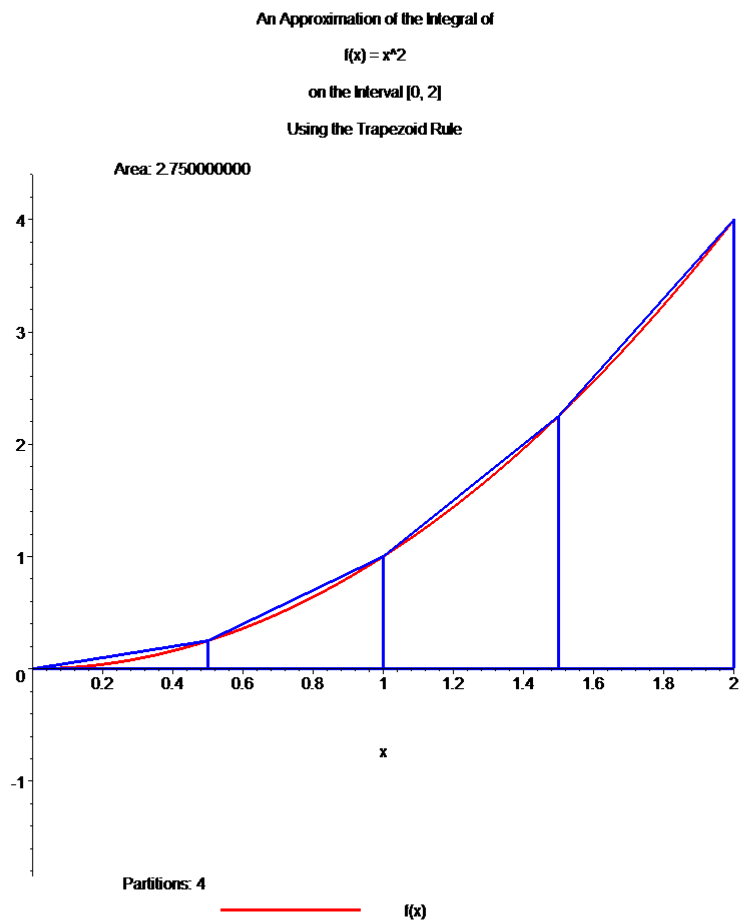
The sum of the areas of the trapezoids reduces to a simple formula,

$$\int_a^b f(x)dx \approx \frac{b-a}{2n}\left[f(a)+2f(x_1)+2f(x_2)+\ldots+2f(x_{n-1})+f(b)\right]$$

Example: **Evaluate $\int_0^2 x^2dx$ using Trapezoidal Rule with 4 and 10 partitions.**

Partitions:4 (n=4)

```
> ApproximateInt(x^2, x=0..2, method = trapezoid,output=plot,partition=4);
```

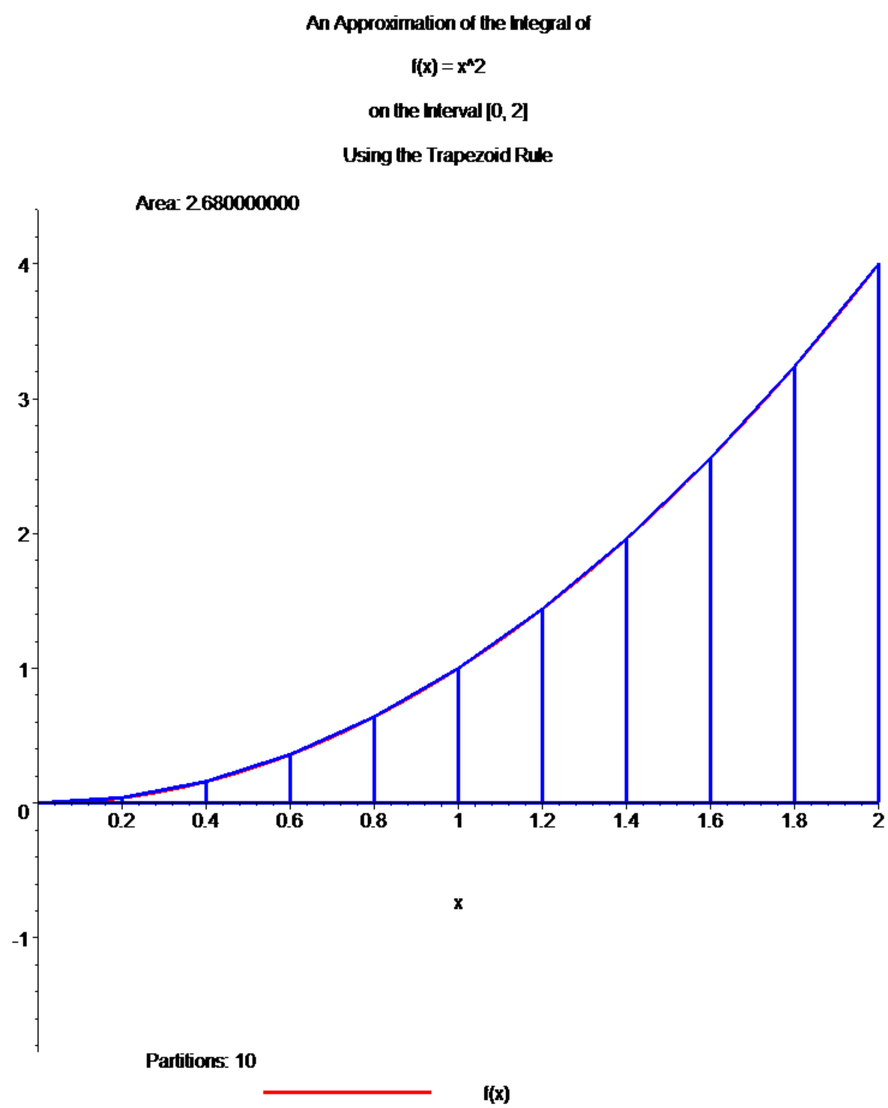


$$\int_a^b f(x)dx \approx \frac{b-a}{2n}[f(a)+2f(x_1)+2f(x_2)+...+2f(x_{n-1})+f(b)]$$

$$\int_0^2 x^2dx \approx \frac{2-0}{2(4)}\left[0+2(0.5^2)+2(1^2)+2(1.5^2)+(2^2)\right]$$

$$=0.25\left[0+2(0.5^2)+2(1^2)+2(1.5^2)+(2^2)\right]=2.75$$

Partitions:10



$$\int_0^2 x^2 dx \approx \frac{2-0}{2(10)}\left[\begin{array}{l}0+2(0.2^2+0.4^2+0.6^2+0.8^2+1^2+\\ 1.2^2+1.4^2+1.6^2+1.8^2)+(2^2)\end{array}\right]$$

$$=0.1[0+2(11.4)+(4)]=0.1[26.8]=2.68$$

Error Estimate For The Trapezoidal Rule

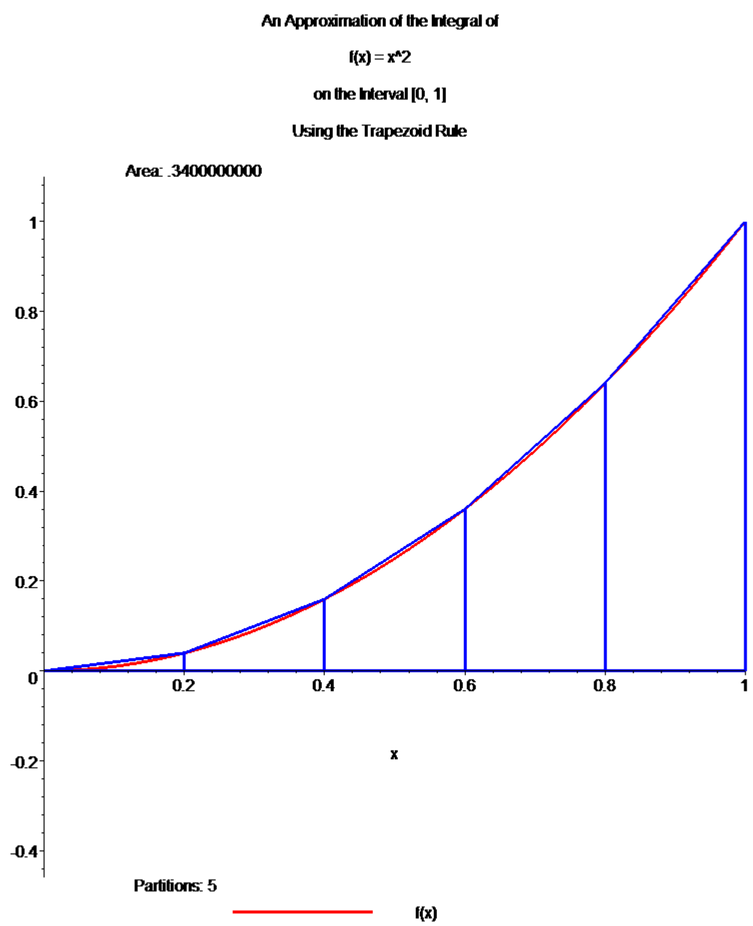
Let T_n be difference between $\int_a^b f(x)dx$ and trapezoid rule estimate for $\int_a^b f(x)dx$ with n intervals. Then T_n is the estimation error. If M_2 is the maximum value of $|f''(x)|$ on [a,b], then

$$|T_n| \le \frac{M_2(b-a)^3}{12n^2}$$

This estimate is <u>valid only when</u> $f''(x)$ is defined on all of [a,b].

Example: **Estimate $\int_0^1 x^2 dx$ using the trapezoidal rule with n=5 partitions (intervals):**

$$\Delta x = \frac{1-0}{5} = \frac{1}{5} \qquad x_0 = 0, x_1 = 1/5, \ldots, x_5 = 1$$



$$\int_0^1 x^2 dx \approx \frac{1}{2(5)}\left[0 + 2(1/25) + 2(4/25)2(9/25) + 2(16/25) + 1\right] = 0.34$$

Example:

Find a bound on the error in estimating $\int_0^2 x^2 dx$ using the trapezoidal rule with 5 partitions.

$f''(x) = 2$ So the maximum value of $|f''(x)|$ on [0, 2] is M_2=2.

Thus

$$|T_5| \le \frac{2(2-0)^3}{12*5^2} = \frac{16}{300} = 0.05333$$

$$\int_0^2 x^2 dx \approx \frac{1}{2(5)}[0 + 2(0.16) + 2(0.64) + 2(1.44) + 2(2.56) + 4] = 2.72$$

True Value: 8/3=2.666 Estimate = 2.72

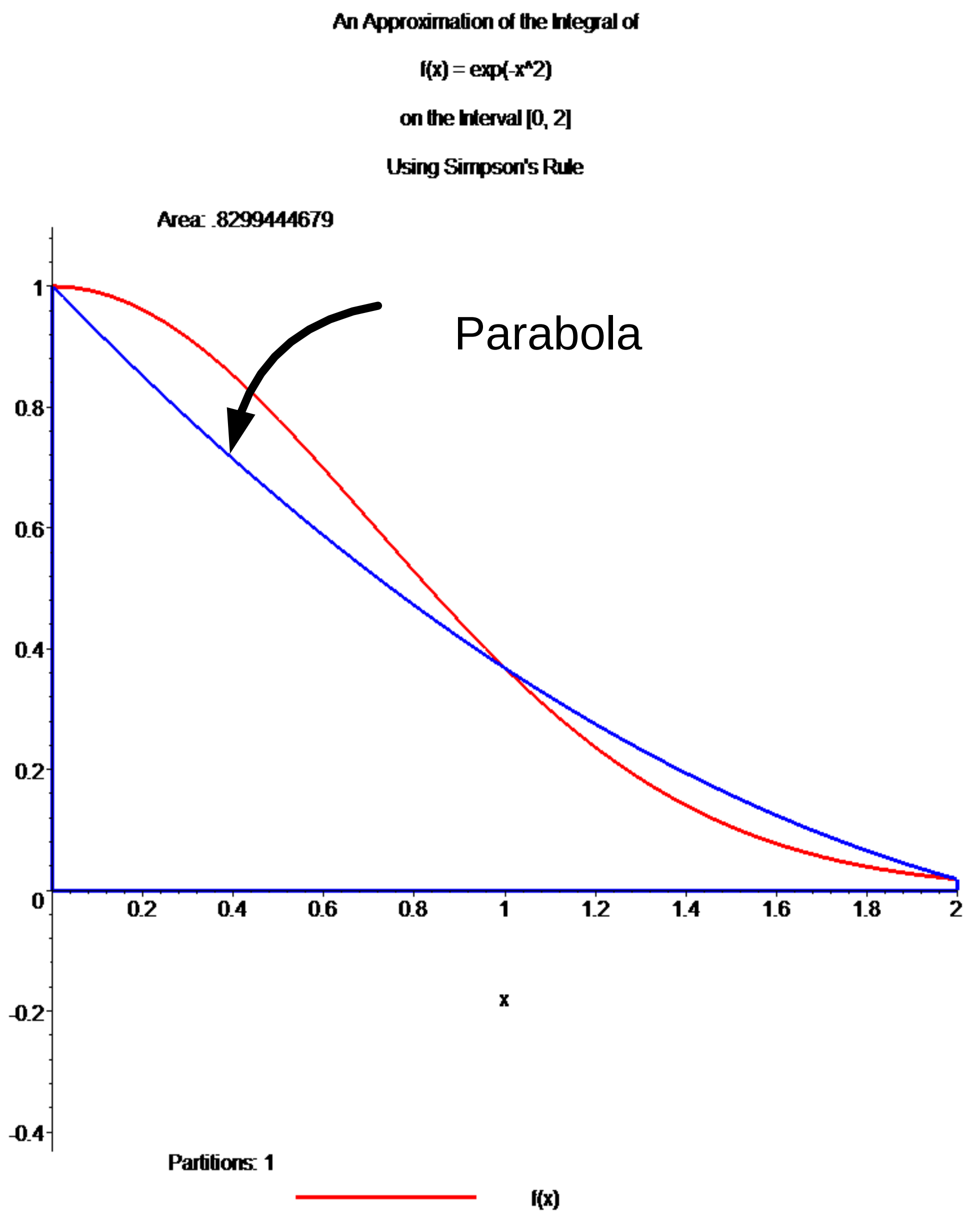
$$Absolute\_Error = |2.666 - 2.72| = 0.05333$$

Matlab (Trapezoidal Rule)

```
function Q = Trap( f, a, b, n)
h = (b-a)/n;
S = feval(f, a);
if n == 1
   Q = (S + feval(f, b))*h/2;
else
   for i = 1 : n-1
      x(i) = a + h*i;
      S = S + 2*feval(f, x(i));
   end
   S = S + feval(f, b);
   Q = h*S/2;
end
```

```
f=inline('x^2')
f =
     Inline function:
     f(x) = x^2
>> Q=Trap(f,0,2,5)
Q =
    2.7200
>> Q=Trap(f,0,2,10)
Q =
    2.6800
```

Simpson's Rule



***Partition**

The area under the parabola may be written in the form

$$\int_{x_{i-1}}^{x_{i+1}} (ax^2 + bx + c)dx = \frac{\Delta x}{3}\left[f(x_{i-1}) + 4f(x_i) + f(x_{i+1}) \right]$$

and this expression is free of a,b and c.

Consider the problem of finding the integral of $\int_0^2 e^{-x^2} dx$

The required values for applying Simpson's rule are

$x_0 = a = 0, x_1 = (b-a)/2 = (2-0)/2 = 1, x_2 = b = 2$ **which gives**

$$\int_0^2 e^{-x^2} dx \approx \frac{1}{3}\left[\exp(-0^2) + 4\exp(-1^2) + \exp(-2^2) \right] = 0.8299$$

Discussion

Simpson's rule is found by integrating the [Lagrange interpolating polynomial]() for f(x)

$$L(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_2-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

$$x_0 = a, x_1 = x_0 + h, x_2 = b, h = \frac{b-a}{2}$$

$$\int_a^b f(x)dx \approx \int_a^b L(x)dx = \frac{h}{3}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b) \right]$$

In general

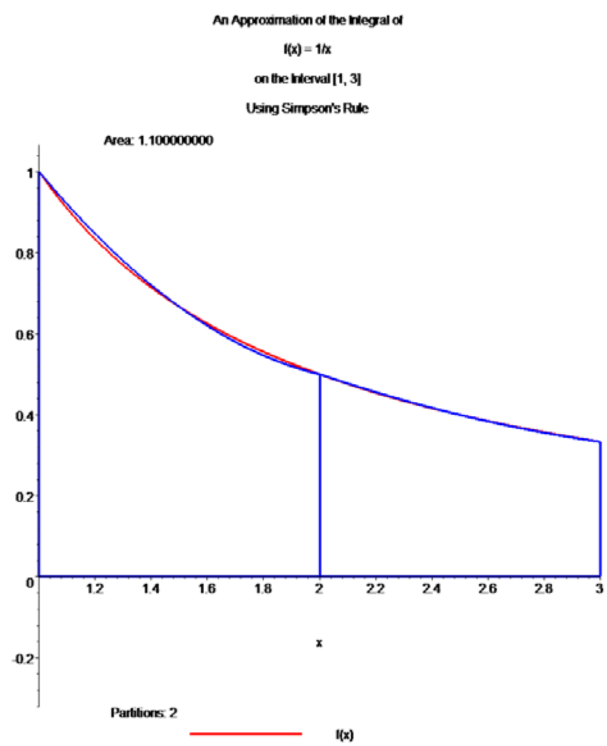
$$\int_a^b f(x)dx \approx \frac{(b-a)}{3n}[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \ldots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)]$$

<u>n must be even number for Simpson's rule</u>.

Example: **Estimate $\int_1^3 \frac{1}{x}\, dx$ using the Simpson's rule with n=4 intervals.**

$$\Delta x = \frac{3-1}{4} = \frac{1}{2} \qquad x_0 = 1, x_1 = 3/2, \ldots, x_4 = 3$$

$$\int_1^3 1/x \, dx \approx \frac{3-1}{3(4)}\left[1 + 4(2/3) + 2(1/2) + 4(2/5) + 1/3\right] = 1.1$$



$$\int_1^3 1/x \, dx = \left|\ln x\right|_1^3 = \ln(3) = 1.09861..$$

Simpson's Rule Error Estimate

$$\left|S_n\right| \le \frac{M_4(b-a)^5}{180n^4}$$

M_4 is the maximum value of $\left|f^{(4)}(x)\right|$ on [a,b]

Find a bound on the error in estimating $\int_1^3 \frac{1}{x}\,dx$ using the Simpson's rule with n=4 intervals.

$f^{(4)}(x) = 24x^{-5}$ So the maximum value of $\left|f^{(4)}(x)\right|$ on [1,3] is $M_4=24$.

Thus

$$\left|S_4\right| \le \frac{24(3-1)^5}{180(4)^2} = 0.01666$$

Exact error:1.1-ln(3)=0.00139

Matlab (Simpson Rule)

```
%metodod de simpson
function Q = Simp( f, a, b, n)
h = (b-a)/n;
S = feval(f, a);
if rem(n,2) != 0
    error('n must be even for Simpson method')
    return
end
if n == 2
    Q = (S + 4*feval(f,a+h) + feval(f, b))*h/3;
else
   for i = 1 : 2 : n-1
      x(i) = a + h*i;
      S = S + 4*feval(f, x(i));
   end
   for i = 2 : 2 : n-2
      x(i) = a + h*i;
      S = S + 2*feval(f, x(i));
   end
S = S + feval(f, b);
Q = h*S/3;
end
```

$$\int_1^3 \frac{1}{x}\,dx \approx \frac{3-1}{3(4)}\left[1+4(2/3)+2(1/2)+4(2/5)+1/3\right]=1.1$$

```
f=inline('1/x')
f =
     Inline function:
     f(x) = 1/x
>> Q=Simp(f,1,3,4)
Q =
    1.1000
```

Length of a Curve

The arc length of the f(x) over the interval $a \le x \le b$ ***is,***

$$Length = \int_a^b \sqrt{1+(f^{'}(x)^2)}dx$$

Surface Area

The solid of revolution obtained by rotating the region under the curve f(x), where $a \le x \le b$ ***,about the axis has surface area given by***

$$SurfaceArea = 2\pi\int_a^b f(x)\sqrt{1+(f^{'}(x)^2)}dx$$

Other Methods

Romberg Integration
Gaussian quadrature