18.330 Lecture Notes: Convergence of Infinite Sums

Homer Reid

February 6, 2014

Consider a convergent infinite sum

$$\mathcal{S} = \sum_{n=1}^{\infty} f(n) \tag{1}$$

We want to know how accurately we can approximate $\mathcal{S}$ by retaining only the first N terms in the sum. That is, if we define the Nth partial sum as

$$\mathcal{S}_N = \sum_{n=1}^{N} f(n) \tag{2}$$

then we want to estimate the error $\mathcal{E}_N$ incurred by approximating $\mathcal{S}$ by $\mathcal{S}_N$. $\mathcal{E}_N$ is of course just the sum of all summands from $N+1$ to infinity:

$$\begin{aligned} \mathcal{E}_N &= \Big| \mathcal{S} - \mathcal{S}_N \Big| \\ &= \Big| \sum_{N+1}^{\infty} f(n) \Big|. \end{aligned} \tag{3}$$

Error estimates for monotonic summands

In the commonly encountered case in which $f(x)$ is positive and *monotonically decreasing* [that is, $y > x$ implies $f(y) < f(x)$], it is easy to estimate the sum in (3) in terms of definite integrals over $f(x)$. To understand the basic idea, consider the following plot of the function $f(x)$ over the interval $[N, M]$ (The particular case we are considering here is $f(x) = 1/x^2$ with $[N, M] = [10, 15]$, but the general principles are valid for any monotonically decreasing function over any interval.)

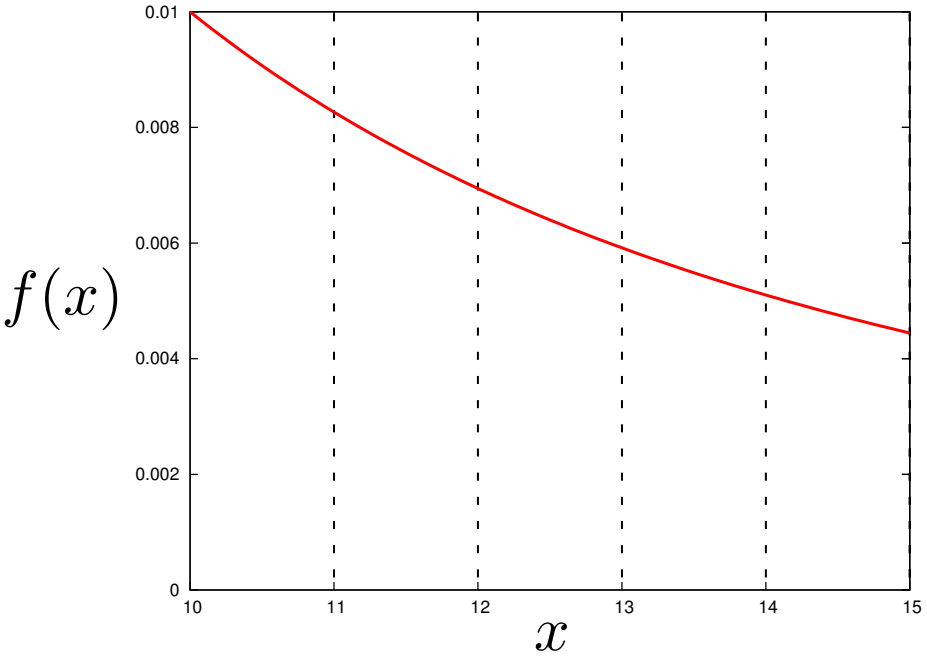

Figure 1: A plot of the function $f(x) = \frac{1}{x^2}$, here considered over the interval $[N, M] = [10, 15]$.

The integral $\int_N^M f(x)\,dx$ gives the area under the curve $f(x)$ between N and M. This is the red-shaped region in Figure 2 below.

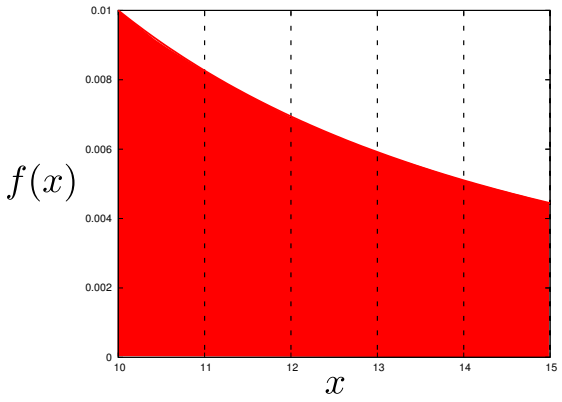

Figure 2: The integral $\int_{10}^{15} f(x)\,dx$ gives the area under the curve $f(x)$ between $x = 10$ and $x = 15$.

On the other hand, the sum $\sum_{n=N}^{M-1} f(n)$ gives the area of the purple-shaded region shown in Figure 3 below.

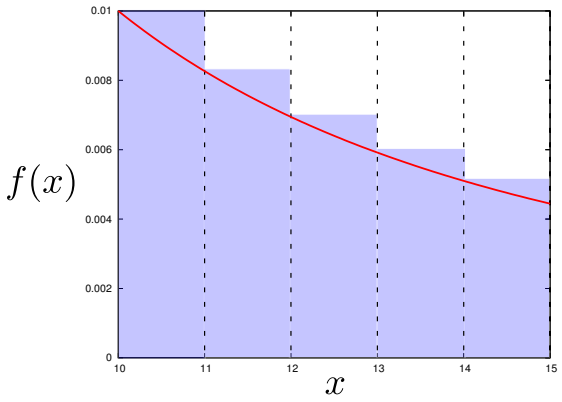

Figure 3: The sum $\sum_{n=N}^{M-1} f(n)$ gives the area of the shape consisting of the blue shaded rectangles. Since $f(x)$ is monotonically decreasing, this area is guaranteed to be *greater* than the area of the red-shaded area in Figure 2.

The purple-shaded region in Figure 3 is a union of rectangles; the rectangle between $x = n$ and $x = n + 1$ has width 1 and height $f(n)$. Since the function $f(x)$ is decreasing, the area of this rectangle is guaranteed to be *greater* than the area under the curve $f(x)$ between n and $n + 1$, and thus the area of the entire purple-shaded region in Figure 3) is *greater* than the red-shaded region

in Figure 2). In other words, we have

$$\sum_{N}^{M-1} f(n) > \int_{N}^{M} f(x)\,dx \tag{4}$$

If we instead take the rectangle between $x = n$ and $x = n+1$ to have height $f(n+1)$ instead of height $f(n)$, we obtain the the green-shaded region depicted in Figure 4 below. In Figure 4, the area of the rectangle between $x = n$ and

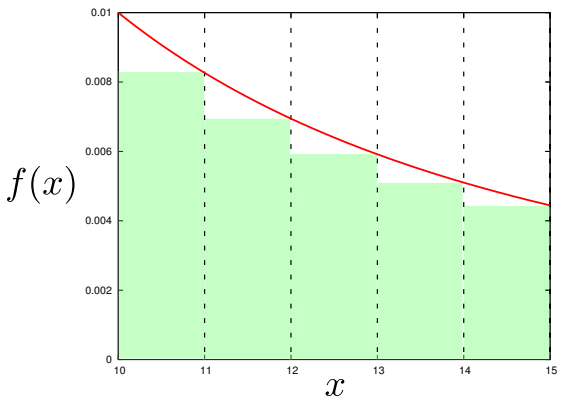

Figure 4: The sum $\sum_{n=N}^{M-1} f(n+1)$ gives the area of the shape consisting of the green shaded rectangles. Since $f(x)$ is monotonically decreasing, this area is guaranteed to be *less* than the area of the red-shaded area in Figure 2.

$x = n+1$ is guaranteed to be *less* than the area under the curve $f(x)$ between n and $n+1$, and thus the area of the entire green-shaded region in Figure 4 is *less* than the red-shaded region in Figure 2. In other words, we have

$$\sum_{N}^{M-1} f(n+1) < \int_{N}^{M} f(x)\,dx \tag{5}$$

which we could alternatively write in the equivalent form

$$\sum_{N+1}^{M} f(n) < \int_{N}^{M} f(x)\,dx. \tag{6}$$

Inequality (6) is the one that will be useful for our purposes. Taking $M \to \infty$, the sum on the RHS is just the quantity that enters the definition of the error (3), and hence we find

$$\mathcal{E}_N = \mathcal{S} - \mathcal{S}_N < \int_{N}^{\infty} f(x)\,dx. \tag{7}$$

(We have dropped the absolute value signs from (3) because $f(x)$ is positive, which means $\mathcal{S} - \mathcal{S}_N$ is always positive.)

Application: Binding energy of a 1D ionic solid

Earlier we considered the sum

$$\mathcal{S} = -\sum_{n=1}^{\infty} \frac{(-1)^n}{n}.$$

The method we discussed earlier cannot be applied to this sum as it stands because the summand is not positive and monotonically decreasing. To rectify this situation, we rewrite the sum as follows:

$$\begin{aligned}\mathcal{S} &= \sum_{n=1}^{\infty} \left(\frac{1}{2n-1} - \frac{1}{2n} \right) \\ &= \sum_{n=1}^{\infty} \frac{1}{2n(2n-1)}.\end{aligned}$$

Now we have a positive and monotonically decreasing summand, so we can apply (7) to estimate the error in the Nth partial sum:

$$\begin{aligned}\mathcal{E}_N = \mathcal{S} - \mathcal{S}_N &< \int_N^{\infty} \frac{dx}{2x(2x-1)} \\ &= -\frac{1}{2} \log \left(1 - \frac{1}{2N} \right)\end{aligned}$$

To find the value of N at which the partial sum becomes correct to 6 digits, we ask for $\mathcal{E}_N$ to be less than 10^{-6} times the exact value of the sum, $\log 2$:

$$-\frac{1}{2} \log \left(1 - \frac{1}{2N} \right) < 10^{-6} \cdot \log 2 \qquad \Longrightarrow \qquad N > 360,674.$$

This corroborates our earlier finding that the 6th digit of the sum stabilized somewhere between $N = 10^5$ and $N = 10^6$.

Estimating the error on the fly

In this case, we estimated the relative error by diving the absolute error by the known value of the exact solution. In general, of course, we won't know *a priori* the exact value of the sum we are computing (otherwise we wouldn't be computing it). So how do we estimate the relative error during the course of a calculation?

Easy: just divide by the current partial sum (that is, our best current approximation to the exact solution) instead of dividing by the exact solution. For a positive, monotonically decreasing summand, the condition $\mathcal{S}_N < \mathcal{S}$ is guaranteed to be satisfied for any N. This means that errors measured relative to $\mathcal{S}_N$ are always *larger* than errors relative to $\mathcal{S}$. In other words, for all N we have

$$\frac{\mathcal{E}_N}{\mathcal{S}_N} > \frac{\mathcal{E}_N}{\mathcal{S}}$$

so $\mathcal{E}_N/\mathcal{S}_N$ gives us an upper bound on the true relative error. (Moreover, in the later stages of a calculation the difference between $\mathcal{S}_N$ and $\mathcal{S}$ is small, so it is a tight upper bound.)

The Euler-Maclaurin Formula

Inequality (6), which we may write in the form

$$\sum_{n=N+1}^{M} f(n) \; - \; \int_N^M f(x)\,dx \quad < \quad 0 \tag{8}$$

is a fairly crude result: It only holds for monotonically decreasing functions, and it really only expresses a particularly obvious geometric statement.

It turns out that it is possible to refine (8) quite dramatically: We can relax the constraint that $f(x)$ be positive and monotonically decreasing, and we can sharpen the inequality into an equality. The result is the *Euler-Maclaurin summation formula*, which reads

$$\sum_{n=N+1}^{M} f(n) \; - \; \int_N^M f(x)\,dx \; = \; \sum_{p=0}^{\infty} C_p \Big[f^{(p)}(M) - f^{(p)}(N) \Big] \tag{9}$$

where $f^{(p)}$ is the pth derivative of f [$f^{(0)}(x)$ is just $f(x)$] and the C_p coefficients decay rapidly with p:

$$C_0 = \frac{1}{2}, \qquad C_1 = \frac{1}{12}, \qquad C_2 = -\frac{1}{720}, \qquad C_3 = \frac{1}{30240}, \qquad C_4 = \frac{1}{1209600}.$$

In contrast to equation (8), equation (9) holds for general smooth functions f, not just functions that are positive and monotonically decreasing.

The Euler-Maclaurin formula is amazing: It says that the difference between the sum and integral may be expressed *entirely in terms of the behavior at the endpoints*. The formula is used extensively by number theorists, who use it to evaluate sums in terms of integrals (which are generally much easier to compute).

The Euler-Maclaurin summation formula is somewhat tedious to derive, and since we won't really use it much in this class we will skip the derivation. (It is derived in many older numerical analysis books, including Stoer&Bulirsch.) However, we want to call your attention to one important property: The error term on the RHS of (8) depends only on the *difference* between the behavior of f at N and the behavior of f at M. This means, in particular, that if f is *periodic* over the interval $[N, M]$ then the entire error term vanishes!

This is our first brush with a general principle of 18.330: *amazing magical things happen when we work with periodic functions.* We will encounter this phenomenon in several places through the remainder of the course.