6-2011

Maxwell's Equations, Part II

David W. Ball
Cleveland State University, d.ball@csuohio.edu

Recommended Citation
Ball, David W. 2011. "Maxwell's Equations, Part II." Spectroscopy 26, no. 6: 14-21

Maxwell's Equations

$$\text{II.}\quad \nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

This is the second part of a multi-part production on Maxwell's equations of electromagnetism. The ultimate goal is a definitive explanation of these four equations; readers will be left to judge how definitive it is. A note: for certain reasons, figures are being numbered sequentially throughout this series, which is why the first figure in this column is numbered 8. I hope this does not cause confusion. Another note: this is going to get a bit mathematical. It can't be helped: models of the physical universe, like Newton's second law F = ma*, are based in math. So are Maxwell's equations.*

Enter Stage Left: Maxwell

James Clerk Maxwell (Figure 8) was born in 1831 in Edinburgh, Scotland. His unusual middle name derives from his uncle, who was the 6th Baronet Clerk of Penicuik (pronounced "penny-cook"), a town not far from Edinburgh. Clerk was, in fact, the original family name; Maxwell's father, John Clerk, adopted the surname Maxwell after receiving a substantial inheritance from a family named Maxwell. By most accounts, James Clerk Maxwell (hereafter referred to as simply "Maxwell") was an intelligent but relatively unaccomplished student.

He began blossoming in his early teens, however, becoming interested in mathematics (especially geometry). He eventually attended the University of Edinburgh and, later, Cambridge University, where he graduated in 1854 with a degree in mathematics. He stayed on for a few years as a Fellow, then moved to Marischal College in Aberdeen. When Marischal merged with another college to form the University of Aberdeen in 1860, Maxwell was laid off (an action for which U. of A. should still be kicking themselves, but who can foretell the future?) and he found another position at King's College London (later the University of London). He returned to Scotland in 1865, only to go back to Cambridge in 1871 as the first Cavendish Professor of Physics. He died of abdominal cancer in November 1879 at the relatively young age of 48; curiously, his mother died of the same ailment and at the same age, in 1839.

Though he had a relatively short career, Maxwell was very productive. He made contributions to color theory and optics (indeed, the first photo in Figure 8 shows Maxwell holding a color wheel of his own invention) and actually produced the first true color photograph as a composite of three images. He made major contributions to the development of the kinetic molecular theory of gases, as the "Maxwell-Boltzmann distribution" is named partially after him. He also made major contributions to thermodynamics, deriving the relations that are named after him and devising a thought experiment about entropy that was eventually called "Maxwell's demon." He demonstrated mathematically that the rings of Saturn could not be solid, but must instead be composed of relatively tiny (relative to Saturn, of course) particles – a hypothesis that was supported spectroscopically in the late 1800s but finally directly observed the first time when the Pioneer 11 and Voyager 1 spacecraft passed through the Saturnian system in the early 1980s (Figure 9).

Maxwell also made seminal contributions to the understanding of electricity and magnetism, concisely summarizing their behaviors with four mathematical expressions known as Maxwell's equations of electromagnetism. He was strongly influenced by Faraday's experimental work, believing that any theoretical description of a phenomenon must be grounded in phenomenological observations. Maxwell's equations essentially summarize everything about classical electrodynamics, magnetism, and optics, and were only supplanted when relativity and quantum mechanics revised our understanding of the natural universe at certain of its limits. Far away from those limits, in the realm of classical physics, Maxwell's equations still rule just as Newton's equations of motion rule under normal conditions.

A Calculus Primer

Maxwell's laws are written in the language of calculus. Before we move forward with an explicit discussion of the first law, here we deviate to a review of calculus and its symbols.

Calculus is the mathematical study of change. Its modern form was developed independently by Isaac Newton and German mathematician Gottfried Leibnitz in the late 1600s. Although Newton's version was used heavily in his influential *Principia Mathematica* (in which Newton used calculus to express a number of fundamental laws of nature), it is Leibnitz's notations that are commonly used today. An understanding of calculus is fundamental to most scientific and engineering disciplines.

Consider a car moving at constant velocity. Its distance from an initial point (arbitrarily set as a position of 0) can be plotted as a graph of distance from zero versus time elapsed. Commonly, the elapsed time is called the independent variable and is plotted on the *x* axis of a graph (called the abscissa) while distance traveled from the initial position is plotted on the *y* axis of the graph (called the ordinate). Such a graph is plotted in Figure 10. The slope of the line is a measure of how much the ordinate changes as the abscissa changes; that is, slope *m* is defined as

$$m = \frac{\Delta y}{\Delta x}$$

For the straight line shown in Figure 10, the slope is constant, so *m* has a single value for the entire plot. This concept gives rise to the general formula for any straight line in two dimensions, which is

$$y = mx + b$$

where y is the value of the ordinate, x is the value of the abscissa, m is the slope, and b is the y-intercept, which is where the plot would intersect with the y axis. Figure 10 shows a plot that has a positive value of m. A plot with a negative value of m; it would be going down, not up, as you go from left to right. A horizontal line has a value of 0 for m; a vertical line has a slope of infinity.

Many lines are not straight. Rather, they are curves. Figure 11 gives an example of a plot that is curved. The slope of a curved line is more difficult to define than that of a straight line because the slope is changing. That is, the value of the slope depends on the point (x, y) of the curve you're at. The slope of a curve is the same as the slope of the straight line that is tangent to the curve at that point (x, y). Figure 11 shows the slopes at two different points. Because the slopes of the straight lines tangent to the curve at different points are different, the slopes of the curve itself at those two points are different.

Calculus provides ways of determining the slope of a curve, in any number of dimensions (Figure 11 is a two-dimensional plot, but we recognize that functions can be functions of more than one variable, so plots can have more dimensions [a.k.a. variables] than two). We have already seen that the slope of a curve varies with position. That means that the slope of a curve is not a *constant*; rather, it is a *function* itself. We are not concerned about the methods of determining the functions for the slopes of curves here; that information can be found in a calculus text. Here, we are concerned with how they are represented.

The word that calculus uses for the slope of a function is *derivative*. The derivative of a straight line is simply m, its constant slope. Recall that we mathematically defined the slope m above using "Δ" symbols, where Δ is the Greek capital letter delta. Δ is used generally to represent "change", as in ΔT (change in temperature) or Δy (change in y coordinate). For straight lines and other simple changes, the change is definite; in other words, it has a specific value.

In a curve, the change Δy is different for any given Δx because the slope of the curve is constantly changing. Thus, it is not proper to refer to a definite change because – to overuse a word – the definite change changes during the course of the change. What we have to do is a thought experiment: we have to imagine that the change is infinitesimally small over both the x and y coordinates. This way, the actual change is confined to an infinitesimally small portion of the curve: a *point*, not a distance. The point involved is the point at which the straight line is tangent to the curve (Figure 11).

Rather than using "Δ" to represent an infinitesimal change, calculus starts by using "d". Rather than using m to represent the slope, calculus puts a prime on the dependent variable as a way to represent a slope (which, remember, is a function and not a constant). Thus, for a curve we have for the slope y′:

$$y' = \frac{dy}{dx}$$

as our definition for the slope of that curve.

We hinted earlier that functions may depend on more than one variable. If that is the case, how do we define the slope? First, we define a *partial derivative* as the derivative of a multi-variable function with respect to only one of its variables. We assume that the other variables are held constant. Instead of using a "d" to indicate a partial derivative, we use the lowercase Greek delta "δ". It is also common to explicitly list the variables being held constant as subscripts to the derivative, although this can be omitted because it is understood that a partial derivative is a one-dimensional derivative. Thus we have

$$f_x' = \left(\frac{\partial f}{\partial x}\right)_{y,z,\ldots} = \frac{\partial f}{\partial x}$$

spoken as "the partial derivative of the function $f(x,y,z,\ldots)$ with respect to x." Graphically, this corresponds to the slope of the multi-variable function f in the x dimension, as shown in Figure 12.

The *total derivative* of a function, df, is the sum of the partial derivatives in each dimension; that is, with respect to each variable individually. For a function of three variables, $f(x,y,z)$, the total derivative is written as

$$\mathrm{d}f = \frac{\partial f}{\partial x}\mathrm{d}x + \frac{\partial f}{\partial y}\mathrm{d}y + \frac{\partial f}{\partial z}\mathrm{d}z$$

where each partial derivative is the slope with respect to each individual variable and dx, dy, and dz are the finite changes in the x, y, and z directions. The total derivative has as many terms as the overall function has variables. If a function is based in three-dimensional space, as is commonly the case for physical observables, then there are three variables and so three terms in the total derivative.

When a function typically generates a single numerical value that is dependent on all of its variables, it is called a scalar function. An example of a scalar function might be

$$F(x,y) = 2x - y^2$$

According to this definition, $F(4,2) = 2 \cdot 4 - 2^2 = 8 - 4 = 4$. The final value of $F(x,y)$, 4, is a scalar: it has magnitude but no direction.

A *vector function* is a function that determines a vector, which is a quantity that has magnitude *and* direction. Vector functions can be easily expressed using unit vectors, which are vectors of length 1 along each dimension of the space involved. It is customary to use the representations **i**, **j**, and **k** to represent the unit vectors in the x, y, and z dimensions, respectively

(Figure 13). Vectors are typically represented in print as boldfaced letters. Any random vector can be expressed as, or decomposed into, a certain number of **i** vectors, **j** vectors, and **k** vectors as is demonstrated in Figure 13. A vector function might be as simple as

$$\mathbf{F} = x\mathbf{i} + y\mathbf{j}$$

in two dimensions, which is illustrated in Figure 14 for a few discrete points. Although only a few discrete points are shown in Figure 14, understand that the vector function is continuous. That is, it has a value at every point in the graph.

One of the functions of a vector that we will have to evaluate is called a *dot product*. The dot product between two vectors **a** and **b** is represented and defined as

$$\mathbf{a}\cdot\mathbf{b} = |a||b|\cos\theta$$

where $|a|$ represents the magnitude (that is, length) of **a**, $|b|$ is the magnitude of **b**, and $\cos\theta$ is the cosine of the angle between the two vectors. The dot product is sometimes called the scalar product because the value is a scalar, not a vector. The dot product can be thought of physically as how much one vector contributes to the direction of the other vector, as shown in Figure 15. A fundamental definition that uses the dot product is that for work, w, which is defined in terms of the force vector **F** and the displacement vector of a moving object, **s**, and the angle between these two vectors:

$$w = \mathbf{F}\cdot\mathbf{s} = |F||s|\cos\theta$$

Thus, if the two vectors are parallel ($\theta = 0^{\circ}$ so $\cos\theta = 1$) the work is maximized, but if the two vectors are perpendicular to each other ($\theta = 00^{\circ}$ so $\cos\theta = 0$), the object does not move and no work is done (Figure 16).

More Advanced Stuff

We have already discussed the derivative, which is a determination of the slope of a function (straight or curved). The other fundamental operation in calculus is *integration*, whose representation is called an *integral*. It is represent as

$$\int_b^a f(x)\,dx$$

where the symbol ∫ is called the *integral sign* and represents the integration operation, $f(x)$ is called the *integrand* and is the function to be integrated, dx is the *infinitesimal* of the dimension of the function, and a and b are the limits between which the integral is numerically evaluated, if it is to be numerically evaluated. (If the integral sign looks like an elongated "s", it should – Leibniz, one of the co-founders of calculus [with Newton], adopted it in 1675 to represent "sum", since an integral is a limit of a sum.) A statement called the fundamental theorem of calculus establishes that integration and differentiation are the opposites of each other, a concept that allows us to calculate the numerical value of an integral. For details of the fundamental theorem of calculus, consult a calculus text. For our purposes, all we need to know is that the two are related and calculable.

The most simple geometric representation of an integral is that it represents the area under the curve given by $f(x)$ between the limits a and b and bound by the x-axis. Look, for example, at Figure 17(a). It is a figure of the line $y = x$ or, in more general terms, $f(x) = x$. What is the area under this function but above the x-axis, shaded gray in Figure 17(a)? Simple geometry indicates that the area is ½ units – the box defined by $x = 1$ and $y = 1$ is 1 unit (1×1), and the right triangle that is shaded gray is one-half of that total area, or ½ unit in area. Integration of the function $f(x) = x$ gives us the same answer. The rules of integration will not be discussed here; it is assumed that the reader can perform simple integration:

$$\int_0^1 f(x)\,dx = \int_0^1 x\,dx = \tfrac{1}{2}x^2\Big|_0^1 = \tfrac{1}{2}(1)^2 - \tfrac{1}{2}(0)^2 = \tfrac{1}{2} - 0 = \tfrac{1}{2}$$

It is a bit messier if the function is more complicated. But, as first demonstrated by Reimann in the 1850s, the area can be calculated geometrically for any function in one variable (most easy to visualize, but in theory this can be extended to any number of dimensions) by using rectangles of progressively narrower widths, until the area becomes a limiting value as the number of rectangles goes to infinity and the width of each rectangle gets infinitely narrow – one reason a good calculus course begins with a study of infinite sums and limits! But I digress. For the function in Figure 17(b), which is $f(x) = x^2$, the area under the curve, now poorly approximated by the shaded triangle, is calculated exactly with an integral:

$$\int_0^1 f(x)\,dx = \int_0^1 x^2\,dx = \tfrac{1}{3}x^3\Big|_0^1 = \tfrac{1}{3} - 0 = \tfrac{1}{3}$$

As with differentiation, integration can also be extended to functions of more than one variable. The issue to understand is that when considering functions, the space you need to use has one more dimension than variables, because the function needs to be plotted in its own dimension. Thus, a plot of a one-variable function requires two dimensions, one to represent the variable and one to represent the value of the function. Figures 10 and 11, thus, are two-dimensional plots. A two-variable function needs to be plotted or visualized in three dimensions, like Figures 12 or 13. Looking at the two-variable function in Figure 18, we see a line across the function's values, with its projection in the (x,y) plane. The line on the surface is parallel to the y axis, so it is showing the trend of the function only as the variable x changes. If we were to integrate this multivariable function with respect only to (in this case) x, we would be evaluating the integral only along this line. Such an integral is called a *line integral*. One interpretation of this integral would be that it is simply the part of the volume under the overall surface that is beneath the given line; that is, it is the area under the line.

If the surface represented in Figure 18 represents a field (either scalar or vector), then the line integral represents the total effect of that field along the given line. The formula for calculating the "total effect" might be unusual, but it makes sense if we start from the beginning. Consider a path whose position is defined by an equation *P*, which is a function of one or more variables. What is the distance of the path? One way of calculating the distance *s* is velocity *v* times time *t*, or

$$s = v \times t$$

But velocity is the derivative of position *P* with respect to time, or dP/dt. Let us represent this derivative as P'. Our equation becomes

$$s = P' \times t$$

This is for finite values of distance and time, and for that matter, for constant P'. (Example: total distance at 2.0 m/s for 4.0 s = 2.0 m/s × 4.0 s = 8.0 m. In this example, P' is 2.0 m/s and *t* is 4.0 s.) For infinitesimal values of distance and time, and for a path whose value may be a function of the variable of interest (in this case, time), the infinitesimal form is

$$ds = P'dt$$

To find the total distance, we integrate between the limits of the initial position *a* and the final position *b*:

$$s = \int_b^a P'dt$$

The point is, it's not the path *P* we need to determine the line integral – it's the change in *P*, denoted as P'. This seems counterintuitive at first, but hopefully the above example makes the point. It's also a bit overkill when one remembers that derivatives and integrals are opposites of each other: the above analysis has us determine a derivative and then take the integral, undoing our original operation, to get the answer. One might have just kept the original equation and determined answer from there. We'll address this issue shortly. One more point: it doesn't have to be a change with respect to time. The derivative involved can be a change with respect to a spatial variable. This allows us to determine line integrals with respect to space as well as time.

Suppose the function for the path *P* is a vector? For example, consider a circle **C** in the (x,y) plane having radius r. Its vector function is $\mathbf{C} = r\cos\theta\mathbf{i} + r\sin\theta\mathbf{j} + 0\mathbf{k}$ (see Figure 19), which is a function of the variable θ, the angle from the positive x axis. What is the circumference of the circle; that is, what is the path length as θ goes from 0 to 2π, the radian measure of the central angle of a circle? According to our formulation above, we need to determine the derivative of our function. But for a vector, if we want the total length of the path, we care only about the *magnitude* of the vector and not its direction. Thus, we'll need to derive the change in the magnitude of the vector. We start by defining the magnitude: the magnitude $|m|$ of a three- (or lesser-) magnitude vector is the Pythagorean combination of its components:

$$|m| = \sqrt{x^2 + y^2 + z^2}$$

For the derivative of the path/magnitude with respect to time, which is the velocity, we have

$$|m'| = \sqrt{(x')^2 + (y')^2 + (z')^2}$$

For our circle, we have the magnitude as simply the **i**, **j**, and/or **k** terms of the vector. These individual terms are also functions of θ. We have:

$$x' = \frac{d(r\cos\theta\mathbf{i})}{d\theta} = -r\sin\theta\mathbf{i}$$

$$y' = \frac{d(r\sin\theta\mathbf{j})}{d\theta} = r\cos\theta\mathbf{j}$$

$$z' = 0$$

From this we have

$$(x')^2 = r^2 \sin^2 \theta \mathbf{i}^2$$

$$(y') = r^2 \cos^2 \theta \mathbf{j}^2$$

(and we will ignore the z part, since it's just zero). For the squares of the unit vectors, we have $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{i}\cdot\mathbf{i} = \mathbf{j}\cdot\mathbf{j} = 1$. Thus, we have

$$s = \int_b^a P'\,dt = \int_0^{2\pi} \left(\sqrt{r^2 \sin^2 \theta + r^2 \cos^2 \theta}\right) d\theta$$

We can factor out the r^2 term from each term and then out of the square root to get

$$s = \int_0^{2\pi} r\sqrt{\sin^2 \theta + \cos^2 \theta}\, d\theta$$

Since, from elementary trigonometry, $\sin^2\theta + \cos^2\theta$ equals 1, we have

$$s = \int_0^{2\pi} (r\sqrt{1})\, d\theta = \int_0^{2\pi} r\, d\theta = r \cdot \theta\Big|_0^{2\pi} = r(2\pi - 0) = 2\pi r$$

This seems like an awful lot of work to show what we all know, that the circumference of a circle is $2\pi r$. But hopefully it will convince you of the propriety of this particular mathematical formulation.

Back to "total effect". For a line integral involving a field, there are two expressions we need to consider: the definition of the field $F[x(q),y(q),z(q)]$ and the definition of the vector path $\mathbf{p}(q)$, where q represents the coordinate along the path. (Note that at least initially, the field F is not necessarily a vector.) In that case, the total effect s of the field along the line is given by

$$s = \int_{\mathbf{p}} F\left[x(q), y(q), z(q)\right] \cdot \left|\mathbf{p}'(q)\right| dq$$

The integration is over the path $\mathbf{p}$, which needs to be determined by the physical nature of the system in interest. Note that in the integrand, the two functions F and $|\mathbf{p}'|$ are multiplying together.

If F is a vector field over the vector path $\mathbf{p}(q)$, denoted $\mathbf{F}[\mathbf{p}(q)]$, then the line integral is defined similarly:

$$s = \int_{\mathbf{p}} \mathbf{F}\left[\mathbf{p}(q)\right] \cdot \mathbf{p}'(q)\, dq$$

Here, we need to take the dot product of the $\mathbf{F}$ and $\mathbf{p}'$ vectors.

A line integral is an integral over one dimension that gives, effectively, the area under the function. We can perform a two-dimensional integral over the surface of a multi-dimensional function, as pictured in Figure 20. That is, we want to evaluate the integral

$$\int_S g(x, y, z)\, dS$$

where $g(x,y,z)$ is some scalar function on a surface S. Technically, this expression is a double integral over two variables. This integral is generally called a *surface integral*.

The mathematical tactic for evaluating the surface integral is to project the functional value into the perpendicular plane, accounting for the variation of the function's angle with respect to the projected plane. The proper variation is the cosine function, which gives you a relative contribution of 1 if the function and the plane are parallel (i.e. $\cos 0^\circ = 1$) and a relative contribution of 0 if the function and the plane are perpendicular (i.e. $\cos 90^\circ = 0$). This automatically makes us think of a dot product. If the space S is being projected into the (x,y) plane, then the dot product will involve the unit vector in the z direction, or $\mathbf{k}$. (If the space is projected into other planes, other unit vectors are involved, but the concept is the same.) If $\mathbf{n}(x,y,z)$ is the unit vector that defines the line perpendicular to the plane marked out by $g(x,y,z)$ [called the *normal vector*], then the value of the surface integral is given by

$$\iint_R \frac{g(x,y,z)}{\mathbf{n}(x,y,z)\cdot\mathbf{k}}\,dx\,dy$$

where the denominator contains a dot product and the integration is over the *x* and *y* limits of the region *R* in the (*x*,*y*) plane of Figure 20. The dot product in the denominator is actually fairly easy to generalize. When that happens, the surface integral becomes

$$\iint_R g(x,y,z)\cdot\sqrt{1+\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2}\,dx\,dy$$

where *f* represents the function of the surface and *g* represents the function you are integrating over. Typically, to make *g* a function of only two variables, you let $z = f(x,y)$ and substitute the expression for *z* into the function *g*, if *z* appears in the function *g*.

If, instead of a scalar function *g* we had a vector function **F**, the above equation gets a bit more complicated. In particular, we are interested in the effect that is normal to the surface of the vector function. Since we previously defined **n** as the vector normal to the surface, we'll use it again: we want the surface integral involving $\mathbf{F}\cdot\mathbf{n}$, or

$$\int_S \mathbf{F}\cdot\mathbf{n}\,dS$$

For a vector function $\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}$ and a surface given by the expression $f(x,y) \equiv z$, this surface integral is

$$\int_S \mathbf{F}\cdot\mathbf{n}\,dS = \iint_R\left[-F_x\frac{\partial f}{\partial x} - F_y\frac{\partial f}{\partial y} + F_z\right]dx\,dy$$

This is a bit of a mess! Is there a better, easier, more concise way of representing this?

A Better Way

There is a better way to represent this last integral, but we need to back up a bit: what exactly is $\mathbf{F}\cdot\mathbf{n}$? Actually, it's just a dot product, but the integral

$$\int_S \mathbf{F} \cdot \mathbf{n}\, dS$$

is called the *flux* of **F**. The word “flux” comes from the Latin word *fluxus*, meaning “flow”. For example, suppose you have some water flowing through the end of a tube, as represented in Figure 21(a). If the tube is cut straight, the flow is easy to calculate from the velocity of the water (given by **F**) and the geometry of the tube. If you want to express the flow in terms of the mass of water flowing, you can use the density of the water as a conversion. But what if the tube isn’t cut straight, as shown in Figure 21(b)? In this case, we need to use some more complicated geometry – vector geometry – to determine the flux. In fact, the flux is calculated using the last integral in the previous section. So, flux is calculable.

Consider an ideal cubic surface with the sides parallel to the axes, as shown in Figure 22, that surrounds the point (x,y,z). This cube represents our function **F**, and we want to determine the flux of **F**. Ideally, the flux at any point can be determined by shrinking the cube until it gets to a single point. We will start by determining the flux for a finite-sized side, then take the limit of the flux as the size of the size goes to zero. If we look at the top surface, which is parallel to the (x,y) plane, it should be obvious that the normal vector is the same as the **k** vector. For this surface by itself, the flux is then

$$\int_S \mathbf{F} \cdot \mathbf{k}\, dS$$

If **F** is a vector function, its dot product with **k** eliminates the **i** and **j** parts (since $\mathbf{i}\cdot\mathbf{k} = \mathbf{j}\cdot\mathbf{k} = 0$) and only the z-component of **F** remains. Thus, the integral above is just

$$\int_S F_z\, dS$$

If we assume that the function F_z has some average value on that top surface, then the flux is simply that average value times the area of the surface, which we will propose is equal to $\Delta x \cdot \Delta y$.

We need to note, though, that the top surface isn't located at z (the center of the cube), but at $z + \Delta z/2$. Thus we have for the flux at the top surface:

$$\text{top flux} \approx F_z\left(x, y, z+\frac{\Delta z}{2}\right)\cdot \Delta x \Delta y$$

where the symbol ≈ means "approximately equal to." It will become "equal to" when the surface area shrinks to zero.

The flux of **F** on the bottom side is exactly the same but for two small changes. First, the normal vector is now –**k**, so there is a negative sign on the expression. Second, the bottom surface is lower than the center point, so the function is evaluated at $z - \Delta z/2$. Thus, we have

$$\text{bottom flux} \approx -F_z\left(x, y, z-\frac{\Delta z}{2}\right)\cdot \Delta x \Delta y$$

The total flux through these two parallel planes is the sum of the two expressions:

$$\text{flux} \approx F_z\left(x, y, z+\frac{\Delta z}{2}\right)\cdot \Delta x \Delta y - F_z\left(x, y, z-\frac{\Delta z}{2}\right)\cdot \Delta x \Delta y$$

We can factor the $\Delta x \Delta y$ out of both expressions. Now, if we multiply this expression by $\Delta z/\Delta z$ (which equals 1), we have

$$\text{flux} \approx \left[F_z\left(x, y, z+\frac{\Delta z}{2}\right) - F_z\left(x, y, z-\frac{\Delta z}{2}\right)\right]\cdot \Delta x \Delta y \frac{\Delta z}{\Delta z}$$

We rearrange:

$$\text{flux} \approx \frac{\left[F_z\left(x, y, z+\frac{\Delta z}{2}\right) - F_z\left(x, y, z-\frac{\Delta z}{2}\right)\right]}{\Delta z}\cdot \Delta x \Delta y \Delta z$$

and recognize that $\Delta x \Delta y \Delta z$ is the change in volume of the cube, ΔV:

$$\text{flux} \approx \frac{\left[F_z\left(x, y, z+\frac{\Delta z}{2}\right)-F_z\left(x, y, z-\frac{\Delta z}{2}\right)\right]}{\Delta z} \cdot \Delta V$$

As the cube shrinks, Δz approaches zero. In the limit of infinitesimal change in z, the first term in the product above is simply the definition of the derivative of F_z with respect to z! Of course, it's a partial derivative, because **F** depends on all three variables, but we can write the flux more simply as

$$\text{flux} = \frac{\partial F_z}{\partial z} \cdot \Delta V$$

A similar analysis can be performed for the two sets of parallel planes; only the dimension labels will change. We ultimately get

$$\text{total flux} = \frac{\partial F_x}{\partial x}\Delta V + \frac{\partial F_y}{\partial y}\Delta V + \frac{\partial F_z}{\partial z}\Delta V = \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}\right)\Delta V$$

(Of course, as Δx and Δy and Δz go to zero, so does ΔV, but this doesn't affect our end result.) The expression in the parentheses above is so useful that it is defined as the *divergence* of the vector function **F**:

$$\text{divergence of } \mathbf{F} \equiv \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} \quad (\text{where } \mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k})$$

Because divergence of a function is defined at a point and the flux, two equations above, is defined in terms of a finite volume, we can also define the divergence as the limit as volume goes to zero of the flux density (defined as flux divided by volume):

$$\text{divergence of } \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = \lim_{\Delta V \to 0} \frac{1}{\Delta V}(\text{total flux}) = \lim_{\Delta V \to 0} \frac{1}{\Delta V} \int_S \mathbf{F} \cdot \mathbf{n}\, dS$$

There are two abbreviations to indicate the divergence of a vector function. One is to simply use the abbreviation "div" to represent divergence:

$$\text{div }\mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$$

The other way to represent the divergence is with a special function. The function ∇ (called "del") is defined as

$$\nabla = \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}$$

If one were to take the dot product between ∇ and F, we would get the following result:

$$\nabla \cdot \mathbf{F} = \left(\mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}\right) \cdot \left(F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k}\right) = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$$

which is the divergence! Note that, although we expect to get nine terms in the dot product above, cross terms between the unit vectors (like $\mathbf{i}\cdot\mathbf{k}$ or $\mathbf{k}\cdot\mathbf{j}$) all equal zero and cancel out, while like terms (that is, $\mathbf{j}\cdot\mathbf{j}$) all equal 1 because the angle between a vector and itself is zero and cos 0 = 1. As such, our nine-term expansion collapses to only three non-zero terms. Alternately, one can think of the dot product in terms of its other definition

$$\mathbf{a}\cdot\mathbf{b} = \sum_i a_i b_i = a_1b_1 + a_2b_2 + a_3b_3$$

where a_1, a_2, etc., are the scalar magnitudes in the x, y, etc., directions. So, the divergence of a vector function $\mathbf{F}$ is indicated by

$$\text{divergence of }\mathbf{F} = \nabla\cdot\mathbf{F}$$

What does the divergence of a function mean? First, note that the divergence is a scalar, not a vector, field. No unit vectors remain in the expression for the divergence. This is not to imply that the divergence is a constant – it may in fact be a mathematical expression whose value varies in space. For example, for the field

$$\mathbf{F} = x^3\mathbf{i}$$

the divergence is

$$\nabla \cdot \mathbf{F} = 3x^2$$

which is a scalar function. Thus, the divergence changes with position.

Divergence is an indication of how quickly a vector field spreads out at any given point; that is, how fast it diverges. Consider the vector field

$$\mathbf{F} = x\mathbf{i} + y\mathbf{j}$$

which we originally showed in Figure 14 and are re-showing in Figure 23. It has a constant divergence of 2 (easily verified), indicating a constant "spreading out" over the plane. However, for the field

$$\mathbf{F} = x^2\mathbf{i}$$

whose divergence is $2x$, the vectors get farther and farther apart as x increases (see Figure 24).

Maxwell's First Equation

If two electric charges were placed in space near each other, as is shown in Figure 25, there would be a force of attraction between the two charges: the charge on the left would exert a force on the charge on the right, and vice versa. That experimental fact is modeled mathematically by Coulomb's law, which in vector form is:

$$\mathbf{F} = \frac{q_1 q_2}{r^2}\mathbf{r}$$

where q_1 and q_2 are the magnitudes of the charges (in elementary units, where the elementary unit is equal to the charge on the electron) and r is the scalar distance between the two charges. The unit vector $\mathbf{r}$ represents the line between the two charges q_1 and q_2. The modern version of Coulomb's law includes a conversion factor between charge units (coulombs, C) and force units (newtons, N), and is written as

$$\mathbf{F} = \frac{q_1 q_2}{4\pi\varepsilon_0 r^2}\mathbf{r}$$

where ε_0 is called the permittivity of free space and has an approximate value of $8.854\ldots\times10^{-12}$ $C^2/N{\cdot}m^2$.

How does a charge cause a force to be felt by another charge? Michael Faraday suggested that a charge had an effect in the surrounding space called an electric field, a vector field, labeled **E**. The electric field is defined as the Coulombic force felt by another charge divided by the magnitude of the original charge, which we will choose to be q_2:

$$\mathbf{E} = \frac{\mathbf{F}}{q_2} = \frac{q_1}{4\pi\varepsilon_0 r^2}\mathbf{r}$$

where in the second expression we have substituted the expression for **F**. Note that **E** is a vector field (as indicated by the bold-faced letter) and is dependent on the distance from the original charge. E too has a unit vector that is defined as the line between the two charges involved, but in this case the second charge has yet to be positioned, so in general E can be thought of as a spherical field about the charge q_1. The unit for an electric field is newton per coulomb, or N/C.

Since **E** is a field, we can pretend it has flux – that is, something is "flowing" through any surface that encloses the original charge. What is flowing? It doesn't matter; all that matters is that we can define the flux mathematically. In fact, we can use the definition of flux given earlier. The electric flux Φ is given by

$$\Phi = \int_S \mathbf{E}\cdot\mathbf{n}\,dS$$

which is perfectly analogous to our previous definition of flux.

Let us consider a spherical surface around our original charge that has some constant radius r. The normal unit vector **n** is simply **r**, the radius unit vector, since the radius unit vector is perpendicular to the spherical surface at any of its points (Figure 26). Since we know the definition of E from Coulomb's law, we can substitute into the expression for electric flux:

$$\Phi = \int_S \frac{q_1}{4\pi\varepsilon_0 r^2} \mathbf{r} \cdot \mathbf{r}\, dS$$

The dot product r·r is simply 1, so this becomes

$$\Phi = \int_S \frac{q_1}{4\pi\varepsilon_0 r^2}\, dS$$

If the charge q_1 is constant, 4 is constant, π is constant, the radius r is constant, and the permittivity of free space is constant, these can all be removed from the integral to get

$$\Phi = \frac{q_1}{4\pi\varepsilon_0 r^2} \int_S dS$$

What is this integral? Well, we defined our system as a sphere, so the surface integral above is the surface area of a sphere. The surface area of a sphere is known: $4\pi r^2$. Thus, we have

$$\Phi = \frac{q_1}{4\pi\varepsilon_0 r^2} \cdot 4\pi r^2$$

The 4, the π, and the r^2 terms cancel. We have left

$$\Phi = \frac{q_1}{\varepsilon_0}$$

Recall, however, that we previously defined the divergence of a vector function as

$$\text{div } \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = \lim_{\Delta V \to 0} \frac{1}{\Delta V} (\text{total flux}) = \lim_{\Delta V \to 0} \frac{1}{\Delta V} \int_S \mathbf{F} \cdot \mathbf{n} dS$$

Note that the integral in the definition has exactly the same form as the electric field flux Φ. Therefore, in terms of the divergence, we have for **E**:

$$\text{div } \mathbf{E} = \operatorname*{im}_{\Delta V \to 0} \frac{1}{\Delta V} \int_S \mathbf{E} \cdot \mathbf{n}\, dS = \lim_{\Delta V \to o} \frac{1}{\Delta V} \Phi = \lim_{\Delta V \to 0} \frac{1}{\Delta V} \frac{q_1}{\varepsilon_0}$$

where we have made the appropriate substitutions to get the final expression. We will rewrite this last expression as

$$\text{div } \mathbf{E} = \frac{\lim_{\Delta V \to o} \dfrac{q_1}{\Delta V}}{\varepsilon_0}$$

The expression $q_1/\Delta V$ is simply the charge density at a point, which we will define as ρ. This last expression becomes simply

$$\text{div } \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

This equation is *Maxwell's first equation of electromagnetism*. It is also written as

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

Maxwell's first equation is also called Gauss' law, after Carl Friedrich Gauss, the German polymath who first determined it but did not publish it. (It was finally published in 1867 after his death by his colleague William Weber; Gauss had a habit of not publishing much of his work, and his many contributions to science and mathematics were only realized posthumously.)

In the next installment, we will expand on our discussion by looking at Maxwell's second equation. In that case, we will be concerned with our old friend magnetism.

References

The following references were extremely helpful in constructing this presentation, along with many hours of researching on the world wide web.

B. Baigrie. *Electricity and Magnetism* Greenwood Press: Westport, CT, 2007.

D. Halliday, R. Resnick, and J. Walker. *Fundamentals of Physics* 6th edition. John Wiley and Sons: New York, 2001.

J. E. Marsden, A. J. Tromba *Vector Calculus* 2nd edition. W. H. Freeman and Company: 1981.

H. M. Schey. *Div, Grad, Curl, and All That: An Informal Text on Vector Calculus* 4th edition. W. W. Norton and Company: New York, 2005.

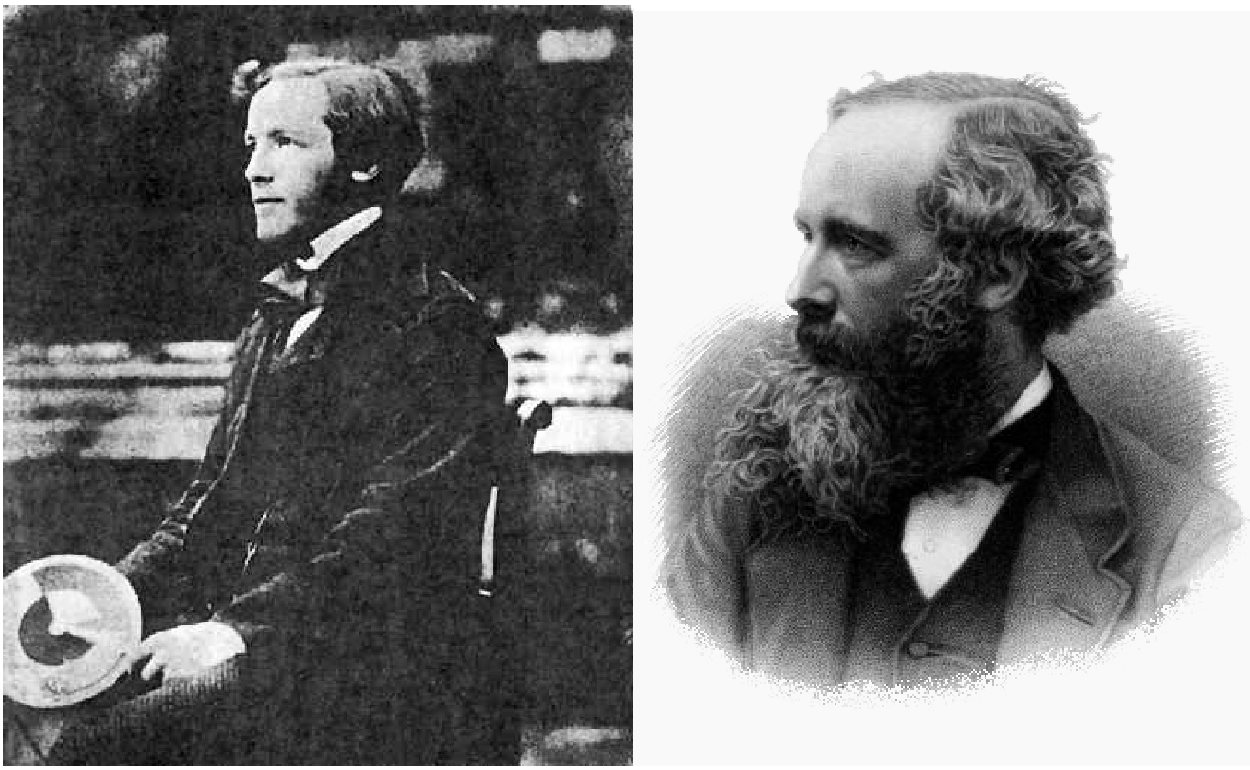


Figure 8. James Clerk Maxwell as a young man and an older man.

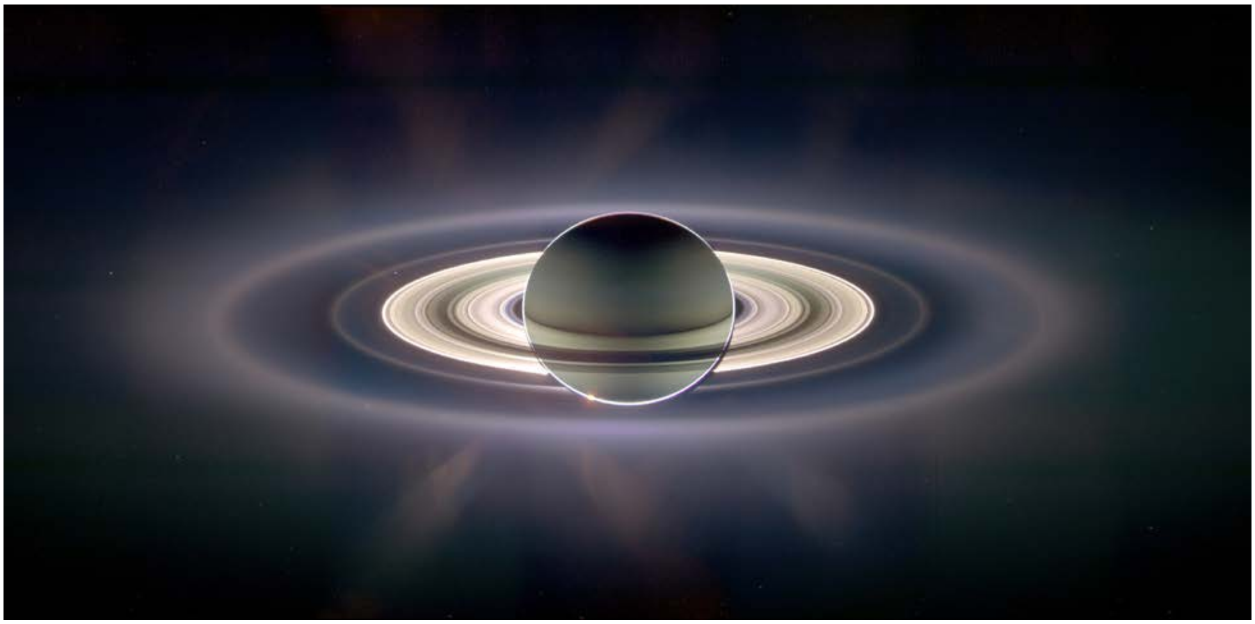


Figure 9. Maxwell proved mathematically that the rings of Saturn couldn't be solid objects, but were likely an agglomeration of smaller bodies. This image of a back-lit Saturn is a composite of several images taken by the Cassini spacecraft in 2006. Depending on the reproduction, you may be able to make out a tiny bluish dot in the 10 o'clock position just inside the second outermost diffuse ring – that's Earth.

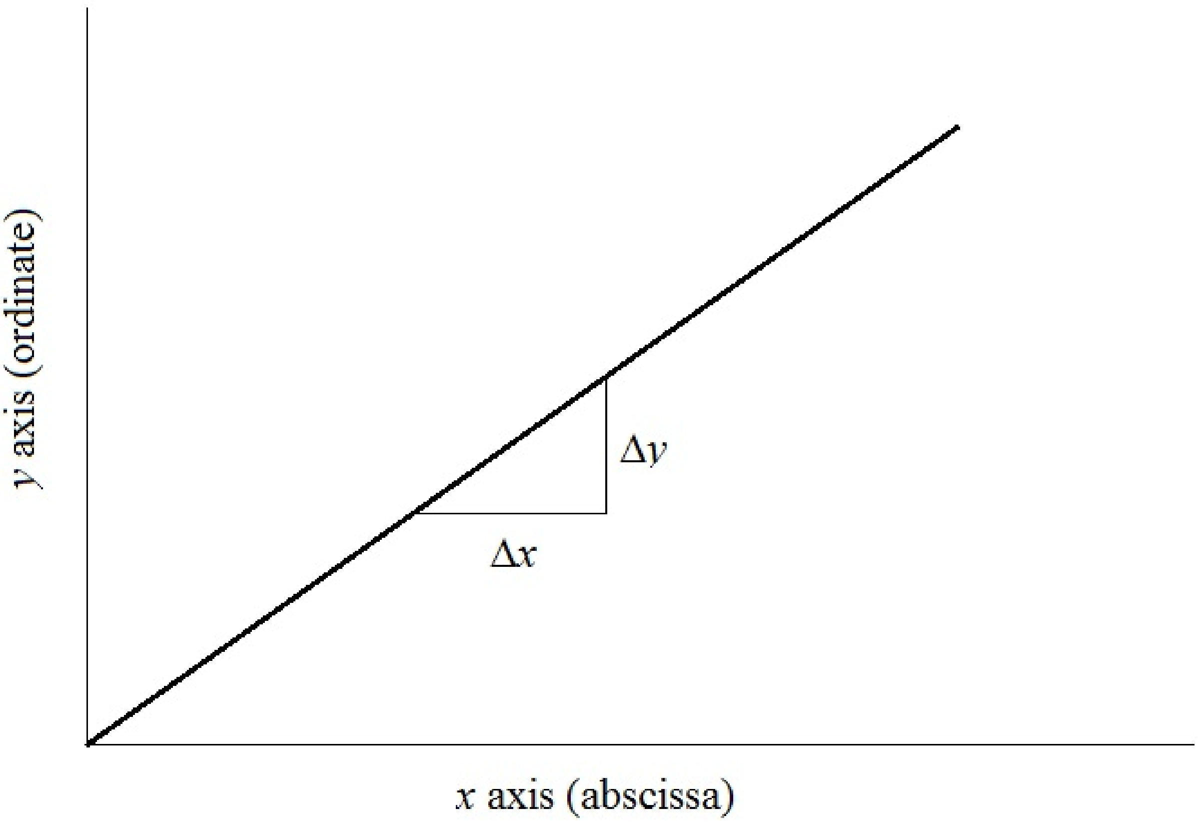


Figure 10. A plot of a straight line, which has a constant slope *m*, given by Δy/Δx.

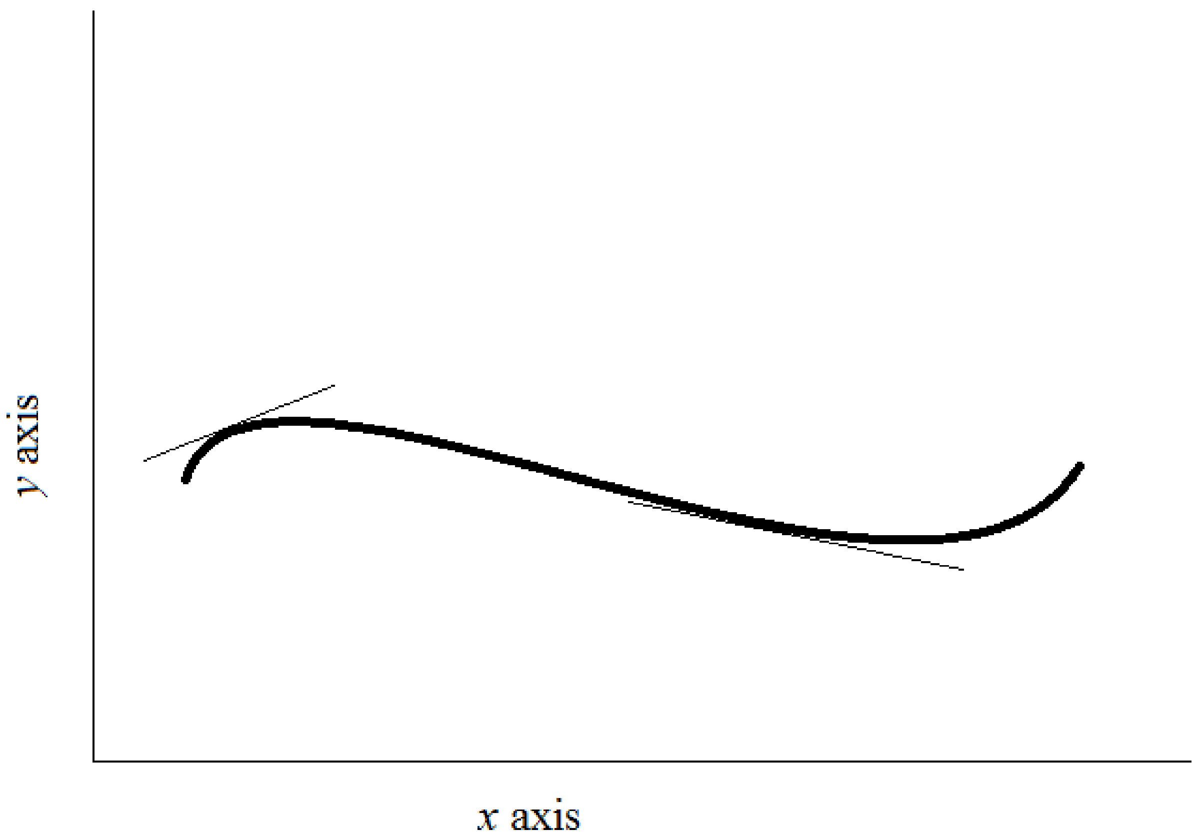


Figure 11. A plot of a curve, showing (with the thinner lines) the different slopes at two different points. Calculus helps us determine the slopes of curved lines.

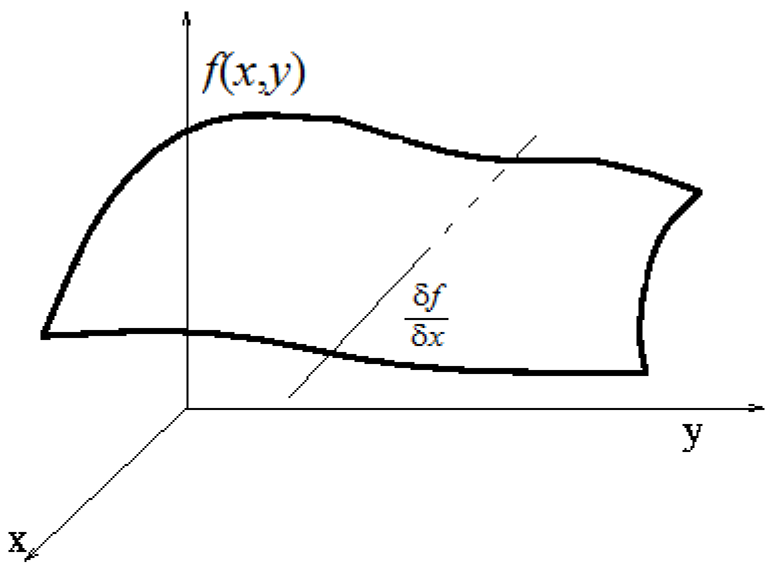


Figure 12. For a function of several variables, a partial derivative is a derivative in only one variable. The line represents the slope in the *x* direction.

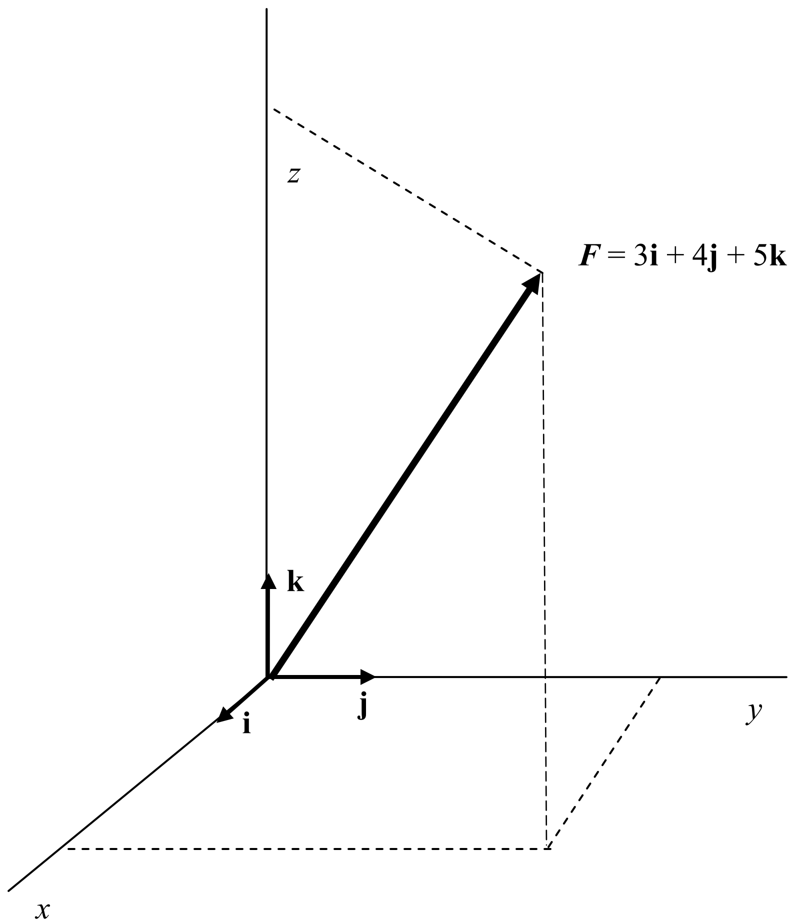


Figure 13. The definition of the unit vectors **i**, **j**, and **k**, and an example of how any vector can be expressed in terms of how many of each unit vector.

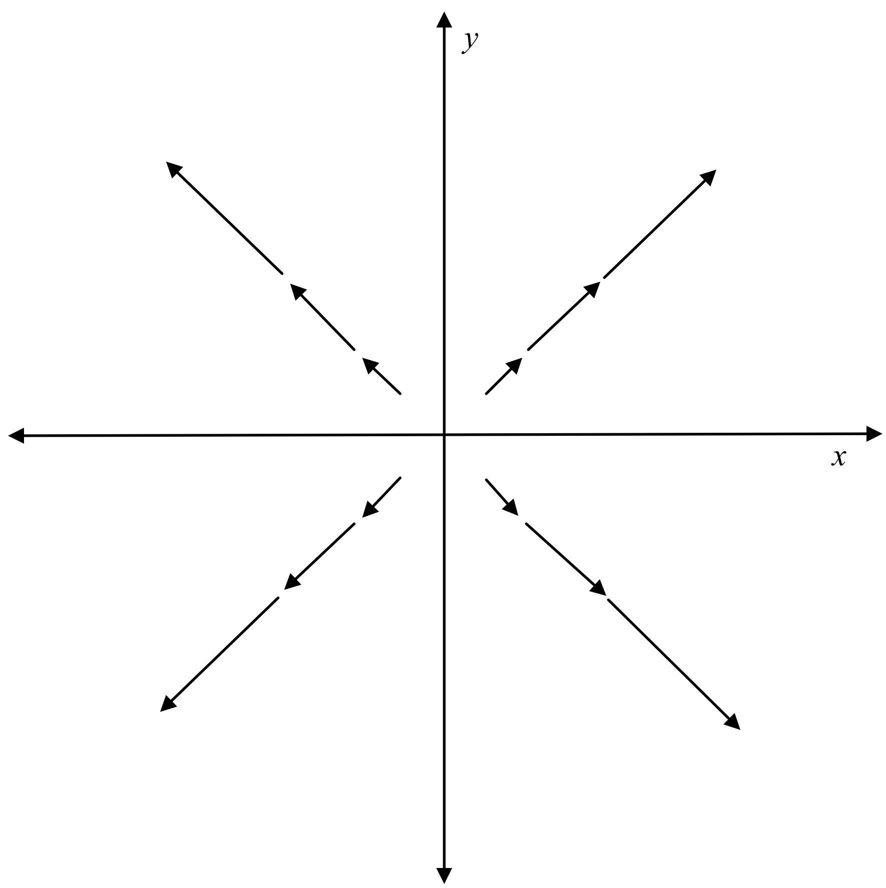


Figure 14. An example of a vector function $\mathbf{F} = x\mathbf{i} + y\mathbf{j}$. Each point in two dimensions defines a vector. Although only twelve individual values are illustrated here, in reality this vector function is a continuous, smooth function on both dimensions.

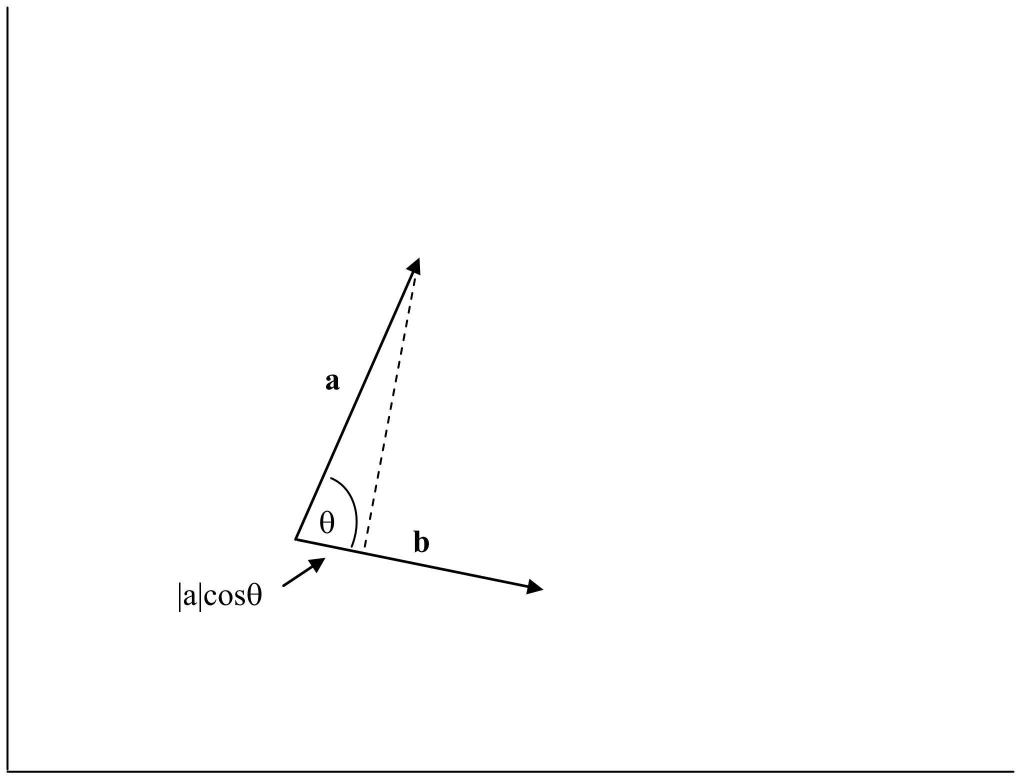


Figure 15. Graphical representation of the dot product of two vectors. The dot product gives the amount of one vector that contributes to the other vector. Understand that an equivalent graphical representation would have the **b** vector projected into the **a** vector. In both cases, the overall scalar results are the same.

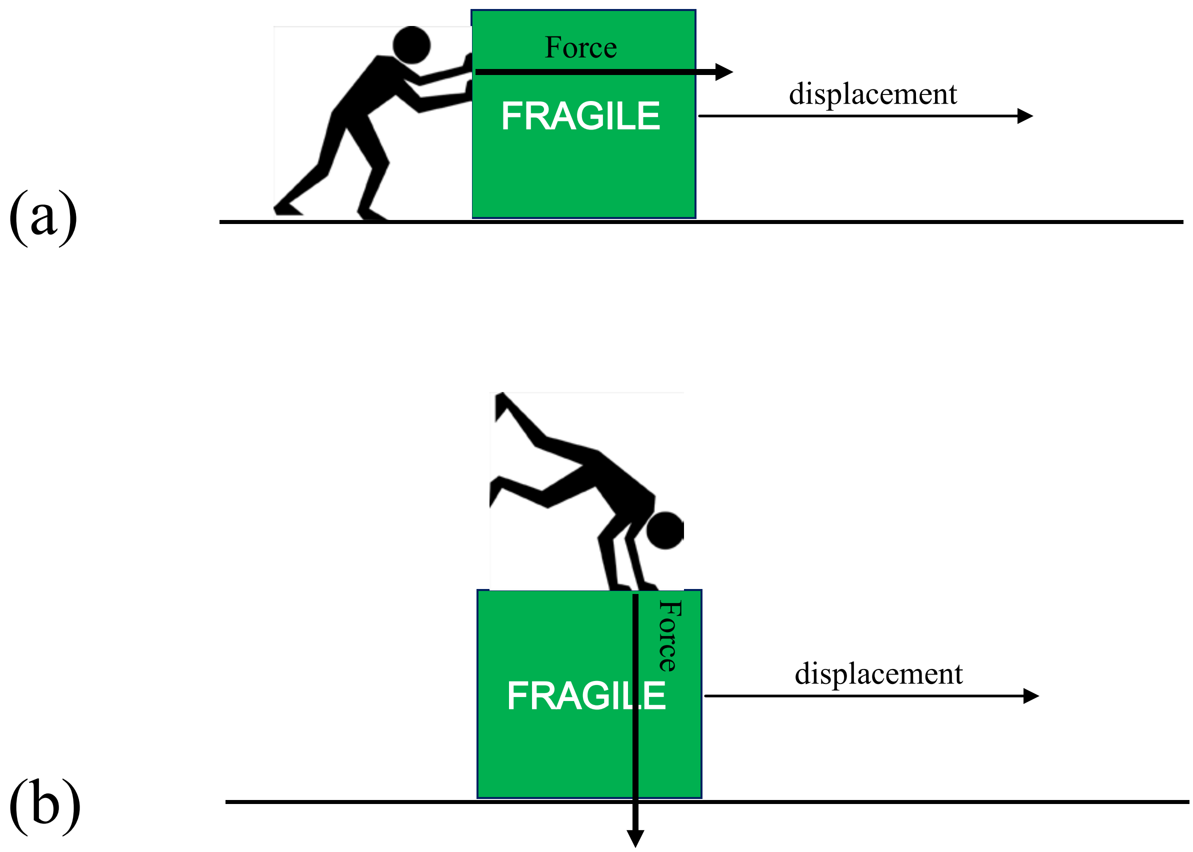
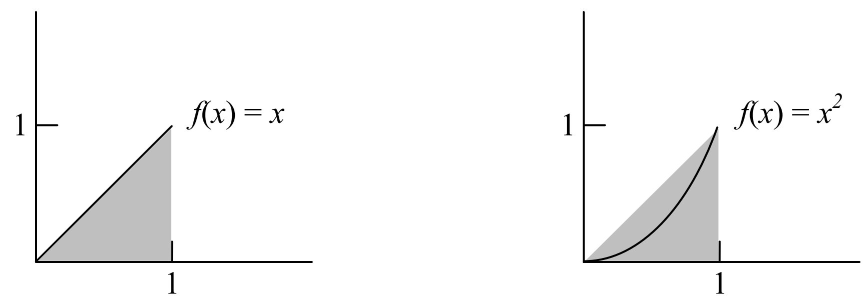


Figure 16. Work is defined as a dot product of a force vector and a displacement vector. (a) If the two vectors are parallel, they reinforce and work is performed. (b) If the two vectors are perpendicular, no work is performed.



(a) (b)

Figure 17. The geometric interpretation of a simple integral is the area under a function and bounded on the bottom by the x-axis (that is, $y = 0$). (a) For the function $f(x) = x$, the areas as calculated by geometry and integration are equal. (b) For the function $f(x) = x^2$, the approximation from geometry is not a good value for the area under the function. A series of rectangles can be used to approximate the area under the curve, but in the limit of an infinite number of infinitesimally-narrow rectangles, the area is equal to the integral.

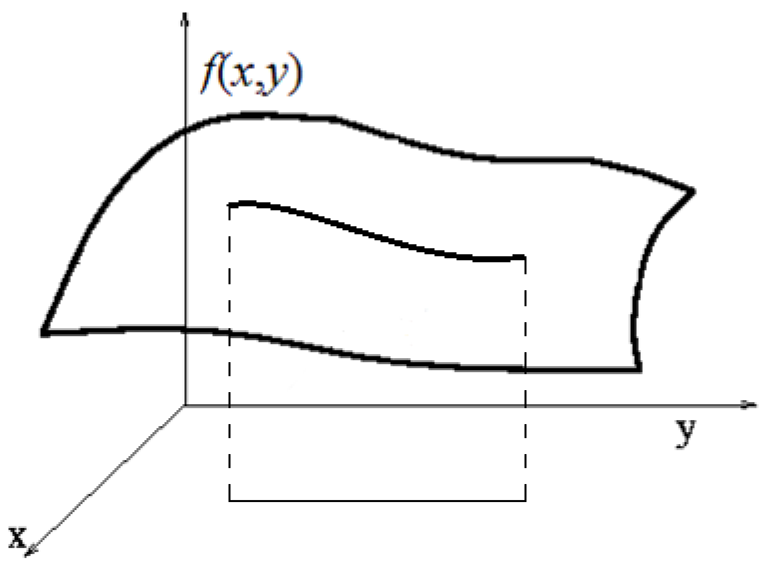


Figure 18. A multivariable function $f(x,y)$ with a line paralleling the y axis.

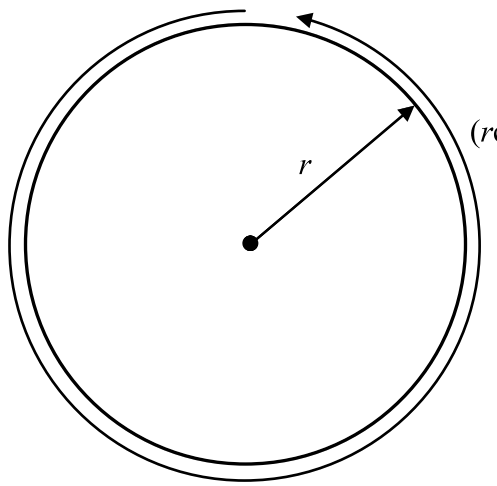


Figure 19. How far is the path around the circle? A line integral can tell us, and it agrees with what basic geometry predicts ($2\pi r$).

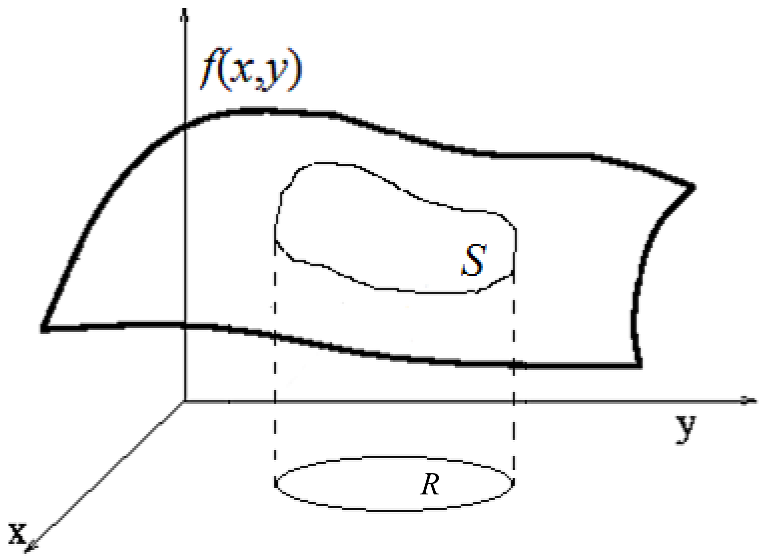


Figure 20. A surface S over which a function $f(x,y)$ will be integrated.

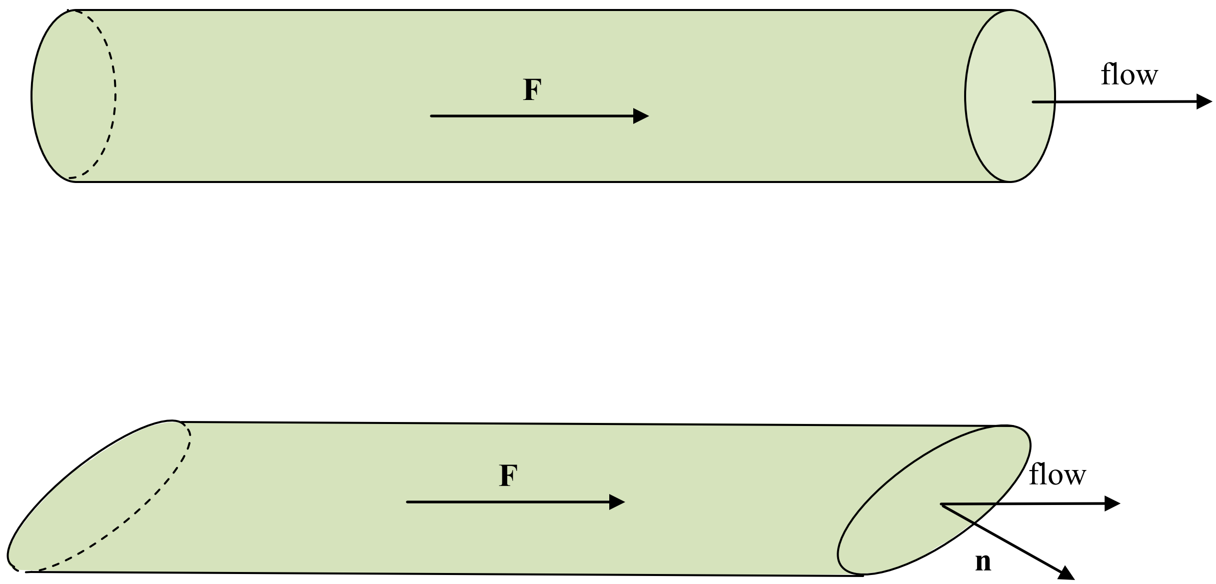


Figure 21. Flux is another word for amount of flow. (a) In a tube that is cut straight, the flux can be determined from simple geometry. (b) In a tube cut at an angle, some vector mathematics is needed to determine flux.

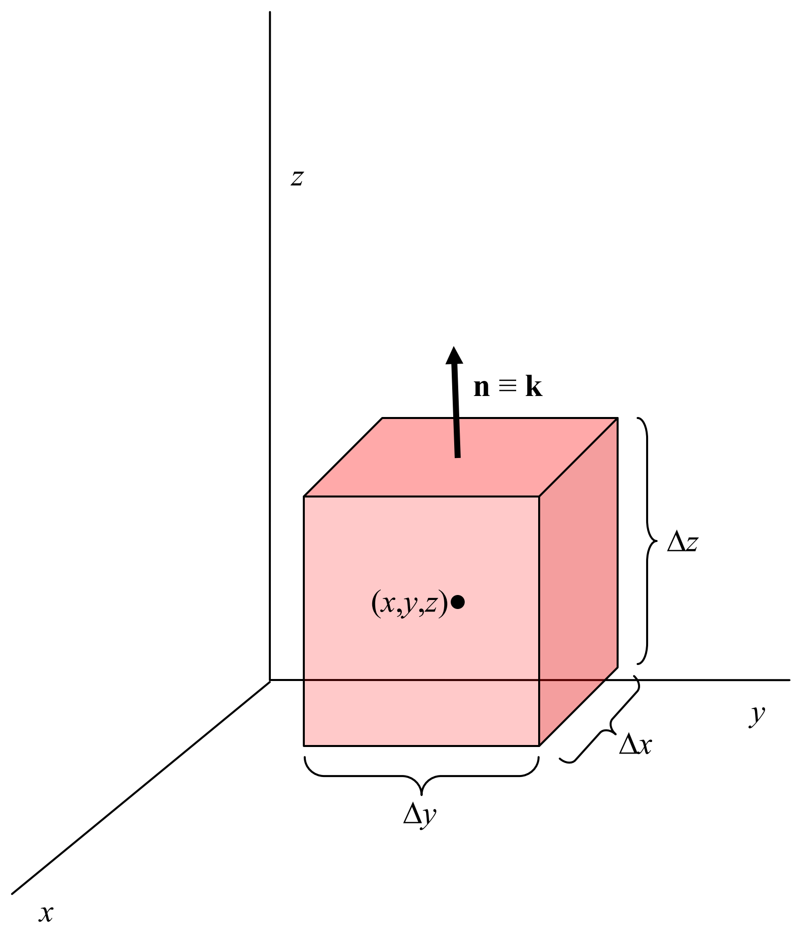


Figure 22. What is the surface integral of a cube as the cube gets infinitely small?

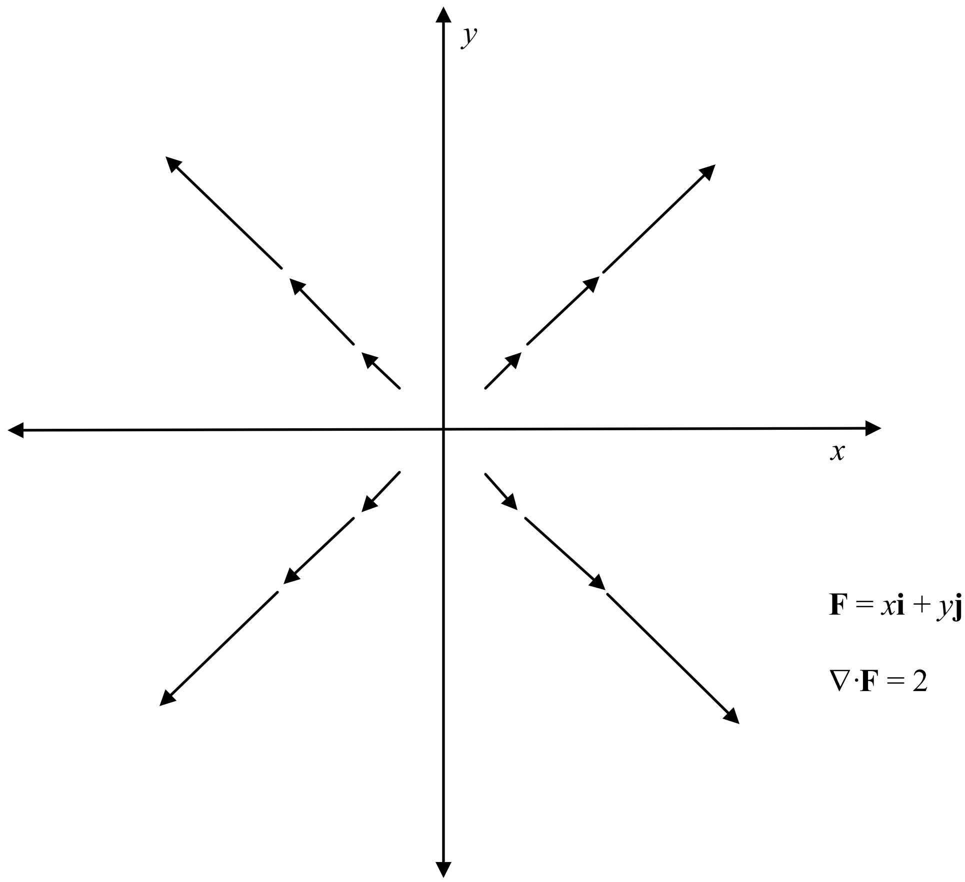


Figure 23. The divergence of the vector field $\mathbf{F} = x\mathbf{i} + y\mathbf{j}$ is 2, indicating a constant divergence, a constant spreading out, of the field at any point in the (x,y) plane.

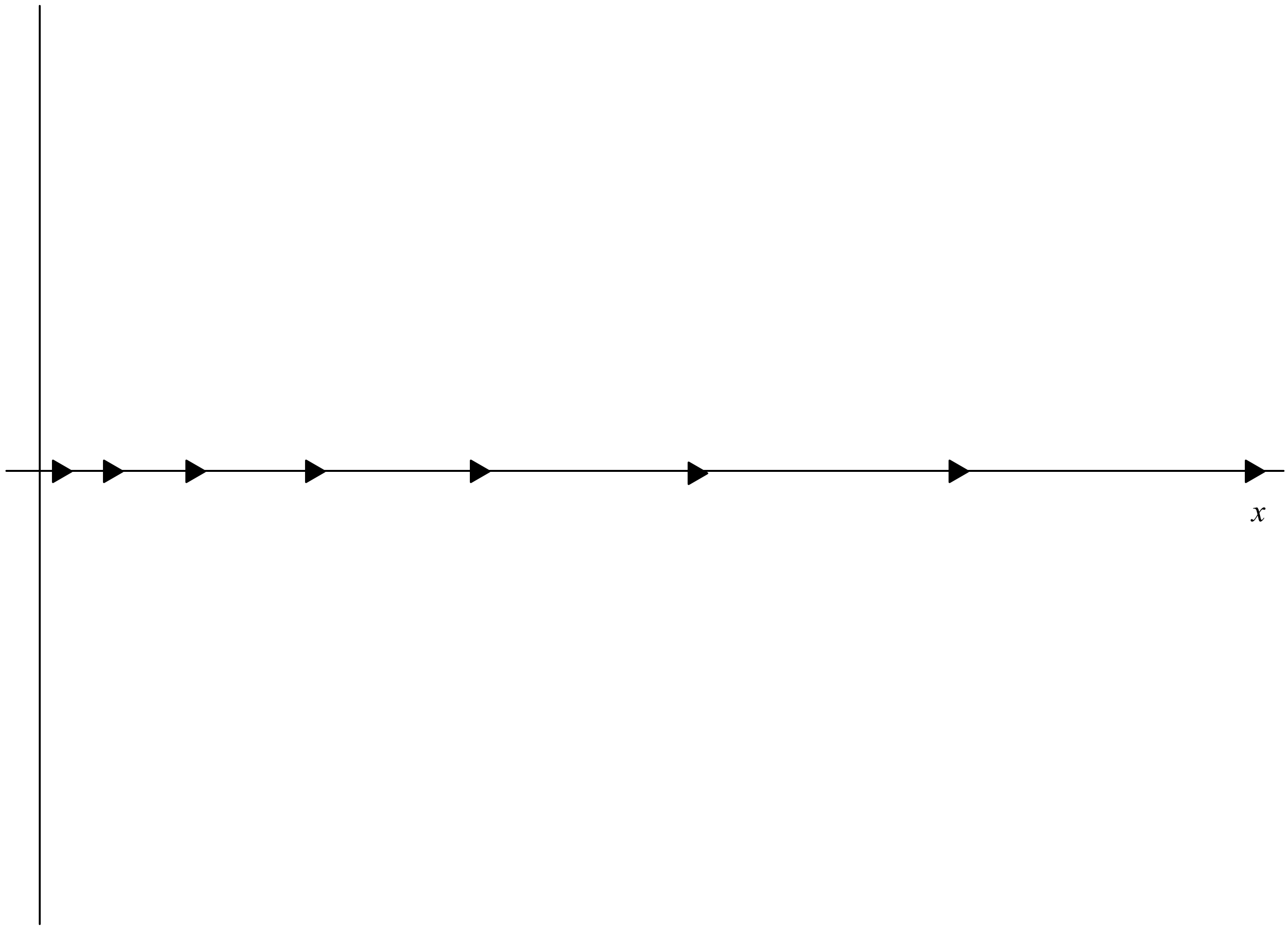


Figure 24. A non-constant divergence is illustrated by this one-dimensional field $\mathbf{F} = x^2\mathbf{i}$ whose divergence is equal to $2x$. The arrowheads represent length of the vector field at values of $x = 1$, 2, 3, 4, etc. The greater the value of x, the farther apart the vectors get – that is, the greater the divergence.

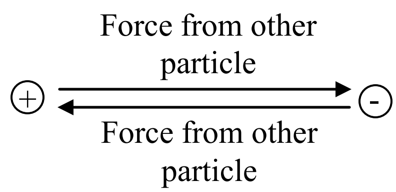


Figure 25. It is an experimental fact that charges exert forces on each other. That fact is modeled by Coulomb's law.

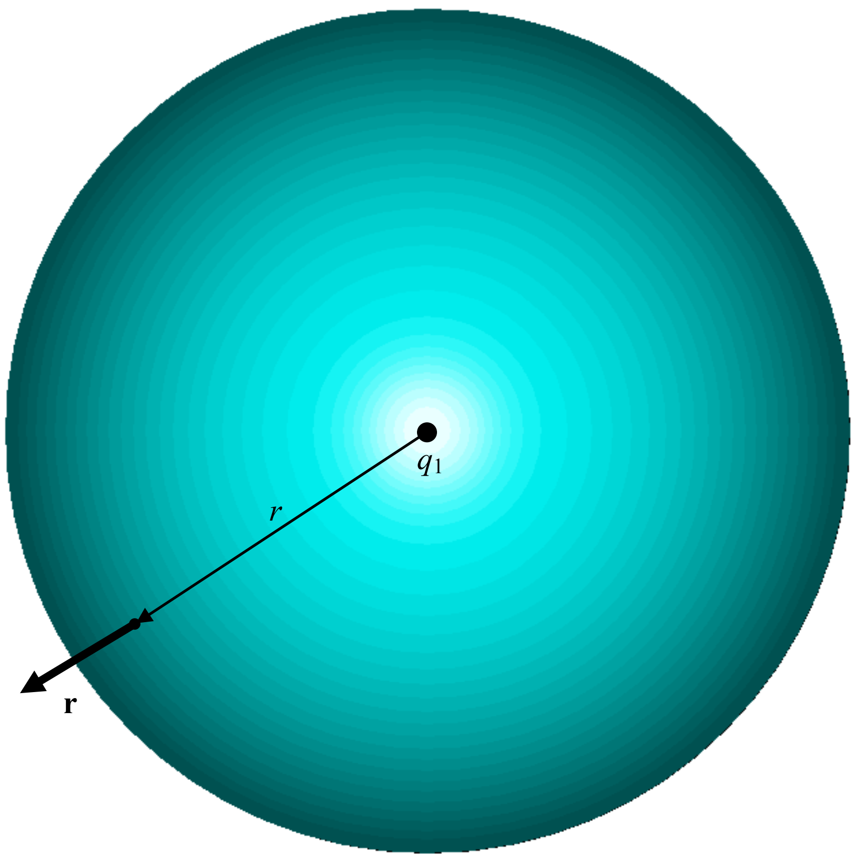


Figure 26. A charge in the center of a spherical shell with radius r has a normal unit vector equal to **r**, in the radial direction and with unit length, at any point on the surface of the sphere.