Electricity, Magnetism and Light

Suppose you are on a camping trip, far from the amenities of modern technology, and that you have brought along only the basic essentials (tent, food, etc.). Imagine that for some reason you are asked to demonstrate the principal forces of nature that we believe to operate at the macroscopic scale (the scale of everyday life). To demonstrate the effects of gravity (at least that of the Earth) would be more or less trivial; to demonstrate those of magnetism would be a little more difficult, but even if you have been so unwise as not to bring a compass with you, you might with luck perhaps find a piece of magnetite and demonstrate its tendency to orient. But how would you demonstrate the effects of electric forces? Unless you happen to have brought along the traditional apparatus of cat's fur and amber, that is not so easy. Probably the most common experience most of us have in everyday life of the effects of *static* electricity (as opposed to those of electric *currents*) is the tendency of shirts or blouses to stick to the tumbler-dryer as you try to remove them; it is ironic that the details of this everyday manifestation of the electric force are still not totally understood!

The basic phenomenon of static electricity was known to the Greeks (the name actually comes from the Greek word for "amber"), but it was not until the late eighteenth century that its fundamental properties really began to be clarified, and the course of the discussion was not always smooth. [Anecdotes from the early days of E & M.] There is one very basic observation that one can make that distinguishes electricity from gravitation: In the case of gravitation, all bodies attract one another independently of their nature. On the other hand, in the case of electricity, we find that if A attracts B and B attracts C, then A always *repels* C; while if D repels E and E repels F, then D always repels F. (In principle it should be possible to verify this statement at the laundromat!) This implies that there are *two and only two* "kinds" of electricity, which we may denote "positive" and "negative", and that "like" charges always repel while "unlike" ones always attract. The definition of "positive" is purely a matter of convention (we now know that the most natural "unit" of electricity, the electron, actually has negative charge).

What is the quantitative law of force between two electric charges? Recall that in the case of gravitation, Newton had postulated the law

$$F_{\text{grav}} = G\,\frac{m_1 m_2}{r_{12}^2} \qquad (\textit{inward})$$

where the masses m_1, m_2 are positive. It is interesting that it was a long time before this law could be directly checked in terrestrial experiments. (In a famous experiment in 1798, Cavendish measured the constant G, but apparently checked neither the r^{-2} dependence nor that on the masses m_1, m_2.)

In the case of electricity, the basic law is known as Coulomb's law and was in fact stated by Coulomb in 1785:

$$F_{\text{el}} = +\ \text{const}\ \frac{q_1 q_2}{r_{12}^2} \qquad (\textit{outward})$$

Actually, Coulomb claimed to have demonstrated only the $1/r^2$ dependence, not that on $q_1 q_2$, and even for that there is a suspicion that he fudged his data. In fact, the law was energetically contested by some of Coulomb's contemporaries. However, we now believe that despite the dubious nature of his original evidence for it, it is in fact correct. [Attempt to replicate in modern times.]

An obvious question is: is the $q_1 q_2$ dependence in Coulomb's law merely in effect a definition of the charges? No, it is actually more than this, at least in principle. Consider four charges q_1, q_2, q_3, q_4 arranged at equal distances from one another (e.g. at the vertices of a tetrahedron). Then it is a nontrivial consequence of Coulomb's law that if F_{ij} denotes the strength of the force between the charges i and j (irrespective of its direction) then

$$F_{12} F_{34} = F_{13} F_{24} \qquad (*)$$

It would be difficult to check this prediction directly, because all that one can normally measure is the *total* force on a charged object; but if one is prepared to assume that under certain circumstances the charge does not change with time (which is actually not a bad approximation under normal conditions for the traditional pith balls, etc., used in this kind of experiment), then one can examine the forces between bodies two at a time and verify the prediction $(*)$.

The constant that occurs in Coulomb's law is, of course, a matter of convention and effectively defines the unit of charge. In the system of units which originally evolved in the nineteenth century, the constant is taken equal to 1, so that a unit charge by definition is such that two such charges placed at unit distance exert on one another unit force. The so called SI system of units which is more commonly used nowadays defines the unit of charge independently of Coulomb's law (just as the unit of mass, the kilogram, is defined independently of Newton's law of gravitation) and therefore there is a constant, analogous to the Cavendish constant G in gravitation, which enters Coulomb's law. The details of all this are not very important: what is important is that whatever system we use we now have, at least in principle, a way of measuring electric charge, *quantitatively*.

Once we have such a quantitative measure, we can verify an extremely important statement: *total electric charge is conserved*, or formally for any closed system of N bodies labeled by i ($i = 1, 2 \ldots N$) we have[†]

$$\sum_i q_i = \text{ const}$$

This "conservation law" is analogous to the conservation of the total mass of a system of bodies, but with the important difference that since charge, unlike mass, can be either positive or negative, an initially uncharged body ($q = 0$) can separate into two oppositely charged ones or vice versa.

[†]Remember that the notation "$\sum_i q_i$" stands for the sum over the charges on the various bodies. (e.g. for 3 bodies the equation given simply means that the quantity $(q_1 + q_2 + q_3)$ is a constant, even though q_1, q_2 and/or q_3 individually may not be).

An obvious question is: *Why* are there two types of electric charge but only one type of mass? From a modern point of view this is actually one of the deepest questions in physics; we will be able to shed some light on it (though not to answer it completely), when we come to general relativity.

The idea of a field

(Einstein, appendix V; Hesse, §VIII; compare the idea of a field in fluid mechanics, etc.: Einstein pp. 144-6, Hesse pp. 189-98.)

Both gravitational and electrostatic forces have a "product" structure ($F_{\rm grav} \propto m_1 m_2$, $F_{\rm el} \propto q_1 q_2$). Hence it is possible to look at it this way: body number 2 produces an "effect" ("field") at the position of 1, and this field exerts a force on 1. E.g. for electrostatics, define an "electric field" E due to 2:

$$E = \text{ const } q_2/r_{12}^2 \qquad \text{(outwards, if } q_2 \text{ is positive)}$$

$$F_1 = q_1 \times (E_2 \text{ at position of 1})$$

However, we could (and must!) equally well regard 1 as producing a field equal to const q_1/r_{12}^2 and 2 as experiencing a force $F_1 = q_2 \times (E_1$ at position of 2). Thus

$$F_1 = -F_2 = \text{ const } q_1 q_2 / r_{12}^2$$

as required by Newton's third law. Now a third body, q_3, will experience the sum of forces from 1 and 2: $F_3 = q_3 \times (E_1$ at 3 plus E_2 at 3), but it will then itself produce its own field. And so on.

Great simplification is introduced if we consider a "test" particle of "infinitesimal" charge, say δq. The advantage is that the field due to δq itself is then negligible. Thus, we can define the total electric field at a given point as simply the (electrostatic) force exerted on an "infinitesimal" charge δq that point divided by δq:

$$E = \lim_{\delta q \to 0} \frac{F \text{ on } \delta q}{\delta q}$$

In a similar way, we can define the gravitational field as the ratio $F_{\rm grav}/\delta m$ for an infinitesimal mass δm. As noted in effect before (lecture 7), the gravitational field (in newtons/kg = m/sec^2) is just the gravitational acceleration g, and therefore is often not introduced explicitly.

We have done something here that is much more fundamental than it looks at first sight. Namely, we have replaced the notion of a direct force between two charged bodies (which can be defined, obviously, only if there happen to actually be two bodies at the points in question!) by the notion of a *field* produced at point 2 by body 1, which exists *whether or not* there is actually a body there at point 2 to feel it. The crucial point is that while in order to *verify* the existence of the field in question we would have to introduce a test charge and measure the force on it, the field is assumed to exist even in the absence of such a test charge. In the language of modern philosophy of science,

its existential status may be that of a "propensity". Eventually, we shall go further and introduce the notion of a "field" that can exist in free space even in the absence of a charged body to produce it.

We must briefly discuss one further concept, that of the electrostatic potential (call it $V_{\rm el}$). This is defined in such a way that if we make a small displacement Δx in the direction of the electric field $E_{\rm el}$ then the charge $\Delta V_{\rm el}$ of the potential is given by the expression

$$\Delta V_{\rm el} = -E_{\rm el}\Delta x$$

(compare the discussion of the "potential energy" in lecture 8).

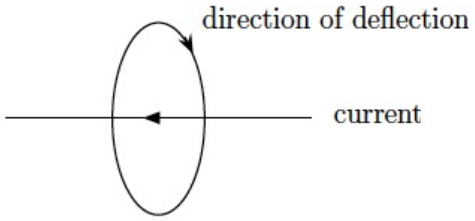
It is not entirely obvious that this permits $V_{\rm el}$ to be uniquely defined as a function of position, but in fact the form of Coulomb's law turns out to guarantee this (the zero of electrostatic potential is however arbitrary). If we take the zero of V at infinity, then it turns out that the potential at point 2 due to a charge q_1 at point 1 is just const q/r_{12}. The contribution of electrostatic effects to the potential energy of a body at a given point in space is just the charge on that body times the value of $V_{\rm el}$ at that point.

[Anecdote re: George III.]

Magnetism

The original "magnetic" effects discovered by the Chinese and the Greeks and systematically investigated by Gilbert (1600) were associated with what we would now call permanent magnets (iron filings, earth's core, etc.). Following the discovery of Coulomb's law of electrostatics, it was tempting to think of magnetism in analogous terms: the forces between permanent magnets were observed to be similar to those between electric *dipoles*, i.e., a positive and negative electric charge placed close to one another $\oplus$–$\ominus$. But *no isolated magnetic poles were ever found.*

Following Coulomb's work, people soon learned to produce (macroscopic) electric *currents* by the use of batteries, etc. (Volta 1775–1800); i.e., they could produce a steady *flow* of charge. A crucial observation was made by Oersted and Ampere in 1820: namely the deflection of a compass needle by an electric current.



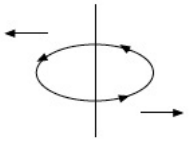
Thus it was clear that magnetism is in some way associated with electric currents. Within a couple of months, Biot and Savart had discovered the quantitative law of force between two currents, which actually has a quite complicated form. For our purposes, it is sufficient to note that two long parallel wires attract if the currents are parallel,

but repel if they are anti parallel. (Thus, a loop is "self-repelling" – an important consideration for energy storage, high-energy accelerators, etc.) But the general structure was similar to Coulomb's law. The Biot-Savart law can be summarized in the following statements:

(a) A current *element* produces a magnetic field H which is proportional to $1/r^2$ (hence, for an infinite straight *wire*, $H \propto 1/r$, where r is the distance from the wire) and is perpendicular to the current and to the direction from the current to the point in question.

(b) A second current element experiences a force that is perpendicular both to itself and to the field, and is proportional to the current and to the field.

(c) A permanent magnet is equivalent to a current *loop*.



Thus, a permanent magnet will tend to orient itself parallel to the magnetic field (so that the forces are only "radial" and hence the system is stable – otherwise there will be a torque, a twisting force). It turns out that magnetic interactions are difficult to discuss in terms of potential energy or "magnetic potential": fortunately, we do not need to go into this point in detail.

Thus, around 1840, one had 3 different kinds of forces which appeared to "act instantaneously at a distance" (gravitation, electrostatic [Coulomb] and magnetic). There was no apparent connection between gravity and either electricity or magnetism, and the connection between the electric and magnetic forces was only through their sources (current = moving charge).

Digression: Why are electrostatic effects so difficult to see by comparison with gravitational and (in modern household machinery) even magnetic effects? According to our current concepts, a single pair of protons interact in all 3 ways, and if they are slowly moving relative to one another the electric force is about 10^{36} times the gravitational one and many times the magnetic one! However, electrostatic effects are typically strongly compensated because of the presence of both positive and negative charge.[‡]

By contrast, gravitational effects are not compensated at all, and magnetic effects need not be (the flow of negative charge is not (usually) compensated by that of positive charge). An interesting question is: how fast would a pair of similar charges have to move in order that their magnetic attraction would compensate their electrostatic repulsion? The answer is: about 3×10^8 m/sec! The significance of this result (which surely must

[‡]Note: To avoid electrostatic effects spoiling Cavendish's experiment, we need a static electrification of < 1 part in 10^{16}!

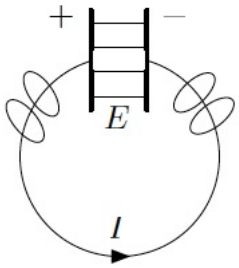
have been worked out in the early nineteenth century) does not seem to have been realized at that time.

Connection between electricity and magnetism

A critical discovery was made by Faraday in 1831: a changing magnetic field through a circuit induces an electric field around the circuit.

$$\boxed{\text{circulating electric field} \propto \text{rate of change of magnetic field}}$$

This field cannot, apparently, be derived from a potential (technically, it is "transverse"), This already seems to show that electricity and magnetism cannot be completely independent phenomena. The final piece of the puzzle was supplied by Maxwell around 1860. Consider the (dis)charging of a pair of capacitor plates:



By Ampere's law, there must be a finite magnetic field around the current carrying wire. What happens *between* the capacitor plates, where there is no current? It can be shown that if the magnetic field were to disappear, then its form in the plates would have to correspond to that of isolated magnetic poles, which as far as we know do not exist. Hence as long as the current is flowing, there must exist a magnetic field also between the plates, and one way of looking at this is that it is produced by the fact that the electric field in this region is changing in time (because the plates are (dis)charging). Quantitative calculation shows this is right, and thus

$$\boxed{\text{(circulating) magnetic field} \propto \text{rate of change of electric field}}$$

Thus: a time-varying magnetic field produces an electric field that "circulates" around it, and a time-varying electric field produces a magnetic field that circulates around it!

Note, however, that so far the argument has always involved actual physical circuits with real charges moving in them. Maxwell's real conceptual leap was to apply the argument *also in empty space*, i.e. make it independent of the sources, *if any*, of E and H.

Electromagnetic waves and light

At this point we need to make Faraday's and Maxwell's laws a bit more quantitative: If we abstract from the physical circuits and imagine the "fields" exist also in "empty"

space, then:

> Faraday: time rate of change of magnetic field = a constant (C_1) × space rate of "transverse" change of electric field.
>
> Maxwell: time rate of change of electric field = another constant (C_2) × space rate of ("transverse") change of magnetic field.

Thus, using D_t as a shorthand for "time rate of change of", and similarly $D_x$for "space rate of change of", we have

$$D_t H = C_1 D_x E \quad \text{and} \quad D_t E = C_2 D_x H$$

Combining these[§]

$$D_t(D_t H) = C_1 D_t(D_x E) = C_1 D_x(D_t E) = C_1 D_x(C_2 D_x H) = C_1 C_2 D_x(D_x H)$$

i.e., second time derivative = const. × second space derivative! Now if we were to regard H as a coordinate, this would read

$$\text{acceleration} \propto \text{rate of change of slope}$$

which is just the equation whose solutions are waves (see lecture 9)! Thus, Maxwell postulated the *existence of electromagnetic waves* in free space, in which E and H each oscillate perpendicular to the direction of propagation of the wave and to one another, and so that the maxima of E coincide with the maxima of H.

What is the speed of these waves? We know C_1 experimentally and C_2 from Maxwell's argument: they are related respectively to the constants that enter the Coulomb and Biot-Savart laws, *provided* that the current is defined as charge/second. And we find

$$\text{speed of EM waves} \approx 3 \times 10^8 \text{m/sec}$$

But this is exactly the speed of light! (The latter was first measured by Roemer as early as 1676, from precise observation of the occultation of Jupiter's moons.) Now though Newton had believed light to be a stream of particles, experiments on interference and diffraction in the early nineteenth century had convinced most people that it was actually some type of wave, and moreover (from polarization experiments) that it had to involve some kind of *transverse* displacement of the medium (unlike, e.g., sound [in air], which is believed to correspond to a *longitudinal* displacement).

Thus, Maxwell's (revolutionary) conclusion was that visible light is simply one part of the electromagnetic wave spectrum, namely that corresponding to frequencies $\sim 10^{16}$ Hz.

audio	radio	microwave	infrared	visible	ultraviolet	X-ray	γ-ray	...

[§]In the second step we use the fact that the operators D_x and D_t "commute", i.e., the result of applying both of them is independent of their order (cf. lecture 8).

This conclusion was soon given strong support by Hertz's experiments in which he showed that spark discharges could indeed produce electric fields at large distances which were considerably greater than any Coulomb fields around (note: wave fields fall off as $1/r$, *not* $1/r^2$!). Since the invention of radio, electromagnetic waves have of course become an everyday matter.

Quite apart from the obvious practical implications of Maxwell's work, it is almost impossible to overestimate its *conceptual* significance.Consider the history of the concept of the electric (or magnetic) "field". Originally it was introduced simply as a convenient shorthand in which to express Coulomb's law of interaction between a "source" charge q_1 and a "test" charge q_2. Then the "test" charge was taken away, and the field became a sort of propensity for a force to be exerted on a charged object, should one happen to be there – but it existed even in the absence of such a charge. However, at that stage it was still necessary for the existence of the field that the "source" charge should be there to produce it. Finally, in the work of Maxwell, the source charge is also removed, and the electric and magnetic fields essentially sustain one another forever as they propagate across empty space. This is a particularly striking example (though certainly not a unique one) of the way in which in the history of physics a concept can be invented initially as little more than a mathematical convenience, and yet eventually as it were come to take on a life of its own.

But if light is a wave, what is it a wave *of*? What, exactly, is oscillating? After all, water waves correspond to a displacement of the surface of the pond, sound waves to that of the air molecules, waves on strings to the transverse displacement of the string... surely the electric and magnetic fields have to represent the transverse displacement of *something*? The answer given by most physicists in the late nineteenth century was that the "something" was "the ether". The concept of the ether had evolved considerably from its late-medieval (Aristotelian) version, and by the end of the nineteenth century it had effectively lost all its properties save that of being the vehicle for the transmission of electromagnetic waves; or, as one commentator put it, "the only function of the ether was to serve as the subject of the verb 'to undulate'." [Modern attempts to resuscitate the ether in quantum-information language.]

[Other major 18th- and 19th-century advances (thermodynamics, chemistry...).]

As the nineteenth century drew to a close, the scientific community as a whole was in a mood of great confidence that all the important principles of physics and chemistry were essentially understood, and that future progress was likely to be a matter of filling in the gaps. ("The end of physics...") These hopes were to be rudely shattered in the first two decades of the twentieth century, and demolished even more finally in the third.