KTH Computer Science
and Communication

Verifying Numerical Convergence Rates

1 Order of accuracy

We consider a numerical approximation of an exact value u. The approximation depends on a small parameter h, such as the grid size or time step, and we denote it by $\tilde{u}_h$. If the numerical method is of order p, we mean that there is a number C independent of h such that

$$|\tilde{u}_h - u| \leq Ch^p, \tag{1}$$

at least for sufficiently small h. We also say that the convergence rate of the method is h^p. (The number C typically depends on the exact solution.) Often the error $u - \tilde{u}_h$ depends smoothly on h. Then

$$\tilde{u}_h - u = Ch^p + O\left(h^{p+1}\right). \tag{2}$$

We will assume this henceforth.

Example 1 *In the* trapezoidal *rule we approximate the exact integral*

$$u = \int_a^b f(x)dx,$$

by a sum

$$\tilde{u}_h = \frac{h}{2}f(a) + h\sum_{j=1}^{N-1} f(a+jh) + \frac{h}{2}f(b), \qquad h = \frac{b-a}{N}.$$

For sufficiently smooth functions $f(x)$ this is a second order method and $\tilde{u}_h - u = Ch^2 + O(h^3)$.

2 Determining the order of accuracy

We are often faced with the problem of how to determine the order p given a sequence of approximations $\tilde{u}_{h_1}, \tilde{u}_{h_2}, \ldots$ This is can be a good check that a method is correctly implemented (if p is known) and also a way to get a feeling for the credibility of an approximation $\tilde{u}_h$ (high p means high credibility). We can either be in the situation that the exact value u is known, or, more commonly, that u is unknown.

2.1 Known u

If the exact value u is known, it is quite obvious how to do this. Then we just check the sequence

$$\log|\tilde{u}_h - u| = \log|C| + p\log h + O(h),$$

for h_1, h_2, $\ldots$ and fit it to a linear function of $\log h$ to approximate p. A quick way to do this is to plot $|\tilde{u}_h - u|$ as a function of h in a loglog plot and determine the slope of the line that

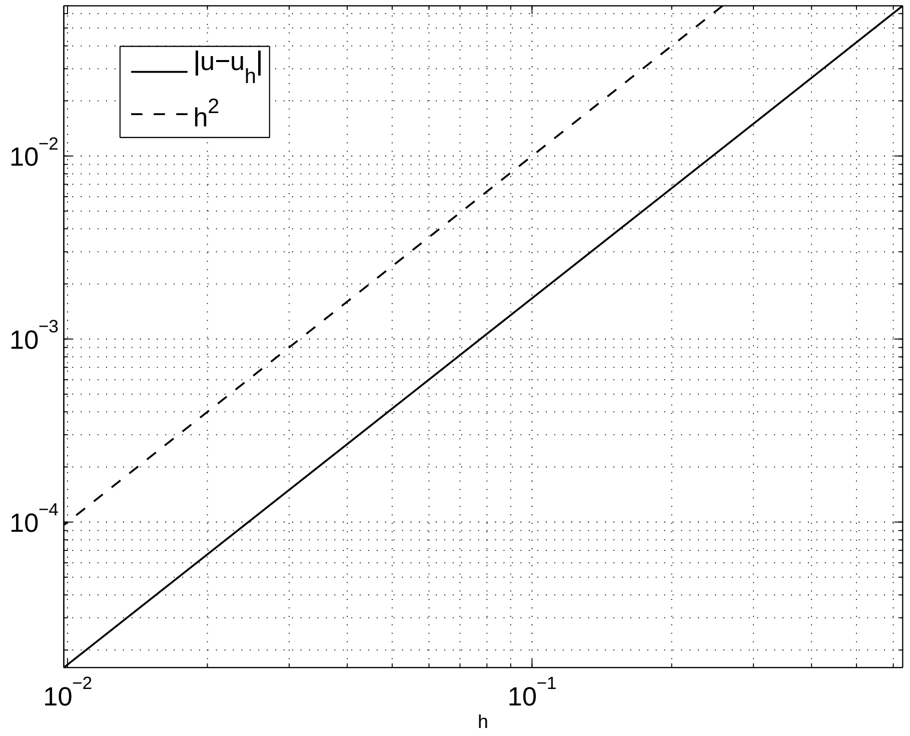


Figure 1. Error in trapezoidal rule for $f(x) = \sin(x)$. The dashed line is h^2 which indicates the slope for a second order method.

appears. The standard way to get a precise number for p is to halve the parameter h and look at the ratios of the errors $u - \tilde{u}_h$ and $u - \tilde{u}_{h/2}$,

$$\frac{\tilde{u}_h - u}{\tilde{u}_{h/2} - u} = \frac{Ch^p + O(h^{p+1})}{C(h/2)^p + O((h/2)^{p+1})} = 2^p + O(h).$$

Hence

$$\log_2 \left| \frac{\tilde{u}_h - u}{\tilde{u}_{h/2} - u} \right| = p + O(h).$$

Example 2 *The exact integral of* $\sin(x)$ *over* $[0, \pi]$ *equals two. Computing* $\tilde{u}_h$ *with the trapezoidal rule and plotting* $|\tilde{u}_h - 2|$ *in a loglog plot we get the result shown in Figure 1.*

2.2 Unknown u

When u is not known there are two main approaches. The first one is to compute a numerical reference solution with a very small h and then proceed as in the case of a known u. This can be quite an expensive strategy if $\tilde{u}_h$ is costly to compute. Using p to gauge the credibility of a $\tilde{u}_h$ is also less relevant when we already have a good reference solution.

The second approach is to look at ratios of differences between $\tilde{u}_h$ computed for different h. Most commonly we compare solutions where h is halved succesively. Then we get

$$\frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}} = \frac{Ch^p - C(h/2)^p + O(h^{p+1})}{C(h/2)^p - C(h/4)^p + O(h^{p+1})} = \frac{1 - 2^{-p} + O(h)}{2^{-p} - 2^{-2p} + O(h)} = 2^p + O(h). \tag{3}$$

h	$\tilde{u}_h$	$\tilde{u}_h - \tilde{u}_{h/2}$	$\frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$	$\log_2 \frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$
$\pi/5$	1.933765598092805	-0.049757939416650	4.024930251575880	2.008963782835339
$\pi/10$	1.983523537509455	-0.012362435199260	4.006184396966857	2.002228827158397
$\pi/20$	1.995885972708715	-0.003085837788350	4.001543117204195	2.000556454557076
$\pi/40$	1.998971810497066	-0.000771161948770	4.000385593360853	2.000139066704584
$\pi/80$	1.999742972445836	-0.000192771904301	4.000096386716427	2.000034763740606
$\pi/160$	1.999935744350136	-0.000048191814813		
$\pi/320$	1.999983936164949			

Table 1. Table of values for the trapezoidal rule for $f(x) = \sin(x)$. The last column is the final approximation of the order of accuracy p.

Hence, after computing $\tilde{u}_h$ for h, $h/2$ and $h/4$ we can evaluate the expression above and get an estimate of p. We can do it similarly for other grid sizes, e.g. h, αh, $\alpha^2 h$ gives

$$\frac{\tilde{u}_h - \tilde{u}_{\alpha h}}{\tilde{u}_{\alpha h} - \tilde{u}_{\alpha^2 h}} = \frac{Ch^p - C(\alpha h)^p + O(h^{p+1})}{C(\alpha h)^p - C(\alpha^2 h)^p + O(h^{p+1})} = \frac{1 - \alpha^p + O(h)}{\alpha^p - \alpha^{2p} + O(h)} = \alpha^{-p} + O(h). \tag{4}$$

Example 3 *Consider again Example 2. If the exact integral value was not known we would look at the values computed by the trapezoidal rule and check the ratios of differences as above. The result is summarized in Table 1.*

3 Asymptotic region

We note that the estimates of p in all the methods above gets better as $h \to 0$ because of the $O(h)$ term. (The precise value is only given in the limit $h \to 0$.) We say that the method is in its asymptotic region of accuracy when h is small enough to give a good estimate of p — then the $O(h^{p+1})$ term in (2) is significantly smaller than Ch^p. This required size of h can, however, be quite different for different problems. To verify that we are indeed in the asymptotic region, it can be valuable to make the estimate of p for several different h and check that we get approximately the same value. Usually one therefore computes $\tilde{u}_h$ not just for three values of h, but for a longer sequence, $h, h/2, h/4, h/8, h/16, \ldots$ and compares the corresponding ratios,

$$\frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}, \quad \frac{\tilde{u}_{h/2} - \tilde{u}_{h/4}}{\tilde{u}_{h/4} - \tilde{u}_{h/8}}, \quad \frac{\tilde{u}_{h/4} - \tilde{u}_{h/8}}{\tilde{u}_{h/8} - \tilde{u}_{h/16}}, \quad \ldots$$

Similarly, if u is known one considers $u - \tilde{u}_h$ for several decreasing values of h when fitting the line.

Example 4 *If we perform the same experiments as in Example 2 and Example 3 above, but with $f(x) = \sin(31x)$ the constant C will be much bigger, meaning that the asymptotic region is shifted to smaller h. The results are shown in Figure 2 and Table 2. It is not until $h < \pi/40 \approx 10^{-1}$ that the numbers start to look reasonable. The general size of the error is also much larger than in Figure 2 because of the bigger C.*

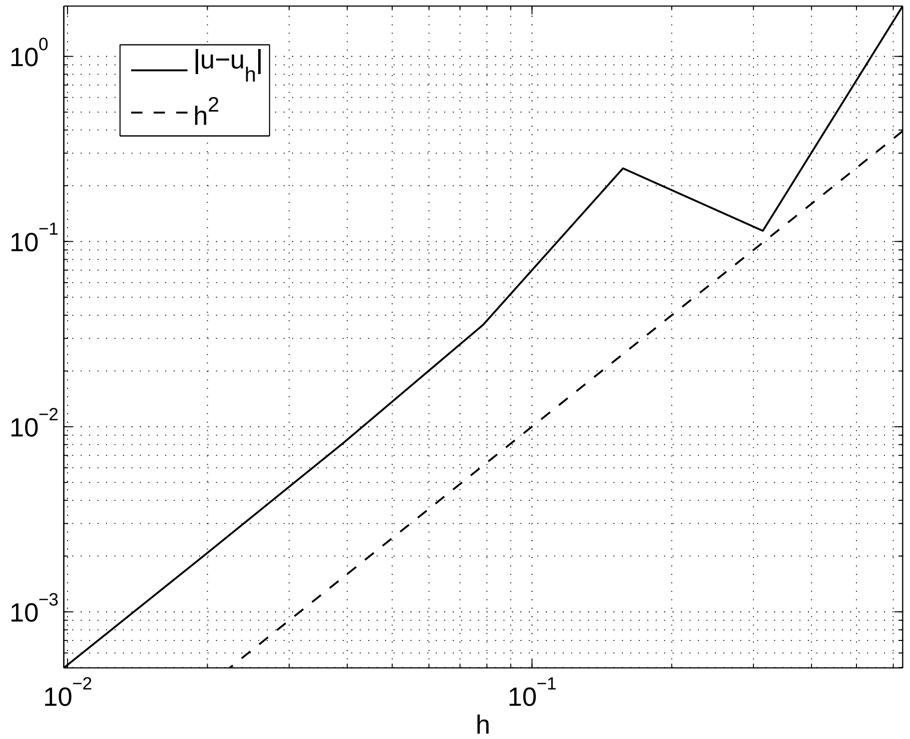
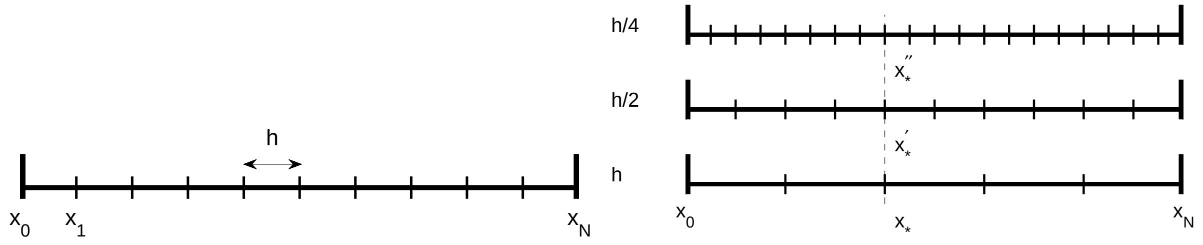


Figure 2. Error in trapezoidal rule for $f(x) = \sin(31x)$. The dashed line is h^2 which indicates the slope for a second order method.

h	$\tilde{u}_h$	$\tilde{u}_h - \tilde{u}_{h/2}$	$\frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$	$\log_2 \frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$
$\pi/5$	1.933765598092808	1.983523537509458	14.784906442999516	3.886053209184444
$\pi/10$	-0.049757939416650	0.134158680351247	-0.630173999781565	—
$\pi/20$	-0.183916619767896	-0.212891487744257	7.778391902691306	2.959471924644287
$\pi/40$	0.028974867976361	-0.027369601635860	4.437830912882666	2.149854700028653
$\pi/80$	0.056344469612220	-0.006167337641551	4.096338487974619	2.034334932805155
$\pi/160$	0.062511807253771	-0.001505573247830		
$\pi/320$	0.064017380501601			

Table 2. Table of values for the trapezoidal rule for $f(x) = \sin(31x)$. The last column is the final approximation of the order of accuracy p.



(a) (b)

Figure 3. Pointwise approximations.

4 Grid functions

When solving differential equations the numerical solution in question is often a grid function u_j which approximates a continuous function on a grid $\{x_j\}$. We assume that the grid is uniform, with $x_j = x_0 + jh$ for some x_0 and fixed h, cf. Figure 3a. The grid size h is restricted to values such that the grid fits the boundary, typically $h = d/N$ for a fixed domain size d and integer N of our choice. We will write $u_j(h)$ to indicate the dependence on h. To check convergence rates for these problems it is very important that we compare with the same thing when we change h. This can be a bit tricky.

4.1 Pointwise values

In a *finite difference* scheme we would approximate pointwise values,

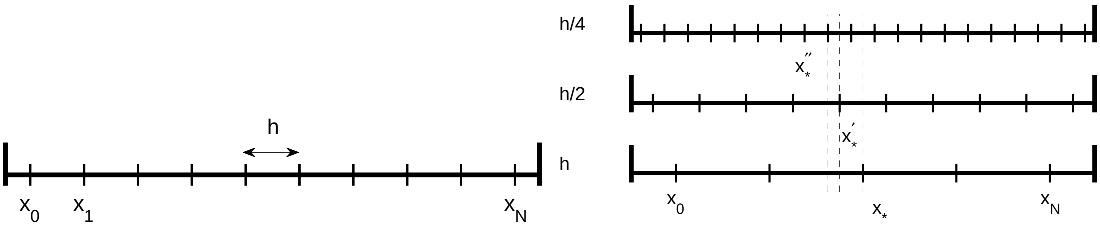
$$u_j(h) = u(x_j) + C(x_j)h^p + O(h^{p+1}),$$

where C now depends on the spatial location x_j. Suppose we want to check pointwise convergence. When computing the ratios of differences in (3) the solutions for different grids must be compared in exactly the same points. Let x_*, x'_* and x''_* be the grid points where the solution is compared for the grid sizes h, h' and h'', respectively. Suppose $x_* = x_0 + j_*h$. If we halve h we must then precisely double j_* to stay at the same grid point. Hence, if $h' = h/2$ and $h'' = h'/2 = h/4$, then $x'_* = x_0 + 2j_*h/2 = x_*$ and $x''_* = x_0 + 4j_*h/4 = x_*$, see Figure 3b. We obtain

$$\frac{u_{j_*}(h) - u_{2j_*}(h/2)}{u_{2j_*}(h/2) - u_{4j_*}(h/4)} = \frac{u(x_*) + C(x_*)h^p - u(x'_*) - C(x'_*)\left(\frac{h}{2}\right)^p + O(h^{p+1})}{u(x'_*) + C(x'_*)\left(\frac{h}{2}\right)^p - u(x''_*) - C(x''_*)\left(\frac{h}{4}\right)^p + O(h^{p+1})} = 2^p + O(h),$$

as before in Section 2.2. However, if we are not careful and the grid points used are just one index off, an error will be introduced which completely ruins the estimate.

Example 5 *To apply Neumann boundary conditions one often shifts the entire grid by half a cell, giving $x_0 = h/2$ as in Figure 4a. Hence, x_0 now depends on h. As before we let the measuring point be $x_* = x_0 + j_*h = h/2 + j_*h$. Then if we do the same thing here, halving h and doubling j, we get three slightly different grid points where the solutions are compared: x_*, $x'_* = h/4 + 2j_*h/2 = x_* - h/4$ and $x''_* = h/8 + 4j_*/4 = x_* - 3h/8$ (see Figure 4b). This gives an*



(a) (b)

Figure 4. Pointwise approximations, shifted grid.

error in the ratio of differences which prevents it from predicting the convergence rate:

$$\begin{aligned}\frac{u_{j_*+1}(h)-u_{2j_*}(h/2)}{u_{2j_*}(h/2)-u_{4j_*}(h/4)} &= \frac{u(x_*)+C(x_*)h^p-u(x'_*)-C(x'_*)\left(\frac{h}{2}\right)^p+O(h^{p+1})}{u(x'_*)+C(x'_*)\left(\frac{h}{2}\right)^p-u(x''_*)-C(x''_*)\left(\frac{h}{4}\right)^p+O(h^{p+1})}\\ &= \frac{u(x_*)-u(x'_*)+C(x_*)h^p(1-2^{-p})+O(h^{p+1})}{u(x'_*)-u(x''_*)+C(x_*)h^p(2^{-p}-2^{-2p})+O(h^{p+1})}\\ &= \frac{(x_*-x'_*)u_x(x_*)+O(h^2)}{(x'_*-x''_*)u_x(x_*)+O(h^2)}=2+O(h).\end{aligned}$$

Hence, regardless of the actual order p, the estimate would just indicate first order convergence.

Example 6 *In time stepping methods for ODEs we often want to check convergence at a fixed time T. The time step Δt must then be chosen such that T is exactly a multiple of Δt. Otherwise we get the same problem as in Example 5, and ratios of differences will always indicate first order convergence. One should therefore avoid setting Δt in the code, but rather set the number of time steps and compute Δt from this.*

One particular pitfall is the common practice of doubling the number of unknowns N in a problem, rather than halving the grid size h. Often $h = d/N$ and the two approaches are equivalent. However, depending on boundary conditions we can also have for instance $h = d/(N+1)$ or $h = d/(N-1)$. Then if N is doubled, we get the wrong order of convergence from our tests.

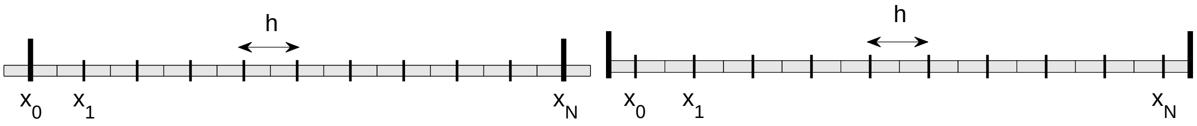
Remark 1 *Using interpolation is one way to avoid the problems associated with choosing the right grid sizes. Then we can evaluate any grid function at an arbitrary point, and these values can then readily be compared. However, it is important that the order of interpolation is at least as high as the order of the method we are studying. Otherwise the interpolation error will dominate.*

4.2 Local averages

In a *finite volume* scheme we would approximate local averages over grid cells $[x_j - h/2, x_j + h/2]$,

$$u_j(h) = \frac{1}{h}\int_{x_j-h/2}^{x_j+h/2} u(x)dx + C(x_j)h^p + O(h^{p+1}).$$

See Figure 5. In this case also the exact value that we compare with changes as we refine the grid. We can deal with this in two ways.



(a)

(b)

Figure 5. Local averages.

First, if $p \leq 2$ we can use the fact that the value in the mid point of the cell is a second order approximation of the average when the function is smooth,

$$\frac{1}{h}\int_{x_j-h/2}^{x_j+h/2} u(x)dx = u(x_j) + C'(x_j)h^2 + O(h^3).$$

and therefore

$$u_j(h) = u(x_j) + C(x_j)h^p + O(h^{p+1}),$$

possibly with a different $C(x)$. Hence, this takes us back to the same considerations as for pointwise values in Section 4.1. Note that also here the grid can be shifted, see Figure 5b.

Second, when $p > 2$ we must really compare local averages instead of pointwise values. This means that $u_j(h)$ must be compared with the average of two computed values when h is halved, $\frac{1}{2}(u_{j'}(h/2) + u_{j'}(h/2))$ for some j', and four values when h is halved again.

In these comparisons it is important then that we consider the same interval in every grid. For instance, let x_*, x_*' and x_*'' be the left edges of three cells in the grids with cell sizes h, $h/2$ and $h/4$. Suppose we have the shifted grid in Figure 5b and that $x_* = x_{j_*} - h/2 = h/2 + j_*h - h/2 = j_*h$. As in the pointwise case we can double the index, taking $2j_*$ and $4j_*$, such that all three points are equal $x_* = x_*' = x_*''$. Our first value I_h for grid size h is then

$$I_h = u_{j_*}(h) = \frac{1}{h}\int_{x_*}^{x_*+h} u(x)dx + C(x_* + h/2)h^p + O(h^{p+1}).$$

For grid size $h/2$ we need two values,

$$\begin{aligned} I_{h/2} &= \frac{u_{2j_*}(h/2) + u_{2j_*+1}(h/2)}{2} \\ &= \frac{1}{h}\int_{x_*}^{x_*+h/2} u(x)dx + \frac{1}{2}C(x_* + h/4)(h/2)^p + O(h^{p+1}) \\ &\quad + \frac{1}{h}\int_{x_*+h/2}^{x_*+h} u(x)dx + \frac{1}{2}C(x_* + 3h/4)(h/2)^p + O(h^{p+1}) \\ &= \frac{1}{h}\int_{x_*}^{x_*+h} u(x)dx + C(x_* + h/2)(h/2)^p + O(h^{p+1}). \end{aligned}$$

Finally, for grid size $h/4$ one can check that

$$I_{h/4} = \frac{1}{4}\sum_{k=0}^{3} u_{4j_*+k}(h/4) = \frac{1}{h}\int_{x_*}^{x_*+h} u(x)dx + C(x_* + h/2)(h/4)^p + O(h^{p+1}).$$

It follows as before in (3) that

$$\frac{I_h - I_{h/2}}{I_{h/2} - I_{h/4}} = 2^p + O(h).$$

h	$\tilde{u}_h$	$\tilde{u}_h - \tilde{u}_{h/2}$	$\frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$	$\log_2 \frac{\tilde{u}_h - \tilde{u}_{h/2}}{\tilde{u}_{h/2} - \tilde{u}_{h/4}}$
0.2	0.302842712474619	0.009289321881345	26.142135623725615	4.708305098603142
0.1	0.293553390593274	0.000355339059327	1.999999999999688	0.999999999999775
0.05	0.293198051533946	0.000177669529664	2.000000000001875	1.000000000001352
0.025	0.293020382004283	0.000088834764832	2.635450714080436	1.398049712285012
0.0125	0.292931547239451	0.000033707617584	12.589489353884787	3.654147861537719
0.00625	0.292897839621867	0.000002677441208		
0.003125	0.292895162180659			

Table 3. Table of values for the trapezoidal rule for $f(x) = |x - \alpha|$ with $\alpha = 1/\sqrt{2}$. The last column is the final approximation of the order of accuracy p, which fails for this case.

5 Non-smooth error

So far we have assumed that the error depends smoothly on the parameter h. Then the error is of the form in (2). This is, however not always the case. The error can, for instance, depend discontinuously on h, eventhough it is bounded as in (1). The reason for this can be discontinuities in the method itself (e.g. case switches) or non-smooth functions in the problem (e.g. solutions, sources, integrands). When the error is non-smooth one cannot check convergence rates by looking at ratios of differences as in Section 2.2. Other methods must be used.

Example 7 *Consider the trapezoidal rule applied to the integral*

$$\int_0^1 |x - \alpha| dx,$$

for some value $0 < \alpha < 1$. The trapezoidal rule is exact everywhere except at the grid cell which contains α. The error there depends crucially on the distance between α and the nearest grid point. More precisely, if $x_j \le \alpha < x_{j+1}$ and $x_{j+1} - x_j = h$,

$$\begin{aligned} u - \tilde{u}_h &= \int_{x_j}^{x_{j+1}} |x - \alpha| dx - h \frac{|x_j - \alpha| + |x_{j+1} - \alpha|}{2} \\ &= \int_{x_j}^{\alpha} (\alpha - x) dx + \int_{\alpha}^{x_{j+1}} (x - \alpha) dx - h \frac{x_{j+1} - x_j}{2} = \beta(\beta - 1) \frac{h^2}{2}, \end{aligned}$$

where $\beta = \beta(h) = (\alpha - x_j)/h$, i.e. the fractional part of α/h, which is a discontinuous function of h. The method is still second order accurate since $|\beta(h)| \le 1$ and (1) therefore holds with $C = 1/8$. However, the results presented in Figure 6 and Table 3 clearly shows the non-smoothness of the error and the failure of the ratios of the differences to predict the order of convergence.

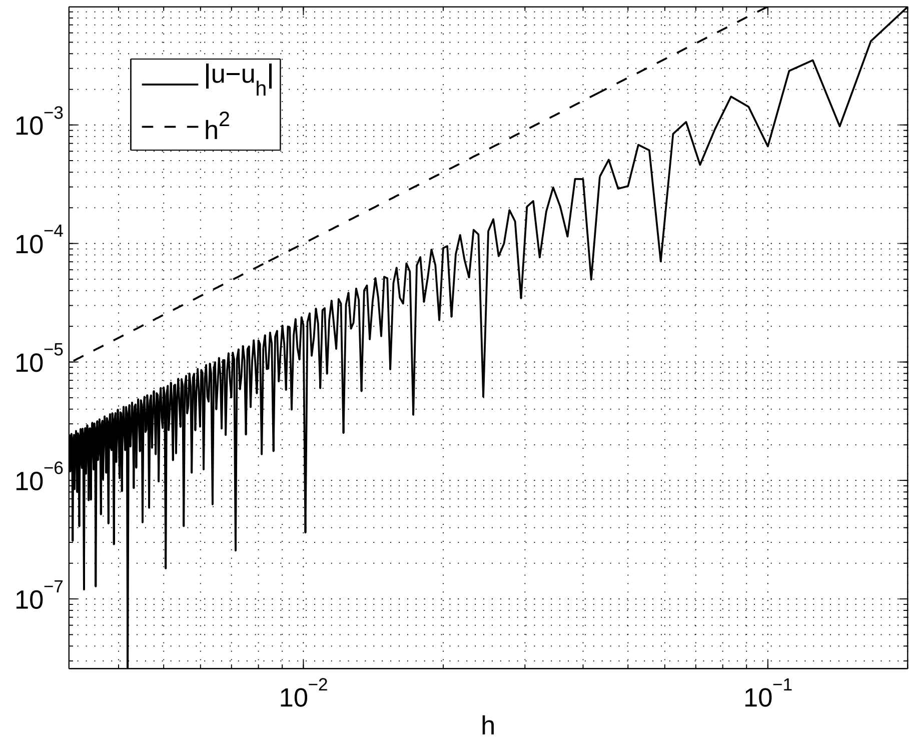


Figure 6. Error in trapezoidal rule for $f(x) = |x - \alpha|$ with $\alpha = 1/\sqrt{2}$. The dashed line is h^2 which indicates the slope for a second order method.