Riemann Integration

D P Patil

Riemann had revolutionized the fields of analysis, geometry and mathematical physics. His ideas concerning geometry of space had a profound effect on the development of modern theoretical physics. Riemannian manifolds, Riemann surfaces, the Cauchy–Riemann equations, the Riemann hypothesis – all these and more are packed into his one-volume collected works. Riemann clarified the notion of integration by defining, a little over 135 years ago, what we now call the *Riemann integral.* This article is devoted to a study of the Riemann integral.

D P Patil got his Ph.D from the School of Mathematics, TIFR and joined IISc in 1992. His interests are commutative algebra, algebraic geometry and algebraic number theory.

The derivative does not display its full strength until allied with the 'integral'. The study of integrals does require a long preparation, but once this preliminary work is completed, integrals will be an invaluable tool for creating new functions and the derivative will appear more powerful than ever.

Although ultimately to be defined in a quite complicated way, the integral formalizes a simple intuitive concept – that of an area. It is no surprise that the definition of an intuitive concept can present great difficulties – 'area' is certainly no exception.

In elementary geometry, formulas are derived for the areas of many plane figures, but a little reflection shows that an acceptable definition of area is seldom given. The area of a region is sometimes defined as the number of squares with sides of length 1, which fit in the region. But this definition is hopelessly inadequate for any but the simplest regions. For example, a circle of radius 1 supposedly has an area equal to the irrational number

Keywords
Area, lower and upper sums, Riemann sums, integrable functions, definite integral, derivative, primitive, anti-derivative.

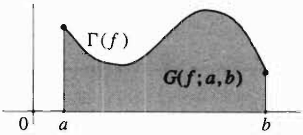


Figure 1.

π, but it is not at all clear what 'π squares' means. Even if we consider a circle of radius $1/\sqrt{\pi}$, which supposedly has area 1, it is hard to say in what way a unit square fits in this circle, since it does not seem possible to divide the unit square into pieces which can be rearranged to form a circle.

In this article we will try to define the area of only some very special regions (*Figure* 1) – those which are bounded by the horizontal axis, the vertical lines through $(a, 0)$ and $(b, 0)$ and the graph $\Gamma(f) := \Gamma_f := \{(x, f(x))| \ x \in [a, b]\}$ of a function $f : [a, b] \to \mathbb{R}$, $a \leq b$, such that $f(x) \geq 0$ for all $x \in [a, b]$. It is convenient to indicate this region by $G(f; a, b)$. Therefore $G(f; a, b) := \{(x, y) \in \mathbb{R}^2| \ a \leq x \leq b, 0 \leq y \leq f(x)\}$.

Notice that these regions include rectangles and triangles as well as many other important geometric figures.

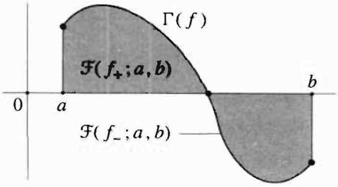
The number $\mathcal{F}(f; a, b)$ which we will eventually assign as the area of $G(f; a, b)$ will be called the *integral* of f on $[a, b]$. Actually, the integral will be defined even for functions f which do not satisfy the condition $f(x) \geq 0$ for all $x \in [a, b]$.

For an arbitrary function $f : [a, b] \to \mathbb{R}$, we put

$$f_+(x) := \text{Max}\ (f(x), 0), \quad f_-(x) := \text{Max}\ (-f(x), 0).$$

Then $f_+ \geq 0$, $f_- \geq 0$, and $f = f_+ - f_-$; f_+ is called the *positive part* and f_- the *negative part* of f. Moreover, if f is continuous, then f_+ and f_- are also continuous. In this case the integral will represent the difference of the area $\mathcal{F}(f_+; a, b)$ of the region $G(f_+; a, b)$ and the area $\mathcal{F}(f_-; a, b)$ of the region $G(f_-; a, b)$ (the 'algebraic area' of the region $G(f; a, b)$; see *Figure* 2).

Figure 2.



The idea behind the prospective definition is indicated in *Figure* 3. The interval $[a, b]$ has been divided into m subintervals $[t_i, t_{i+1}]$, $i = 0, \ldots, m-1$ by means of numbers $a = t_0, t_1, \ldots, t_{m-1}, t_m = b$ with $a = t_0 < t_1 <$

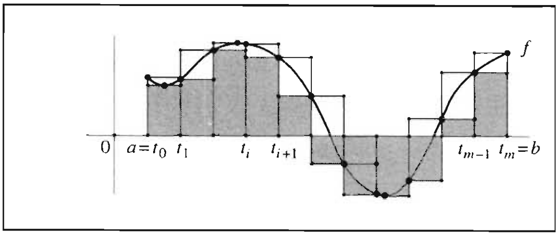


Figure 3.

$\cdots < t_m = b$. Let m_i and M_i denote the minimum and maximum values of f on the i-th interval $[t_i, t_{i+1}]$. Then the sum $s := \sum_{i=0}^{m-1} m_i(t_{i+1}-t_i)$ represents the total area of rectangles lying inside the region $G(f; a, b)$, while the sum $S := \sum_{i=0}^{m-1} M_i(t_{i+1} - t_i)$ represents the total area of rectangles containing the region $G(f; a, b)$.

The guiding principle of our attempt to define the area of $G(f; a, b)$ is the observation that $\mathcal{F}(f; a, b)$ should satisfy

$$s \leq \mathcal{F}(f; a, b) \leq S,$$

and that this should be true, *no matter how the interval* $[a, b]$ *is subdivided.* It is to be hoped that these requirements will determine $\mathcal{F}(f; a, b)$. The following definitions formalize and eliminate some of the implicit assumptions in the above discussion.

1. The Lower and Upper Sums A *partition* of the interval $[a, b]$ is a finite collection of points in $[a, b]$, one of which is a and one of which is b. The points in a partition can be numbered $t_0, t_1, \ldots, t_{m-1}, t_m$ so that

$$a = t_0 < t_1 < \cdots t_{m-1} < t_m = b;$$

we shall always assume that such a numbering has been assigned.

Suppose that a function $f : [a, b] \to \mathbb{R}$ is bounded and $P = \{t_0, t_1, \ldots, t_{m-1}, t_m\}$ is a partition of $[a, b]$. Then, for each $i = 0, \ldots, m-1$, the infimum and the supremum

$$\begin{aligned} m_i : &= \inf\{f(x) | t_i \leq x \leq t_{i+1}\} \quad \text{and} \\ M_i : &= \sup\{f(x) | t_i \leq x \leq t_{i+1}\} \end{aligned}$$

of f on $[t_i, t_{i+1}]$ exist. The *lower sum* and the *upper sum* of f with respect to P, are denoted by $L(f, P)$ and $U(f, P)$ respectively, and are defined by

$$L(f,P): = \sum_{i=0}^{m-1} m_i(t_{i+1} - t_i) \quad \text{and}$$
$$U(f,P): = \sum_{i=0}^{m-1} M_i(t_{i+1} - t_i).$$

The lower and upper sums are supposed to represent the total areas of rectangles lying below and above the graph of f. Notice that despite the geometric motivation, these sums have been defined precisely, without any appeal to a concept of 'area'.

Note that the requirement that f be bounded on $[a, b]$ is essential in order that all the m_i and M_i be defined. Further, note that it was necessary to define the numbers m_i and M_i as infimums and supremums rather than as minima and maxima, since f was not assumed to be continuous.

Since $m_i(t_{i+1}-t_i) \le M_i(t_{i+1}-t_i)$ for each $i = 0, \quad ,m - 1$ we have

$$L(f,P) \le U(f,P).$$

On the other hand, something less obvious *ought* to be true: *If P, Q are any two partitions of $[a,b]$, then* $L(f,P) \le U(f,Q)$, (because $L(f,P)$ should be $\le$ area $G(f;a,b)$ and $U(f,P)$ should be $\ge$ area $G(f;a,b)$. This proves nothing, since the 'area of $G(f;a,b)$' has not been defined yet), but it indicates that if there is to be any hope of defining area of $G(f;a,b)$, a proof that $L(f,P) \le U(f,P)$ should come first. The proof which we give depends on a lemma concerning the behavior of lower and upper sums when more points are included in a partition. More precisely:

2. Lemma. *Let P and Q be two partitions of the interval $[a,b]$. If Q is a refinement of P, i.e. $P \subseteq Q$, then*

A *Riemann sum* of the function $f : [a,b] \to \mathbb{R}$ for the partition $P = \{t_0, t_1, \dots, t_{m-1}, t_m\}$ of $[a,b]$ is a sum of the form $\sum_{i=0}^{m-1} f(\tau_i)(t_{i+1} - t_i)$, where $\tau_i \in [t_i, t_{i+1}]$. The minimum (respectively, maximum) value of the Riemann sum for f corresponding to the given partition P of $[a,b]$ is precisely the lower (respectively, upper) sum of f with respect to P.

$L(f,P) \leq L(f,Q)$ *and* $U(f,P) \geq U(f,Q)$.

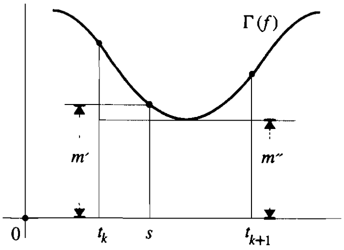


Figure 4.

Proof. Consider the special case in which Q contains just one more point than P, i.e. $P = \{t_0, t_1, \qquad t_{m-1}, t_m\}$ and $Q = \{\{t_0, t_1, \qquad t_k, s, t_{k+1}, \qquad , t_{m-1}, t_m\}$, where $a = t_0 < t_1 < \qquad < t_k < s < t_{k+1} < \qquad < t_{m-1} < t_m = b$. Let (see *Figure* 4)

$$\begin{aligned} m' : &= \inf\{f(x) | t_k \leq x \leq s\} \quad \text{and} \\ m'' : &= \inf\{f(x) | s \leq x \leq t_{k+1}\}. \end{aligned}$$

Then $L(f,P) = \sum_{i=0}^{m-1} m_i(t_{i+1} - t_i)$ and

$$L(f,Q) = \sum_{i=0}^{k-1} m_i(t_{i+1} - t_i) + m'(s - t_k) + m''(t_{k+1} - s) + \sum_{i=k+1}^{m-1} m_i(t_{i+1} - t_i).$$

Therefore to prove the inequality $L(f,P) \leq L(f,Q)$, it is enough to prove that

$$m_k(t_{k+1} - t_k) \leq m'(s - t_k) + m''(t_{k+1} - s).$$

Now, since $\{f(x) | t_k \leq x \leq s\} \subseteq \{f(x) | t_k \leq x \leq t_{k+1}\}$, we have

$$m_k = \inf\{f(x) | t_k \leq x \leq t_{k+1}\} \leq \inf\{f(x) | t_k \leq x \leq s\} = m'$$

Similarly, $m_k \leq m''$ Therefore

$$m_k(t_{k+1} - t_k) = m_k(s - t_k) + m_k(t_{k+1} - s) \leq m'(s - t_k) + m''(t_{k+1} - s).$$

This proves, in this special case, that $L(f,P) \leq L(f,Q)$. The proof of the inequality $U(f,P) \geq U(f,Q)$ is similar.

The general case can now be deduced easily. The partition Q can be obtained from P by adding one point at a time, i.e., there is a sequence of partitions $P =$

$P_1, P_2, \quad , P_r = Q$ such that P_{j+1} contains just one more point than P_j. Then

$$L(f,P) = L(f,P_1) \leq \qquad \leq L(f,P_r) = L(f,Q)$$

and

$$U(f,P) = U(f,P_1) \geq \qquad \geq U(f,P_r) = U(f,Q).$$

This proves the lemma. □

The theorem we wish to prove is a simple consequence of this lemma.

3. Corollary. *Let P_1 and P_2 be two partitions of the interval $[a,b]$ and let $f : [a,b] \to \mathbb{R}$ be a bounded function on $[a,b]$. Then $L(f,P_1) \leq U(f,P_2)$.*

Proof. There is a partition Q which contains both P_1 and P_2, namely, $Q := P_1 \cup P_2$. Then by lemma 2, we have $L(f,P_1) \leq L(f,Q) \leq U(f,Q) \leq U(f,P_2)$.

It follows from the above Corollary that any upper sum $U(f,P')$ is greater than or equal to the *least upper bound* of all lower sums:

$$\sup\{L(f,P)|P \text{ a partition of } [a,b]\} \leq U(f,P')$$

$$\text{for every partition } P'\text{of } [a,b].$$

This in turn, means that the supremum $\sup\{L(f,P)|P$ a partition of $[a,b]\}$ is a lower bound for the set of all upper sums of f:

$$\sup\{L(f,P)|P \text{ a partition of } [a,b]\} \leq$$

$$\inf\{U(f,P)|P \text{ a partition of } [a,b]\}.$$

It is clear that both these numbers are in between the lower sum and upper sum of f for *all* partitions P of $[a,b]$. Therefore

$$L(f,P') \leq \sup\{L(f,P)|P \text{ a partition of } [a,b]\} \leq U(f,P')$$

and

$L(f,P') \leq \inf\{U(f,P)|P$ a partition of $[a,b]\} \leq U(f,P')$,

for all partitions P' of $[a,b]$.

It may very well happen that

$$\sup\{L(f,P)|P \text{ a partition of } [a,b]\} =$$

$$\inf\{U(f,P)|P \text{ a partition of } [a,b]\};$$

in this case, that is the *only* number between the lower sum and upper sum of f for all partitions and this number is consequently an ideal candidate for the area of $G(f;a,b)$. On the other hand, if

$$\sup\{L(f,P)|P \text{ a partition of } [a,b]\} <$$

$$\inf\{U(f,P)|P \text{ a partition of } [a,b]\},$$

then every number x between $\sup\{L(f,P)|P$ a partition of $[a,b]\}$ and $\inf\{U(f,P)|P$ a partition of $[a,b]\}$ will satisfy

$$L(f,P') \leq x \leq U(f,P')$$

for all partitions P' of $[a,b]$.

It is not at all clear just when such an embarrassment of riches will occur. The following two examples show that both phenomena are possible.

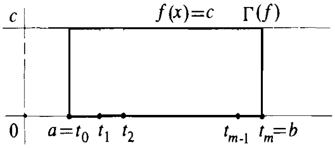
4. Example. Let $f : [a,b] \to \mathbb{R}$ be a constant function, i.e., $f(x) = c$ for all $x \in [a,b]$. (See *Figure* 5.) If $P = \{t_0, t_1, \quad t_{m-1}, t_m\}$ is any partition of $[a,b]$, then $m_i = M_i = c$ for all $i = 0, 1, \quad , m-1$ and so

$$L(f,P) = \sum_{i=0}^{m-1} c(t_{i+1} - t_i) = c(b-a)$$

and

$$U(f,P) = \sum_{i=0}^{m-1} c(t_{i+1} - t_i) = c(b-a).$$

Figure 5.



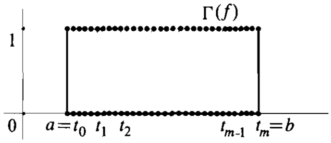


Figure 6.

Therefore

$$\sup\{L(f,P)|P \text{ a partition of } [a,b]\} =$$

$$\inf\{U(f,P)|P \text{ a partition of } [a,b]\} = c(b-a).$$

5. Example. Let $f:[a,b]\to\mathbb{R}$ be the function defined by (see *Figure* 6)

$$f(x)=\begin{cases}0, & \text{if } x \text{ is irrational,}\\ 1 & \text{if } x \text{ is rational.}\end{cases}$$

If $P=\{t_0,t_1, \quad t_{m-1},t_m\}$ is any partition of $[a,b]$, then $m_i=0$, since there is a irrational number in $[t_i,t_{i+1}]$ and $M_i=1$ since there is an rational number in $[t_i,t_{i+1}]$.

Therefore $L(f,P)=\sum_{i=0}^{m-1}0(t_{i+1}-t_i)=0$ and $U(f,P)=\sum_{i=0}^{m-1}1(t_{i+1}-t_i)=(b-a)$; hence $0=\sup\{L(f,P)|P$ a partition of $[a,b]\}\neq\inf\{U(f,P)|P$ a partition of $[a,b]\}$ $=(b-a)$.

The principle upon which the definition of area was to be based provides insufficient information to determine a specific area for $G(f;a,b)$ – any number between 0 and $b-a$ seems equally good. On the other hand, region $G(f;a,b)$ is so weird that we might with justice refuse to assign it any area at all. In fact, we can maintain more generally that whenever $\sup\{L(f,P)|P$ a partition of $[a,b]\}\neq\inf\{U(f,P)|P$ a partition of $[a,b]\}$, the region $G(f;a,b)$ is too unreasonable to deserve having an area. As our appeal to the word 'unreasonable' suggests, we are about to cloak our ignorance in terminology.

6. Definition. A function $f:[a,b]\to\mathbb{R}$ which is bounded is said to be *Riemann-integrable* or R-*integrable* or just *integrable* on $[a,b]$ if $\sup\{L(f,P)|P$ a partition of $[a,b]\}=\inf\{U(f,P)|P$ a partition of $[a,b]\}$. In this case, the common number is called the *Riemann-integral* or just *integral* of f on $[a,b]$ and is denoted by

$$\int_a^b f.$$

Riemann integrals are also called *definite integrals.*

(The symbol $\int$ is called an *integral sign* and was originally an elongated s, for 'sum'; the numbers a and b are called the *lower* and *upper limits of integration.*)

In future, for brevity, we shall say that a function is *integrable* on a closed interval, rather than *Riemann-integrable* and speak of its *integral* instead of its *Riemann integral* It should be borne in mind however that there are other integration processes than that of Riemann, and for these processes our results may or may not be true. For example, the most commonly used integral after that of Riemann is that of Lebesgue. A given real-valued function on $[a, b]$ may or may not be Lebesgue integrable; if it is, then its Lebesgue integral is a certain real number. If a function is Riemann integrable then it is also Lebesgue integrable and the two integrals are the same (and hence the notation $\int_a^b f(x)dx$). But many functions that are not Riemann integrable are Lebesgue integrable and therefore the Lebesgue integral can be of greater use. For example, the function of Example 5 is Lebesgue integrable, but not Riemann integrable ; as a matter of fact its Lebesgue integral is 0, in line with the fact that in some sense the points of the interval $[a, b]$ that are rational are relatively few in comparison with those that are not.

For a partition $P = \{t_0, t_1, \ldots, t_{m-1}, t_m\}$ of $[a, b]$, the maximum $max\{\Delta t_i = (t_{i+1} - t_i) \mid i = 0, 1, \ \ldots, m-1\}$ is called the *norm* or *mesh* of P. So, to say that Δt_i tends to 0 is equivalent to saying that the mesh of P tends to 0. Therefore we may say that *a Riemann integral is the limit of a Riemann sum as the mesh of the partition of $[a, b]$ tends to* 0.

The integral $\int_a^b f$ is also called the *area* of $G(f; a, b)$ when $f(x) \geq 0$ for all $x \in [a, b]$.

Therefore, if $f : [a, b] \to \mathbb{R}$ is integrable, then by definition,

$$L(f, P) \leq \int_a^b f \leq U(f, P) \quad \text{for every partition } P \text{ of } [a, b].$$

Moreover, $\int_a^b f$ is the *unique* number with this property.

We do not know which functions are integrable nor do we know how to find the integral of f on $[a, b]$ when f is integrable. We know that the constant function

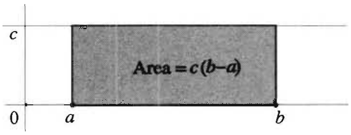


Figure 7.

$f(x) = c$ is integrable on $[a, b]$ and $\int_a^b f = c(b-a)$ (note that this integral assigns the expected area to a rectangle. See *Figure* 7.)

The following useful simple criterion for integrability is nothing more than a restatement of the definition of integrability. It concerns not the character of the function f but the nature of lower and upper sums. Nevertheless, it is a very convenient restatement because there is no mention of sup's and inf's which are often difficult to work with.

7. Theorem. *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. Then f is integrable on $[a, b]$ if and only if for every $\varepsilon > 0$ there exists a partition P of $[a, b]$ such that*

$$U(f, P) - L(f, P) < \varepsilon.$$

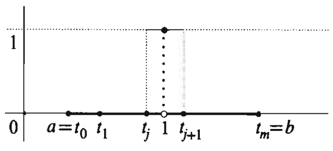
The next example illustrates the point mentioned above and also serves as a good introduction to the type of reasoning which the complicated definition of the integral necessitates, even in very simple situations.

8. Example. Let $f : [0, 2] \to \mathbb{R}$ be the function defined by (see *Figure* 8)

$$f(x) = \begin{cases} 0, & \text{if } x \neq 1, \\ 1 & \text{if } x = 1. \end{cases}$$

Let $P = \{t_0, t_1, \ldots, t_{m-1}, t_m\}$ be a partition of $[a, b]$ with $t_j < 1 < t_{j+1}$. Then $m_i = M_i = 0$ if $i \neq j$, but $m_j = 0$ and $M_j = 1$. Therefore $L(f, P) = \sum_{i=0}^{m-1} m_i(t_{i+1} - t_i) = 0$ and $U(f, P) = \sum_{i=0}^{m-1} M_i(t_{i+1} - t_i) = M_j(t_{j+1} - t_j) = t_{j+1} - t_j$ and so $U(f, P) - L(f, P) = t_{j+1} - t_j$.

Figure 8.



Therefore to show that f is integrable it is only necessary to choose a partition with $t_j < 1 < t_{j+1}$ and $t_{j+1} - t_j < \varepsilon$. Moreover, it is clear that $L(f, P) \leq 0 \leq U(f, P)$ for every partition P of $[0, 2]$. Since f is integrable, there is only one number between all lower and upper sums, namely, the integral of f, therefore $\int_0^2 f = 0$.

Although the discontinuity of f at $x = 1$ was responsible for the difficulties in the above example, worse problems arise even for very simple continuous functions. See Examples 9 and 10 below.

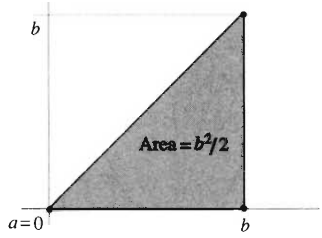


Figure 9.

9. Example. Let $f : [0, b] \to \mathbb{R}$ be the identity function, i.e., $f(x) = x$ for all $x \in [0, b]$. (See *Figure* 9.) If $P = \{t_0, t_1, \ldots, t_{m-1}, t_m\}$ is any partition of $[0, b]$, then $m_i = t_i$ and $M_i = t_{i+1}$ for all $i = 0, 1, \ldots, m-1$. (See *Figure* 10.) Therefore

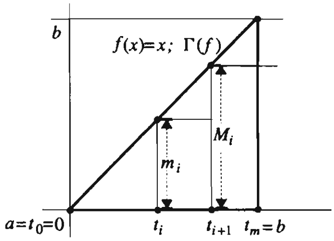


Figure 10.

$$L(f, P) = \sum_{i=0}^{m-1} t_i(t_{i+1} - t_i) = t_0(t_1 - t_0) + t_1(t_2 - t_1) + \cdots + t_{m-1}(t_m - t_{m-1}),$$

$$U(f, P) = \sum_{i=0}^{m-1} t_{i+1}(t_{i+1} - t_i) = t_1(t_1 - t_0) + t_2(t_2 - t_1) + \cdots + t_m(t_m - t_{m-1}).$$

Neither of these formulas is particularly appealing, but both simplify considerably for partitions $P_m = \{t_0, t_1, \ldots, t_{m-1}, t_m\}$ into m *equal* subintervals, i.e. the length $t_{i+1} - t_i$ of each subinterval $[t_{i+1}, t_i]$ is b/m. Therefore in this case $t_0 = 0, t_1 = b/m, t_2 = 2b/m, \ldots, t_i = ib/m, \ldots, t_{m-1} = (m-1)b/m, t_m = b$ and so

$$L(f, P_n) = \sum_{i=0}^{m-1} t_i(t_{i+1} - t_i) = \sum_{i=0}^{m-1} \frac{ib}{m} \cdot \frac{b}{m} = \left(\sum_{i=0}^{m-1} i\right) \frac{b^2}{m^2} = \frac{(m-1)m}{2} \cdot \frac{b^2}{m^2} = \frac{m-1}{m} \cdot \frac{b^2}{2}.$$

Remember the formula $1 + 2 + \ldots + k = \dfrac{k(k+1)}{2}$.

Similarly,

$$U(f, P_n) = \sum_{i=0}^{m-1} t_{i+1}(t_{i+1} - t_i) = \left(\sum_{i=0}^{m-1} (i+1)\right) \frac{b^2}{m^2} = \frac{m(m+1)}{2} \cdot \frac{b^2}{m^2} = \frac{m+1}{m} \cdot \frac{b^2}{2}.$$

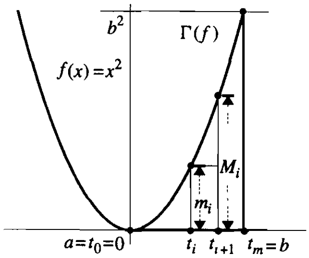


Figure 11.

If m is very large, then both $L(f,P_m)$ and $U(f,P_m)$ are close to $b^2/2$ and this makes it easy to show that f is integrable, for the equality

$$U(f,P_m)-L(f,P_m)=\frac{2}{m}\ \frac{b^2}{2},$$

and hence there are partitions P_m of $[0,b]$ with $U(f,P_m)-L(f,P_m)$ as small as desired. Therefore the function f is integrable. Moreover, $\int_0^b f$ may now be found with only little work : It is clear that

$$L(f,P_m)\le\frac{b^2}{2}\le U(f,P_m)\quad\text{for all } m.$$

This inequality shows only that $b^2/2$ lies between certain special upper and lower sums, but we have seen that the difference $U(f,P_m)-L(f,P_m)$ can be made as small as desired and so there is only one number with this property, namely the integral $\int_0^b f$ and hence we conclude that

$$\int_0^b f=\frac{b^2}{2}.$$

Note that this equation assigns area $b^2/2$ to the right-angled triangle with base and altitude b. Using more involved calculations or by using Theorem 12, it can be shown that

$$\int_a^b f=\frac{b^2}{2}-\frac{a^2}{2}.$$

10. Example. Let $f:[0,b]\to\mathbb{R}$ be the function defined by $f(x)=x^2$ for all $x\in[0,b]$. (See *Figure* 11.) If $P=\{t_0,t_1,\quad t_{m-1},t_m\}$ is any partition of $[0,b]$, then $m_i=f(t_i)=t_i^2$ and $M_i=f(t_{i+1})=(t_{i+1})^2$ for all $i=0,1,\quad,m-1$. Choosing once again a partition $P_m=\{t_0,t_1,\quad,t_{m-1},t_m\}$ into m equal subintervals, the lower and upper sums become

$$L(f,P_n)=\sum_{i=0}^{m-1}(t_i)^2(t_{i+1}-t_i)=\left(\sum_{i=0}^{m-1}i^2\right)\frac{b^2}{m^2}\ \frac{b}{m}=$$

Recall the formula $1^2+2^2+\ldots+k^2=\frac{k(k+1)(k+2)}{6}$.

$$\frac{(m-1)m(2m-1)}{6}\ \frac{b^3}{m^3}.$$

Similarly,

$$U(f,P_n)=\sum_{i=0}^{m-1}(t_{i+1})^2(t_{i+1}-t_i)=\left(\sum_{i=0}^{m-1}(i+1)^2\right)\frac{b^2}{m^2}\cdot\frac{b}{m}$$

$$=\frac{m(m+1)(2m+1)}{6}\ \frac{b^3}{m^3}.$$

Now it is not too hard to show that $L(f,P_m)\le\frac{b^3}{3}\le U(f,P_m)$ for all m and that the difference $U(f,P_m)-L(f,P_m)$ can be made as small as desired by choosing m sufficiently large. The same sort of reasoning as before then shows that (see *Figure* 12)

$$\int_0^b f=\frac{b^3}{3}.$$

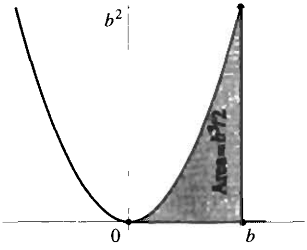


Figure 12.

Note that this calculation already represents a nontrivial result – the area of the region bounded by a parabola is not usually derived in elementary geometry. Nevertheless, the result was known to Archimedes, who derived it in essentially the same way.

Using more involved calculations, it can be shown that (for example by using Theorem 12)

$$\int_a^b f=\frac{b^3}{3}-\frac{a^3}{3}.$$

Some of our investigations are summarized as follows :

$$\int_a^b f=c(b-a)\qquad\text{if } f(x)=c \text{ for all } x\in[a,b],$$

$$\int_a^b f=\frac{b^2}{2}-\frac{a^2}{2}\qquad\text{if } f(x)=x \text{ for all } x\in[a,b],$$

$$\int_a^b f=\frac{b^3}{3}-\frac{a^3}{3}\qquad\text{if } f(x)=x^2 \text{ for all } x\in[a,b].$$

This list already reveals that the notation $\int_a^b f$ is not convenient. For this reason an alternative notation[1] analogous to the notation $\lim\limits_{x\to a} f(x)$ is very useful. Therefore we use

$$\int_a^b f(x)dx \qquad \text{instead of} \qquad \int_a^b f.$$

With this notation we have the formulas:

$$\int_a^b cdx = c(b-a), \quad \int_a^b xdx = \frac{b^2}{2}-\frac{a^2}{2}, \quad \int_a^b x^2dx = \frac{b^3}{3}-\frac{a^3}{3}$$

Note that as in the notation $\displaystyle\int_a^b f(x)dx$, the symbol x can be replaced by any other letter (except f, a, or b, of course) ;

$$\int_a^b f(x)dx = \int_a^b f(t)dt = \int_a^b f(y)dy = \int_a^b f(\alpha)d\alpha.$$

The symbol dx has no meaning in isolation, any more than the symbol $x \to a$ has any meaning, except in the context $\lim\limits_{x\to a} f(x)$. In the equation

$$\int_a^b x^2dx = \frac{b^3}{3} - \frac{a^3}{3},$$

the entire symbol x^2dx may be regarded as an abbreviation for:

the function f such that $f(x) = x^2$ for all $x \in [a,b]$.

The computations of $\displaystyle\int_a^b xdx$ and $\displaystyle\int_a^b x^2dx$ may suggest that evaluating integrals is generally difficult or impossible. As a matter of fact, the integrals of most functions are impossible to determine exactly (*although they may be computed to any degree of accuracy desired by calculating lower and upper sums*). Nevertheless, the integral of many functions can be computed very easily.

[1] The notation $\displaystyle\int_a^b f(x)dx$ is actually very old and was for many years the only symbol for the integral. Leibniz used this symbol because he considered the integral to be the sum (denoted by $\int$) of infinitely many rectangles with height $f(x)$ and "infinitely small" (or "infinitesimal") width dx. Later writers used x_0, x_1,..., x_m to denote the points of a partition and abbreviated $x_{i+1}-x_i$ by Δx_i. The integral was defined as the limit as Δx_i approaches 0 of the sums $\sum_{i=0}^{m-1} f(x_i)\Delta x_i$ (analogous to lower and upper sums). The fact that the limit is obtained by changing Σ to $\int$, $f(x_i)$ to $f(x)$ and Δx_i to dx, delights many people.

Even though most integrals cannot be computed exactly, it is of interest and important at least to know when a function f is integrable, i.e., to give necessary and sufficient conditions for integrability that concerns the character of the function f directly and are not simply reflected in the properties of lower and upper sums based on it.

It is clear that in the foregoing definitions (see for example, Definition 6) it was essential that f be bounded, for otherwise the set of lower sums could have no supremum or the set of upper sums have no infimum. Boundedness, therefore, is a necessary condition for integrability. Example 5 shows however, that it is not sufficient.

In another vein, it may be shown that continuity of f, although not necessary (see Example 8) to the existence of a integral, is nevertheless sufficient. Therefore *every continuous function on* $[a, b]$ *is integrable.*

Effectively integrability places a limitation on the set of points at which the function may be discontinuous. To make this statement more precise, first we introduce the concept of a *zero set or a set of measure zero*; a subset $A \subseteq \mathbb{R}$ is called a *zero set* if for every each $\varepsilon > 0$, there exists a sequence I_ν, $\nu \in \mathbb{N}^*$ of open intervals which covers A, i.e. $A = \cup_{\nu=1}^{\infty} I_\nu$ and is such that $\sum_{\nu=1}^{\infty} \ell(I_\nu) < \varepsilon$, where $\ell(I)$ denotes the length of the interval I. It immediately follows that any countable subset of real numbers is a zero set. However, we must not leap to the conclusion that every zero set is countable as the Cantor's perfect set[2] is an example of a non-countable zero set. Historically, it was the first set of this character to be constructed. With this one can prove that: *If a function* $f : [a, b] \to \mathbb{R}$ *is integrable, then the set of discontinuities of* f *is a zero set.*

It is quite clear that the possession of a zero set of discontinuities is not sufficient to integrablity since unbounded

[2] The *Cantor's perfect set* is the intersection $\bigcap_{n=0}^{\infty} F_n$ of a collection of closed sets F_n, $n = 0,1,2,\ldots$ where for each n, the set F_n is the union of 2^n closed intervals, each of length $1/3^n$.

sets exist with this property. However, boundedness and possession of a zero set of discontinuities are both necessary to the existence of an integral, but separately neither is sufficient. What is interesting is that together these two properties do suffice to assure integrability. With these concepts and results we now state a criterion for integrability: *A function $f : [a, b] \to \mathbb{R}$ is integrable if and only if f is bounded and the set of discontinuities of f is a zero set.*

Although the class of continuous functions provides many integrable functions, it will be more satisfying to have a somewhat larger supply of integrable functions. For this first note that the set $\mathcal{R}([a, b])$ of integrable functions on $[a, b]$ is a vector space over the field $\mathbb{R}$ of real numbers, i.e. *the sum of two integrable functions and a scalar multiplication of an integrable function are again integrable.* Moreover, $\mathcal{R}$ $([a, b])$ is a subspace of the vector space $\mathcal{C}_{\mathbb{R}}([a, b])$ of the continuous real-valued functions on $[a, b]$. Further, it is worthwhile to mention the *linearity property: the function $f \mapsto \int_a^b f$ is a linear functional on the vector space $\mathcal{R}([a, b])$.*

Proofs of the above theorems usually use Theorem 7; as some of our previous demonstrations illustrate, the details of the argument often conspire to obscure the point of the proof. It is a good idea to attempt proofs of your own, this will probably clarify the proofs and will certainly give good practice in the techniques used in some problems.

As a simple application of these theorems, recall that if f is 0 except at one point, where its value is 1 (see Example 8), then f is integrable. Multiplying this function by a constant c, it follows that the same is true if the value of f at the exceptional point is c. Adding such function to an integrable function, we see that the value of an integrable function may be changed arbitrarily at one point without destroying integrablity. By breaking up the interval into many subintervals, we see that the value can be changed at finitely many points.

So far we have acquired only one complicated definition, a few simple theorems with intricate proofs. This is not because integrals constitute a more difficult topic than derivatives. The most significant discovery of calculus is the fact that the integral and the derivative are intimately related – once we learn this connection, the integral will become as useful as derivative and as easy to use. The striking connection between derivatives and integrals is given by Theorem 11 – known as the *fundamental theorem of calculus* which we state here without proof.

A function $F : [a,b] \to \mathbb{R}$ is called a *primitive* or an *antiderivative* of a function $f : [a,b] \to \mathbb{R}$ if F is differentiable on $[a,b]$ and if the derivative $F'(x) = f(x)$ for all $x \in [a,b]$. It is clear, of course, that if a primitive of a function exists, then that primitive is continuous on $[a,b]$ since it is differentiable on $[a,b]$. It is easily seen, too, that a primitive is not unique, for if F is a primitive of f, then so is any function that differs from F by a constant. Conversely, if F_1 and F_2 are primitives of f, then $(F_1 - F_2)' = F_1' - F_2' = 0$, whence F_1 and F_2 differ by a constant. Since a differentiable function may have a discontinuous derivative, it follows that a discontinuous function may possess a primitive.

11. Example. For the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(x) = \begin{cases} x^{4/3}\sin(1/x), & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases}$$

it is easy to show that

$$h'(x) = \begin{cases} \frac{4}{3}x^{1/3}\sin(1/x) - x^{-2/3}\cos(1/x), & \text{if } \ x \neq 0, \\ 0 & \text{if } \ x = 0. \end{cases}$$

Further, since $h'(1/2n\pi) = -(2n\pi)^{2/3}$ for any integer n, it follows that h' is unbounded and hence is not integrable on any interval containing 0. However, a primitive, that is, h exists.

Suppose now that the function $f : [a,b] \to \mathbb{R}$ is integrable on $[a,b]$. We can define a new function $F : [a,b] \to \mathbb{R}$ by

$$F(x) = \int_a^x f = \int_a^x f(t)dt.$$

We have seen that f may be integrable even if it is not continuous. The behavior of F is therefore a very pleasant surprise.

For a continuous function $f : [a,b] \to \mathbb{R}$, the function $x \to \int_a^x f$ is an antiderivative of f on $[a,b]$. In symbols, for all $x \in [a,b]$, we have $\frac{d}{dx}\left(\int_a^x f(t)dt\right) = f(x)$.

12. Theorem. (The Fundamental Theorem of Calculus.) *Let $f : [a,b] \to \mathbb{R}$ be an integrable function on $[a,b]$ and that $F : [a,b] \to \mathbb{R}$ be defined by*

$$F(x) = \int_a^x f.$$

Then F is continuous on $[a,b]$. Moreover, if f is continuous on $[a,b]$, then F is differentiable on $[a,b]$ and $F'(x) = f(x)$ for all $x \in [a,b]$, i.e. *F is a primitive of f on $[a,b]$.* (If $c = a$ or b, then $F'(c)$ is understood to mean the right- or left-hand derivative of F.)

The Fundamental Theorem of Calculus (Theorem 12) establishes a link between derivatives and integrals. This link is used so heavily that it gives the convenient but conceptually wrong impression that integration is just reverse of differentiation. Also, the problem of finding an antiderivative of a given function becomes very important. Therefore, the antiderivatives began to be (and to some extent still are) popularly called *indefinite integrals* and written using the integral sign without lower and upper limits. For example, $\int x^2dx = x^3/3$.

The following simple Corollary to Theorem 12 frequently reduces computations of integrals to a triviality.

13. Corollary *Let $f : [a,b] \to \mathbb{R}$ be a continuous function on $[a,b]$ and that $f = g'$ for some function $g : [a,b] \to \mathbb{R}$. Then*

$$\int_a^b f = g(b) - g(a).$$

Proof. Let $F(x) := \int_a^x f$. Then $F' = f = g'$ on $[a,b]$ and hence $F = g + c$ for some constant number c. The number c can be evaluated easily : $0 = F(a) = g(a) + c$, i.e., $c = -g(a)$; and hence $F(x) = g(x) - g(a)$. In particular, putting $x = b$, we get $\int_a^b f = F(b) = g(b) - g(a)$.□

At first sight, the Corollary 13 seem useless : after all, what good is it to know that $\int_a^b f = g(b) - g(a)$ if g is, for example, $g(x) = \int_a^x f$? The point, of course, is that one might happen to know a quite different function g with this property. For example :

Corollary 13 is called the *first form* of the Fundamental Theorem of Calculus. A definite integral is very rarely evaluated by actually computing all its Riemann sums. The most common method of evaluating definite integrals is based on Corollary 13.

14. Example. If $g(x) = x^3/3$ and $f(x) = x^2$, then $g'(x) = f(x)$. Therefore we obtain (by Corollary 13), without ever computing lower and upper sums

$$\int_a^b x^2 dx = b^3/3 - a^3/3.$$

Other powers are treated similarly ; if n is a natural number and $g(x) = x^n/(n+1)$, then $g'(x) = x^n$ and so

$$\int_a^b x^n dx = b^{n+1}/(n+1) - a^{n+1}/(n+1).$$

For any natural number n, the function $f(x) = x^{-n}$ is not bounded on any interval containing 0, but if a and b are both positive or both negative, then

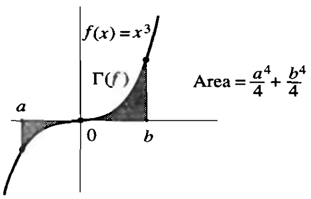
$$\int_a^b x^{-n} dx = b^{-n+1}/(-n+1) - a^{-n+1}/(-n+1).$$

Naturally this formula is only true for $n \neq -1$. *We do not know a simple expression for the integral* $\int_a^b \frac{1}{x} dx$. The problem of computing this integral provides a good opportunity to warn against a serious error.

The following example shows that the integral does not always represent the area bounded by the graph of the function, the horizonal axis and the vertical lines passing through the points $(a, 0)$ and $(b, 0)$.

15. Example. Let $a < 0 < b$ be real numbers. Then the integral $\int_a^b x^3 dx$ does not represent the area of the region shown in *Figure* 13 which is given instead by

Figure 13.



$$-\left(\int_a^0 x^3dx\right)+\int_0^b x^3dx = -\left(\frac{0^4}{4}-\frac{a^4}{4}\right)+\left(\frac{b^4}{4}-\frac{0^4}{4}\right) = \frac{a^4}{4}+\frac{b^4}{4}.$$

The conclusion of the above Corollary 13 is often confused with the definition of integrals – many students think that $\int_a^b f$ is defined as: '$g(b) - g(a)$, where g is a function whose derivative is f'. This 'definition' is wrong ; one reason is that a function f may be integrable without being the derivative of another function. For example, if $f(x) = 0$ for $x \neq 1$ and $f(1) = 1$, then f is integrable, but f cannot be a derivative (why not?). There is also another reason that is much more important : if f is continuous, then $f = g'$ for some function; but we know this *only because of* Theorem 12. The function $f(x) = 1/x$ provides an excellent illustration: if $x > 0$, then $f(x) = g'(x)$, where $g(x) = \int_0^x \frac{1}{t}dt$, and we know of no simpler function g with this property.

The following somewhat stronger result than Corollary 13 is still true. The proof, however, must be entirely different.

16. Theorem. *Let $f : [a,b] \to \mathbb{R}$ be a function on $[a,b]$. If f is integrable on $[a,b]$ and that $f = g'$ for some function $g : [a,b] \to \mathbb{R}$, Then*

$$\int_a^b f = g(b) - g(a).$$

It should be noted that in the statement of Theorem 16 it is not redundant to require that f be integrable and possess a primitive. It is quite possible for a function to be integrable and yet not have a primitive. Moreover, a function which has a primitive need not be integrable. The above Examples 8 and 11 attest to the accuracy of these statements.

Suggested Reading

[1] Tom M Apostol, *Mathematical Analysis*, Addition-Wesley Publishing, World Student Series Edition, 1973.

[2] R Courant, Differential and Integral Calculus, Vol.I, second edition, 1937; Vol.II, first edition, Wiley (Interscience) New York, 1936.

[3] Walter Rudin, *Principles of Mathematical Analysis*, Third Edition, McGraw-Hill International Editions, 1976.

Address for Correspondence
D P Patil
Department of Mathematics
Indian Institute of Science
Bangalore 560 012, India.
Email:
patil@math.iisc.ernet.in