Math 1B, lecture 4: Error bounds for numerical methods

Nathan Pflueger

14 September 2011

1 Introduction

The five numerical methods descried in the previous lecture all operate by the same principle: they approximate the integral $\int_a^b f(x)dx$ by dividing the interval $[a, b]$ into n slices of equal width, and approximate the area under the curve in each slice by assuming that the curve has a certain shape (flat, linear, or quadratic), and approximating the area by what the area would be in this case. Using any of these numerical methods, it is necessary to understand the error that may result from these assumptions. This lecture presents a method for computing bounds on the error of each numerical method.

The objective of this topic is not that you should remember these error bounds after the course; in practice, any time you need to the value of an integral, you will use a machine that implements these methods (or even more sophisticated methods which we do not discuss). Rather, the objective is to familiarize you with how these methods work, so that you will be aware of the basic qualitative aspects of the methods that machines use. Furthermore, we wish to emphasize that *an approximation is useless without knowledge of its possible error,* so that you will always remember to demand error bounds whenever you use approximate methods.

As a secondary objective, we wish to use these methods and their error bounds to illustrate how polynomials are, in some sense, the fundamental functions in calculus; each of these numerical methods has the feature that it is exact for polynomials of a certain degree, and its error can be found by considering polynomials of higher degree. The basic philosophy that a function can be understood by means of the polynomials it resembles is a central feature of calculus, which will be made especially vivid by the development of Taylor series in a couple weeks.

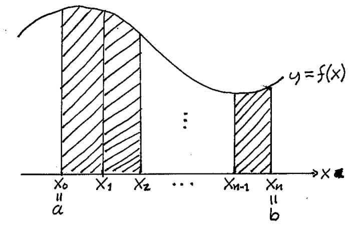
The reading for today is Gottlieb §26.2. The homework is to do problems $1, 3, 4, 5$ from §26.2, as well as a Topic Outline.

2 Summary of results

We first recall the definitions of the five numerical methods discussed last time. We are evaluating the integral:

$$\int_a^b f(x)dx.$$

In order to do this, divide the interval $[a, b]$ on which the integral is being computed into n pieces (where n can be chosen to be anything), thus dividing the value of the integral into n slices, with endpoints $x_0, x_1, \ldots, x_n$, as shown.



$$\begin{aligned}\int_a^b f(x)dx &= \sum_{k=1}^{n} \big(\text{area under the curve in the } k^{th} \text{ slice}\big)\\ &= \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} f(x)dx\end{aligned}$$

Here we make the following definitions.

$$\begin{aligned}x_0 &= a\\ x_n &= b\\ \Delta x &= (b-a)/n\\ x_k &= a + k\cdot\Delta x\end{aligned}$$

Thus $[x_0, x_1]$ is the interval corresponding to the first slice, $[x_1, x_2]$ corresponds to the seconds slice, and in general $[x_{k-1}, x_k]$ corresponds to the k^{th} slice.

Using this notation, the five numerical methods that we consider can be summarized as follows.

$$\begin{aligned}\text{Left approximation } L_n &= \sum_{k=1}^{n} f(x_{k-1})\Delta x\\ \text{Right approximation } R_n &= \sum_{k=1}^{n} f(x_k)\Delta x\\ \text{Midpoint approximation } M_n &= \sum_{k=1}^{n} f\left(\frac{x_{k-1}+x_k}{2}\right)\Delta x\\ \text{Trapezoid approximation } T_n &= \frac{1}{2}(L_n + R_n)\\ &= \sum_{k=1}^{n} \frac{1}{2}\left[f(x_{k-1}) + f(x_k)\right]\Delta x\\ \text{Simpson's rule } S_{2n} &= \frac{2}{3}M_n + \frac{1}{3}T_n\\ &= \sum_{k=1}^{n} \frac{1}{6}\left[f(x_{k-1} + 4f\left(\frac{x_{k-1}+x_k}{2}\right) + f(x_k)\right]\Delta x\end{aligned}$$

Note that both the trapezoid approximation and Simpson's rule can be defined either in terms of earlier approximations or directly in terms of values of the function f. Note also that each approximation is a sum of estimates for individual slices, each of which consists of the width of the slice times an estimate for the average value of f on that slice. The estimate for this average value depends on the assumed shape of the function on the slice.

The error bounds which we present will include a constant factor determined by the size of some derivative of $f(x)$. We shall use the following notation.[1]

[1] The symbol $\mathfrak{M}$ is called a "Gothic M," or a "German M." It is frequently used by mathematicians in situations where the ordinary letter M would be ambiguous, as it certainly would be here.

$$
\begin{aligned}
\mathfrak{M}_1 &= \max\{|f'(x)| : \ x \text{ in } [a,b]\} \\
\mathfrak{M}_2 &= \max\{|f''(x)| : \ x \text{ in } [a,b]\} \\
\mathfrak{M}_4 &= \max\left\{|f^{(4)}(x)| : \ x \text{ in } [a,b]\right\}
\end{aligned}
$$

Using this notation, the error bounds that we shall use are expressed by the following theorem (which will not be proved in class).

Theorem 2.1. *If the integral $\int_a^b f(x)dx$ is approximated using the methods above, then the following bounds hold.*

$$
\begin{aligned}
\left| L_n - \int_a^b f(x)dx \right| &\leq \frac{1}{2}\mathfrak{M}_1(b-a)^2/n \qquad \left(= \frac{1}{2}\mathfrak{M}_1(\Delta x)^2 \cdot n\right) \\
\left| R_n - \int_a^b f(x)dx \right| &\leq \frac{1}{2}\mathfrak{M}_1(b-a)^2/n \qquad \left(= \frac{1}{2}\mathfrak{M}_1(\Delta x)^2 \cdot n\right) \\
\left| M_n - \int_a^b f(x)dx \right| &\leq \frac{1}{24}\mathfrak{M}_2(b-a)^3/n^2 \qquad \left(= \frac{1}{24}\mathfrak{M}_2(\Delta x)^3 \cdot n\right) \\
\left| T_n - \int_a^b f(x)dx \right| &\leq \frac{1}{12}\mathfrak{M}_2(b-a)^3/n^2 \qquad \left(= \frac{1}{12}\mathfrak{M}_2(\Delta x)^3 \cdot n\right) \\
\left| S_{2n} - \int_a^b f(x)dx \right| &\leq \frac{1}{180}\mathfrak{M}_4(b-a)^5/(2n)^4 \qquad \left(= \frac{1}{180}\mathfrak{M}_4(\Delta x/2)^5 \cdot n\right)
\end{aligned}
$$

The expressions in parentheses are included simply to show how much each of the n slices contributes to the total error of the approximation.

The proof of the first two bounds is not difficult; we will sketch the argument in section 4. The latter three bounds are somewhat more technical, and require the use of integration by parts.

The particular forms of these bounds are not important (and you will almost certainly forget them after the exam). However, the following features are important.

- The error terms have constant factors coming from the maximum value of some derivative of $f(x)$. The better the approximation, the higher the derivative which governs the error.
- The error bound shrinks as n grows. The better the approximation, the faster the error bound will shrink.

Theorem 2.1, together with the qualitative descriptions discusses in the last lecture, are summarized in the following table. Remember that the conditions under the columns “overestimates” and “underestimates” must hold at *all* points in the interval $[a,b]$ in order to conclude whether the method gives an overestimate or underestimate.

Method	**Overestimates**	**Underestimates**	**Error bound**
Left L_n	$f'(x) < 0$	$f'(x) > 0$	$\frac{1}{2}\mathfrak{M}_1(b-a)^2/n$
Right R_n	$f'(x) > 0$	$f'(x) < 0$	$\frac{1}{2}\mathfrak{M}_1(b-a)^2/n$
Midpoint M_n	$f''(x) < 0$	$f''(x) > 0$	$\frac{1}{24}\mathfrak{M}_2(b-a)^3/n^2$
Trapezoid T_n	$f''(x) > 0$	$f''(x) < 0$	$\frac{1}{12}\mathfrak{M}_2(b-a)^3/n^2$
Simpson S_{2n}	$f^{(4)}(x) > 0$	$f^{(4)}(x) < 0$	$\frac{1}{180}\mathfrak{M}_4(b-a)^5/(2n)^4$

3 Examples

Example 3.1. How accurate will Simpson's rule be for n slices be in computing $\int_0^2 x^3 dx$? The error bound is $\frac{1}{180}\mathfrak{M}_4(b-a)^5/(2n)^4$. For this particular function, the fourth derivative is 0, and therefore $\mathfrak{M}_4 = 0$. Therefore the error bound is 0, no matter which n is used. So Simpson's rule should be exactly correct for this integral. This may seem too good to be true, so here is a computation, for $n = 1$ (i.e. a computation of S_2). By definition of Simpson's rule,

$$\begin{aligned} S_4 &= \frac{1}{6}(0^3 + 4 \cdot 1^3 + 2^3) \cdot 2 \\ &= \frac{1}{6}(0 + 4 + 8) \cdot 2 \\ &= 4. \end{aligned}$$

Indeed, the standard calculation shows that $\int_0^2 x^3 dx = 4$.

Example 3.2. Consider the integral $\displaystyle\int_0^2 e^{-x^2} dx$. Determine a sufficient number of slices for the midpoint approximation to be accurate within 0.001.

The error bound for M_n is $\frac{1}{24}\mathfrak{M}_2(b-a)^3/n^2$. Since the length of the interval is $(b-a) = 2$, this is equal to $\frac{1}{3}\mathfrak{M}_2/n^2$ in this case. We need to select an n large enough that this bound is no more that 0.001, i.e. such that $\frac{1}{3}\mathfrak{M}_2/n^2 \leq 0.001$. Equivalently, $n^2 \geq \frac{1000}{3}\mathfrak{M}_2$, or $n \geq \sqrt{\frac{1000}{3}\mathfrak{M}_2}$. So we simply need to find $\mathfrak{M}_2$. In fact, it is not necessary to find $\mathfrak{M}_2$ exactly; an upper bound on $\mathfrak{M}_2$ will do.

To find such an upper bound, begin by computing derivatives of the function (details of the computation are omitted).

$$\begin{aligned} f(x) &= e^{-x^2} \\ f'(x) &= (-2x)e^{-x^2} \\ f''(x) &= (4x^2 - 2)e^{-x^2} \end{aligned}$$

It would be a hassle to actually explicitly find the maximum of this second derivative, so simply get an upper bound. For this purpose, observe that for all x, $-x^2 \leq 0$, hence $e^{-x^2} \leq e^0 = 1$. Therefore $|f''(x)| = |(4x^2-2)e^{-x^2}| \leq |4x^2 - 2|$. Now, the function $4x^2 - 2$ is an increasing function on $[0, 2]$, with values ranging from -2 to 14. Thus an upper bound on its size is 14. Taken together, this implies that $|f''(x)| \leq 14$ on $[0, 2]$, and thus $\mathfrak{M}_2 \leq 14$.

Now, since this is an upper bound on $\mathfrak{M}_2$, if we can ensure that $n \geq \sqrt{\frac{1000}{3} \cdot 14}$, then we will certainly have that $n \geq \sqrt{\frac{1000}{3}\mathfrak{M}_2}$, and in turn that the error of M_n is less than 0.001. Now $\sqrt{\frac{1000}{3} \cdot 14} \approx 68.3$. Thus 69 slices will be sufficient to ensure error of less than 0.001 for the midpoint approximation.

Example 3.3. Consider the integral $\int_1^2 \frac{dx}{x}$, whose exact value is equal to $\ln 2$. Numerical methods give one way to calculate this value to arbitrary accuracy (better methods will come from the notion of Taylor series). How many slices are needed in order to compute this integral with error at most one millionth (that is, 10^{-6}), using the five numerical methods discussed?

To begin, we need the maximum values $\mathfrak{M}_1, \mathfrak{M}_2$ and $\mathfrak{M}_4$, for which we need the derivatives of the function.

$$\begin{aligned} f(x) &= 1/x \\ f'(x) &= -1/x^2 \\ f''(x) &= 2/x^3 \\ f'''(x) &= -6/x^4 \\ f^{(4)}(x) &= 24/x^5 \end{aligned}$$

Each of these functions decrease in magnitude as x increases, so all of them have maximum absolute value at $x = 1$. From this we obtain the following values.

$$\begin{aligned} \mathfrak{M}_1 &= 1 \\ \mathfrak{M}_2 &= 2 \\ \mathfrak{M}_4 &= 24 \end{aligned}$$

From this, as well as the fact that the length of the interval is $(b - a) = 1$, the following bounds for the error of each approximation can be found.

$$\begin{aligned} |L_n - \ln 2| &\leq \frac{1}{2}\mathfrak{M}_1(b-a)^2/n = \frac{1}{2n} \\ |R_n - \ln 2| &\leq \frac{1}{2}\mathfrak{M}_1(b-a)^2/n = \frac{1}{2n} \\ |M_n - \ln 2| &\leq \frac{1}{24}\mathfrak{M}_2(b-a)^3/n^2 = \frac{1}{12n^2} \\ |T_n - \ln 2| &\leq \frac{1}{12}\mathfrak{M}_2(b-a)^3/n^2 = \frac{1}{6n^2} \\ |S_{2n} - \ln 2| &\leq \frac{1}{180}\mathfrak{M}_4(b-a)^5/(2n)^4 = \frac{1}{120n^4} \end{aligned}$$

Now, observe that $\frac{1}{2n} \leq 10^{-6} \Leftrightarrow n \geq 500{,}000$, thus $500{,}000$ slices will ensure that L_n and R_n are both within 10^{-6} of the true value.

For the midpoint approximation, $\frac{1}{12n^2} \leq 10^{-6} \Leftrightarrow n^2 \geq 10^6/12 \Leftrightarrow n \geq \sqrt{10^6/12} \approx 288.7$. So 289 slices will ensure that M_n is within 10^{-6} of the true value.

For the trapezoid approximation, $\frac{1}{6n^2} \leq 10^{-6} \Leftrightarrow n^2 \geq 10^6/6 \Leftrightarrow n \geq \sqrt{10^6/6} \approx 408.2$. So 409 slices will ensure that T_n is within 10^{-6} of the true value.

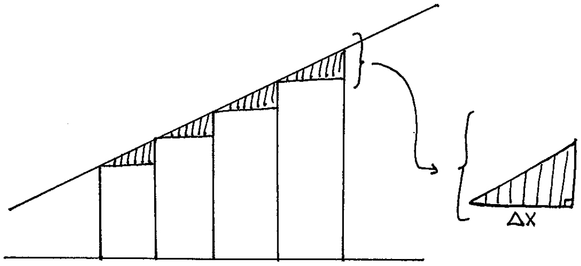
For Simpson's rule, $\frac{1}{120n^4} \leq 10^{-6} \Leftrightarrow n^4 \geq 10^6/120 \Leftrightarrow n \geq \sqrt[4]{10^6/120} \approx 9.6$. So 10 slices will ensure that S_{2n} is within 10^{-6} of the true value.

As this result indicates, the error bounds guarantee much faster convergence for Simpson's rule than the less sophisticated approximations.

In fact, explicit computation reveals that, in order to achieve error of less than one millionth, only 7 slices are needed for Simpson (the bounds ensure that 10 will work), 177 slices are needed for midpoint approximation (the bounds ensure that 289 will work), and 250 slices are needed for trapezoid approximation (the bounds ensure that 409 will work). This illustrates the fact that the error bounds, while obviously not exact, give a good rough idea for how many slices are needed to achieve a desired accuracy of computation.

4 Error of L_n and R_n

In order to understand the error bound for left and right approximation, we begin by considering the error of these approximations in the special case where $f(x)$ is a linear function, with slope m. That is, $f'(x) = m$ for all x, where m is a constant. Then we can see the error of the left approximation L_n in the following picture.



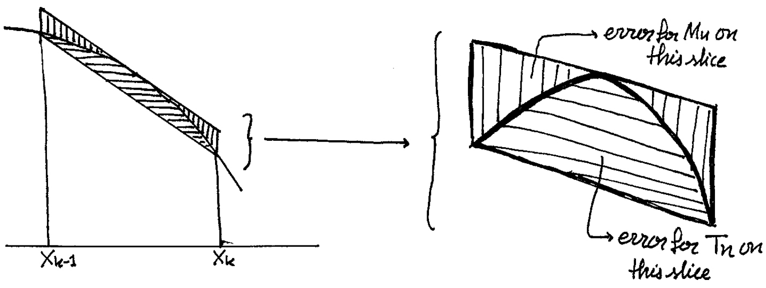
The left approximations L_n undershoots the area of each slice by precisely the area of one of the shades triangles. Observe that the base of each triangle is equal to the width of the slice, namely Δx, and the height must be equal to $m\Delta x$ (where m is the slope of the function, i.e. the constant value of its first derivative). Therefore the area of each shaded triangle is $\frac{1}{2}(\Delta x)(m\Delta x) = \frac{1}{2}m(\Delta x)^2$. Since there are n such triangles, the total error is precisely $n \cdot \frac{1}{2}m(\Delta x)^2 = \frac{1}{2}mn\left(\frac{b-a}{n}\right)^2 = \frac{1}{2}m\frac{(b-a)^2}{n}$. Because m is the value of $f'(x)$ at every point, $|m|$ is equal to the value $\mathfrak{M}_1$, and so in fact the error is precisely equal to the bound $\frac{1}{2}\mathfrak{M}_1\frac{(b-a)^2}{n}$.

Now suppose that $f(x)$ is any function, not necessarily linear. Then although we can no longer view the error as the areas of right triangles, essentially the same technique will apply: since the function never grows at a rate faster than $\mathfrak{M}_1$, the error of approximation in each slice is still bounded by the area of a triangle with base Δx and height $\mathfrak{M}_1\Delta x$. A complete proof of the error bound can be given by using the mean value theorem; we omit the details. The result is, as stated in section 2,

$$\left|\int_a^b f(x)dx - L_n\right| \leq \frac{1}{2}\mathfrak{M}_1(b-a)^2/n.$$

5 Error of M_n and T_n

In this section, we give only a rough sketch of the principles behind the error bounds for M_n and T_n. To give complete details would require somewhat more technical notation than seems appropriate at the moment. Consider the following image, depicting the error of both the midpoint approximation and trapezoid approximation on a single slice.



What can make these errors as large as possible? First of all, if f is linear in this slice, then both errors will be 0. So the error is governed by the second derivative. The largest possible error will result when the function begins, at the lest end of the slice, by increasing (respectively, decreasing) very quickly, turns around in the middle, and then decreases (respectively, increases) rapidly to the other endpoint. Such behavior involves a dramatic change in the first derivative, i.e. a large second derivative.

Upon formalizing this rough idea and doing some (not entirely trivial) computation, the following result can be obtained.

$$\begin{aligned}\text{error in a single slice for } M_n &\leq \frac{1}{24}\mathfrak{M}_2(\Delta x)^3 \\ \text{error in a single slice for } T_n &\leq \frac{1}{12}\mathfrak{M}_2(\Delta x)^3\end{aligned}$$

Replacing Δx by $(b-a)/n$ as usual, and multiplying by n (since there are n slices, each contributing some error) gives the error bounds from section 2.

The get an idea for where the constants $\frac{1}{24}$ and $\frac{1}{12}$ come from, you may evaluate the error in the case of a quadratic function, where these error bounds are exact. From this it follows (among other things), the classical fact (known to the Greeks by much more elementary methods) that the area of a parabolic segment is equal to $\frac{2}{3}$ of the base times the height.