The Skeptic's Guide to Physics

by *Jess H. Brewer*

September 13, 2018

Contents

Why Am I Doing This?

Once upon a time I wrote a book to go with Physics 340, a course for Arts students at the University of British Columbia. After several experiments with existing textbooks, I decided to start my own, based on the usual collection of handwritten lecture notes. My reasons did not include any conviction that I could do a better job than anyone else; rather that I hadn't found any text that set out to do quite the same thing that I wanted to do, and I was too stubborn to revise my intentions to fit the literature. I have gotten worse with age.

What *do* I want to do? The impossible. Namely, to take you on a whirlwind tour of Physics from classical mechanics through modern elementary particle physics, without any patronizing appeals to faith in the experts. I especially want to avoid any hint of phrases like, "scientific tests prove..." that are employed with such poisonous efficiency by media manipulators. I want to treat you like a savvy graduate student auditing a course outside your specialty, not like a woodenheaded ignoramus who has no intellect to appeal to. In particular, I believe that smart Arts people are as smart as (maybe smarter than!) smart Science people, and a good deal more eclectic on average. So I will be addressing you as if you were in the Humanities, though you may just as well be a Nobel laureate chemist or a short-order cook at a fast food restaurant. What do I care what you do for a living? I do want you to see Physics the way I see it, not some edited-for-television version. A tall order? You bet. I'm asking a lot? That's what I'm here for.

Another point I ought to make clear immediately is that this is not a presentation "for people who hate math." That would be like teaching a Mathematics course "for people who hate words." Anyone who hates a tool is suffering from a neurosis; it may be sensible to hate one or more of the ways the tool is *used*, but the tool itself is just a thing. I do propose to craft this resource "for people who hate boredom."

My idol, Richard Feynman, is reputed to have said, "Science is the belief in the ignorance of experts." I love that phrase. It sums up the bare essence of the intellectual arrogance, the willingness to believe in one's own reasoning regardless of what "experts" say, that makes original science (and art) possible. In my opinion, it also makes democracy and justice possible; consider Stanley Milgram's famous research on obedience.... But I digress. It is also true that, while experts may be ignorant, they are rarely stupid; and that a person who wants to trust his or her own judgement above that of any authority has some obligation to hone said judgement to a razor edge. With arrogance comes responsibility. So I am not just setting out to encourage people to disregard or denigrate experts; merely to recognize their *ignorance* and to realize that we all have so much more ignorance than knowledge that in that regard we are almost perfect equals.

Chapter 1

Art and Science

There seems to be an ancient struggle in human conceptual evolution between what might be called the *yin* and *yang* of epistemology (the study of learning and knowing): on the *yin* (receptive, peaceful) side is what I would call *knowledge of the Particular*, or the primitive, intimate knowledge of an instant's experience of reality, without words or explanations or internal dialogue. There are many names (ironically) for this form of knowing, some popular today, such as "Being Here Now," or "Surrender to the *Tao*." There is no denying that a wise person seeks this form of knowledge. The other, *yang* (creative, aggressive) side of knowing I call *knowledge of the Abstract*, which is intrinsically verbal — it is the passion for *naming* which sets humans apart (for better or worse) from other reasonably intelligent animals. And it is the answer to "What's in a name?" — namely, everything we know *that can be communicated* about the thing named. This side has numerous hazards for us, but it is essential for the existence of communication or the improvement of "comprehension."

Here is a tidy example of the distinction between these two forms of knowing: suppose you are walking in the woods and come upon a tiny flower growing in the shade of a large tree; suppose you have never seen a flower like this one before. On the one hand, your *experience* of *this particular flower* can be deepened and explored: smell the flower, study it from all sides, touch it, lie down in the pine needles and look up through the branches to get the flower's viewpoint on things, etc. In all this you are best served by a *lack of words* and a receptive spirit. On the other hand, you can tell by the structure of the stamen, etc., that this is an *orchid* and probably (since it is on a red stem with no leaves) a specied of "coralroot" — perhaps a new variety of *corallorhiza maculata*. And so on. There is real satisfaction in finding a verbal "box" to put this experience in for classification, categorization, filing and retrieval. If we were dealing with a brightly coloured *snake*, rather than a flower, the practical value of the *yang* form of knowing would be more obvious.

Physics, like most philosophy, is devoted to knowledge of the Abstract. This is not to say that physicists are disinterested in knowledge of the Particular, either in their personal lives or in the laboratory; but I believe they agree almost unanimously upon the *yang* principle as the æsthetic basis for their work. All sciences are not necessarily so devoted to Abstraction; a more *empirical* science will attach more significance to Particular information, and this is neither good nor bad — it is merely in æsthetic discord with the "spirit" of Physics.

Such conflict can grow more acute at the ill-defined interface between "science" (æsthetically yang-based pursuits) and "art" (æsthetically yin-based pursuits), and this sometimes leads to unpleasant misunderstandings in which

an insecure scientist will label all artists as ignorant buffoons or an insecure artist will lash out at all scientists as callous androids. (Brilliant members of both species rarely need to elevate their own importance by downgrading others.) From the silly coffee-room dispute between "pure" and "applied" physicists over what constitutes valid or "legitimate" science to the total alienation of a culture from the technology on which it depends for survival, all such conflicts are pitiful stupidity. To be human involves an integration of both ways of knowing, and neither a poet nor a physicist can perform competently without this integration.

This interdependence is nowhere as obvious as in the *tools* used by physicists and poets. How, for instance, does either devise a means for expressing a truly new idea? (For surely the goal of poetry is to say what has never before been said in quite the same way — *i.e.* to create a new idea/feeling for the reader/listener.) One seemingly logical answer is that there *is* no way; that language includes a finite number of ideas and images which can be expressed by a finite number of words or combinations of words, and that this large but finite space of old ideas can never be escaped through language. This notion is the source of the pessimistic aphorism "There's nothing new under the sun." It is patently absurd, inasmuch as all languages were once nonexistent and were built up gradually — are still in the process of being created today, mostly by poets and their close relatives. This process is called *Emergence* by Michael Polanyi, my favorite modern philosopher, who used to be a physical chemist. As he carefully points out, the same is true of Physics, the poetry of nature: new ideas are always Emerging as older ideas become familiar and "tacit."

To return to the original question, how does this happen? What is the essential mechanism for Emergence in both science and art? The answer, I believe, is that *metaphor* (and its less ambitious ally, *simile*) is the vehicle for all Emergence of ideas and feelings, whether we are explicitly aware of it or not. Half the descriptive idioms in our language involve explicitly metaphorical images ("leaf" through a book?) which vividly convey the desired idea and at the same time add to the connotative richness of the individual words; these images were originally created by poets (for my purposes a "poet" is defined as one who creates new language through such images). Similarly, in Physics we speak of "isospin" as a particle property, even though it certainly has nothing to do with rotation in normal space, because this esoteric quantity seems to have transformation properties analogous to those of angular momentum. The metaphor is a little more explicit and a little less tangible to everyday experience than "leafing," but the same process is at work.

Thus today's Physics rests, like today's language, on a monumental pyramid of metaphors and similes, leading back to our most primitive notions of space and time and force, which are ultimately indefinable. When I subtitled this *HyperReference* "Physics *as* Poetry" I was being most literal-minded!

Table 1.1 The Great False Dichotomy

KNOWLEDGE OF THE PARTICULAR	*vs.*	**KNOWLEDGE OF THE ABSTRACT**
YIN the Receptive	← **THEME** →	*YANG* the Creative
Perceptual Private Intimate Wordless Accepting Wondering Intuitive	**QUALITIES and ACTIVITIES**	Analytical Extrovert Impersonal Communicative Cataloguing Naming Logical
Calm Peaceful Integrated Mystical	**EFFECTS**	Impatient Agressive Alienated Egotistical
Vast but Unreliable & Inconsistent	**POWERS**	Circumscribed but Reliable & Predictable
Aristotle (details = essence)	**Classical Protagonists**	Plato, Galileo (ideal = essence)
ART & MAGIC	**MODERN POLITICAL DIVISION**	SCIENCE & TECHNOLOGY

Chapter 2

Poetry of *physics vs.* "doing" *Physics*

2.1 Poetry as "Language Engineering"

Communication requires a consensus about language. We have dictionaries to help stabilize that consensus; we have poets to help keep it evolving. I am not much of a poet, but I identify with their part of the task: I use the dictionary words (making up "new" words like *quark* has always seemed a little on the tachy side to me; why break rules if they are fair?) but I sometimes try to decorate their meanings with a lot of connotations and allusions and specific details *in a given context* that are not in any dictionary and would be inappropriate in another context. This is a fun ego trip; it is also necessary whenever one is trying to make a point that goes a little beyond where existing language leaves off – which isn't far from where we live daily.

Unlike most poets, however, I will do my best to spoil the mystery of my private terminology: whenever I realize that I am using a word in a specific sense that transcends the dictionary meaning and its colloquial connotations, I will try to call attention to it and explain as much as I can about the differences. Poets don't do this for a very good reason: part of the magic of poetry is its ambiguity. Not just random ambiguity like dictionary words out of context, but coherently ambiguous; a good poet is offended by the question, "What exactly did you mean by that?" because *all* the possible meanings are intended. Great poetry does not highlight one meaning above all, but rather manipulates the interactions between the several possible interpretations so that each enriches the others and all unite to form a whole greater than the sum of its parts. Unfortunately, the reader/listener can only appreciate this subtlety *after* mastering the nuances of the language in which the poet writes or speaks. Those who have mastered the language of Physics do indeed rely upon the same sort of "coherent ambiguities" to get their points across, or else no one would be able to discuss quantum mechanics at all (to give the prime example); this is why I have given the subtitle *Physics as Poetry* to this collection of *HyperReferences*. But at the beginning we are learning "science as a second language" and it is best to minimize ambiguity where possible.

The first and obvious example is the word *PHYSICS*. If I mean the (hypothetical) orderly behaviour of the (hypothetical) objective physical universe, I will write "physics." If I mean the sociopolitical human activity, the consensual reality prescribed by a set of conventional paradigms and accepted models about said universe, I will write "Physics." Unlike some deconstructionist sociologists, I believe the former exists independently of the latter. Or at least I have a commitment to that *æsthetic*....

2.2 Understanding *physics*

First let's examine some of the assumptions with which a physicist tries to comprehend the universe. The most important of these is the assumption that there *is* a universe. That is, that there is a real, substantial, external "physical" reality[1] which is the same for everyone, which we interact with directly and perceive directly through our senses, which are usually fairly trustworthy as far as they go. In other words, the opposite of Solipsism (look it up if it's unfamiliar; you should know your enemy). This could be wrong, of course, but if you are really God in the universe of your own imagination, why not imagine an objective, consistent universe with other people in it so we can get on with this? I did, heh heh.

Given that assumption, we physicists go on to postulate that the universe obeys the same rules in all places and at all times. Yes, yes, there are lots of speculations about changes in the "laws of physics" *as we know them now*, such as Inflation in the Early Universe and all that, but if that was how it happened and if there was a good reason for it then those *are* the laws of physics; we just (once again) accept that what seem like laws today are just a local or temporary approximation or special case of something more general and more subtle. This happens all the time (on a scale of decades or centuries) in Physics.[2] Whatever we observe, we have an unshakeable conviction that there is a perfectly sensible reason for it. That does *not* mean that we *know* the reason, or ever will, or are even *capable* of understanding it, but we try to.

These are the personality traits that make a physicist. First was the æsthetic commitment to the idea of a "real world." Second is the urge to understand why things behave the way they do (or just are the way they are); this could be labelled *curiosity*, I suppose, but the physicist's trait is usually a bit more obsessive-compulsive than connoted by that innocuous word. Third is the *arrogance* to assume that we *can* understand virtually anything. There are examples of systems which can be proven to be *intrinsically unpredictable*, but that doesn't faze the physicist; we are smugly satisfied with our understanding of the unpredictability itself.

So how does this make us like poets? It's hard to explain, but for both physicists and poets there's a thrill in the moment of "Aha!" when all the grotty little details finally come together in our presumptuous little heads and synthesize a sense that we "get it" at last.[3] And for both poets and physicists, the most common vehicle for this epiphany is the **metaphor**.

Therefore be not surprised when I haul out one bizarre image after another with great pride to show yet another way of looking at angular momentum, or waves, or Relativity. And remember, you don't have to be a *good* poet to love poetry. . . .

2.3 "Doing *Physics*"

There is more to this story, of course. Whether for some excellent, deep reason or just because of the practical benefits to society, professional Physicists are also almost always selected and trained to enjoy "doing Physics." You will hear this phrase used frequently among Physicists. What does it mean? How do you "do" the underlying principles governing the behaviour of the universe? You don't, of course; when we use this phrase we are talking about capital-P Physics, the human enterprise.

There are several aspects to "doing Physics." I will list them in what is, for me, today, ascend-

[1] Boy, what a bunch of loaded terms! For now I will have to fall back on the old standby, "*You* know what I mean. . . ."

[2] There, did you notice the distinction between *physics* and *Physics* in that long sentence? Watch carefully!

[3] Whether we actually *do* "get it" accurately is not terribly important, as long as those other traits keep bringing us back to the real world to *test* our newfound understanding.

ing order of "enjoyability." There is no reason why anyone else should agree with this order, but I believe in full disclosure.

2.3.1 Politics

— explicitly sociopolitical activities usually involving distasteful compromises.

- **Applying for grants**: Mercifully, novices are spared the dirty work of grantsmanship for the first few years of their involvement with Physics.
- **Getting papers published** as distinguished from *writing* papers, which (along with giving lectures) falls more into the "fun" category. If a novice writes a publishable paper there will usually be some mentor willing to do the dirty political work of getting it published (usually in return for co-authorship).
- **Managing equipment**: The ugly part of experimental science is bound up in the politics of getting money to buy equipment, organizing it and finding places to set it up, keep it running *etc.* so that the novice experimenter can focus on actually getting the apparatus to *work*, which is (relatively speaking) the fun part.
- **Managing people**: Although the practice of Physics has an intrinsically solitary aspect, many projects can only reach fruition when many people join in a common effort; in these cases it is arguable that the most important people involved are those who provide leadership and organization. Fortunately, in Physics such positions are rarely occupied by those who just like telling others what to do. Physics has room for an astonishing variety of personal styles, which makes it a rewarding field in which to be an administrator, providing of course that one enjoys people generally.

I am not a very enthusiastic manager, as you may have surmised, but even in politics there is room for real satisfaction. There can be quite a thrill in obtaining a few billion dollars for the construction of the world's greatest accelerator or managing a huge army of Ph.D. physicists to accomplish a spectacularly ambitious task taking hundreds of person-years of intense effort; however, like all forms of satisfaction related to power, these fade with familiarity and eventually demand greater and greater achievements to maintain the glamour. If you get aboard this vehicle, be sure to plan carefully where you want to get off.

2.3.2 Craftsmanship

— the fulfillment of the artisan.

- **Tinkering with the apparatus**: Before experimental equipment or theoretical models can be used to conduct a conversation with Nature, they have to be working properly. Achieving this state is nontrivial. In fact it takes most of the effort; once the apparatus it working and configured for the desired task, "getting the answer" can be just a matter of "turning the crank" and watching the results pour out. But first you must get to know the equipment intimately, and there is only one way to do that: by using it.
- **Problem-solving**: This is an *absolutely essential* aspect of "doing Physics" that is often neglected by novices, with catastrophic consequences. It is one thing to understand physics and quite another to be able to put that understanding to work. A good metaphor is the difference between a brilliant automotive mechanic and a great driver. It will help a lot if you

know how your car works, but winning the Molson Indy takes something else. Driving experience will also help you be a better mechanic, and that's an aspect of this metaphor I want to explore later. But for now I can't emphasize strongly enough that *most of the hard work* in a Physics apprenticeship is in learning how to *solve problems* — and the *only* way to learn that is by *doing it* — a *lot* of it. This puts most people off at first. I know it did me.

- **Engineering**: Once you know how to solve problems, you pick the ones you want to solve and you learn how to put the solutions to work in the real world. This is what I call Engineering, the art of making Technology work. Lots of people will be offended by the fact that I placed this rather extensive field of endeavour so far toward the "not so enjoyable" end of my ordered list of Physics activities; they should not be. For one thing, this is just a list of my personal tastes. For another, just because I don't *enjoy* Engineering as much as (for instance) writing does not mean I don't *appreciate* it; in fact, some of the most satisfying work I have ever done would fall into this category. Just as the most enjoyable activities can be made unpleasant by excess (writing a Ph.D. thesis is rarely a *pleasant* experience, but it is almost always a *satisfying* one), drudgery in the service of an inspiring goal can leave very pleasant memories.

Not surprisingly, I like an athletic metaphor for Craftsmanship in Physics: competing in the World Championships may be the ultimate experience for the athlete, but it represents a very tiny fraction of the athletic experience, most of which consists of endless gruelling workouts that are rarely pleasant but always rewarding, both in terms of the final goal and in terms of hard-won accomplishment. There is only one way to find out what you can do, and that's by doing it.

2.3.3 Teaching

— sharing your understanding.

- **Lecturing**: finding a really nice way to get across to others what I have just figured out myself.

- **Writing**: same as lecturing except one gets more time to perfect one's delivery. Here I include the electronic version(s) of "writing" as a natural extension of words on paper; the Web also offers an opportunity to use more tools similar to those one might employ in lectures, like sound and images.

I am not counting the "political" aspect of professional teaching — organizing lectures, preparing and marking homework and exams, making judgements about other people's performance and submitting those evaluations in the form of marks. This has little to do with the fun part of teaching except insofar as the one makes a place for the other to happen.

2.3.4 Learning

— the interface between *Physics* and *physics*.

- **The glimpse of Nature**: When you finally finish fiddling with the apparatus (whether theoretical or experimental) and it seems to be working, it makes a sort of conduit through which a shy Nature can reveal her secrets;[4] such moments are

[4] If anyone is offended by my gender-specific reference to Nature, tough. That's the metaphor that works for me. If I were a different gender myself, maybe I would prefer a different one.

rather rare, and too often occur when the experimenter (or theorist) is dead tired, but one glimpse is usually all it takes to make it all seem worthwhile.

- **The epiphany**: After you have assembled all you know about a new subject and stirred the mix long enough, something starts to congeal and the primal "Aha!" bursts through all the layers of confusion to enlighten you for a while. For me this almost always takes the form of a **metaphor** that lifts my comprehension from the realm of *Physics* and plants it in the Platonic ideal world of *physics*. (Or so it seems; but after all, Reality is what we make it. . . .)

Chapter 3

Representations

In *Art and Science* we pondered the distinction between intuitive knowledge of the particular and analytical knowledge of the abstract. The former governs intimate personal experience — about which, however, nothing further can be said without the latter, since all communication relies upon abstract symbolism of one form or another. We can *feel* without symbols, but we can't *talk*.

Moreover, before two people can communicate they must reach a *consensus* about the symbolic *representation* of reality they will employ in their conversation. This is so obvious that we usually take it for granted, but few experiences are so unsettling as to meet someone whose personal symbolic representation differs drastically from consensual reality.

How was this consensus reached? How arbitrary are symbolic conventions? Do they continue to evolve? They never represent quite the same things for different people; how do we know if there is a reality "out there" to be represented? These are questions that have perplexed philosophers for thousands of years; we are not going to find final answers to them here. But within the oversimplified context of Physics (the social enterprise, the human consensus of paradigmatic conventions, as opposed to *physics*, the actual workings of the universe) we may find some instructive lessons in the interactions between tradition, convention, consensus and analytical logic. This is the focus of the present Chapter.

Each word in a dictionary plays the same role in writing or speech (or in "verbal" thought itself) as the hieroglyphic-looking symbols play in algebraic equations describing the latest ideas in Physics. The big difference is ... well, in truth there *isn't* really a big difference. The *small* differences are in compactness and in the degree to which ambiguity depends upon context. Obviously an algebraic symbol like t is rather compact relative to a word composed of several letters, like *time*. This allows storage of more information in less space, which is practical but not always pleasing.

As for ambiguity in context, words are designed to have a great deal of ambiguity until they are placed in sentences, where the context partially dictates which meaning is intended. *But never entirely*. Part of the magic of poetry is its ambiguity; a good poet is offended by the question, "What exactly did you mean by that?" because *all* the possible meanings are intended. Great poetry does not highlight one meaning above all, but rather manipulates the interactions between the several possible interpretations so that each enriches the others and all unite to form a whole greater than the sum of its parts. As a result, no one ever knows for certain what another person is talking about; we merely learn to make good guesses.[1]

[1] This seems to be holding up progress in Artificial Intelligence (AI) research, where people trying to teach computers to understand "natural language" (human speech) are stymied by the impossibility of reaching a unique logical interpretation of a typical sentence. Methinks they are trying

In Mathematics, some claim, every symbol must be defined exhaustively and explicitly prior to its use. I will not comment on this claim, but I will pounce on anyone who tries to extend it to Physics. A meticulous physicist will *try* to provide an unambiguous definition of every *unusual* symbol introduced, but there are many symbols that are used so often in Physics to mean a certain thing that they have a well-known "default" meaning as long as they are used in a familiar *context*.

For instance, if $F(t)$ is written on a blackboard in a Physics classroom, it is a good bet that F stands for some *force*, t almost certainly represents *time*, especially when appearing in this form, and the parentheses () *always* denote that F (whatever that is) is *a function of* t (whatever it may be). This will be discussed further below and in later Chapters. The point is, algebraic notation follows a set of conventions, just like the grammar and syntax of verbal language, that defines the context in which each symbol is to be interpreted and thus provides a large fraction of the meaning of a given expression.

It is tempting to try to distinguish the dictionary from the Physics text by pointing out that every word in the former is defined in terms of the other words, so that the dictionary (plus the grammar of its language) form a perfectly closed, self-reference universe; while all the symbols of Physics refer to entities in the *real* world of *physics*. However, any such distinction is purely æsthetic and has no rigourous basis. Ordinary words are also meant to refer to *things* (*i.e.* personal experiences of reality) or at least to abstract classes of particular experiences. If there is a noteworthy difference, it consists of the potency of the æsthetic commitment to the notion of an external reality. "Natural" language can be applied as effectively in the service of solipsism as materialism, but Physics was designed exclusively to describe a reality independent of human perception, "out there" and immutable, that admits of analytical dissection and conforms to its own hidden laws with absolute consistency. The physicist's task is to discover those laws by ingenuity and patience, and to find ways of expressing them so that other humans can understand them as well.

This may be a big mistake, of course. There may not *be* any external reality; *physics* may be just the consensual symbolic representation of Physics and physicists; or there may not be any physicists other than myself, nor students in my class nor readers of this text, other than in my vivid imagination. But who cares? Solipsism cannot be proven wrong, but it can be proven boring. And since Physics lies at the opposite end of the æsthetic spectrum, no wonder it is so exciting!

3.1 Units & Dimensions

3.1.1 Time & Distance

Two of the most important concepts in Physics are "length" and "time." As is often the case with the most important concepts, neither can be defined except by example — *e.g.* "a meter is this long. . . ." or, "a second lasts from now . . . to now." Both of these "definitions" completely beg the question, if you consider carefully what we are after; they merely define the *units* in which we propose to *measure* distance and time. Except for analogic reinforcements they do nothing at all to explain the "meaning" of the concepts "space" and "time."

Modern science has replaced the standard platinum-iridium reference **meter** (m) stick with the indirect prescription, ". . . the distance travelled by light in empty space during a time of 1/299,792,458 of a second," where a **second** (s) is now defined as the time it takes a certain frequency of the light emitted by ce-

too hard.

sium atoms to oscillate 9,192,631,770 times.[2] This represents a significant improvement inasmuch as we no longer have to resort to carrying our meter stick to the International Bureau of Weights and Measures in Sèvres, France (or to the U.S. National Bureau of Standards in Boulder, Colorado) to make sure it is the same length as the Standard Meter. We can just build an apparatus to count oscillations of cesium light and mark off how far light goes in 30.663318988 or so oscillations [well, it's easy if you have the right tools. . . .] and make our own meter stick independently, confident that it will come out the same as the ones in France and Colorado, because our atoms are guaranteed to be just like theirs. We can even send signals to neighbors on Tau Ceti IV to tell them what size to make screwdrivers or crescent wrenches for export to Earth, since there is overwhelming evidence that their atoms also behave exactly like ours. This is quite remarkable, and unprecedented before the discovery of quantum physics; but unfortunately it does not make much difference to the dilemma we face when we try to define "distance." Nature has kindly provided us with an unlimited supply of accurate meter sticks, but it is still just a name we give to something.

To learn the properties of that "something" which we call "distance" requires first that we believe that there is truly a physical entity, with intrinsic properties independent of our perceptions, to which we have given this name. This is extremely difficult to prove. Maybe not impossible, but I'll leave that to the philosophers. For the physicist it is really a matter of æsthetics to enter into conversations with Nature as if there were really a partner in such conversations. In other words, I cannot tell you what "distance" is, but if you will allow me to assume that the word refers to something "real," I can tell you a great deal about its properties, until at some point you feel the partial satisfaction of intimate familiarity where perfect comprehension is denied.

How do we begin to talk about time and space? The concepts are so fundamental to our language that all the words we might use to describe them have them built in! So for the moment we will have to give up and say, "Everyone knows pretty much what we mean by time and distance." This is always where we have to begin. Physics is just like poetry in this respect: you start by accepting a "basis set" of images, without discussion; then you work those images together to build new images, and after a period of refinement you find one day, miraculously, that the new images you have created can be applied to the ideas you began with, giving a new insight into their meaning. This "bootstrap" principle is what makes thinking profitable.

Later on, then, when we have learned to manipulate time and space more critically, we will acquire the means to break down the concepts and take a closer look.

3.1.2 Choice of Units

All choices of units are completely arbitrary and are made strictly for the sake of convenience. If you were a surveyor in 18th-Century England, you would consider the **chain** (66 feet by our standards) an extremely natural unit of length, and the **meter** would seem a completely artificial and useless unit, because people were shorter then and the **yard** (1 yard = 3600/3937 of a meter) was a better approximation to an average person's stride. **Feet** and **hands** were even better length units in those days; and if you hadn't noticed, an **inch** is just about the length of the middle bone in a small person's index finger.

If you couldn't get your hands on a timepiece

[2] This is only the latest in a long sequence of redefinitions of the meter. Today's version reflects our recognition of the speed of light as a universal constant. (Here is a trick question for you: if the speed of light *were* different in one time and place from another, how could we tell?)

with a second hand, the utility of **seconds** would seem limited to the (non-coincidental) fact that they are about the same as a resting heartbeat period. **Years** and **days** might seem less arbitrary to us, but we would have trouble convincing our friends on Tau Ceti IV.[3] Remember, our perspective in Physics is universal, and in that perspective all units are arbitrary.

We choose all our measurement conventions for convenience, often with monumental shortsightedness. The decimal number system is a typical example. At least when we realize this we can feel more forgiving of the clumsiness of many established systems of measurement. After all, a totally arbitrary decision is always wrong. (Or always right.)

Physicists are fond of devising "natural units" of measurement; but as always, what is considered "natural" depends upon what is being measured. Atomic physicists are understandably fond of the **Angstrom** (Å), which equals 10^{-10} m, which "just happens" to be roughly the diameter of a hydrogen atom. Astronomers measure distances in **light years**, the distance light travels in a year ($365 \times 24 \times 60 \times 60 \times 2.99 \times 10^8 = 9.43 \times 10^{15}$ m), **astronomical units** (a.u.), which I think have something to do with the Earth's orbit about the sun, or **parsecs**, which I seem to recall are related to seconds of arc at some distance. [I am not biased or anything. . . .]

Astrophysicists and particle physicists tend to use units in which the velocity of light (a fundamental constant) is dimensionless and has magnitude 1; then times and lengths are both measured in the same units. People who live near New York City have the same habit, oddly enough: if you ask them how far it is from Hartford to Boston, they will usually say, "Oh, about three hours." This is perfectly sensible insofar as the velocity of turnpike travel in New England is nearly a fundamental constant. In my own work at TRIUMF, I habitually measure distances in **nanoseconds** (billionths of seconds: 1 ns = 10^{-9} s), referring to the distance (29.9 cm) covered in that time by a particle moving at essentially the velocity of light.[4]

In general, physicists like to make *all* fundamental constants dimensionless; this is indeed economical, as it reduces the number of units one must use, but it results in some oddities from the practical point of view. A nuclear physicist is content to measure distances in *inverse pion masses*, but this is not apt to make a tailor very happy.

3.1.3 Perception Through Models

The upshot of all this is that you can't trust any units to carry lasting significance; all is vanity. Each and every choice of units represents essentially *a model of what is significant*. What is vitally relevant to one observer may be trivial and ridiculous to another. Lest this seem a depressing appraisal, consider that the same is true of all our means of perception, even including the physical sensing apparatus of our own bodies: our eyes are sensitive to an incredibly tiny fraction of the spectrum of electromagnetic radiation; what we miss is inconceivably vast compared to what we detect. And yet we see a lot, especially under the light of Sol, which at the Earth's surface happens to peak in just the region of our eyes' sensitivity. Our eyes are simply a model of what is important locally, and well adapted for the job.

The only understanding you can develop that

[3] This is a recurring problem in science fiction novels: will our descendents on other planets use a "local" definition of years, [months,] days, hours and minutes or try to stick with an Earth calendar despite the fact that it would mean the local sun would come up at a different time every day? Worse yet, how will a far-flung Galactic Empire reckon *dates*, especially considering the conditions imposed by Relativity? [The *Star Trek* solution is, of course, to ignore the laws of physics entirely.]

[4] Inasmuch as a *ns* is a roughly "person-sized" distance unit, it could actually be used rather effectively in place of feet and meters, which would get rid of at least *one* arbitrary unit. Oh well.

is independent of units has to do with how dimensions can be combined, juxtaposed, *etc.* — their *relationships* with each other. The notion of a velocity as a ratio of distance to time is a concept which will endure all vagaries of fashion in measurment. This is the sort of concept that we try to pick out of the confusion. This is the sort of understanding for which the physicist searches.

3.2 Number Systems

We have seen that *units* of measurement and indeed the very nature of the *dimensions* of measurement are arbitrary models of what is significant, constructed for the practical convenience of their users. If this causes you some frustration or disappointment, you are not alone; most students of Physics initially approach the subject in hope of finding, at last, some rigor and reliability in an increasingly insubstantial and malleable reality. Sorry.

What most disillusioned Physics students do next is to seek refuge in mathematics. If physical reality is subject to politics, at least the rarefied abstract world of numbers is intrinsically absolute.

Sorry again. Higher mathematics relies on pure logic, to be sure, but the *representation* used to describe all the practically useful examples (*e.g.* "arithmetic") is intrinsically arbitrary, based once again on rather simpleminded models of what is significant in a practical sense. The decimal number system, based as it is upon a number whose only virtue is that most people have that number of fingers and thumbs, is a typical example. If we had only thought to distinguish between fingers and thumbs, using thumbs perhaps for "carrying," we would be counting in *octal* and be able to count up to twenty-four on our hands. Better yet, if we assigned significance to the *order* of which fingers we raised, as well as the *number* of fingers, we could count in *binary* up to 31 on one hand, and up to 1023 using both hands! However, we have already made use of that information for other communication purposes. . . .

Is mathematics then arbitrary? Of course not. We can easily understand the distinction between the *representation* (which is arbitrary) and the *content* (which is not). Ten is still ten, regardless of which number system we use to write it. Much more sophisticated notions can also be expressed in many ways; in fact it may be that we can only achieve a deep understanding of the concept by learning to express it in many alternate "languages."

The same is true of Physics.

3.3 Symbolic Conventions

In Physics we like to use a very compact notation for things we talk about a lot; this is æsthetically mandated by our commitment to making complicated things look [and maybe even *be*] simpler. Ideally we would like to have a single character to represent each paradigmatic "thing" in our lexicon, but in practice we don't have enough characters[5] and we have to re-use some of them in different contexts, just like English!

In principle, any symbol can be used to represent any quantity, or even a non-quantity (like an "*operator*"), as long as it is explicitly and carefully defined. In practice, life is easier with some "default" *conventions* for what various symbols should be assumed to mean *unless otherwise specified*. On the next pages are some that I will be using a lot.[6]

[5] The wider availability of nice typesetting languages like LaTeX, in which this manuscript is being prepared, offers us the opportunity to add new symbols like ℵ, ϖ and ♡, but this won't change the qualitative situation.

[6] (You will want to refer to these occasionally when trying to guess what I am trying to say with formulae. Don't worry if some are incomprehensible initially; for completeness, the list includes lots of "advanced" stuff.)

Table 3.1 Roman symbols commonly used in Physics

ROMAN LETTERS:

A = an *area*; Ampere(s).	a = *acceleration*; a general constant.
B = magnetic field.	b = a general constant.
C = heat capacity; Coulomb(s).	c = *speed of light*; a gen. constant.
D = a form of the electric field.	d = *differential* operator; *diameter*.
E = *energy*; electric field.	e = 2.71828...; electron's charge.
F = *force*; a general *function*.	f = a fraction; a *function* as in $f(x)$.
G = grav. constant; prefix *Giga-*.	g = *accel. of gravity* at Earth's surface.
H = magnetic field; Hamiltonian op.	h = Planck's constant; a height.
I = electric current.	$i = \sqrt{-1}$; an index (subscript).
J = *Joules*; spin; angular momentum.	j = a common integer index.
K = degrees Kelvin.	k = an integer index; a gen. constant; *kilo-*.
L = *angular momentum*; length.	l = an integer index; a *length*.
M = magnetization; mass; *Mega-*.	m = *metre(s)*; *mass*; an integer index.
N = *Newton(s)*; a large number.	n = a small number; prefix *nano-*.
$\mathcal{O}$ = "order of" symbol as in $\mathcal{O}(\alpha)$.	o = rarely used (looks like a 0).
P = probability; pressure; power.	p = *momentum*; prefix *pico-*.
Q = *electric charge*.	q = *elec. charge*; "canonical coordinate".
R = radius; electrical *resistance*.	r = *radius*.
S = *entropy*; surface area.	s = *second(s)*; distance.
T = *temperature*.	t = *time*.
U = potential energy; internal energy.	u = an abstract variable; a velocity.
V = *Volts*; *volume*; *potential energy*.	v = *velocity*.
W = *work*; weight.	w = a small weight; a width.
X = an abstract function, as $X(x)$.	x = *distance*; any *independent variable*.
Y = an abstract function, as $Y(y)$.	y = an abstract *dependent variable*.
Z = atomic number; $Z(z)$.	z = an abstract *dependent variable*.

Table 3.2 Greek symbols commonly used in Physics

GREEK LETTERS: (Capital Greek letters that look the same as Roman are omitted.)

	α = fine structure constant; an angle.
	$\beta = v/c$; an angle.
Γ = *torque*; a rate.	$\gamma = E/mc^2$; an angle.
Δ = "change in...", as in Δx.	δ = an infinitesimal; same as Δ.
	ϵ = an infinitesimal quantity.
$\mathcal{E}$ = "electromotive force".	ε = an energy.
	ζ = a general parameter.
	η = index of refraction.
Θ = an angle.	θ = an *angle* (most common symbol).
	ι = rarely used (looks like an i).
	κ = arcane version of k.
Λ = a rate; a type of baryon.	λ = *wavelength*; a rate.
	μ = reduced mass; muon; prefix *micro-*.
	ν = *frequency* in cycles/s (Hz); neutrino.
Ξ = a type of baryon.	ξ = a general parameter.
Π = *product* operator.	$\pi = 3.14159\ldots$; pion (a meson).
	ρ = *density* per unit volume; resistivity.
Σ = *summation* operator.	σ = cross section; area density; conductivity.
	τ = a *mean lifetime*; tau lepton.
Υ = an elementary particle.	υ = rarely used (looks like v).
Φ = a *wave function*; an angle.	ϕ = an angle; a wave function.
	χ = susceptibility.
Ψ = a *wave function*.	ψ = a *wave function*.
	ω = *angular frequency* (radians/s).

Table 3.3 Mathematical symbols commonly used in Physics

OPERATORS:

$\rightarrow$ = "...approaches in the limit..." (as in $\Delta t \rightarrow 0$).

∂ = *partial derivative* operator (as in $\frac{\partial F}{\partial x}$).

∇ = *gradient* operator (as in $\nabla\phi = \hat{x}\frac{\partial\phi}{\partial x} + \hat{y}\frac{\partial\phi}{\partial y} + \hat{z}\frac{\partial\phi}{\partial z}$).

$\int$ = *integral* operator as in $\int y(x)dx$

LOGICAL SYMBOLS: (Handy shorthand that I use a lot!)

$\therefore$ = "Therefore..." $\Rightarrow$ = "...implies..." $\equiv$ = "...is *defined* to be..."

$\exists$ = "there exists..." $\ni$ = "...such that..."

/ [a *slash* through any logical symbol] = *negation*; *e.g.* $\not\Rightarrow$ = "...does *not* imply..."

3.4 Functions

Mathematics is often said to be the language of Physics. This is not the whole truth, but it is part of the truth; one ubiquitous characteristic of Physics (the human activity), if not physics (the supposed methodology of nature), is the expression of relationships between measurable quantities in terms of mathematical formulae. The advantages of such notation are that it is concise, precise and "elegant," and that it allows one to calculate quantitative predictions which can be compared with measured experimental results to test the validity of the description.

The nearly-universal image used in such mathematical descriptions of nature is the FUNCTION, an abstract concept symbolized in the form $y(x)$ [read "y of x"] which formally represents *mathematical shorthand* for a *recipé* whereby a value of the "dependent variable" y can be calculated for any given value of the "independent variable" x.

The *explicit* expression of such a *recipé* is always in the form of an *equation*. For instance, the answer to the question, "What is $y(x)$?" may be "$y = 2 + 5x^2 - 3x^3$." This tells us how to get a numerical value of y to "go with" any value of x we might pick. For this reason, in Mathematics (the human activity) it is often formally convenient to think of a function as a *mapping* — *i.e.* a collection of pairs of numbers (x, y) with a concise prescription to tell us how to find the y which goes with each x. In this sense it is also easier to picture the "inverse function" $x(y)$ which tells us how to find a value of x corresponding to a given y. [There is not always a unique answer. Consider $y = x^2$.] On the other hand, whenever we go to *use* an explicit formula for $y(x)$, it is essential to think of it as a *recipé* — *e.g.* for the example described above, "Take the quantity inside the parentheses (whatever it is) and do the following arithmetic on it: first cube whatever-it-is and multiply by 3; save that result and subtract it from the result you get when you multiply 5 by the square of whatever-it-is; finally add 2 to the difference and *voila!* you have the value of y that goes with $x =$ whatever-it-is."

This is most easily understood by working through a few examples, which we will do shortly.

3.4.1 Formulae *vs* Graphs

In Physics we often prefer the image of the GRAPH, because the easiest way to compare *data* with a theoretical function in a holistic manner is to plot both on a common graph. (The right hemisphere is best at holistic perception, so we go right in through the visual cortex.) Fortunately, the issue of whether a graph or an equation is "better" is entirely subjective, because *for every function there is a graph* — although sometimes the interesting features are only obvious when small regions are blown up, or when one or the other variable is plotted on a logarithmic scale, or suchlike.

Nevertheless, this process of translating between left and right hemispheres has far-reaching significance to the practice of Physics. When we draw a *graph*, we cathect the *pattern recognition* skills of our visual cortex, a large region of the brain devoted mainly to forming conceptual models of the "meaning" of visual stimuli arriving through the optic nerve. This is the part that learned to tell the difference between a leaf fluttering in the breeze and the tip of a leopard's tail flicking in anticipation; it performs such pattern recognition without our conscious intervention, and thus falls into the "intuitive" realm of mental functions. It is fantastically powerful, yet not entirely reliable (recall the many sorts of "optical illusions" you have seen).

The mere fact that many (not all) physicists like to display their results in graphical form offers a hint of our preferred procedure for hy-

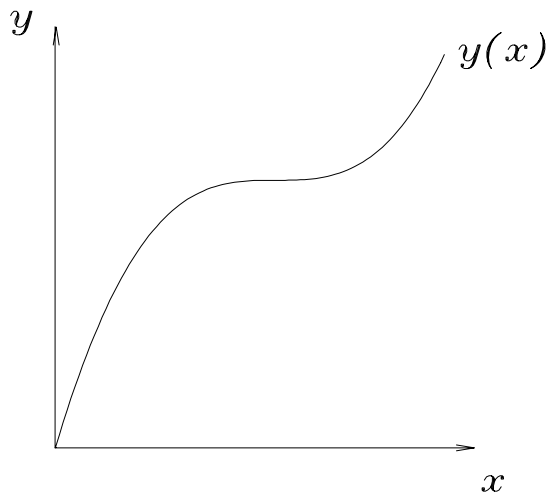


Figure 3.1 A typical graph of $y(x)$ [read "y as a function of x"].

pothesis formation (Karl Popper's *conjectures*). Namely, the data are "massaged" [not the same as "fudged" — massaging is strictly legitimate and all the steps are required to be explained clearly] until they can be plotted on a graph in a form that "speaks for itself" — *i.e.* that excites the strongest pattern-recognition circuit in the part of our visual cortex that we use on science — namely, the straight line. Then the author/speaker can enlist the collaboration of the audience in forming the hypothesis that there is a linear relationship between the two "massaged" variables.

For a simple example, imagine that a force F actually varies inversely with the square of distance r: $F(r) = k/r^2$ with k some appropriate constant. A graph of measured values of F *vs.* r will not be very informative to the eye except to show that, yes, F sure gets smaller fast as r increases. But if the ingenious experimenter discovers by hook or by crook that a plot of F *vs.* $1/r^2$ (or $1/F$ *vs.* r^2 or $\sqrt{F}$ *vs.* $1/r$ or...) comes out looking like a straight line, you can be sure that the data will be presented in that form in the ensuing talk or paper. The rigourous validity of this technique may be questionable, but it works great.

You may have perceived an alarmingly liberal use of *algebra* (or at least algebraic notation) in this last section. I have "pulled no punches" here, showing the "proper" Physics notation for functions and derivatives right at the beginning, for several reasons. First is simple intellectual honesty: this is the mathematical notation used in Physics; why pretend otherwise? Eventually you want to be able to translate this notation into your own favourite representation (words, graphs, whatever) so why not start getting used to it as soon as possible? Second, this is a sort of "implosion therapy" whereby I treat any math phobias by saturating the fear response: once you know it can't get any worse, it starts getting better. Be advised that we will spend the next few chapters (off and on) getting used to algebraic representations and their graphical counterparts.

Chapter 4

The Language of Math

Soon we will tackle the problem of *measurement*, with all its pitfalls and practical tricks. You may then sympathize with Newton, who took such delight in retreating into the Platonic ideal world of pure mathematics, where relationships between "variables" are not fraught with messy errors, but defined by simple and elegant prescriptions. No matter that we are unable to measure these perfect relationships directly; this is merely an unfortunate consequence of our imperfect instruments. (Hmmm....) But first we need to describe the notational conventions to be used in this book for the language of Mathematics, without which Physics would have remained mired in the rich but confusing ambiguities of natural language. Here is where we assemble the *symbols* into *structures* that express (in some conventional idiom) the *relationships* between the "things" the symbols represent.

Please do not feel insulted if the following review seems too elementary for someone at your level. I have always found it soothing to review material that I already know well, and am usually surprised to discover how much I forgot in such a short while. Also, I think you'll find it picks up a bit later on.

4.1 Arithmetic

We have already dwelt upon the formalism of Number Systems in a previous Chapter, where we reminded ourselves that just counting to ten on paper involves a rather sophisticated and elaborate representational scheme that we all learned as children and which is now *tacit* in our thought processes until we go to the trouble to dismantle it and consider possible alternatives.

Arithmetic is the basic algebra of Numbers and builds upon our tacit understanding of their conventional representation. However, it would be emphatically wrong to claim that, "Arithmetic is made up of Numbers, so there is nothing to Arithmetic but Numbers." Obviously Arithmetic treats a new *level* of understanding of the properties of (and the relationships between) Numbers — something like the Frank Lloyd Wright house that was not there *in* the bricks and mortar of which it is built. [One can argue that in fact the conceptual framework of Number Systems implicitly contains intimations of Arithmetic, but this is like arguing that the properties of atoms are implicit in the behaviour of electrons; let's leave that debate for later.]

We learn Arithmetic at two levels: the *actual* level ("If I have two apples and I get three more apples, then I have five apples, as long as nothing happens to the first two in the meantime.") and the *symbolic* level ("2+3=5"). The former level is of course both *concrete* (as in all the *examples*) and profoundly *abstract* in the sense that one learns to understand that two of anything added to three of the same sort of thing will make five of them, independent of words or numerical symbols. The latter level is

more for *communication* (remember, we have to adopt and adapt to a notational convention in order to express our ideas to each other) and for *technology* — *i.e.* for developing *manipulative tricks* to use on Numbers.

Skipping over the simple Arithmetic I assume we all know tacitly, I will use *long division* as an example of the conventional technology of Arithmetic.[1] We all know (today) how to do long division. But can we *explain how it works*? Suppose you were Cultural Attaché to Alpha Centauri IV, where the local intelligent life forms were interested in Earth Math and had just mastered our ridiculous decimal notation. They understand addition, subtraction, multiplication and division perfectly and have developed the necessary skills in Earth-style gimmicks (carrying, *etc.*) for the first three, but they have no idea how we actually go about dividing one multi-digit number by another. Try to imagine how you would explain the long division trick. Probably by example, right? That's how most of us learn it. Our teacher works out *beaucoup* examples on the blackboard and then gives us *beaucoup* homework problems to work out ourselves, hopefully arrayed in a sequence that sort of leads us through the process of *induction* (not a part of Logic, according to Karl Popper, but an important part of human thinking nonetheless) to a bootstrap grasp on the method. Nowhere, in most cases, does anyone give us a full rigourous derivation of the method, yet we all have a deep confidence in its universality and reliability — which, I hasten to add, I'm sure *can* be rigourously derived if we take the trouble. Still, we are awfully trusting....

The point is, as Michael Polanyi has said, *"We know more than we can tell."* The *tacit* knowledge of Arithmetic that you possess represents an enormous store of

- sophisticated abstract understanding
- arbitrary conventions of representational notation
- manipulative technology

that have already coloured your thought processes in ways that neither you nor anyone else will ever be able to fathom. We are all brainwashed by our Grammar school teachers![2] This book, if it is of any use whatsoever, will have the same sort of effect: it will "warp" your thinking forever in ways that cannot be anticipated. So if you are worried about being "contaminated" by Scientism (or whatever you choose to label the paradigms of the scientific community) then stop reading immediately before it is too late! (While you're at it, there are a few other activities you will also have to give up....)

4.2 Geometry

In Grammar school we also learn to recognize (and learn the grammar of) geometrical shapes. Thus the Right Hemisphere also gets early training. Later on, in High School, we

[1] No doubt the useful lifetime of this example is only a few more years, since many students now learn to divide by punching the right buttons on a hand calculator, much to the dismay of their aged instructors. I am not so upset by this — one arithmetic manipulation technology is merely supplanting another — except that "long division" is *in principle* completely understood by its user, whereas few people have any idea what actually goes on inside an electronic calculator. This dependence on mysterious and unfamiliar technology may have unpleasant long-term psychological impact, perhaps making us all more willing to accept the judgements of authority figures without question.... But in Mathematics, as long as you have *once* satisfied yourself completely that some technology is indeed trustworthy and reliable, of course you should make use of it! (Do you *know* that your calculator *always* gives the right answers...?)

[2] It occurs to me that Grammar school is called Grammar school because it is where we learn *grammar* — *i.e.* the *conventional representations* for things, ideas and the relationships between them, whether in verbal language, written language, mathematics, politics, science or social behaviour. These are usually called "rules" or even (when a particularly heavy-handed emphasis is desired) "laws" of notation or manipulation or behaviour. We also pick up a little *technology*, which in this context begins to look pretty innocuous!

get a bit more insight into the *intrinsic* properties of Euclidean space (*i.e.* the "flat" kind we normally *seem* to be occupying).

4.2.1 Areas of Plane Figures

- The area A of a *square* is the *square* of the length ℓ of any one of its 4 sides: $A = \ell^2$. In fact the question of which word "square" is named after which is a sort of chicken *vs.* egg problem for which there is no logical resolution (even though there may be an historically correct etymological answer).
- The area A of a *rectangle* (a bit more general) is the product of the length b of a long side ("base") and the length h of a short side ("height"): $A = bh$.
- The area A of a *triangle* with base b and height h (measured from the opposite vertex down perpendicular to the base) is $A = \frac{1}{2}bh$. (This is easy to see for a *right* triangle, which is obviously half a rectangle, sliced down the diagonal. You may want to convice yourself that it is also true for "any old triangle.")
- The area A of a *circle* of radius r is given by $A = \pi r^2$ where π is a number, approximately 3.14159 [it takes an infinite number of decimal digits to get it exactly; this is because π is an *irrational number*[3] — *i.e.* one which cannot be expressed as a ratio of integers], defined in turn to be the ratio of the *circumference* ℓ of a circle to its *diameter* d: $\pi = \ell/d$ or $\ell = \pi d$.

Were you able to visualize all these simple plane (2-dimensional) shapes "in your head" without resort to actual drawings? If so, you may have a "knack" for geometry, if not Geometry. If it was confusing without the pictures, they are provided in Fig. 4.1 with the appropriate labels.

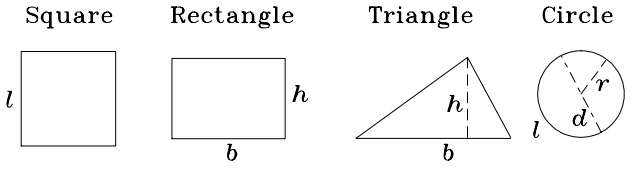


Figure 4.1 A few plane geometrical shapes, with labels.

4.2.2 The Pythagorean Theorem:

> The square of the length of the hypotenuse of a right triangle is equal to the sum of the squares of the lengths of the two shorter sides.

I.e. for the Left Hemisphere we have

$$c^2 = a^2 + b^2 \tag{1}$$

where a, b and c are defined by the labelled picture of a right triangle, shown in Fig. 4.2, which cathects the Right Hemisphere and gets the two working together.

4.2.3 Solid Geometry

Most of us learned how to calculate the *volumes* of various solid or 3-dimensional objects even before we were told that the name for the system of conventions and "laws" governing such topics was "Solid Geometry." For instance, there is the *cube*, whose volume V is the *cube* (same chicken/egg problem again) of the length ℓ of one of its 8 *edges*: $V = \ell^3$. Similarly, a *cylinder* has a volume V equal to the product of its cross-sectional area A and its height h perpendicular to the base: $V = Ah$. Note that this works just as well for *any shape* of the cross-section — square, rectangle, triangle, circle or even some irregular oddball shape.

[3] I do not know the proof that π is an irrational number, but I have been told by Mathematicians that it is, and I have never had any cause to question them. In principle, this is reprehensible (shame on me!) but I am not aware of any practical consequences one way or the other; if anyone knows one, please set me straight!

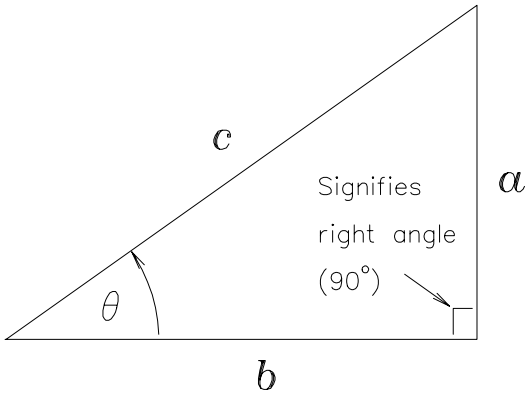


Figure 4.2 A right triangle with hypotenuse c and short sides a and b. The right angle is indicated and the angle θ is defined as shown. Note that a is always the (length of the) side "across from" the *vertex* forming the angle θ. This convention is essential in the *trigonometric* definitions to follow.

If you were fairly advanced in High School math, you probably learned a bit more abstract or general stuff about solids. But the really deep understanding that (I hope) you brought away with you was an awareness of the *qualitative* difference between 1-dimensional *lengths*, 2-dimensional *areas* and 3-dimensional *volumes*. This awareness can be amazingly powerful even without any "hairy Math details" if you consider what it implies about how these things change with *scale*.[4]

4.3 Algebra 1

A handy trick for introducing Algebra to young children (who have not yet learned that it is supposed to be too hard for them) is to phrase a typical Algebra problem in the following way: "I'm thinking of a number, and its name is 'x' ...so if $2x + 3 = 7$, what is x?" (You may have to spend a little time explaining the notational conventions of equations and that $2x$ means 2 *times* x.) Most 7-year-olds can then solve this problem by inspection (my son and daughter both could!) but they may not be able to tell you *how* they solved it. This suggests either that early Arithmetic has already sown the seeds of algebraic manipulation conventions or that there is some understanding of such concepts "wired in" to our brains. We will never know how much of each is true, but certainly neither is entirely false!

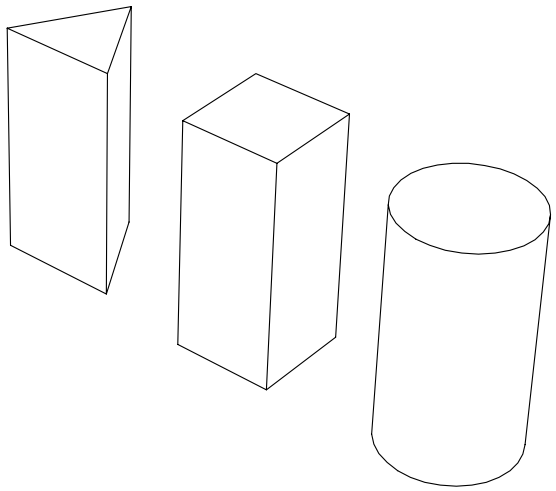


Figure 4.3 Triangular, square and circular right cylinders.

What we learn in High School Algebra is to examine *how* we solve problems like this and to refine these techniques by adapting ourselves to a particular formalism and technology. Unfortunately our intuitive understanding is often trampled upon in the process — this happens when we are actively discouraged from treating the technology as a convenient representation for what we already understand, rather than a definition of correct procedure.

In Algebra we learn to "solve" equations. What does that mean? Usually it means that we are to take a (relatively) complicated equation that

[4] For instance, it explains easily why the largest animals on Earth have to live in the sea, why insects can lift so many times their own weight, why birds have an easier time flying than airliners, why bubbles form in beer and how the American nuclear power industry got off to a bad start. All in due time....

has the "unknown" (often but not always called "x") scattered all over the place and turn it into a (relatively) simple equation with x on the left-hand side by itself and a bunch of other symbols (*not* including x) on the right-hand side of the "=" sign. Obviously this particular *format* is "just" a convention. But the *idea* is independent of the representation: "solve" for the "unknown" quantity, in this case x.

There are a few basic rules we use to "solve" problems in Algebra; these are called "laws" by Mathematicians who want to emphasize that you are not to question their content or representation.

- **Definition of Zero:**

$$a - a = 0 \tag{2}$$

- **Definition of Unity:**

$$\frac{a}{a} = 1 \tag{3}$$

- **Commutative Laws:**[5]

$$a + b = b + a \tag{4}$$

$$\text{and} \qquad ab = ba \tag{5}$$

- **Distributive Law:**

$$a(b + c) = ab + bc \tag{6}$$

- **Sum or Difference of Two Equations:** Adding (or subtracting) the same thing from both sides of an equation gives a new equation that is still OK.

$$\begin{array}{rrcl} & x & -a = & b \\ + (& & a = & a \quad) \\ & x & = & b + a \end{array} \tag{7}$$

$$\begin{array}{rrcl} & x & +c = & d \\ - (& & c = & c \quad) \\ & x & = & d - c \end{array} \tag{8}$$

- **Product or Ratio of Two Equations:** Multiplying (or dividing) both sides of an equation by the same thing also gives a new equation that is still OK.

$$\begin{array}{rrcl} & x/a = & & b \\ \times (& a = & & a \quad) \\ & x = & & ab \end{array} \tag{9}$$

$$\begin{array}{rrcl} & cx = & & d \\ \div (& c = & & c \quad) \\ & x = & & d/c \end{array} \tag{10}$$

These "laws" may seem pretty trivial (especially the first two) but they define the rules of Algebra whereby we learn to manipulate the form of equations and "solve" Algebra "problems." We quickly learn equivalent *shortcuts* like "moving a factor from the bottom of the left-hand-side [often abbreviated LHS] to the top of the right-hand side [RHS]:"

$$\frac{x - a}{b} = c + d \quad \Rightarrow \quad x - a = b(c + d) \tag{11}$$

and so on; but each of these is just a well-justified concatenation of several of the fundamental steps. *(Emergence!)*

You may ask, "Why go to so much trouble to express the obvious in such formal terms?" Well, as usual the obvious is not necessarily the truth. While the real, imaginary and complex numbers may all obey these simple rules, there are perfectly legitimate and useful fields of "things" (usually some sort of *operators*) that do *not* obey all these rules, as we shall see much

[5] Note that *division* is *not* commutative: $a/b \neq b/a$! Neither is *subtraction*, for that matter: $a - b \neq b - a$. The Commutative Law for *multiplication*, $ab = ba$, holds for ordinary numbers (real and imaginary) but it does *not* necessarily hold for all the mathematical "things" for which some form of "multiplication" is defined! For instance, the *group* of *rotation operators* in 3-dimensional space is *not* commutative — think about making two successive rotations of a rigid object about perpendicular axes in different order and you will see that the final result is different! This seemingly obscure property turns out to have fundamental significance. We'll talk about such things later.

later in the course (probably). It is generally a good idea to know your own assumptions; we haven't the time to keep reexamining them constantly, so we try to state them as plainly as we can and keep them around for reference "just in case. . . ."

4.4 Trigonometry

Trigonometry is a specialized branch of Geometry in which we pay excruciatingly close attention to the properties of *triangles*, in particular *right* triangles. Referring to Fig. 4.2 again, we define the *sine* of the angle θ (abbreviated $\sin\theta$) to be the ratio of the "far side" a to the hypotenuse c and the *cosine* of θ (abbreviated $\cos\theta$) to be the ratio of the "near side" b to the hypoteneuse c:

$$\sin\theta \equiv \frac{a}{c} \qquad \cos\theta \equiv \frac{b}{c} \tag{12}$$

The other trigonometric functions can easily be defined in terms of the sin and cos:

tangent: $\tan\theta \equiv \dfrac{a}{b} = \dfrac{\sin\theta}{\cos\theta}$

cotangent: $\cot\theta \equiv \dfrac{b}{a} = \dfrac{\sin\theta}{\cos\theta} = \dfrac{1}{\tan\theta}$

secant: $\sec\theta \equiv \dfrac{c}{b} = \dfrac{1}{\cos\theta}$

cosecant: $\csc\theta \equiv \dfrac{c}{a} = \dfrac{1}{\sin\theta}$

For the life of me, I can't imagine why anyone invented the *cotangent*, the *secant* and the *cosecant* — as far as I can tell, they are totally superfluous baggage that just slows you down in any actual calculations. Forget them. [Ahhhh. I have always wanted to say that! Of course you are wise enough to take my advice with a grain of salt, especially is you want to appear clever to Mathematicians. . . .]

The *sine* and *cosine* of θ are our trigonometric workhorses. In no time at all, I will be wanting to think of them as *functions* — *i.e.* when you see "$\cos\theta$" I will want you to say, "cosine of theta" and think of it as $\cos(\theta)$ the same way you think of $y(x)$. Whether as simple ratios or as functions, they have several delightful properties, the most important of which is obvious from the Pythagorean Theorem:[6]

$$\cos^2\theta + \sin^2\theta = 1 \tag{13}$$

where the notation $\sin^2\theta$ means the *square* of $\sin\theta$ — *i.e.* $\sin^2\theta \equiv (\sin\theta) \times (\sin\theta)$ — and similarly for $\cos\theta$. This convention is adopted to avoid confusion, believe it or not. If we wrote "$\sin\theta^{\,2}$" it would be impossible to know for sure whether we meant $\sin(\theta^2)$ or $(\sin\theta)^2$; we could always put parentheses in the right places to remove the ambiguity, but in this case there is a convention instead. (People always have conventions when they are tired of thinking!)

I will need other trigonometric identities later on, but they can wait — why introduce math until we need it? [I have made an obvious exception in this Chapter as a whole only to "jump start" your Mathematical language (re)training.]

4.5 Algebra 2

"I'm thinking of a number, and its name is 'x' . . . " So if

$$ax^2 + bx + c = 0, \tag{14}$$

what is x? Well, we can only say, "It depends." Namely, it depends on the values of a, b and c, whatever they are. Let's suppose the *dimensions* of all these "parameters" are mutually consistent[7] so that the equation makes sense.

[6] Surely you aren't going to take my word for this! *Convince yourself* that this formula is really true!

[7] In Mathematics we never worry about such things; all our symbols represent *pure numbers*; but in Physics we *usually* have to express the value of some physical quantity in units which make sense and are consistent with the units of other physical quantities symbolized in the same equation!

Then "it can be shown" (a classic phrase if there ever was one!) that the "answer" is *generally*[8]

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \qquad (15)$$

This formula (and the preceding equation that defines what we mean by a, b and c) is known as the *Quadratic Theorem*, so called because it offers "the answer" to *any* quadratic equation (*i.e.* one containing powers of x up to and including x^2). The power of such a *general* solution is prodigious. (Work out a few examples!) It also introduces an interesting new way of looking at the relationship between x and the *parameters* a, b and c that determine its value(s). Having x all by itself on one side of the equation and no x's anywhere on the other side is what we call a "solution" in Algebra. Let's make a simpler version of this sort of equation: "I'm thinking of a number, and its name is 'y' ..." So if $y = x^2$, what is y? The answer is again, "It depends!" (In this case, upon the value of x.) And that leads us into a new subject....

4.6 Calculus

In a *stylistic* sense, Algebra starts to become Calculus when we write the preceding example, $y = x^2$, in the form

$$y(x) = x^2$$

which we read as "y of x equals x squared." This is how we signal that we mean to think of y as a *function* of x, and right away we are leading into the terminology of Calculus. Recall the final sections of the preceding Chapter.

However, Calculus really begins when we start talking about the *rate of change* of y as x varies.

[8]The $\pm$ symbol means that *both* signs (+ and −) should represent legitimate answers.

4.6.1 Rates of Change

One thing that is easy to "read off a graph" of $y(x)$ is the *slope* of the curve at any given point x. Now, if $y(x)$ is quite "curved" at the point of interest, it may seem contradictory to speak of its "slope," a property of a *straight* line. However, it is easy to see that as long as the curve is *smooth* it will always *look like a straight line* under sufficiently high *magnification*. This is illustrated in Fig. 4.4 for a typical $y(x)$ by a process of successive magnifications.

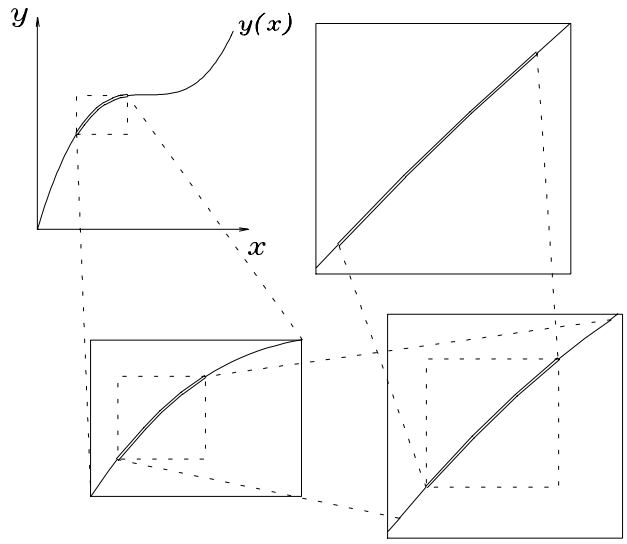


Figure 4.4 A series of "zooms" on a segment of the curve $y(x)$ showing how the *curved* line begins to look more and more like a *straight* line under higher and higher magnification.

We can also prescribe an algebraic method for *calculating* the slope, as illustrated in Fig. 4.5: the *definition* of the "slope" is the ratio of the increase in y to the increase in x on a vanishingly small interval. That is, when x goes from its initial value x_0 to a slightly larger value $x_0+\Delta x$, the curve carries y from its initial value $y_0 = y(x_0)$ to a new value $y_0+\Delta y = y(x_0+\Delta x)$, and the slope of the curve at $x = x_0$ is given by $\Delta y/\Delta x$ for a vanishingly small Δx. When a small change like Δx gets *really* small (*i.e.* small enough that the curve looks like a straight

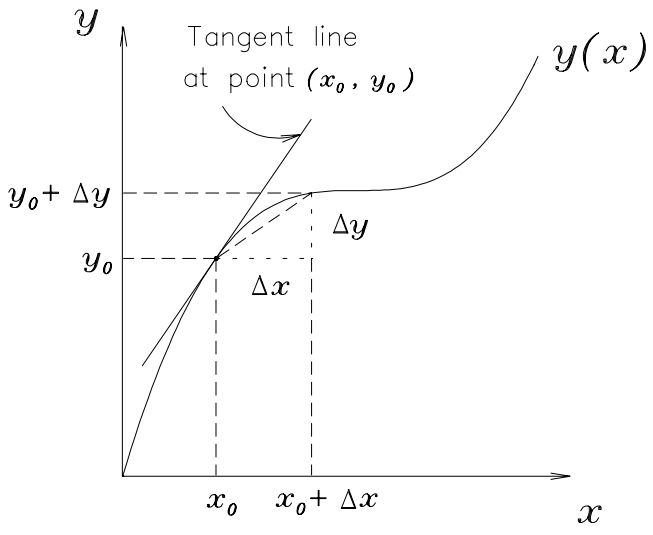


Figure 4.5 A graph of the function $y(x)$ showing how the *average* slope $\Delta y/\Delta x$ is obtained on a *finite* interval of the curve. By taking smaller and smaller intervals, one can eventually obtain the slope at a *point*, dy/dx.

line on that interval, or "small enough to satisfy whatever criterion you want," then we write it differently, as dx, a "*differential*" (vanishingly small) change in x. Then the exact definition of the SLOPE of y with respect to x at some particular value of x, written in conventional Mathematical language, is

$$\frac{dy}{dx} \equiv \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \equiv \lim_{\Delta x \to 0} \frac{y(x+\Delta x) - y(x)}{\Delta x} \tag{16}$$

This is best understood by an example: consider the simple function $y(x) = x^2$. Then

$$y(x+\Delta x) = (x+\Delta x)^2 = x^2 + 2x\Delta x + (\Delta x)^2$$

$$\text{and} \quad y(x+\Delta x) - y(x) = 2x\Delta x + (\Delta x)^2.$$

Divide this by Δx and we have

$$\frac{\Delta y}{\Delta x} = 2x + \Delta x.$$

Now let Δx shrink to zero, and all that remains is

$$\frac{\Delta y}{\Delta x} \underset{\Delta x \to 0}{\longrightarrow} \frac{dy}{dx} = 2x.$$

Thus the slope [or *derivative*, as mathematicians are wont to call it] of $y(x) = x^2$ is $dy/dx = 2x$. That is, the slope increases linearly with x. The slope *of* the slope — which we call[9] the *curvature*, for obvious reasons — is then trivially $d(dy/dx)/dx \equiv d^2y/dx^2 = 2$, a constant. Make sure you can work this part out for yourself.

We have defined all these algebraic solutions to the geometrical problem of finding the slope of a curve on a graph in completely abstract terms — "x" and "y" indeed! What *are* x and y? Well, the whole idea is that they can be anything you want! The most common examples in Physics are when x is the *elapsed time*, usually written t, and y is the *distance travelled*, usually (alas) written x. Thus in an elementary Physics context the function you are apt to see used most often is $x(t)$, the position of some object as a function of time. This particular function has some very well-known derivatives, namely $dx/dt = v$, the *speed* or (as long as the motion is in a straight line!) *velocity* of the object; and $dv/dt \equiv d^2x/dt^2 = a$, the *acceleration* of the object. Note that both v and a are themselves (in general) functions of time: $v(t)$ and $a(t)$. This example so beautifully illustrates the "meaning" of the slope and curvature of a curve as first and second derivatives that many introductory Calculus courses and virtually all introductory Physics courses use it as *the* example to explain these Mathematical conventions. I just had to be different and start with something a little more formal, because I think you will find that the idea of one thing being a *function* of another thing, and the associated ideas of graphs and slopes and curvatures, are handy notions worth putting to work far from their traditional realm of classical kinematics.

[9] This differs from the conventional mathematical definition of *curvature*, $\kappa \equiv d\phi/ds$, where ϕ is the tangential angle and s is the arc length, but I like mine better, because it's simple, intuitive and useful. (OK, I'm a Philistine. So shoot me. ;-) Thanks to Mitchell Timin for pointing this out.

Chapter 5

Measurement

Earlier we discussed the tactical problems associated with *describing* quantitative measurements, reminding ourselves that the tools we use (numbers, dimensions, units) are almost perfectly arbitrary in isolation but embody a functional or *relational* truth in the "grammar" of their use. Accepting these tools provisionally, we turn now to the far messier problem of actually *performing* measurements.

5.1 Tolerance

(Advertising Your Uncertainty)

Virtually all [I could follow the consensus and say *all*, but I feel like hedging] "scientific" procedures involve *measurement* of experimental parameters such as distance, time, velocity, mass, energy, temperature, ... *etc.* Virtually all measurements are subject to *error*; that is, they may be *inaccurate* (wrong) by some unknown amount due to effects ranging from errors in recording ["I said 3.32, not 3.23!"] to miscalibrated instruments ["I thought these tic marks were *centimetres*!"]. Such "systematic errors" are embarrassing to the experimenter, as they imply poor technique, and are always hard to estimate; but we are honour-bound to try. An entirely different source of error that conveys *no* negative connotations on the experimenter is the fact that all measurements have limited *precision* or "tolerance" — limited by the "marks" on the generalized "ruler" used for measuring-by-comparison. (*E.g.*, the distance your measure with a micrometer is more precisely known than the distance you measure with a cloth tape measure.)

Knowing this, most scientists and virtually all physicists have an æsthetic about measured values of things: they are *never* to be reported without an explicit estimation of their *uncertainty*. That is, measurements must always be reported in the form

(VALUE ± UNCERTAINTY) UNITS

or equivalent notation (sometimes a shorthand version), such as 3.1416(12) radians, meaning (3.1416 ± 0.0012) radians. [The (12) means the uncertainty in the last two digits is ± 12.] This shorthand form is convenient for long strings of digits with only the last 1 or 2 digits uncertain, but the explicit form with the ± is more pleasing to the æsthetic mentioned above.

When, as in some elementary particle physics experiments lasting many years and costing millions of dollars, a great deal of effort has gone into measuring a single number, it is common practice to make a clear distinction between "statistical errors" (the *precision* of our instrumentation) and suspected "systematic errors" (mistakes). In most situations, however, both are lumped together or "added in quadrature" (the total uncertainty is the square root of the sum of the squares of the uncertainties due to all the independent sources of error).[1] It is

[1] More on this later....

considered poor form to cavalierly overestimate one's uncertainty to reduce the significance of deviations from expectations.

To write a measured value without its *tolerance* (uncertainty, "possible error," *etc.*) is as bad form as leaving out the *units* of the measurement. The significance of your measurement is lost. To do this in the presence of physicists is like ordering Ripple with your meal at Maxim's. Sadly, values are slipping throughout society, and otherwise respectable scientists can often be heard to quote numbers without specifying uncertainties. The best we can do is to be sure we do not contribute to this decay.

5.1.1 Sig Figs

In other disciplines (Chemistry and Engineering especially) some people like to stick with the explicit notational conventions described above but others are fond of the supposed economy introduced by the so-called "**significant figures**" (or "sig figs") convention, in which uncertainties are not explicitly expressed but are *implicit* in the last significant digit of the number written down. I have listened to so many long explanations of this convention in attempts to clarify its proper use and interpretation for confused students that I prefer to adopt the short form:

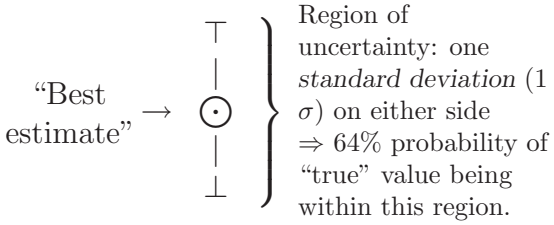
Forget "sig figs" in Physics.

Whenever you make a Physics measurement, express your result *and its uncertainty* in one of the *explicit* forms described earlier. The issue of how many significant digits to write down is then one of common sense rather than convention. It is silly to write down $\mathbf{0.123456 \pm 0.03}$, but it is not wrong.

One reason I encourage you to eschew "sig figs" is that even silly-looking results like $\mathbf{0.1234 \pm 0.0321}$ (why would anyone express an *uncertainty* to three significant digits?) actually make sense in special circumstances. For instance, if thousands of professors from all over the world spend half a billion dollars over twenty years just to measure one really important property of elementary particles, you can hardly blame them for expressing the result as painstakingly as possible. However, even then it makes no sense to show more significant digits in the result than in its uncertainty; that's just common sense. So use yours.

5.1.2 Graphs &Error Bars

When plotting points on a *graph*, the uncertainty is included in the form of "error bars" which look like this:



5.1.3 Vector Tolerance

Allow me to slip into something a little more formal. . . .

Usually this topic would be called "Error Propagation in Functions of Several Variables" or something like that; I have used the term "vector tolerance" because (*a*) the word "error" has these perjorative connotations for most people, whereas "tolerance" is usually considered a *good* thing;[2] (*b*) when our final result is calculated in terms of several other quantities, each of which is uncertain by some amount, and when those uncertainties are *independent* of each other, we get a situation much like trying to define the overall length of a *vector* with several independent perpendicular components. Each contribution to the overall uncertainty can be positive or negative, and on av-

[2] "Uncertainty" is somewhere in between.

erage you would not expect them to all add up; that would be like assuming that if one were positive they all must be. So we *square* each contribution, add the squares and take the square root of the sum, just as we would do to find the length of a vector from its components.

The way to do this is easily prescribed if we use a little calculus notation: suppose the "answer" A is a function of several variables, say x and y. We write $A(x, y)$. So what happens to A when x changes by some amount δx?[3] Simple, we just write $\delta A_x \approx (\partial A/\partial x)\ \delta x$ where the $_x$ subscript on δA_x reminds us that this is just the contribution to the change in A from that little change in x, not from any changes in y; the $\approx$ sign acknowledges that this doesn't get exact until $\delta x \to dx$, which is *really* small; and the ∂ symbols are like derivatives except they remind us that *we are treating y as if it were a constant* when we take this derivative.

The same trick works for changes in y, of course, so then we have two "orthogonal" shifts of the result to combine into one *uncertainty* in A. I have already given the prescription for this above. The formula reads

$$(\delta A)^2 \approx \left(\frac{\partial A}{\partial x}\ \delta x\right)^2 + \left(\frac{\partial A}{\partial y}\ \delta y\right)^2 \qquad (1)$$

This can be extended to a function of N variables $\{x_1, x_2, \cdots x_i \cdots x_N\}$:

$$(\delta A)^2 \approx \sum_{i=1}^{N} \left(\frac{\partial A}{\partial x_i}\ \delta x_i\right)^2 \qquad (2)$$

where the $\sum$ symbol means "sum over all terms of this form, with the index i running from 1 to N."

The treatment above is a little too "advanced" mathematically for some people (or for anyone on a bad day), so here are a few special cases that the enthusiast may wish to derive from the general form in Eq. (2):

- **Uncertainty in a Sum:** If $A(x, y) = a\ x + b\ y$, with constants a and b, then

  $$(\delta A)^2 \approx (a\ \delta x)^2 + (b\ \delta y)^2. \qquad (3)$$

  That is, just *add the uncertainties in quadrature.*

- **Uncertainty in a Product:** If $A(x, y) = a\ x\ y$, with constant a, then

  $$\left(\frac{\delta A}{A}\right)^2 \approx \left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2. \qquad (4)$$

  That is, just *add the fractional uncertainties in quadrature.*

- **Uncertainty in a Quotient:** If $A(x, y) = a\ x/y$, with constant a, then

  $$\left(\frac{\delta A}{A}\right)^2 \approx \left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2. \qquad (5)$$

  That is, just *add the fractional uncertainties in quadrature*, just like for a *product.*

- **Uncertainty in a Product of Power Laws:** If $A(x, y) = a\ x^p\ y^q$, with constant a, p and q, then

  $$\left(\frac{\delta A}{A}\right)^2 \approx \left(p\ \frac{\delta x}{x}\right)^2 + \left(q\ \frac{\delta y}{y}\right)^2 \qquad (6)$$

  which includes simple products and quotients.

These should get you through almost anything, if applied wisely.

5.2 Statistical Analysis

It's all very well to say that one should always report the results of measurements with uncertainties (or "errors" as they are often misleadingly called) specified; but this places a burden

[3] Notational convention: we use Δx to denote "a change in x, not necessarily tiny" whereas δx usually means "a little bitty change in x, but definitely finite!" and dx means "a change in x that is so teensy that it can be neglected relative to anything else but another really teensy thing." That last one (dx) is called a "differential" — Mathematicians don't like it much but Physicists use it all the time.

of judgement on the experimenter, who must estimate uncertainties in a manner fraught with individual idiosyncracies. Wouldn't it be nice if there were a way to *measure* one's uncertainty in a rigourous fashion?

Well, there is. It is a little tedious and complicated, but easily understood: one must make a *large number of repeated measurements of the same thing* and analyze the "scatter" of the answers!

Suppose we are trying to determine the "true" value of the quantity x. (We usually refer to unspecified things as "x" in this business.) It could be your pulse rate or some other simple physical observable.

We make N *independent* measurements x_i ($i = 1, 2, 3, \ldots, N$) under as close to identical conditions as we can manage. Each measurment, we suspect, is not terribly precise; but we don't know just how imprecise. (It could be largely due to some factor beyond our control; pulse rates, for instance, fluctuate for many reasons.)

Now, the x_i will "scatter" around the "true" x in a *distribution* that will put some x_i smaller than the true x and others larger. We *assume* that whatever the cause of the scatter, it is basically *random — i.e.* the exact value of one measurement x_{i+1} is not directly influenced by the value x_i obtained on the previous measurement. (Actually, perfect randomness is not only hard to define, but rather difficult to arrange in practice; it is sufficient that *most* fluctuations are random *enough* to justify the treatment being described here.) It is intuitively obvious (and can even be rigorously proved in most cases) that our *best estimate* for the "true" x is the average or *mean* value, $\bar{x}$, given by:[4]

$$\bar{x} \equiv \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{7}$$

But what is the *uncertainty* in $\bar{x}$? Let's call it $\bar{\sigma}_x$.

How can we *find* $\bar{\sigma}_x$ mathematically from the data? Well, if we assume that each individual measurement x_i has the same *single-measurement uncertainty* σ_x, then the distribution of x_i should look like a "bell-shaped curve" or *gaussian distribution*:

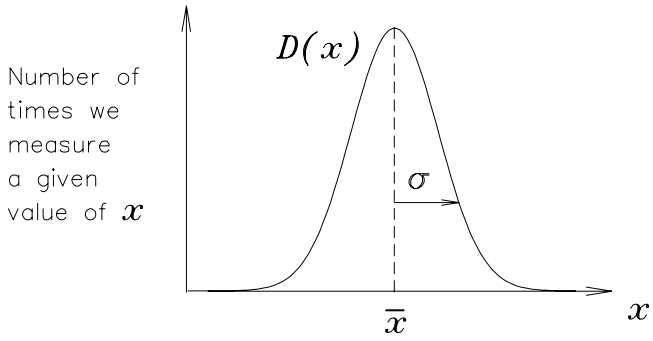


Figure 5.1 A typical graph of $\mathcal{D}(x)$, the *distribution* of x, defined as the relative frequency of occurence of different values of x from successive measurements. The "centre" of the distribution is at $\bar{x}$, the average or *mean* of x. The "width" of the distribution is 2σ (one σ on either side of the mean.

Obviously, $\Delta x_i \equiv x_i - \bar{x}$ is a measure of the "error" in the i^{th} measurement, but we cannot just find the average of Δx_i, since by definition the

[4] The symbol $\sum_{i=1}^{N}$ represents an *operator* called "summation" — it means that {the stuff to the right of the Σ}, which will always have a subscript $_i$ in one or more places, is to be thought of as the "i^{th} term" and all such terms with i values running from 1 to N are to be *added together* to form the desired result. So, for instance, $\sum_{i=1}^{N} x_i$ means $\{x_1 + x_2 + x_3 + \ldots + x_{N-1} + x_N\}$, or (to be more specific) if $N = 3$, just $\{x_1 + x_2 + x_3\}$. This may seem a little arcane, but it is actually a very handy compact notation for the rather common *summation* operation.

sum of all Δx_i is zero (there are just as many negative errors as positive errors). The way out of this dilemma is always to take the average of the *squares* of Δx_i, which are all positive. This "mean square" error is called the *variance*, s_x^2:

$$s_x^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{8}$$

and its square root, the "root mean square error", is called the *standard deviation* — which can be shown (rigorously, in many cases, although not without a good deal of math) to be the best possible estimate for the single-measurement uncertainty σ_x.

So we actually have a way of "calculating" our uncertainty *directly from the data*! This is quite remarkable. But wait. We have not just measured x once; we have measured it N times. Our instincts (?) insist that our *final best estimate* of x, namely the *mean*, $\bar{x}$, is determined more precisely than we would get from just a single measurement. This is indeed the case. The uncertainty in the mean, $\bar{\sigma}_x$, is smaller than σ_x. By how much? Well, it takes a bit of math to *derive* the answer, but you will probably not find it implausible to accept the result that $\bar{\sigma}_x^2$ is smaller than σ_x^2 by a factor of $1/N$. That is,

$$\bar{\sigma}_x = \frac{\sigma_x}{\sqrt{N}}. \tag{9}$$

Thus 4 measurements give an average that is twice as precise as a single measurement, 9 give an improvement of 3, 100 give an improvement of 10, and so on. This is an extremely useful principle to remember, and it is worth thinking about its implications for a while.

COMMENT:

> The above analysis of statistical uncertainties explains how to find the best estimate (the *mean*) from a number N of independent measurements with unknown but similar individual uncertainties. Sometimes we can estimate the uncertainty σ_{x_i} in each measurement x_i by some independent means like "common sense" (watch out for that one!). If this is the case, and if the measurements are not all equally precise (as, for instance, in combining all the world's best measurements of some esoteric parameter in elementary particle physics), then it is wrong to give each measurement equal *weight* in the average. There is then a better way to define the average, namely the "*weighted mean*":
>
> $$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$
>
> where $w_i \equiv 1/\sigma_{x_i}^2$. If the reader is interested in the proper way to estimate the *uncertainty* $\bar{\sigma}_x$ in the mean under these circumstances, it is time to consult a statistics text; the answer is not difficult, but it needs some explanation that is beyond the scope of this *HyperReference*.

Chapter 6

Falling Bodies

Now that we have mastered all sorts of Algebra and Calculus skills, it is time to get on with Newtonian Mechanics, Gravitation, Cosmology and all that, right?

Gee, I sure wish it were true.

Although (I hope you will agree) there are some interesting historical and perceptual lessons to be learned from Newton's Mechanics, it is generally rated as one of the more boring topics in Physics;[1] worse yet, we are not yet ready for Newton — "You have to creep before you can crawl," as it were. And in this business "creeping" is the business of *Kinematics* — the study of *motion per se*.

Besides, before we go on to expound Newton's "Laws" in their modern form we will need to have a chapter on *vectors*, since *forces* are classic examples of vectors — *i.e.* they have both *magnitude* and *direction*.[2]

[1]This is partly because everyone is so anxious to "get on to the good stuff" that they are predisposed to give a rather superficial treatment to Mechanics; and partly because most beginning Physics courses are expected to produce graduates who can actually calculate tensions in wires, whether boxes will slide off trucks and other practical things like that. Fortunately, I don't care whether you can do that stuff or not, except for a few simple examples for the sake of illustration and familiarization. This book *may* help you build a bridge in your back yard, but honestly I think there are much more useful study aids for developing such skills. What I am after is just to get you familiar enough with the *paradigms* of Mechanics to allow bootstrapping on to the next stage.

[2]This is also true of *distance*, *velocity* and *acceleration*, which are the topics of this Chapter; but we have to start *somewhere*.

6.1 Galileo

As I warned the reader in several places earlier, I am no historian. However, I do have many traits in common with real Historians; in particular, I like to construct theories of "what probably *really* happened" to fit my own interpretation of the historical "data." Physicists also like this sort of revisionism, but I think we are mercifully more shameless and direct about it. ["Yeah, OK, I lied; but it was a *good* lie — doesn't it make everything easier to understand?"] With this *caveat*, I will relate a bit of Brewer's History of Classical Mechanics.

Galileo Galilei (1564-1642) was a clever Italian megalomaniac who took pleasure in publicly ridiculing his intellectual opponents and regarded the authorities as annoying buffoons to be manipulated by any means available in order to obtain funding for his pet projects. He thus epitomized a fine tradition which continues to this day. Galileo is widely credited with being "the Father of Modern Science" because of the experimental æsthetics he championed[3] and because of the impact of his major work, *Two New Sciences* [mechanics and the strength of materials], published in 1636. I am inclined to think that his distinctive personality and style had just as much to do with his deserving this title; today these traits are still apt to improve the bearer's chances for distinction by various

[3]Often referred to as the "Scientific Method," about which I will have more to say later on.

prizes and accolades.

6.1.1 Harvard?

Rather than reproduce the list of Galileo's adventures available in any textbook with even a pretense of historical perspective, I will mention one amusing claim that I heard somewhere:[4] when Galileo got into trouble with the Church over his heretical views[5] he was offered a faculty position at a new University in another country where the Roman Church was not all that popular — the school in question was Harvard.[6]

6.1.2 Weapons Research: Telescopes & Trajectories

Ever the Modern Physicist, Galileo recognized clearly that the big money and prestige were in military applications of science. In those days the new weapons technology was *cannons* and how to aim them more accurately at targets. His contributions to this art took two main forms: the first was his invention of the *magnifying telescope*, with which it was possible to identify targets at great range and assess the damage done to them by one's cannonballs. To be fair, I should point out that this invention was warmly received by seafarers and astronomers as well as generals; in fact, with it Galileo himself made famous and wonderful observations of the Moon, the "Galilean" moons (named after guess whom) of Jupiter and numerous other objects in our Solar System, thereby initiating the modern pastime of Planetology that recently culminated in the fantastic close-up views of the outer planets and their satellites by Terran space probes. One can easily imagine how ridiculous the Church's Ptolemaic ergocentric model of the Heavens must have seemed to Galileo after watching so many other planets execute their orbits as clearly visible globes lit on the Sun side.[7] There are two sides to every coin.

Galileo's second contribution to the art of artilliery was his formal explication of the behaviour of *falling bodies*, of which cannon and musket balls were oft-mentioned examples. Galileo "showed"[8] that the velocity of a falling body increases by equal increments in equal times (in the absence of friction), which is the definition of a state of constant acceleration.

Constant Acceleration

In terms of our newly-acquired left hemisphere skills, if we use y to designate *height* [say, above sea level] and t to designate *time*, then the *upward velocity* v_y [where the subscript tells us explicitly that this is the *upward* velocity as opposed to the *horizontal* velocity which would probably be written v_x][9] is given by

$$v_y = v_{y_0} - gt \tag{1}$$

[4] You real Historians go check this out!

[5] Actually they would probably have left him alone if he hadn't been so obnoxious about publicly rubbing their noses in it.

[6] One imagines Galileo's response was, "I'm not *that* desparate." In those days Harvard had presumably not yet acquired much of a reputation. It is amusing to speculate on how much *more* classic an example of the Modern Physicist he would have made had Galileo accepted this offer of a New World professorship.

[7] The astronomical observations of Tycho Brahe and Johannes Kepler empirically obliterated the Ptolemaic system in favour of a correct heliocentric model of the Solar system at about the same time as Galileo took on the Church in Italy; I am not certain how much interaction there was between these apparently separate battles. More on this later.

[8] There is room for argument over whether he really "showed" this, both from a Popperian purist's point of view [you can never *verify* a conjecture, only *refute* it] and from the point of view of the very æsthetic he helped to popularize — namely, that you shouldn't "fudge" your results and that other people should be able to *reproduce* them. It is, however, certainly true that he *made a very persuasive case* for the economy and utility of this confessed overidealization; and this is, after all, the true measure of any theory!

[9] Why not just call it v, if I am not going to be talking about any of the horizontal stuff? Well, this is a pretty simple equation, so I am going to "stack" it with lessons in *notation* which will serve to make its meaning absolutely unambiguous (subject to all these explanations) and to introduce fine points I will be needing shortly anyway.

where v_{y_0} is the initial[10] upward velocity (*i.e.* the upward velocity at $t = 0$), if any,[11] and g is the downward[12] acceleration of gravity, $g \approx 9.81$ m/s^2 on average at the Earth's surface.[13] Another way of writing the same equation is in terms of the *derivative* of the velocity with respect to time,

$$a_y \equiv \frac{dv_y}{dt} \equiv \dot{v}_y = -g, \tag{2}$$

where I have introduced yet *another* notational convention used by Physicists: a little dot above a symbol means the *time derivative* of that symbol — *i.e.* the rate of change (per unit time) of the quantity represented by that symbol.[14] And since v_y is itself the time derivative of the height y [*i.e.* $v_y \equiv dy/dt \equiv \dot{y}$], if we like we can write the original equation as

$$\dot{y} = v_{y_0} - gt. \tag{3}$$

All these notational gymnastics have several purposes, one of which is to make you appreciate the simple clarity of the declaration, "The vertical speed increases by equal increments in equal times," as originally stated by Galileo himself. But I also want you to see how Physicists like to *condense* their notation until a very compact equation "says it all."

Principles of Inertia and Superposition

Galileo was actually the first to write down "Newton's" celebrated First Law, in a form slightly different from Newton's but just as good:[15]

Galileo's Principle of Inertia:

> *A body moving on a level surface will continue in the same direction at constant speed unless disturbed.*

Note the term "body" employed in order to be deliberately vague about what *sort* of entities the Principle is meant to apply to. This term is retained in the language of modern Mechanics. It means, more or less, "a massive thing that hangs together." Note also the other ringers, "level surface" and "unless disturbed." *Perfectly* level surfaces are mighty hard to come by, but Galileo means, of course, a *hypothetical* perfectly level surface. More serious is the vagueness of "unless disturbed." This can easily be used to make the argument circular: if

[10] Note: generally any symbol with a subscript $_0$ (read "nought" as in x_0 = "x nought") designates an *initial value* of the subscripted symbol — *i.e.* the value at $t = 0$. (We stop short of writing t_0 for the initial time, in most cases, because we usually don't need any further redundancy to make the the description completely general.) Thus x may be a variable, a function of time $x(t)$, but its initial value $x_0 \equiv x(0)$ is a constant, a *parameter* of its evolution in time. Since we will often talk about the *final* value of some variable (*e.g.* x_f) at time t_f (at the end of some process), using the subscript $_f$ to designate "final," it is equally logical to use a subscript $_i$ for "initial," so that the value of $x(t)$ at $t = 0$ would be written x_i — this notation is *perfectly synonymous* with the "nought" notation: $x_0 \equiv x_i$ and the two may be used interchangably according to taste.

[11] Lots of people leave out the v_{y_0} in order to keep it simpler, but of course that would be tantamount to assuming that we were starting *from rest*, which ain't necesarily so! Why *over*simplify an already simple equation?

[12] Note that the conventional choice of "up" as being the *positive* y direction forces us to put the acceleration of gravity into the equation with a minus sign, since it is in the "down" direction. Sometimes people try to make this look simpler for beginners by defining *down* as the $+y$ direction, but I like to get across as early as possible that a negative acceleration simply means an acceleration in the direction opposite to the one we arbitrarily defined to be positive. The same is true of any quantity (*e.g.* the velocity or the position) that has a direction as well as a magnitude; this idea is vital to an understanding of *vectors*, which are coming up soon!

[13] What?! How come I don't give g to a huge number of significant figures, with an uncertainty specified, as one is supposed to do for fundamental constants? Because g is neither fundamental nor constant! Far from it. More on this later.

[14] I will soon need the analogous notation $\ddot{x} \equiv d^2x/dt^2$ to signify the *second* time derivative of x, so that $a_y \equiv dv_y/dt \equiv d^2y/dt^2 \equiv \ddot{y}$. The "double-dot" form is the preferred Physics notation for acceleration, mainly for reasons of economy (it takes so few strokes to write).

[15] This is translated from the Italian by someone else; I can't vouch for the translation but I am confident that it gets the right idea across and I am not much interested in quibbles over the exact wording or what it might have meant about Galileo's "authentic originality."

the body's velocity changes direction or magnitude, it is because it is "disturbed." Well.... Newton invented a new concept to make "disturbance" a little more specific.

The other important insight Galileo saw fit to enshrine as a Principle was

Galileo's Principle of Superposition:

> *If a body is subjected to two separate influences, each producing a characteristic type of motion, it responds to each without modifying its response to the other.*

This, like the other Principle, seems transparently obvious to Modern eyes,[16] but without it one would never know how to start applying Galileo's simplified kinematics to the practical problem of trajectories. Again there is a little sloppiness to the Principle that allows for counterexamples; no doubt Galileo had to rely regularly on the most honest of all appeals: "*You* know what I mean."

Calculating Trajectories

Applied to the case of *trajectories* close to the Earth's surface,[17] the equations governing constant horizontal velocity *superimposed* upon constant downward acceleration take the form

$$\ddot{x} = 0 \quad (4)$$

$$\dot{x} = v_{x_0} \quad \text{(constant)} \quad (5)$$

$$x = x_0 + v_{x_0} t \quad (6)$$

and (7)

$$\ddot{y} = -g \quad (8)$$

$$\dot{y} = v_{y_0} - g t \quad (9)$$

$$y = y_0 + v_{y_0} t - \frac{1}{2} g t^2 \quad (10)$$

where

$$\ddot{x} \equiv \frac{d^2x}{dt^2} \equiv \frac{dv_x}{dt} \equiv \dot{v_x} \equiv a_x, \quad (11)$$

$$\dot{x} \equiv \frac{dx}{dt} \equiv v_x, \quad (12)$$

$$\ddot{y} \equiv \frac{d^2y}{dt^2} \equiv \frac{dv_y}{dt} \equiv \dot{v_y} \equiv a_y \quad (13)$$

and $$\dot{y} \equiv \frac{dy}{dt} \equiv v_y \quad (14)$$

Hold it! Before you bolt for the door, take a moment to casually read through all these horrible-looking equations. I have made them look long and hirsute *on purpose*, for two reasons: first, because this way they are in their *most general form* — *i.e.* we can be confident that these equations will correctly describe *any* trajectory problem, but for any actual problem the equations will usually *simplify*; and second, because this is a sort of practical joke — if you look carefully you will see that the equations are really pretty simple! All those "$\equiv$" symbols just mean, "...another way of putting it, which amounts to exactly the same thing, is...." That is, they just indicate *equivalent notations* — or, in the language of linguistics, *synonyms*. So the latter batch of equations is just reminding you of the convention Physicists use for writing time derivatives: "dot" and "double-dot" notation. The first batch of equations tells you (in this notation) *everything there is to know* about the motion: the horizontal [x] motion is not under any acceleration [$a_x \equiv \ddot{x} = 0$] so the horizontal velocity [$v_x \equiv \dot{x}$] is constant [$\dot{x} = v_{x_0}$] and the distance travelled horizontally [$x(t)$] is just increasing linearly with time t relative to its initial value x_0 — *i.e.* $x = x_0 + v_{x_0} t$. The vertical motion differs only in that it includes a constant downward acceleration [$a_y \equiv \ddot{y} = -g$] which adds a term [$-gt$] to $\dot{y}$ and another familiar term [$-\frac{1}{2}gt^2$] to $y(t)$. Note that in every case the whole idea is to get the quantity on the left-hand side [lhs] of the equation equal to an *explicit function of* t on the right-hand side [rhs].

[16] This may well be a good measure of the brilliance of an insight.

[17] *E.g.*, cannonballs! This sort of "techno doubletalk" is not *always* used for obfuscation [I, for instance, am simply trying to be general!] but Pentagon aides trying to be Generals are very fond of it too.

Let's do a problem to illustrate how these equations work: Suppose we fire a cannon horizontally from the top of a 19.62 m high bluff, imparting an initial velocity $v_{x_0} = 10$ m/s to the cannonball. [By the definition of "horizontal," $v_{y_0} = 0$.] Where does the ball hit? [We neglect air friction and assume level (horizontal) ground at the bottom of the bluff.] For simplicity we can take $x = 0$ at

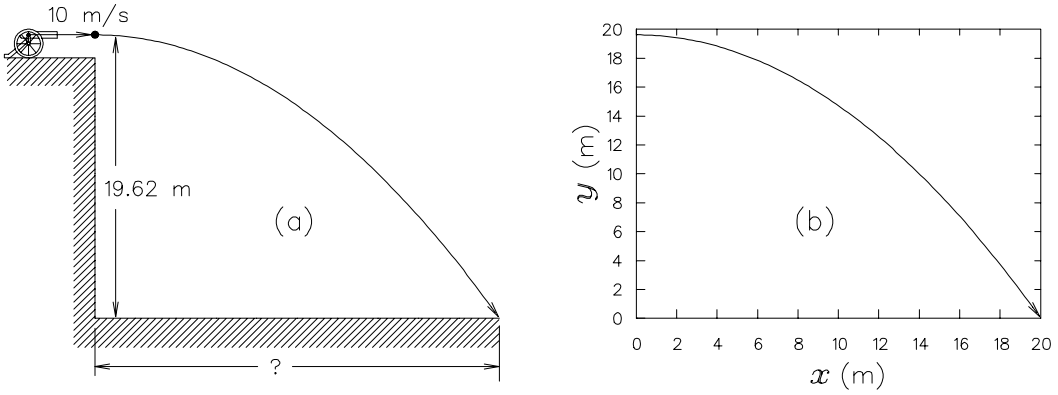


Figure 6.1 (a) Sketch of a trajectory problem in which the initial height [$y_0 = 19.62$ m] and the initial (horizontal) velocity [$v_{x_0} = 10$ m/s] are given and we want to calculate the horizontal distance [x_f] at which the cannonball hits the ground [$y_f = 0$]. (b) Corresponding plot of $y(x)$, the *trajectory* followed by the cannonball.

the muzzle of the cannon;[18] similarly, we (naturally enough) take $t = 0$ to be the instant at which the ball leaves the muzzle of the cannon. Our *general* equations now "reduce" to a more *particular* set of equations for this specific example:

$$x = v_{x_0}\, t \qquad \text{and} \qquad y = y_0 \; - \; \frac{1}{2}\, g\, t^2$$

or, since $v_{x_0} = 10$ m/s and $y_0 = 19.62$ m,

$$x = (10\text{m/s})\, t \qquad \text{and} \qquad y = (19.62\text{m}) \; - \; \frac{1}{2}\, (9.81\text{m/s}^2)\, t^2$$

We now have a choice between working out the algebra in the first pair of equations or working out the arithmetic in the second pair. The former is preferable partly because we don't have to "juggle units" while we work out the equations (a clumsy process which is usually neglected, leading to equations with numbers but no units, which in turn can lead to considerable confusion) and because solving for x_f in terms of the two "parameters" y_0 and v_{x_0} [g is also a parameter, although we usually treat it as if it were a *constant* of Nature] gives an "answer" to *any such problem* with *qualitatively* similar conditions. Here's the algebra:

$$x = v_{x_0}\, t \qquad \Longrightarrow \qquad t = \frac{x}{v_{x_0}}$$

which can be substituted for t in the second equation, giving

$$y = y_0 \; - \; \frac{1}{2}\, g \, \left[\frac{x}{v_{x_0}} \right]^2 .$$

[18] (This is typical — we always make as many simplifications as the arbitrariness of the notation allows!)

We are interested in the value of x_f at the end of the trajectory — *i.e.* when $y_f = 0$:

$$y_f = 0 = y_0 \; - \; \frac{1}{2}\, g \left[\frac{x_f}{v_{x_0}}\right]^2 \qquad \Longrightarrow \qquad y_0 = \frac{1}{2}\, g\, \frac{x_f^2}{v_{x_0}^2}$$

$$\Longrightarrow \qquad \frac{2y_0}{g} = \frac{x_f^2}{v_{x_0}^2} \qquad \Longrightarrow \qquad \frac{2y_0\, v_{x_0}^2}{g} = x_f^2 \qquad \Longrightarrow \qquad x_f = \sqrt{\frac{2y_0\, v_{x_0}^2}{g}}.$$

Now we "plug in" $y_0 = 19.62$ m, $v_{x_0} = 10$ m/s and $g = 9.81$ m/s^2, giving

$$x_f = \sqrt{\frac{2 \times 19.62\text{m} \times [10\text{m/s}]^2}{9.81\text{m/s}^2}} = \sqrt{\frac{2 \times 2 \times 9.81 \times 100\text{m}^3/\text{s}^2}{9.81\text{m/s}^2}} = \sqrt{400\text{m}^2} = 20\text{m}.$$

And that's the answer: $x_f = 20$ m. Simple, huh?

6.2 The Scientific Method

One often hears that "the modern Scientific Method" can be traced back to Galileo, who first prescribed the panacea of "Observe, Hypothesize, Experiment and Confirm." This is complete nonsense.[19]

First of all, people have been doing more or less the same thing since before the Dawn of Recorded History;[20] Galileo just grabbed the headlines when there was first Good Press to get! He was a hero, true, in that he championed the arrogance of *thinking for oneself* against formidable odds and outlined a procedure for doing it successfully (*i.e.* getting away with it) for which we all are in his debt. But he could hardly claim a patent on the idea.

Second, Galileo's Scientific Method, like his Mechanics, was an *idealization* of an imperfect experimental reality. As discussed earlier, we cannot Observe without relying upon our *repertoire* of *models* through which we interpret our sense data; the phrase, "Seeing is believing," betrays a profound *naiveté* if we consider carefully what we know about the retina, the optic nerve and the visual cortex. We may Hypothesize freely, but only the most righteous scientists are actually honest about *when* their hypotheses were formed — before or after the experiment![21] The one part of Galileo's prescription that we truly took to heart was the exhortation to *Experiment* — *i.e.* to go directly to Nature with our questions about "what will happen if we. . . ?" Asking such questions in a form that Nature will deign to answer unambiguously is a profound art indeed; a lifetime is too short to learn it in. Finally, Galileo can be considered charmingly naive in his expectation that Experimentation will be able to Confirm any Hypothesis. As Karl Popper has pointed out, there is no logical basis upon which any "general explanatory theory" can be proven correct by any finite number of experiments; the best we can hope for is a Conjecture which is "not yet Refuted" by the evidence, and this is impressive only if there is a *lot* of non-contradictory evidence!

[19] By now, you no longer need to be reminded that such comments are "in my humble opinion."

[20] Why not the *Sunset* of Recorded History, I sometimes wonder?

[21] Newton, whom we often picture as the gardener who brought Galileo's seeds to flower, is also famous for his arrogant statement [a blatant lie], "*Hypotheses non fingo*," or "I do not make conjectures." (What a jerk!)

So the revised version of the "Scientific Method" should read something like this:

1. Based on a lifetime of experience, form a Hunch.

2. Using a trained analytical mind, refine the Hunch into a well-posed Hypothesis.[22]

3. Think of a few Consequences of the Hypothesis that lead to Predictions that can be tested by Experiment.[23]

4. Perform a *Gedankenexperiment*[24] to visualize the results you should expect to get under different circumstances.

5. Design a real Experiment, if possible, to produce the most clear and unambiguous results[25] possible.

6. Descend to the level of grubby sociopoliticoeconomic reality to seek funding, recruit personnel, fight battles for priority, coordinate with engineers, construct several versions of the apparatus (all but the last of which do not work), tinker with balky equipment, coax plausible results out of partially recorded data, argue with collaborators about procedure and interpretation, *etc.*, for as long as it takes to get the Experiment done [which may exceed your lifespan in certain disciplines].

7. Publish a Result (or Results) — often determined by "consensus" [*i.e.* politics] among Collaborators — and let the Community decide what it means.

8. Go back to Step 1, if you did not already do so earlier.

Of course, these are the rules for a Professional Scientist; if you are content to remain an Amateur, the Scientific Method is a little simpler:

Think for yourself.

In all the above arguments, there is an implicit assumption that we usually do not discuss: namely, that there is an "external" Real World independent of our perceptions and models that behaves the way it does regardless of our expectations or observations — that we can, at least in spirit, set ourselves apart from The World as mere *observers* of its behaviour. Even in Classical Mechanics this is an obvious idealization, but perhaps a conscionable one. In Quantum Mechanics (as we shall see) this basic view of the Experimenter as Observer is challenged at its roots! Nevertheless there are things we can do which *seem like* Observations and which we will have to use to "pull ourselves up by the bootstraps" if we are to even grasp what Quantum Mechanics has to tell us. So, for the time being, I encourage you to steep yourself in the traditional æsthetic of Experimental Science and try to be as "objective" and "non-interfering" as possible in making (or imagining) your Experimental Observations.

6.3 The Perturbation Paradigm

Galileo "demonstrated" the phenomenon of constant acceleration using a water clock and a ball rolling down an inclined groove. In my experience, even with modern equipment it is dif-

[22] This is not as easy as it sounds. Most Hunches do not survive close examination; they usually contain irreducible internal inconsistencies or self-contradictions that may, at best, lead the Scientist back to a completely new Hunch.

[23] This is also harder than it sounds. Many Hypotheses have no testable Consequences at all; most of the rest could be tested in principle but might require manipulation of galaxies or reenactments of the Big Bang to produce unambiguous experimental results.

[24] *I.e.*, a "thought experiment." This term was invented by Albert Einstein, I believe, but the *technique* is as old as Humanity — this was the approved methodology of Aristotelian science, and is still a great boon to research funding agencies!

[25] *I.e.*, those most commensurate with conventional models and paradigms, either *pro* or *con* the Predictions of the Hypothesis.

ficult to obtain decent data on this sort of phenomenon; and even these data are typically *not* consistent with a true state of constant acceleration! There is no doubt that Galileo was quite aware of these flaws in his description; he was also quite happy to consign them to the realm of the "non-ideal" — *i.e.* the deviations from his predictions were due to *imperfections* in the ramp and the *disturbance* of the motion by the presence of *air*. Galileo argued that the results of a falling-body experiment performed *underwater* would be a lot *worse* than those of his experiments in *air*, so that one merely needed to extrapolate to *no medium at all* (*i.e.* perfect vacuum) to obtain results in perfect agreement with his predictions!

This overtly Platonic idealism was not new; but Galileo had hit upon a "good" approximation — one which actually *did* work better and better as the circumstances got closer and closer to a well-defined ideal case. The corrections could be regarded as negligible *perturbations* upon an "essentially correct" idealization, to be beaten into submission either by improvement of the apparatus or by laborious calculations.

Thus began what I call the "Perturbation Paradigm" of Physics. This simple prescription — find a nice simple model that does "pretty well" and then "fix up" its inadequacies with a series of corrections or "perturbations" — is so powerful that we Physicists use it on almost everything. The recent history of elementary particle physics gives a particularly poignant example of how a problem that was seemingly intractable by this perturbative method (and which promised for a while to lead us into genuinely new ways of thinking, which might have been nice for a change) was finally recast into a form that allowed application of the Perturbation Paradigm after all. I will suppress the urge to tell you about it now. But just wait!

Chapter 7

The Exponential Function

Suppose the newspaper headlines read, "The cost of living went up 10% this year." Can we translate this information into an *equation*? Let "V" denote the value of a dollar, in terms of the "real goods" it can buy — whatever economists mean by that. Let the elapsed time t be measured in years (y). Then suppose that V is a function of t, $V(t)$, which function we would like to know explicitly. Call *now* "$t = 0$" and let the initial value of the dollar (now) be V_0, which we could take to be \$1.00 if we disregard inflation at earlier times.[1]

Then our news item can be written

$$V(0) = V_0 \quad \& \quad V(1\,\mathrm{y}) = (1 - 0.1)\,V_0 = 0.9\,V_0.$$

This formula can be rewritten in terms of the *changes* in the dependent and independent variables, $\Delta V = V(1\,\mathrm{y}) - V(0)$ and $\Delta t = 1\,\mathrm{y}$:

$$\frac{\Delta V}{\Delta t} = -0.1\,V_0, \tag{1}$$

where it is now to be *understood* that V is measured in "1998 dollars" and t is measured in years. That is, the average *time rate of change* of V is proportional to the value of V at the beginning of the time interval, and the constant of proportionality is $-0.1\ \mathrm{y}^{-1}$. (By y^{-1} or "inverse years" we mean the *per year* rate of change.)

This is almost like a derivative. If only Δt were infinitesimally small, it would *be* a derivative. Since we're just trying to describe the qualitative behaviour, let's make an *approximation*: assume that $\Delta t = 1$ year is "close enough" to an infinitesimal time interval, and that the above formula (1) for the inflation rate can be turned into an *instantaneous* rate of change:[2]

$$\frac{dV}{dt} = -0.1\,V. \tag{2}$$

This means that the dollar in your pocket right now will be worth only \$0.99999996829 in one second.

Well, this is interesting, but we cannot go any further with it until we ask a crucial question: "What will happen if this goes on?" That is, suppose we assume that equation (2) is not just a temporary situation, but *represents a consistent and ubiquitous property* of the function $V(t)$, the "real value" of your dollar bill as a function of time.[3]

Applying the d/dt "operator" to both sides of Eq. (2) gives

$$\frac{d}{dt}\left(\frac{dV}{dt}\right) = \frac{d}{dt}(-0.1\,V) \quad \text{or} \quad \frac{d^2V}{dt^2} = -0.1\,\frac{dV}{dt}. \tag{3}$$

But dV/dt is given by (2). If we substitute that formula into the above equation (3), we get

$$\frac{d^2V}{dt^2} = (-0.1)^2\,V = 0.01\,V. \tag{4}$$

[1] Since our dollar will be worth *less* a year from now, we should really call it **de**flation!

[2] The error introduced by this approximation is not very serious.

[3] Banks, insurance companies, trade unions, and governments all pretend that they don't assume this, but they would all go bankrupt if they *didn't* assume it.

That is, the rate of change of the rate of change is always positive, or the (negative) rate of change is getting *less* negative all the time.[4] In general, whenever we have a *positive second derivative* of a function (as is the case here), the *curve is concave upwards*. Similarly, if the second derivative were *negative*, the curve would be concave *downwards*.

So by noting the initial value of V, which is formally written V_0 but in this case equals \$1.00, and by applying our understanding of the "graphical meaning" of the first derivative (slope) and the second derivative (curvature), we can visualize the function $V(t)$ pretty well. It starts out with a maximum downward slope and then starts to level off as time increases. This general trend continues indefinitely. Note that while the function always decreases, it *never reaches zero*. This is because, the closer it gets to zero, the slower it decreases [see Eq. (2)]. This is a very "cute" feature that makes this function especially fun to imagine over long times.

We can also apply our analytical understanding to the formulas (2) and (4) for the derivatives: every time we take still another derivative, the result is still proportional to V — the constant of proportionality just picks up another factor of (-0.1). This is a *really neat* feature of this function, namely that we can write down *all its derivatives* with almost no effort:

$$\frac{dV}{dt} = -0.1\,V \tag{5}$$

$$\frac{d^2V}{dt^2} = (-0.1)^2\,V = +0.01\,V \tag{6}$$

$$\frac{d^3V}{dt^3} = (-0.1)^3\,V = -0.001\,V \tag{7}$$

$$\frac{d^4V}{dt^4} = (-0.1)^4\,V = +0.0001\,V \tag{8}$$

$$\vdots$$

$$\frac{d^nV}{dt^n} = (-0.1)^n\,V \qquad \text{for } \textit{any } n. \tag{9}$$

[4] A politician trying to obfuscate the issue might say, "The rate of decrease is decreasing."

This is a pretty nifty function. What *is* it? That is, can we write it down in terms of familiar things like t, t^2, t^3, and so on?

First, note that Eq. (9) can be written in the form

$$\frac{d^nV}{dt^n} = k^n\,V, \qquad \text{where} \qquad k = -0.1 \tag{10}$$

A simpler version would be where $k = 1$, giving

$$\frac{d^nW}{dt^n} = W, \tag{11}$$

$W(t)$ being the function satisfying this criterion. We should perhaps try figuring out this simpler problem first, and then come back to $V(t)$.

Let's try expressing $W(t)$, then, as a linear combination[5] of such terms. For starters we will try a "third order polynomial" (*i.e.* we allow terms up to t^3):

$$W(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3.$$

Then

$$\frac{dW}{dt} = a_1 + 2a_2 t + 3a_3 t^2$$

follows by simple "differentiation" [a single word for "taking the derivative"]. Now, these two equations have similar-looking right-hand sides, provided that we pretend not to notice that term in t^3 in the first one, and provided the constants a_n obey the rule $a_{n-1} = n a_n$ [*i.e.* $a_0 = a_1$, $a_1 = 2a_2$ and $a_2 = 3a_3$]. But we can't really neglect that t^3 term! To be sure, its "coefficient" a_3 is smaller than any of the rest, so if we had to neglect anything it might be the best choice; but we're trying to be precise, right? How precise? Well, precise enough. In that case, would we be precise enough if we added a term $a_4 t^4$, preserving the rule about coefficients [$a_3 = 4a_4$]? No? Then how about $a_5 t^5$? And so on. No matter how precise an

[5] "Linear combination" means we multiply each term by a simple constant and add them up.

agreement with Eq. (11) we demand, we can always take enough terms, using this procedure, to achieve the desired precision. Even if you demand infinite precision, we just [just?] take an infinite number of terms:

$$W(t) = \sum_{n=0}^{\infty} a_n\, t^n, \quad \text{where} \quad a_{n-1} \;=\; n\, a_n \tag{12}$$

$$\text{or} \quad a_n \;=\; \frac{a_{n-1}}{n}. \tag{13}$$

Now, suppose we give $W(t)$ the initial value 1. [If we want a different initial value we can just multiply the whole series by that value, without affecting Eq. (11).] Well, $W(0) = 1$ tells us that $a_0 = 1$. In that case, $a_1 = 1$ also, and $a_2 = \frac{1}{2}$, and $a_3 = \frac{1}{2} \times \frac{1}{3}$, and $a_4 = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{4}$, and so on. If we define the *factorial* notation,

$$n! \;\equiv\; n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1 \tag{14}$$

(read, "*n factorial*") and define $0! \equiv 1$, we can express our function $W(t)$ very simply:

$$W(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \tag{15}$$

We could also write a more abstract version of this function in terms of a generalized variable "x":

$$W(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{16}$$

Let's do this, and then define $x \equiv k\, t$ and set $V(t) = V_0\, W(x)$. Then, by the Chain Rule for derivatives,[6]

$$\frac{dV}{dt} \;=\; V_0\, \frac{dW}{dx}\, \frac{dx}{dt} \tag{17}$$

and since $\frac{d}{dt}(k\, t) = k$, we have

$$\frac{dV}{dt} \;=\; k\, V_0\, W \;=\; k\, V. \tag{18}$$

[6]The Chain Rule for derivatives says that if z is an explicit function of y, $z(y)$, and y is an explicit function of x, $y(x)$, then z is an *implicit* function of x and its derivative with respect to x is given by

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}.$$

By repeating this we obtain Eq. (10). Thus

$$V(t) \;=\; V_0\, W(kt) \;=\; V_0 \sum_{n=0}^{\infty} \frac{(kt)^n}{n!} \tag{19}$$

where $k = -0.1$ in the present case.

This is a nice description; we can always calculate the value of this function to any desired degree of accuracy by including as many terms as we need until the change produced by adding the next term is too small to worry us.[7] But it is a little clumsy to keep writing down such an unwieldy formula every time you want to refer to this function, especially if it is going to be as popular as we claim. After all, mathematics is the art of precise abbreviation. So we give $W(x)$ [from Eq. (16)] a special name, the "**exponential**" function, which we write as either[8]

$$\exp(x) \qquad \text{or} \qquad e^x. \tag{20}$$

In FORTRAN it is represented as `EXP(X)`. It is equal to the number

$$e \;=\; 2.71828182845904509\cdots \tag{21}$$

raised to the x^{th} power. In our case we have $x \equiv -0.1\, t$, so that our "answer" is

$$V(t) \;=\; V_0\, e^{-0.1\, t} \tag{22}$$

which is a lot easier to write down than Eq. (19).

Now, the choice of notation e^x is not arbitrary. There are a lot of rules we know how to use on a number raised to a power. One is that

$$e^{-x} \;\equiv\; \frac{1}{e^x} \tag{23}$$

You can easily determine that this rule also works for the definition in Eq. (16).

The "inverse" of this function (the power to which one must raise e to obtain a specified

[7]This is exactly what a "scientific" hand calculator does when you push the function key whose name will be revealed momentarily.

[8]Now you know which key it is on a calculator.

number) is called the "**natural logarithm**" or "ln" function. We write

$$\text{if} \qquad W = e^x, \qquad \text{then} \qquad x = \ln(W)$$

or

$$x \;=\; \ln(e^x) \tag{24}$$

A handy application of this definition is the rule

$$y^x = e^{x\ln(y)} \quad \text{or} \quad y^x = \exp[x\ln(y)]. \tag{25}$$

Before we return to our original function, is there anything more interesting about the "natural logarithm" than that it is the inverse of the "exponential" function? And what is so all-fired special about e, the "base" of the natural log? Well, it can easily be shown[9] that the *derivative* of $\ln(x)$ is a very simple and familiar function:

$$\frac{d[\ln(x)]}{dx} \;=\; \frac{1}{x}. \tag{26}$$

This is perhaps the most useful feature of $\ln(x)$, because it gives us a direct connection between the exponential function and a function whose derivative is $1/x$. [The handy and versatile rule $\frac{d(x^r)}{dx} = rx^{r-1}$ is valid for any value of r, including $r = 0$, but it doesn't help us with this task. Why?] It also explains what is so special about the number e.

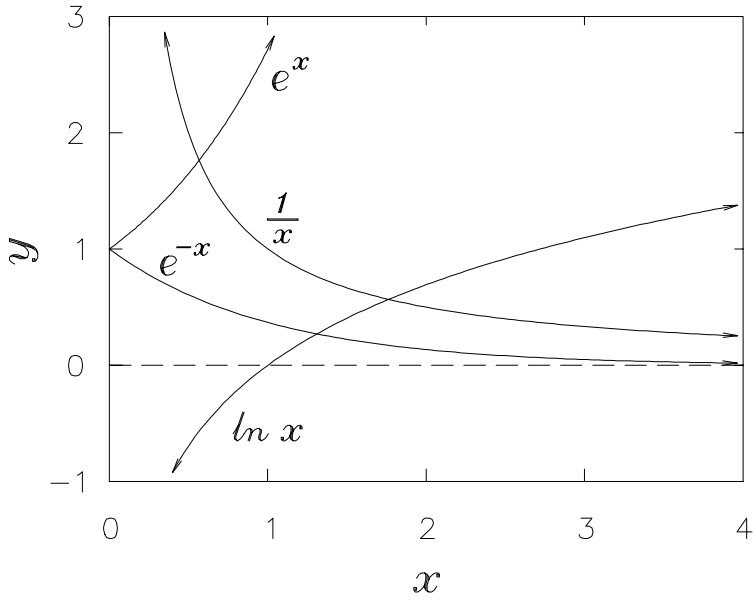


Figure 7.1 The functions e^x, e^{-x}, $\ln(x)$ and $1/x$ plotted on the same graph over the range from $x = 0$ to $x = 4$. Note that $\ln(0)$ is undefined. [There is no finite power to which we can raise e and get zero.] Similarly, $1/x$ is undefined at $x = 0$, while $1/(-x) = -1/x$. Also, $\ln(1) = 0$ [because any number raised to the zeroth power equals 1 — you can easily check this against the definitions] and $\ln(\xi)$ [where ξ any positive number less than 1] is negative. However, there is no such thing as the natural logarithm of any *negative* number.

Summary: Exponential Functions

Our formula (22) for the real value of your dollar as a function of time is the *only* function which will satisfy the differential equation (2) from which we started. The *exponential* function is one of the most useful of all for solving a wide variety of differential equations. For now, just remember this:

> Whenever you have $\frac{dy}{dx} = k\,y$, you can be sure that $y(x) = y_0\,e^{kx}$ where y_0 is the "initial value" of y [when $x = 0$]. Note that k can be either positive or negative.

Finally, note the property of the *second* derivative:

$$\frac{d^2y}{dx^2} \;=\; k^2\,y. \tag{27}$$

We will see another equation almost like this when we talk about Simple Harmonic Motion.

[9] Watch for this phrase! Whenever someone says "It can easily be shown. . . ," they mean, "This is possible to prove, but I haven't got time; besides, I might want to assign it as homework."

Mechanics Example: Damping

We should really work out at least one example applying the exponential function to a real Mechanics problem. The classic example is where an object (mass m) is moving with an initial velocity v_0, starting from an initial position x_0, and experiences a *frictional damping force* F_d which is *proportional to the velocity* and (as always, for frictional forces) in the direction opposite to the velocity: $F_d = -\kappa\, v$. The equation of motion then reads $a = -(\kappa/m)\, v$ or

$$\frac{d^2x}{dt^2} \;=\; -k\,\frac{dx}{dt} \qquad (28)$$

where we have combined κ and m into the constant $k \equiv \kappa/m$. This can also be written in the form

$$\frac{dv}{dt} \;=\; -k\,v$$

which should ring a bell! The solution (for the velocity v) is

$$v(t) \;=\; v_0\, e^{-k\,t} \qquad (29)$$

To obtain the solution for $x(t)$, we switch back to the notation

$$\frac{dx}{dt} = v_0\, e^{-k\,t} \quad \Longrightarrow \quad \int_{x_0}^{x} dx = v_0 \int_0^t e^{-k\,t}\, dt$$

and note that the function whose time derivative is $e^{-k\,t}$ is $-\frac{1}{k}\, e^{-k\,t}$, giving

$$x - x_0 \;=\; -\frac{v_0}{k}\left[e^{-k\,t}\right]_0^t$$

where the $[\cdots]_0^t$ notation means that the expression in the square brackets is to be "evaluated between 0 and t" — *i.e.* plug in the upper limit (just t itself) for t in the expression and then *subtract* the value of the expression with the lower limit (0) substituted for t. In this case the lower limit gives $e^{-0} = e^0 = 1$ (anything to the zeroth power gives one) so the result is

$$x(t) \;=\; x_0 \;+\; \frac{v_0}{k}\left(1 - e^{-k\,t}\right) \qquad (30)$$

The qualitative behaviour is plotted in Fig. 7.2. Note that $x(t)$ approaches a fixed "asymptotic" value $x_{\max} = x_0 + v_0/k$ as $t \to \infty$. The generic function $(1 - e^{-k\,t})$ is another useful addition to your pattern-recognition repertoire.

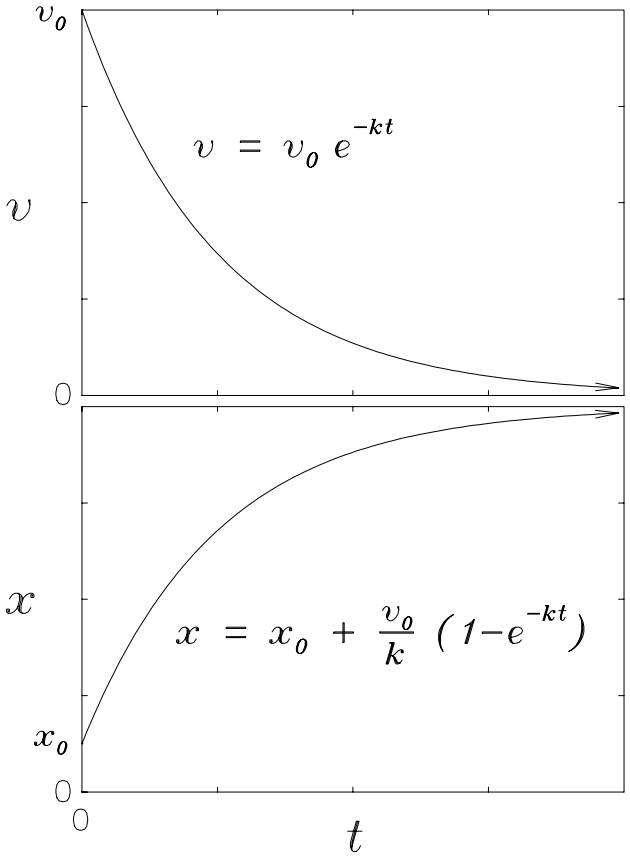


Figure 7.2 Solution to the *damping force* equation of motion, in which the frictional force is proportional to the velocity.

Chapter 8

Vectors

A Generalization of "Orthogonality"

The definition of a *vector* as an entity with both magnitude and direction can be generalized if we realize that "direction" can be defined in more dimensions than the usual 3 spatial directions, "up-down, left-right, and back-forth," or even other dimensions *excluding* these three. The more general definition would read,

> *Definition:* a **vector** quantity is one which has *several independent attributes* which are *all measured in the same units* so that "transformations" are possible. (This last feature is only essential when we want the advantages of mathematical manipulation; it is not necessary for the concept of multi-dimensional entities.)

We can best illustrate this generalization with an example of a vector that has nothing to do with 3-D space:

> Example: the **Cost of Living**, $\vec{\mathbf{C}}$, is in a sense a true vector quantity (although the Cost of Living *index* may be properly thought of as a *scalar*, as we can show later).

To construct a simple version, the Cost of Living can be taken to include:

- C_1 = **housing** (e.g., monthly rent);
- C_2 = **food** (e.g., cost of a quart of milk);
- C_3 = **medical** service (e.g., cost of a bottle of aspirin);
- C_4 = **entertainment** (e.g., cost of a movie ticket);
- C_5 = **transportation** (e.g., bus fare);
- $C_6 \ldots C_7 \ldots$ etc. (a finite number of "*components*.")

Thus we can write $\vec{\mathbf{C}}$ as an ordered sequence of numbers representing the values of its respective "components":

$$\vec{\mathbf{C}} = (C_1, C_2, C_3, C_4, C_5, \ldots) \qquad (1)$$

We would normally go on until we had a reasonably "complete" list – i.e., one with which the cost of any additional item we might imagine could be expressed *in terms of* the ones we have already defined. The technical mathematical term for this condition is that we have a "*complete basis set*" of components of the Cost of Living.

Now, we can immediately see an "inefficiency" in the way $\vec{\mathbf{C}}$ can been "composed:" As recently as 1975, it was estimated to take approximately one pound of gasoline to grow one pound of food in the U.S.A.; therefore the cost of **food** and the cost of **transportation** are obviously *not independent*! Both are closely tied to the

cost of **oil**. In fact, a large number of the components of the cost of living we observe are intimately connected to the cost of oil (among other things). On the other hand (before we jump to the fashionable conclusion that these two components should be replaced by oil prices alone), there is *some* measure of independence in the two components. How do we deal with this quantitatively?

To reiterate the question more formally, how do we quantitatively describe the extent to which certain components of a vector are superfluous (in the sense that they merely represent combinations of the other components) *vs.* the extent to which they are truly "independent?" To answer, it is convenient to revert to our old standby, the (graphable) analogy of the **distance** vector in *two dimensions*.

Suppose we wanted to describe the position of any point P in the "$x-y$ plane." We could draw the two axes "a" and "b" shown above. The position of an arbitrary point P is uniquely determined by its (a, b) coordinates, defined by the prescription that to change a we move parallel to the a-axis and to change b we move parallel to the b-axis. This is a unique and quite legitimate way of specifying the position of any point (in fact it is often used in crystallography where the orientation of certain crystal axes is determined by nature); yet there is something vaguely troubling about this choice of coordinate axes. What is it? Well, we have an intuitive sense of "up-down" and "sideways" as being *perpendicular*, so that if something moves "up" (as we normally think of it), in the above description the values of both a and b will change. But isn't our intuition just the result of a well-entrenched convention? If we got used to thinking of "up" as being in the "b" direction shown, wouldn't this cognitive dissonance dissolve?

No. In the first place, nature provides us with an unambiguous characterization of "down:" it is the direction in which things fall when released; the direction a string points when tied to a plumb bob. "Sideways," similarly, is the direction defined by the surface of an undisturbed liquid (as long as we neglect the curvature of the Earth's surface). That is, gravity fixes our notions of "appropriate" geometry. But is this in turn arbitrary (on nature's part) or is there some good reason why "independent" components of a vector should be perpendicular? And what exactly do we *mean* by "perpendicular," anyway? Can we define the concept in a way which might allow us to generalize it to other kinds of vectors besides space vectors?

The answer is bound up in the way Euclid found to express the geometrical properties of the world we live in; in particular, the "metric" of space – the way we define the **magnitude** (*length*) of a vector. Suppose you take a ruler and turn it at many angles; your idea of the *length* of the ruler is *independent* of its *orientation*, right? Suppose you used the ruler to make off distances along two perpendicular axes, stating that these were the horizontal and vertical components (x, y) of a distance vector. Then you use the usual "parallelogram rule" to locate the tip of the vector, draw in a line from the origin to that point, and put an arrowhead on the line to indicate that you have a vector. Call it "$\vec{\mathbf{r}}$". *You can use the same ruler, held at an angle, to measure the length r of the vector.* Pythagoras gave us a formula for this length. It is

$$r = \sqrt{x^2 + y^2}. \tag{2}$$

This formula is the key to Euclidean geometry, and is the working definition of perpendicular axes: x and y are perpendicular if and only if Eq. (2) holds. It does *not* hold for "a" and "b" described earlier!

You may feel that this "metric" is obvious and necessary from first principles; it is not. If you treat this formula as correct using the Earth's

surface as the "$x - y$ plane" you will get good results until you start measuring off distances in the thousands of miles; then you will be 'way off! Imagine for instance the perpendicular lines formed by two longitudes at the North Pole: these same "perpendicular" lines *cross again* at the South Pole!

Well, of course, you say; that is because the Earth's surface is *not* a plane; it is a sphere; it is *curved*. If we didn't feign ignorance of that fact, if we did our calculations in *three* dimensions, we would always get the right answers. Unfortunately not. The space we live in is actually *four*-dimensional, and it is *not* flat, *not* "Euclidean," in the neighborhood of large masses. Einstein helped open our eyes to this fact, and now we are stuck with a much more cognitively complex understanding.

But we have to start somewhere, and the space we live in from day to day in "pretty Euclidean," and it is only in the violation of sensible approximations that modern physics is astounding, so we will pretend that only Euclidean vector spaces are important. (Do you suppose there is a way to generalize our definition of "perpendicularity" to include non-Euclidean space as well?)

Finally returning to our original example, we would like to have $\vec{\mathbf{C}}$ expressed in an "orthogonal, complete basis", $\vec{\mathbf{C}} = (C_1, C_2, C_3, C_4, C_5, \ldots)$, so that we can define the **magnitude** of $\vec{\mathbf{C}}$ by

$$C = \mid \vec{\mathbf{C}} \mid = \sqrt{C_1^2 + C_2^2 + C_3^2 + \ldots} \qquad (3)$$

("Orthogonal" and "normal" are just synonyms for "perpendicular.") We could call $\vec{\mathbf{C}}$ the "Cost of Living Index" if we liked. There is a problem now. Our intuitive notion of "independent" components is tied up with the idea that one component can change without affecting another; yet as soon as we attempt to be specific about it, we find that we cannot even define a criterion for formal and exact independence (orthogonality) without generating a new notion: the idea of a magnitude as defined by Eq. (3). Does this definition agree with out intuition, the way the "ruler" analogy did? Most probably we *have* no intuition about the "magnitude" of the "cost of living vector." So we have created a new concept – not an arbitrary concept, but one which is guaranteed to have a large number of "neat" consequences, one we will be able to do calculations with, make transformations of, and so on. In short, a "rich" concept.

There is another problem, though; while we can easily test our space vectors with a ruler, there is no unambiguous "ruler" for the "cost of living index." Furthermore, we may make the *approximation* that the cost of tea bags is orthogonal to the cost of computer maintenance, but in so "messy" a business as economics we will never be able to prove this rigorously. There are too many "hidden variables" influencing the results in ways we do not suspect. This is too bad, but we can still live with the imperfections of an approximate model if it serves us well.

Chapter 9

Force *vs.* Mass

> *"If I have seen further than other men, it is because I stood on the shoulders of giants."* - Isaac Newton

Isaac Newton (1642-1727) published his masterwork,
Philosophiae Naturalis Principia Mathematica ("Mathematical Principles of Natural Philosophy") in 1687. In this tome he combined the individually remarkable conceptual achievements of calculus, vectors and an elegant expression of the simple relationship between *force* and *inertia* (which in effect gave definition to those entities for the first time) to produce an integrated description of the interactions between objects and exactly how they produce different kinds of motion. This was the true beginning of the science of *dynamics*, for it marked the adoption of the *descriptive paradigms* that are still used universally to describe dynamics, even after Quantum Mechanics has exposed Newtonian Mechanics as fundamentally inadequate.[1] Newton, like most great thinkers, had a variety of ludicrous foibles and was often a jerk in his dealings with others. I will not attempt to document his personal life, though many have done so [you can consult their work]; although it is interesting and revealing, it doesn't matter to our understanding of the conceptual edifice he built in the *Principia*. Moreover, I will make no attempt to introduce concepts in the order that Newton did, nor will I hesitate to use a more modern notation or even an updated version of a paradigm, with the rationale that (a) what matters most is getting the idea across clearly; and (b) we may have actually achieved a more elegant, compact understanding than Newton in the intervening centuries. This is one of the endearing (to me) traditions of Physics – and indeed of all genuine pursuit of truth[2] – we treasure an *æsthetic* of searching for a better, more elegant, more reliable, more accurate (with regard to predicting the results of experiments), *truer* model of the world and rooting out the demonstrably wrong parts of existing models. A frightening number of people who claim to *know* the Truth share no such æsthetic and in fact are dedicated to suppressing such activities when they threaten their most cherished and unexamined Truths. Grrrr....

Before we go on to expound Newton's "Laws" in their modern form it is useful to examine the "self-evident" [oh, yeah?] concepts of *force* and *mass* and their relationship with that relatively rigorously defined kinematic quantity, the *ac-*

[1] Note that Quantum Mechanics does *not* "prove Newtonian Mechanics wrong;" it merely reveals its shortcomings and the limits of its straightforward applicability. All paradigms have such shortcomings and limits, even Quantum Mechanics! Bridges did not fall down when Quantum Mechanics was "discovered," nor did engines or electromagnetic devices cease to function; we simply learned that Newtonian Mechanics and electromagnetic theory were *approximations* to a more fundamentally accurate picture furnished by Quantum Mechanics and Relativity, and where the approximation was no longer adequate to give a qualitatively correct description of the actual behaviour of matter.

[2] I should say "truth" or otherwise indicate that I don't mean there *is* some sort of ultimate Truth that we can discover and then relax.

celeration.

9.1 Inertia *vs.* Weight

Prior to Newton, people who thought about such things observed that objects which had lots of *inertia* [i.e. were hard to get moving by pushing on them, even where a nearly frictionless horizontal motion was possible] were also invariably *heavy* [i.e. were pulled down toward the centre of the Earth with great force]. It was therefore understandable for them to have *equated* inertia with *weight*, the magnitude of the force of attraction to the Earth.[3] Newton was among the first to suggest that *inertia* and *weight* were not necessarily the same thing, but that in fact the Earth's gravity *just happened* to pull down on objects with a force proportional to their inertial factor or "mass" (m) which was actually defined in terms of their resistance to horizontal acceleration by some force other than gravity.

9.1.1 The Eötvös Experiment

Is there any way to *test* Newton's conjecture that "inertial mass" (the quantitative measure of an objects resistance to acceleration by an applied force) is different from "gravitational mass" (the factor determining the weight of said object)? Certainly. But first we must make the proposition more explicit:

- Inertial mass m_I is an *additive property* of matter. That is, two identical objects, when combined, will have twice the inertial mass of either one by itself.[4]
- When subjected to a given force $\vec{\mathbf{F}}$ [a vector quantity, since it certainly has both magnitude and direction], an object will be accelerated in the direction of $\vec{\mathbf{F}}$ at a rate $\vec{\mathbf{a}}$ which is *inversely proportional*[5] *to its inertial mass* m_I. Mathematically,

$$\vec{\mathbf{a}} \propto \frac{\vec{\mathbf{F}}}{m_I}. \tag{1}$$

- Gravitational mass m_G is also an additive property of matter.
- The force of gravity $\vec{\mathbf{W}}$ pulling an object "down" toward the centre of the Earth (i.e. its *weight*) is proportional to its gravitational mass m_G. Let's write the constant of proportionality "g" so that $W = g\, m_G$ (where $W \equiv |\vec{\mathbf{W}}|$ is the *magnitude* of the weight, which is usually all we need, knowing as we do which way is "down") – or, in full vector notation,

$$\vec{\mathbf{W}} = -g\, m_G\, \hat{\mathbf{r}} \tag{2}$$

(where $\hat{\mathbf{r}}$ is the unit vector pointing *from* the centre of the Earth *to* the object in question).

The *combination* of the last two postulates is easy to check using a simple *balance*. However, it is not so easy to *separately* check these two propositions. See why? Fortunately, we don't have to.

If we put together the two equations $\vec{\mathbf{a}} \propto \vec{\mathbf{F}}/m_I$ and $\vec{\mathbf{W}} = -g\, m_G\, \hat{\mathbf{r}}$, noting that, in the case of the force of gravity *itself*, $\vec{\mathbf{F}} \equiv \vec{\mathbf{W}}$, we get

$$\vec{\mathbf{a}} \propto -\hat{\mathbf{r}}\, g\, \frac{m_G}{m_I} \tag{3}$$

[3] It is of course easy for us to see the error of such thinking, because we are privy to Newton's paradigms; this should not delude us into scorning the efforts of the "giants" on whose shoulders Newton stood to "see further than other men."

[4] This may seem absurdly self-evident, but in fact there *are* physical properties that are *not* additive! So we want to explicitly point out this assumption as a point of vulnerability of the model, in case it is found to break down later on. This sort of "full disclosure" is characteristic of any enterprise designed to get at the truth rather than to win an argument.

[5] This can be checked by applying a force to two identical objects stuck together and seeing if they accelerate exactly half as fast as either one individually subjected to the same force.

– i.e. the acceleration *due to gravity* is in the $-\hat{\mathbf{r}}$ direction (towards the centre of the Earth), and is proportional to the *ratio* of the gravitational mass to the inertial mass. *So...* if the gravitational mass is *proportional to* the inertial mass, then *all objects should experience the same acceleration when falling due to the force of gravity*, at least in the absence of any other forces like air friction. Wait! Isn't this just what Galileo was always trying to tell us? Yep. *But was he right?*

Clearly the answer hangs on the proportionality of m_G and m_I. As we shall see, any nontrivial constant of proportionality can be absorbed into the definition of the *units* of force; thus instead of $\vec{\mathbf{a}} \propto \vec{\mathbf{F}}/m_I$ we can write $\vec{\mathbf{a}} = \vec{\mathbf{F}}/m_I$ and the question becomes, "Are inertial mass and gravitational mass *the same thing*?" The experimental test is of course to actually *drop* a variety of objects in an evacuated chamber where there truly is no air friction (nor, we hope, any other more subtle types of friction) and measure their accelerations *as accurately as possible*. This was done by Eötvös to an advertised accuracy of 10^{-9} (one part per billion – often written 1 *ppb*) who found satisfactory agreement with Galileo's "law."[6] Henceforth I will therefore drop the $_G$ and $_I$ subscripts on mass and assume there is only one kind, *mass*, which I will write m.

9.1.2 Momentum

René Descartes and Christian Huygens together introduced the concept of *momentum* as the combination of an object's *weight* with its *velocity*, developing a rather powerful scheme for "before and after" analysis of isolated collisions and similar messy processes. I will be unfaithful to the historical sequence of conceptual evolution in this case primarily because I want to introduce the "impulse and momentum conservation law" later on as an example of the "emergence" of new paradigms from a desire to invent shortcuts around tedious mathematical calculations. Nevertheless, Newton actually formulated his Second Law in terms of momentum, so it would be too much of a distortion to omit at least a definition of momentum at this point, to wit:

$$\vec{\mathbf{p}} \equiv m\,\vec{\mathbf{v}} \tag{4}$$

I.e., the *momentum* of an object, a vector quantity which is almost always written $\vec{\mathbf{p}}$ (magnitude $|\vec{\mathbf{p}}| \equiv p$), is the *product* of the object's *mass* m and its vector *velocity* $\vec{\mathbf{v}}$.

9.2 Newton's Laws

We are now ready to state Newton's three "Laws" of motion, in Newton's own words:

1. First Law: *Every body continues in its state of rest, or of uniform motion in a right* [straight] *line, unless it is compelled to change that state by a force impressed on it.*
2. Second Law: *The change in motion* [rate of change of momentum with time] *is proportional to the motive force impressed; and is made in the direction of the right line in which that force is impressed.*
3. Third Law: *To every action there is always opposed an equal reaction; or, the mutual actions of two bodies are always equal, and directed to contrary parts.*

[6] Recent re-measurements by Dicke *et al.* challenged Eötvös' ability to measure so accurately; they tentatively reported *deviations* from the expected results, suggesting that there might be an incredibly weak "fifth force" between the Earth and other matter that is different for protons than for neutrons. This was hot news for a while, but the excitement seems to have died down now, presumably due to new measurements that once again agree with Galileo and Eötvös.

Now, Newton's language was fairly precise, but to our modern ears it sounds a bit stilted and not very concise. We also imagine that, with the benefit of several centuries of practice, we have achieved a clearer understanding of these Laws than Newton himself. Regardless of the validity of this conceit, we like to express the Laws in a more modern form with a little mathematical notation thrown in:

1. First Law: *A body's velocity* $\vec{\mathbf{v}}$ [which might be zero] *will never change unless and until a force* $\vec{\mathbf{F}}$ *acts on the body.*
2. Second Law: *The time rate of chage of the momentum of a body is* equal *to the force acting on the body.* That is,
$$\frac{d\vec{\mathbf{p}}}{dt} = \vec{\mathbf{F}}. \qquad (5)$$
3. Third Law: *Whenever a force* $\vec{\mathbf{F}}_{BA}$ *is applied to A by B, there is an equal and opposite reaction force* $\vec{\mathbf{F}}_{AB}$ *on B due to A.* That is,
$$\vec{\mathbf{F}}_{AB} = -\vec{\mathbf{F}}_{BA}, \qquad (6)$$
where the subscript $_{AB}$ (for instance) indicates the force *from A to B.*

As long as the mass m is constant[7] we have

$$\frac{d\vec{\mathbf{p}}}{dt} \;=\; \frac{d}{dt}(m\vec{\mathbf{v}}) \;=\; m\,\frac{d\vec{\mathbf{v}}}{dt} \;=\; m\,\vec{\mathbf{a}}$$

since the derivative of a constant times a variable is the constant times the derivative of the variable. Then the Second Law takes the more familiar form,

$$\vec{\mathbf{F}} \;=\; m\,\vec{\mathbf{a}}. \qquad (7)$$

[7] Counterexamples are not as rare as you might think! Consider for instance a *rocket*, which is constantly losing mass as the motor burns fuel. In such cases the original form of the Second Law is essential.

This famous equation is often written in scalar form,

$$\dot{p} \;\equiv\; \frac{dp}{dt} \;=\; F \qquad \text{or} \qquad F \;=\; m\,a$$

because $\dot{\vec{\mathbf{p}}}$ and $\vec{\mathbf{F}}$ are always in the same direction.

9.3 *What* Force?

The Third Law is a real ringer. It looks so trivial, yet it warns us of a leading cause of confusion in mechanics problems: *There are always two forces* for every interaction! When A exerts a force $\vec{\mathbf{F}}_{AB}$ on B there is always an equal and opposite force $\vec{\mathbf{F}}_{BA} = -\vec{\mathbf{F}}_{AB}$ exerted back on A *by* B. The latter is arbitrarily designated the "reaction force," but of course this is only because we first started talking about the former; both forces have equal intrinsic status. So if you say, "The force between A and B is…." I don't know which force you are talking about! Never talk about "the force" unless you mean "the Force" from *Star Wars*. Always make up a *sentence* describing the *action* taking place: "The force *exerted on* [A] *by* [B] is…."

9.3.1 The Free Body Diagram

A good way to keep track of this (and cathect the right hemisphere in the process) is to draw what is universally known in Physics as a *Free Body Diagram* [FBD]. When you need to analyze the forces acting on a body [there are usually more than one!] the first step is to decide upon the *boundary* of "the body" – i.e. an imaginary surface that separates "the body" from "the outside world" so that we can talk unambiguously about who is applying which force to whom. Having done this in our imagination, it is usually wise to actually *draw* a little sketch of "the body" isolated from the rest of the world; it needn't be a good sketch, just a blob of approximately the right shape so

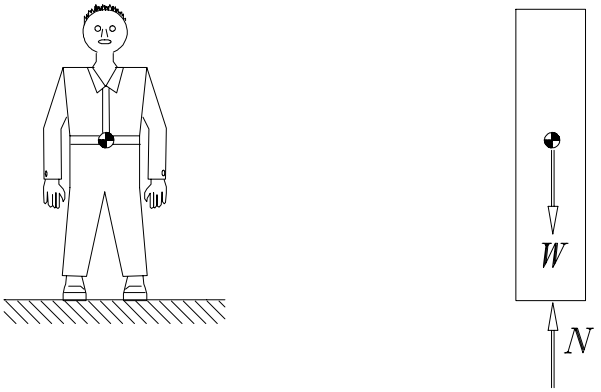


Figure 9.1 A man standing on the Earth (left) and his FBD (right). The man is pulled downward by the force of gravity W which is spread out over all his individual atoms but can be treated as if it were concentrated at his *centre of gravity* [CG] indicated on the diagram at about belt-buckle position. He is prevented from accelerating [falling] toward the centre of the Earth by the "normal force" N exerted upwards by the ground against his feet. These are the only two forces we need to consider to treat the problem of his *equilibrium* – i.e. the fact that he is not accelerating. The FBD on the right is perhaps a rather extreme example of a "simplified sketch" but it does serve the purpose, which is to show just the object in question and the forces acting on it *from outside*.

we know what we are talking about. Then we draw in each of the vector *forces* acting *on* the body *from* other entities in the outside world; forces are always pictured as little arrows pointing in the direction of application of the force.[8] A rather trivial example is shown in Fig. 9.1. We call N a "normal" force because it is *normal* (perpendicular) to the horizontal surface on which he stands; this terminology (and the N symbol) will be extended to describe any force exerted by a *frictionless surface* [yes, I know, another idealization...], which can *only* be perpendicular to that surface. Think about it.

Atwood's Machine:

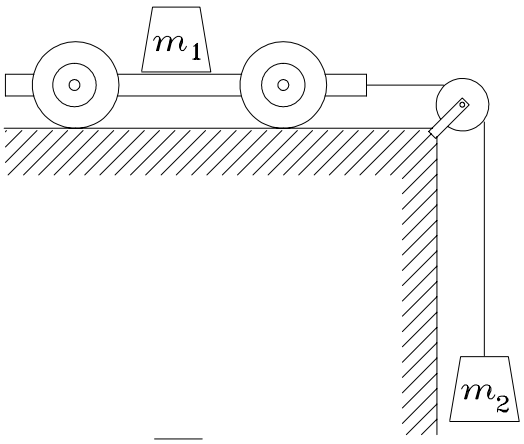


Figure 9.2 Atwood's Machine – one object labelled m_1 is glued to a massless cart with massless wheels that roll without friction on a perfectly horizontal surface. The cart is attached to a massless, unstretchable string that runs over yet another massless, frictionless pulley and is attached at the other end to a second object labelled m_2 that is pulled downward by the force of gravity. [You can see that a real experiment might involve a few corrections!] At the right are pictured the two separate FBD's for m_1 and m_2, showing all the external forces acting on each. Here W_1 is the *weight* of m_1 and N is the *normal force* exerted on m_1 by the horizontal surface (through the cart) to keep it from falling. Since it does *not* fall, N must exactly balance W_1. The only *unbalanced* force on m_1 is the *tension* T in the string, which accelerates it to the right. The tension in a string is the same everywhere, so the same T pulls *up* on m_2, partly counteracting its *weight* W_2.

To illustrate the use of the FBD in nontriv-

[8] If we mess up and draw the force in the opposite direction from its actual direction of application, we needn't worry, as the mathematics will automatically deliver up a result with a $-$ sign as if to say, "This force is in the opposite direction from the way you drew it, dummy!"

ial mechanics problems we can imagine another series of measurements[9] with a simple device known as Atwood's Machine. The apparatus is pictured in Fig. 9.2.

It is easy to see that the two vertical forces (W_1 and N) acting on m_1 must cancel. The rest is less trivial. The weight of m_2 is given by $W_2 = m_2\ g$; thus for m_1 and m_2, respectively, we have the "equations of motion"

$$a_1 \ = \ \frac{T}{m_1} \qquad \text{(to the right)}$$

and

$$a_2 \ = \ \frac{m_2\ g - T}{m_2} \qquad \text{(downward)}.$$

But we have here three unknowns (a_1, a_2 and T) and only two equations. The rules of linear algebra say that we need at least as many equations as unknowns to find a solution! Our salvation lies in recognition of the *constraints* of the system: Because the string does not stretch or go limp, both masses are *constrained* to move exactly the same distance (though in different directions) and therefore both experience the same *magnitude* of acceleration a. Thus our third equation is $a_1 = a_2 = a$ and we can equate the right sides of the two previous equations to get

$$\frac{T}{m_1} = \frac{m_2\ g - T}{m_2}$$

which we multiply through by $m_1\ m_2$ to get

$$m_2\ T \ = m_1\ m_2\ g - m_1\ T$$

$$\text{or} \qquad T\ [m_1 + m_2] = m_1\ m_2\ g$$

$$\text{or} \qquad T = \frac{m_1\ m_2\ g}{m_1 + m_2}.$$

Plugging this back into our first equation gives

$$a \ = \ g\ \frac{m_2}{m_1 + m_2}.$$

A quicker, simpler, more intuitive (and thus riskier) way of seeing this is to picture the pair of constrained masses *as a unit*. Let's use this approach to replace the distinction bewteen gravitational and inertial mass, just to see how it looks. The accelerating force is provided by the *weight* W_2 of m_2 which is given by $W_2 = g\, m_{2_G}$, where m_{2_G} is the gravitational mass of m_2. However, this force must accelerate *both* objects at the same rate because the string *constrains* both to move together (though in different directions). Thus the net inertia to be overcome by W_2 is the *sum* of the inertial masses of m_1 and m_2, so the acceleration is given by

$$a = \frac{W_2}{m_{1_I} + m_{2_I}} \ = \ g\ \frac{m_{2_G}}{m_{1_I} + m_{2_I}}$$

$$\text{or} \qquad \frac{a}{g} \ = \ \frac{m_{2_G}}{m_{1_I} + m_{2_I}}.$$

The latter form expresses the acceleration explicity in units of g, the acceleration of gravity, which is often called "one gee."

Suppose we have *three identical objects*, each of which has the *same* inertial mass m_I and the *same* gravitational mass m_G. [This can easily be checked using a balance and a standard force like a spring.] Then we use two of them for m_1 and m_2, set the apparatus in motion and measure the acceleration in "gees." The result will be $a/g = m_G/2m_I$. Next we put *two* of the objects on the *cart* and leave the third hanging. This time we should get $a/g = m_G/3m_I$. Finally we hand two and leave one on the cart, for $a/g = 2m_G/3m_I$. If the measured accelerations are actually in the ratios of $\frac{1}{2} : \frac{1}{3} : \frac{2}{3}$ then it must be true that m_G/m_I is constant – i.e. that m_G is proportional to m_I or that in fact they are really basically the same thing (in this case)! Unfortunately we have only confirmed

[9] Aha! another *Gedankenexperiment*! But this time we can actually imagine performing it in our basement – or in a teaching lab at the University (where in fact it is almost always one of the required experiments in every first year Physics course). Of course, the actual experiment is beset by numerous annoying imperfections that interfere with our cherished idealizations and require tedious and ingenious corrections. Even simple experiments are hard in real life!

this *for these three identical objects*. In fact all we have really demonstrated is that our original postulates are not trivially wrong. To go further we need to repeat the Eötvös experiment.

Chapter 10

Celestial Mechanics

One of the triumphs of Newton's Mechanics was that he was able, using only his Laws of Motion and a postulated Universal Law of Gravitation, to explain the empirical Laws of Planetary Motion discovered by Johannes Kepler. [Clearly there was a great deal more respect for Law in those days than there is now!] Although the phenomenology of *orbits* (circular, elliptical and hyperbolic) would appear to be rather esoteric and applicable *only* to astronomy [and, today, astrogation], in fact the paradigm of *uniform circular motion* (*i.e.* motion in a circle at constant speed) is one of the most versatile in Physics. Let us begin, therefore, by deriving its essential and characteristic features.

10.1 Circular Motion

Although no real orbit is ever a perfect circle, those of the inner planets aren't too far off and in any case it is a convenient idealization. Besides, we aren't restricted to planetary orbits here; the following derivation applies to *any* form of uniform circular motion, from tether balls on ropes to motorcycles on a circular track to charged particles in a cyclotron.[1]

10.1.1 Radians

In Physics, angles are measured in *radians*. There is no such thing as a "degree," although Physicists will sometimes grudgingly admit that π is equivalent to 180°. The angle θ shown in Fig. 10.1 is *defined* as the *dimensionless ratio* of the distance ℓ travelled along the circular arc to the radius r of the circle. There is a good reason for this. The trigonometric functions $\cos(\theta) \equiv x/r$, $\sin(\theta) \equiv y/r$, $\tan(\theta) \equiv y/x$ *etc.* are themselves defined as dimensionless ratios and their *argument* (θ) ought to be a dimensionless ratio (a "pure number") too, so that these functions can be expressed as *power series* in θ:

$$\cos(\theta) = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots$$

$$\sin(\theta) = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots$$

Why would anyone want to do this? You'll see, heh, heh....

10.1.2 Rate of Change of a Vector

The derivative of a *vector* quantity $\vec{\boldsymbol{A}}$ with respect to some independent variable x (of which it is a function) is defined in exactly the

[1] You could even imagine examples from "outside Physics," in which the *radius* and *speed* were purely metaphorical; but I can't think of one....

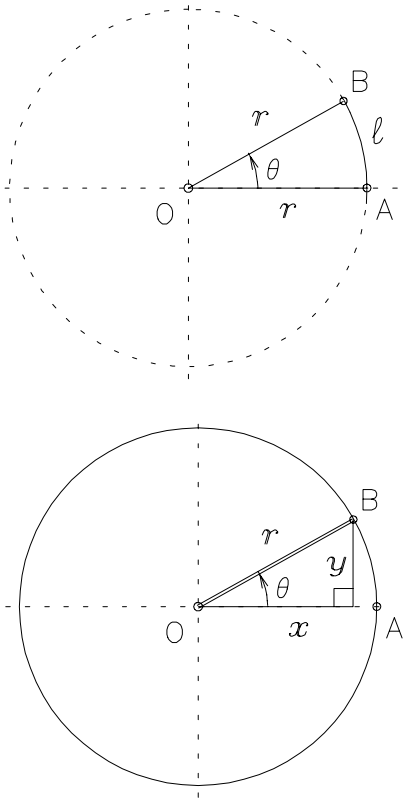


Figure 10.1 [top] Definition of the angle $\theta \equiv \ell/r$. [bottom] Illustration of the trigonometric functions $\cos(\theta) \equiv x/r$, $\sin(\theta) \equiv y/r$, $\tan(\theta) \equiv y/x$ *etc.* describing the position of a point B in circular motion about the centre at O.

---

same way as the derivative of a *scalar* function:

$$\frac{d\vec{\boldsymbol{A}}}{dx} \equiv \lim_{\Delta x \to 0} \frac{\vec{\boldsymbol{A}}(x + \Delta x) - \vec{\boldsymbol{A}}(x)}{\Delta x} \qquad (1)$$

There is, however, a dramatic difference between scalar derivatives and vector derivatives: the latter can be nonzero *even if the* **magnitude** A of the vector $\vec{\boldsymbol{A}}$ *remains constant*. This is a consequence of the fact that vectors have *two* properties: magnitude and direction. If the *direction* changes, the derivative is nonzero, even if the *magnitude* stays the same!

This is easily seen using a sketch in two dimensions:

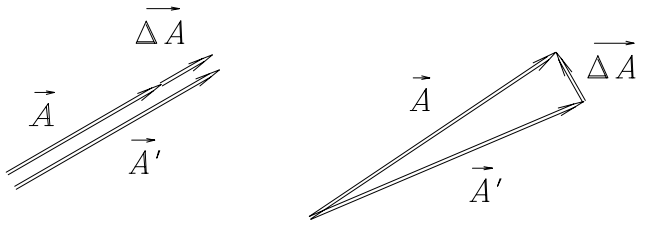


Figure 10.2 Note that the notation $\vec{\boldsymbol{A}}'$ does not denote the *derivative* of $\vec{\boldsymbol{A}}$ as it might in a Mathematics text.

In the case on the left, the vector $\vec{\boldsymbol{A}}'$ is in the same direction as $\vec{\boldsymbol{A}}$ but has a different *length*. [The two vectors are drawn side by side for visual clarity; try to imagine that they are on top of one another.] The difference vector $\Delta\vec{\boldsymbol{A}} \equiv \vec{\boldsymbol{A}}' - \vec{\boldsymbol{A}}$ is *parallel* to both $\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{A}}'$.[2] If we divide $\Delta\vec{\boldsymbol{A}}$ by the change Δx in the independent variable (of which $\vec{\boldsymbol{A}}$ is a function) and let $\Delta x \to 0$ then we find that the *derivative* $\dfrac{d\vec{\boldsymbol{A}}}{dx}$ is also $\parallel \vec{\boldsymbol{A}}$.

In the case on the right, the vector $\vec{\boldsymbol{A}}'$ has the same *length* (A) as $\vec{\boldsymbol{A}}$ but is *not* in the same direction. The difference $\Delta\vec{\boldsymbol{A}} \equiv \vec{\boldsymbol{A}}' - \vec{\boldsymbol{A}}$ formed by the "tip-to-tip" rule of vector subtraction is also no longer in the same direction as $\vec{\boldsymbol{A}}$. In fact, it is useful to note that for these conditions (constant magnitude A), as the difference $\Delta\vec{\boldsymbol{A}}$ becomes *infinitesimally small* it also becomes *perpendicular* to both $\vec{\boldsymbol{A}}$ and $\vec{\boldsymbol{A}}'$.[3] Thus the *rate of change* $\dfrac{d\vec{\boldsymbol{A}}}{dx}$ of a **vector** $\vec{\boldsymbol{A}}$ *whose magnitude A is constant* will always be *perpendicular to the vector itself*: $\dfrac{d\vec{\boldsymbol{A}}}{dx} \perp \vec{\boldsymbol{A}}$ if A is *constant*.

10.1.3 Centripetal Acceleration

From Fig. 10.3 we can see the relationship between the change in position $\Delta\vec{\boldsymbol{r}}$ and the change

[2] We write this $\Delta\vec{\boldsymbol{A}} \parallel \vec{\boldsymbol{A}} \parallel \vec{\boldsymbol{A}}'$ in standard notation.
[3] We write this $\Delta\vec{\boldsymbol{A}} \perp \vec{\boldsymbol{A}}$ in standard notation.

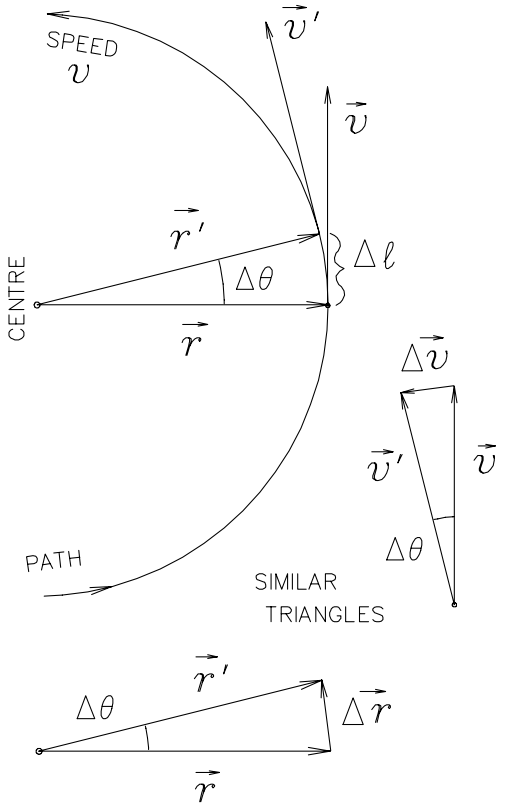


Figure 10.3 Differences between vectors at slightly different times for a body in uniform circular motion.

in velocity $\Delta\vec{\boldsymbol{v}}$ in a short time interval Δt. As all three get smaller and smaller, $\Delta\vec{\boldsymbol{v}}$ gets to be more and more exactly in the *centripetal direction* (along $-\hat{\boldsymbol{r}}$) and its scalar magnitude Δv will always (from *similar triangles*) be given by

$$\frac{|\Delta\vec{\boldsymbol{v}}|}{v} \;=\; \frac{|\Delta\vec{\boldsymbol{r}}|}{r}$$

where I have been careful to write $|\Delta\vec{\boldsymbol{r}}|$ rather than Δr since the *magnitude* of the radius vector, r, does not change! Now is a good time to note that, for a tiny sliver of a circle, there is a vanishingly small difference between $|\Delta\vec{\boldsymbol{r}}|$ and the actual distance $\Delta\ell$ travelled along the arc, which is given exactly by $\Delta\ell = r\Delta\theta$. Thus

$$\frac{\Delta\vec{\boldsymbol{v}}}{v} \;\approx\; -\hat{\boldsymbol{r}}\,\frac{r\Delta\theta}{r} \;=\; -\hat{\boldsymbol{r}}\,\Delta\theta.$$

If we divide both sides by Δt and then take the limit as $\Delta t \to 0$, the approximation becomes arbitrarily good and we get

$$\frac{1}{v}\left(\frac{d\vec{\boldsymbol{v}}}{dt}\right) \;=\; -\hat{\boldsymbol{r}}\left(\frac{d\theta}{dt}\right).$$

We can now combine this with the definitions of acceleration ($\vec{\boldsymbol{a}} \equiv d\vec{\boldsymbol{v}}/dt$) and angular velocity ($\omega \equiv d\theta/dt$) to give (after multiplying both sides by v) $\vec{\boldsymbol{a}} = -\hat{\boldsymbol{r}}\;\omega\;v$. We need only divide the equation $\Delta\ell = r\Delta\theta$ by Δt and let $\Delta t \to 0$ to realize that $v = r\omega$. If we substitute this result into our equation for the acceleration, it becomes

$$\vec{\boldsymbol{a}} \;=\; -\hat{\boldsymbol{r}}\,\frac{v^2}{r} \;=\; -\vec{\boldsymbol{r}}\,\omega^2 \qquad (2)$$

which is our familiar result for the *centripetal acceleration* in explicitly vectorial form.

10.2 Kepler

10.2.1 Empiricism

We are often led to believe that new theories are derived in order to explain fresh data. In actuality this is never the case. Theories are proposed to explain *experimental results*, which are always reported in an intermediate state of digestion somewhere between the raw data and the general explanatory theory. *Data* are merely meaningless bits of information and are often disregarded entirely unless and until their custodian (usually the Experimenter who collected them) translates them into some *empirical shorthand* that allows their *essential features* to be easily appreciated by other people. This is not always a simple task. Kepler, for instance, accumulated a large body of information in the form of observations of the positions of planets and stars as a function of time. In that form the data were incomprehensible to anyone, including Kepler. First he had to extract the *interesting part*, namely the positions

of the planets *relative to the Sun*, from raw data complicated by the *uninteresting* effects of the Earth's rotation and its own annual trip around Sol, which required both a good model of what was basically going on and a *lot* of difficult calculations. Then, with these "reduced" data in hand, he had to draw pictures, plot different combinations of the variables against each other, and generally mull over the data (presumably scratching his head and thinking, "Now what the hell does *this* mean?" or his contemporary equivalent) until he began to notice some interesting *empirical generalizations* that could be made about his results. Of course I don't know exactly how Kepler went about this, but I do know the experience of turning new data over and over in my mind and on paper until some consistent empirical relationship between the variables "leaps out at me." And I am *very* impressed with the depth and delicacy of Kepler's observations.

Note that the Empiricist[4] has not *explained* the observed behaviour at this point, merely *described* it.[5] But a good description goes a long way! One should never underestimate the importance of this intermediate step in experimental science.

10.2.2 Kepler's Laws of Planet Motion

1. Elliptical Orbits: The orbits of the planets are[6] *ellipses*[7] with the Sun at one of the foci.

2. Constant Areal Velocity: The *area* swept out *per unit time* by a line joining the Sun to the planet in question is constant throughout the orbit.[8]

3. Scaling of Periods: The *square* of the *period* T of the orbit is proportional to the *cube* of the length of the semi-major axis (or, in the case of a circular orbit, the *radius* r) of the orbit:

$$T^2 \propto r^3$$

10.3 Universal Gravitation

By a process of logic that I will not attempt to describe, Newton deduced that the force F between two objects with masses m and M separated by a distance r was given by

$$F = \frac{GmM}{r^2} \tag{3}$$

where
$G = (6.67259 \pm 0.00085) \times 10^{-11}\ \mathrm{m^3 \cdot kg^{-1} \cdot s^{-2}}$ is the *Universal Gravitational Constant*. Actually, Newton didn't know the value of G; he only postulated that it was *universal* — *i.e.* that it was the same constant of proportionality for *every* pair of masses in this universe.

[4] (who may or may not be the same person as the Experimentalist and/or the Theoretician — these are just different "hats" that a Physicist may put on)

[5] Of course, as in Kepler's case, the empirical description is always *in terms of* some preselected *model* or paradigm; but the paradigm in question is generally a familiar and widely accepted one, otherwise it is not very helpful in communicating the results to others. Besides, *the data themselves* are "collected" within the context of the Experimenter's paradigms and models about the world. The "simple" act of vision employs an enormous amount of "processing" in the visual cortex, as discussed earlier....

[6] (neglecting perturbations from the other planets, as is assumed in all Kepler's laws)

[7] Note that a *circle* is just a special case of an ellipse in which the major and semimajor axes are both equal to the *radius* and both *foci* are at the centre of the circle.

[8] This feature, unlike the other two Laws, is true for *any* "central force" (a force attracting the body back toward the centre, in this case the sun). The other two are only true for *inverse square laws*, $F \propto 1/r^2$.

The actual determination of the value of G was first done by Cavendish in an experiment to be described below.

We should also express this equation in vector form to emphasize that the force on either mass acts in the direction of the other mass: if $\vec{\boldsymbol{F}}_{12}$ denotes the force acting on mass m_2 due to its gravitational attraction by mass m_1 then

$$\vec{\boldsymbol{F}}_{12} = -\frac{G\,m_1\,m_2}{r_{12}^2}\,\hat{\boldsymbol{r}}_{12} \tag{4}$$

where $\hat{\boldsymbol{r}}_{12}$ is a unit vector in the direction of $\vec{\boldsymbol{r}}_{12}$, the vector distance from m_1 to m_2, and r_{12} is the scalar magnitude of $\vec{\boldsymbol{r}}_{12}$. Note that the reaction force $\vec{\boldsymbol{F}}_{21}$ on m_1 due to m_2 is obtained by interchanging the labels "1" and "2" which ensures that it is equal and opposite because $\vec{\boldsymbol{r}}_{21} \equiv -\vec{\boldsymbol{r}}_{12}$ by definition.

10.3.1 Weighing the Earth

Suppose you know your own mass m, determined not from your weight but from experiments in which you are accelerated horizontally by known forces. Then from your weight W you can calculate the mass of the Earth, M_E, if only you know G, the universal gravitational constant, and R_E, the radius of the Earth. The trouble is, you cannot use the same measurement (or any other combination of measurements of the weights of objects) to determine G. So how do we know G? If we can measure G then we can use our own weight-to-mass ratio (*i.e.* the acceleration of gravity, g) with the known value of $R_E = 6.37\times10^6$ m to determine M_E. How do we do it?

The trick is to measure the gravitational attraction between two masses m_1 and m_2 that are *both* known. This seems simple enough in principle; the problem is that the attractive force between two "laboratory-sized" masses is *incedibly tiny*.[9] Cavendish devised a clever method of measuring such tiny forces: He hung a "dumbbell" arrangement (two large spherical masses on opposite ends of a bar) from the ceiling by a long thin wire and let the system come completely to rest. Then he brought another large spherical mass up close to each of the end masses so that the gravitational attraction acted to *twist the wire*. By careful tests on shorter sections of the same wire he was able to determine the *torsional spring constant* of the wire and thus translate the angle of twist into a torque, which in turn he divided by the moment arm (half the length of the dumbbell) to obtain the force of gravity F between the two laboratory masses M_1 and m_2 for a given separation r between them. From this he determined G and from that, using

$$g = \frac{G\,M_E}{R_E^2} \tag{5}$$

he determined $M_E = 5.965\times10^{24}$ kg for the first time. We now know G a bit better (see above) but it is a hard thing to measure accurately!

10.3.2 Orbital Mechanics

Let's pretend for now that all orbits are simple circles. In that case we can easily calculate the orbital radius r at which the centripetal force of gravitational attraction F is just right to produce the centripetal acceleration a required to maintain a steady circular orbit at a given speed v. For starters we will refer to a light object (like a communication satellite) in orbit about the Earth.

Orbital Speed

The force and the acceleration are both *centripetal* (*i.e.* back towards the centre of the

[9] If the Earth attracts a 1 kg mass with a force of 9.81 N, the gravitational force between two 1 kg masses separated by R_E would be smaller by a factor equal to the number of kilograms in M_E, which is a *large* number. Fortunately the smaller masses can be placed much closer together; this helps quite a bit, but the force is still miniscule!

Earth, so we can just talk about the *magnitudes* of $\vec{\boldsymbol{F}}$ and $\vec{\boldsymbol{a}}$:

$$F = \frac{GmM_E}{r^2} \qquad \text{and} \qquad a = \frac{v^2}{r}.$$

$$\text{but} \qquad F = ma, \qquad \text{so}$$

$$\frac{GmM_E}{r^2} = \frac{mv^2}{r} \Longrightarrow \frac{G\,M_E}{r} = v^2.$$

We can "solve" this equation for v in terms of r,

$$v = \sqrt{\frac{G\,M_E}{r}}, \tag{6}$$

or for r in terms of v:

$$r = \frac{G\,M_E}{v^2}. \tag{7}$$

You can try your hand with these equations. See if you can show that the orbital velocity at the Earth's surface (*i.e.* the speed required for a frictionless train moving through an Equatorial tunnel to be in free fall all the way around the Earth) is 7.905 km/s. For a more practical example, try calculating the radius and velocity of a *geosynchronous satellite* — *i.e.* a signal-relaying satellite in an Equatorial orbit with a period of exactly one day, so that it *appears* to stay at exactly the same place in the sky all the time.[10]

Changing Orbits

The first thing you should notice about the above equations is that satellites move *slower* in *higher orbits*. This is slightly counterintuitive in that they go slower when they have further to go to get all the way around, which has a dramatic effect on the period (see below). However, that's the way it is. Consequently, if you are in a *low* orbit and you want to transfer into a *higher* orbit, you eventually want to end up going *slower*. Nevertheless, the first thing you do to initiate such a change is to *speed up*! See if you can figure out why.[11]

Periods of Orbits

We can now explain (at least for circular orbits) Kepler's Third Law. The period T of an orbit is the circumference $2\pi r$ divided by the speed of travel, v. Using the equation above for v in terms of r gives

$$T = \frac{2\pi r}{\sqrt{\frac{GM_E}{r}}}$$

$$= \frac{2\pi}{\sqrt{GM_E}} r^{\frac{3}{2}}$$

$$\text{or} \qquad T^2 \propto r^3$$

as observed by Kepler. Newton explained *why*.

10.4 Tides

Here on the surface of the Earth, we have little occasion to notice that the force of gravity drops off inversely as the square of the distance from the centre of the Earth.[12] This is fortunate, since otherwise Galileo would not have been able to do his experiments demonstrating the (approximate) constancy of the acceleration of gravity, g; moreover, scales and other mass-measuring technology based on uniform gravity would not work well enough for commerce of engineering to have evolved as it did. So we

[10] If you have a TV satellite dish, it is pointing at such a satellite; note that (if you live in the Northern Hemisphere) it is tipped toward the South. Why?

[11] (The most intuitive explanation for this involves the concepts of *kinetic and potential energy*, which we will watch *emerge* from Newton's Mechanics in succeeding Chapters.

[12] Surely by now you have gotten skeptical of my repeated declarations that the mass of the Earth can be treated as if it were all concentrated at the Earth's centre of gravity (*i.e.* the centre of the Earth). What about all the bits right next to us? They have a much smaller r^2 and thus contribute far more "pull" than those 'way on the other side. Well, hang on to that skepticism! I'm not leading you astray (promise!) but a little later on I will be in a better position to use *Gauss' Law* to explain in a few quick steps why this works. You should only provisionally accept this notion until you have seen a convincing argument with your own eyes.

don't notice any effects of the inverse square law "here at home," right? Well, let's not be hasty.

The Moon exerts an infinitesimal force on every bit of mass on Earth. At a distance of $R_ME = 380,000$ km, the Moon's mass of $M_M = 7.4 \times 10^{22}$ kg generates a gravitational acceleration of only $g_ME = 3.42 \times 10^{-5}$ m/s^2; in other words, our gravitational attraction to the Moon is 3.5×10^{-6} of our Earth weight. Moreover, the Moon's gravitational acceleration *changes* by only -1.8×10^{-13} m/s^2 for every metre further away from the Moon we move — a really *tiny* gravitational *gradient*. Nevertheless, the fact that the water in the oceans on the side of the Earth facing the Moon is attracted *more* and that on the side away from the Moon is attracted *less* leads to a slight *bulge* of the water on *both sides* and a concomitant *dip* around the middle. As the Earth turns under these bulges and dips, we experience (normally) *two* high tides and *two* low tides every day.

The consequences of these tides are nontrivial, as we all know. Even though they are the result of an incredibly small gravitational gradient, they represent enormous energies that have been tapped for power generation in a few places like the Bay of Fundy where resonance effects generate huge movements of water. More importantly in the long run (but of negligible concern in times of interest to humans) is the fact that the "friction" generated by these tides is gradually sapping the kinetic energy of the Earth's rotation and at the same time causing the Moon to drift slowly further away from the Earth so that in a few billion years the Earth will be "locked" as the Moon is now, with its day the same length as a month (which will then be twice as long as it is now) and the same side always facing its partner. *"Sic transit gloria Mundi,"* indeed! Let's enjoy our spin while we can.

A less potent source of tidal forces (gravitational gradients) on Earth is the Sun, with a mass of about 3×10^{40} kg at a distance of about 93 million miles or 1.5×10^{11} m. You can calculate for yourself the Sun's gravitational acceleration at the Earth: small but not entirely negligible. The Sun's gravitational *gradient*, on the other hand, is truly miniscule; yet various species of fish seem to have feeding patterns locked to the relative positions of the Sun and the Moon, even at night when the more obvious effects of the Sun are absent. The so-called "solunar tables" are an essential aid to the fanatically determined fisherman! Yet, so far as I know, no one has any plausible explanation for how a fish (or a bird or a shellfish, which also seem to know) can detect these minute force gradients.

A more dramatic example of tidal forces is the gravitational field near a *neutron star*, which has a large enough gradient to dismember travellers passing nearby even though their orbits take them safely past.[13] Near a *small black hole* the tidal forces can literally rip the vacuum apart into matter and antimatter, causing the black hole to explode with unmatched violence; this in fact limits how small black holes can be and still remain stable.[14]

[13] This *motif* has been used in several delightful science fiction stories, notably "Neutron Star" by Larry Niven. and ? Egg ? by ? .

[14] Bill Unruh, of the UBC Physics Department, is one of the world's leading experts on this subject.

Chapter 11

The Emergence of Mechanics

What *use* are Newton's "Laws" of Mechanics? Even a glib answer to that question can easily fill a 1-year course, if you really want to know. My purpose here is merely to offer some hints of how people learned to apply Newton's Laws to different types of Mechanics problems, began to notice that they were repeating certain calculations over and over in certain wide classes of problems, and eventually thought of cute shortcuts that then came to have a life of their own. That is, in the sense of Michael Polanyi's *The Tacit Dimension*, a number of new paradigms *emerged* from the technology of *practical application* of Newton's Mechanics.

The mathematical process of emergence generally works like this: we take the SECOND LAW and transform it using a formal *mathematical identity* operation such as "Do the same thing to both sides of an equation and you get a new equation that is equally valid." Then we think up names for the quantities on both sides of the new equation and *presto!* we have a new paradigm. I will show three important example of this process, not necessarily the way they first were "discovered," but in such a way as to illustrate how such things can be done. But first we will need a few new mathematical tools.

11.1 Some Math Tricks

11.1.1 Differentials

We have learned that the symbols df and dx represent the *coupled changes* in $f(x)$ and x, in the limit where the change in x (and consequently also the change in f) become infinitesimally small. We call these symbols the **differentials** of f and x and distinguish them from Δf and Δx only in this sense: Δf and Δx can be any size, but df and dx are always *infinitesimal* — *i.e.* small enough so that we can treat $f(x)$ as a straight line over an interval only dx wide.

This does not change the interpretation of the representation $\dfrac{df}{dx}$ for the *derivative* of $f(x)$ with respect to x, but it allows us to think of these *differentials* df and dx as "normal" algebraic symbols that can be manipulated in the usual fashion. For instance, we can write

$$df = \left(\frac{df}{dx}\right) dx$$

which looks rather trivial in this form. However, suppose we give the derivative its own name:

$$g(x) \equiv \frac{df}{dx}$$

Then the previous equation reads

$$df = g(x)\, dx \qquad \text{or just} \qquad df = g\, dx$$

which can now be read as an expression of the *relationship between the two differentials* df and dx. Hold that thought.

As an example, consider our familiar kinematical quantities

$$a \equiv \frac{dv}{dt} \qquad \text{and} \qquad v \equiv \frac{dx}{dt}.$$

If we treat the differentials as simple algebraic symbols, we can invert the latter definition and write

$$\frac{1}{v} = \frac{dt}{dx}.$$

(Don't worry too much about what this "means" for now.) Then we can multiply the left side of the definition of a by $1/v$ and multiply the right side by dt/dx and get an equally valid equation:

$$\frac{a}{v} = \frac{dv}{dt} \cdot \frac{dt}{dx} = \frac{dv}{dx}$$

or, multiplying both sides by $v\,dx$,

$$a\,dx \;=\; v\,dv \tag{1}$$

which is a good example of a *mathematical identity*, in this case involving the *differentials* of distance and velocity. Hold *that* thought.

11.1.2 Antiderivatives

Suppose we have a function $g(x)$ which we know is the derivative [with respect to x] of some other function $f(x)$, but *we don't know which* — *i.e.* we know $g(x)$ explicitly but we don't know [yet] what $f(x)$ it is the derivative *of*. We may then ask the question, "What is the function $f(x)$ whose derivative [with respect to x] is $g(x)$?" Another way of putting this would be to ask, "What is the *antiderivative* of $g(x)$?"[1]

Another name for the *antiderivative* is the *integral*, which is in fact the "official" version, but I like the former better because the name suggests how we go about "solving" one.[2]

For a handy example consider $g(x) = k\,x$. Then the *antiderivative* [integral] of $g(x)$ with respect to x is $f(x) = \frac{1}{2}\,k\,x^2 + f_0$ [where f_0 is some *constant*] because the derivative [with respect to x] of x^2 is $2x$ and the derivative of any constant is zero. Since any *combination* of constants is also a constant, it is equally valid to make the arbitrary constant term of the same *form* as the part which actually varies with x, *viz.* $f(x) = \frac{1}{2}\,k\,x^2 + \frac{1}{2}\,k\,x_0^2$. Thus f_0 is the same thing as $\frac{1}{2}\,k\,x_0^2$ and it is a matter of taste which you want to use.

Naturally we have a shorthand way of writing this. The *differential* equation

$$df = g(x)\,dx$$

[1] This is a lot like knowing that 6 is some number n multiplied by 2 and asking what n is. We figure this out by asking ourselves the question, "What do I have to multiply by 2 to get 6?" Later on we learn to call this "division" and express the question in the form, "What is $n = 6/2$?" but we might just as well call it "anti-multiplication" because that is how we solve it (unless it is too hard to do in our heads and we have to resort to some complicated technology like long division).

[2] Any introductory Calculus text will explain what an integral "means" in terms of visual pictures that the right hemisphere can handle easily: whereas the *derivative* of $f(x)$ is the *slope* of the curve, the *integral* of $g(x)$ is *the area under the curve*. This helps to visualize the integral as the limiting case of a *summation*: imagine the area under the curve of $g(x)$ from x_0 to x being divided up into N rectangular *columns* of equal width $\Delta x = \frac{1}{N}(x - x_0)$ and height $g(x_n)$, where $x_n = n\,\Delta x$ is the position of the n^{th} column. If N is a small number, then $\sum_{n=1}^{N} g(x_n)\,\Delta x$ is a crude approximation to the area under the smooth curve; but as N gets bigger, the columns get skinnier and the approximation becomes more and more accurate and is eventually (as $N \to \infty$) exact! This is the meaning of the integral sign:

$$\int_{x_0}^{x} g(x)\,dx \equiv \lim_{N\to\infty} \sum_{n=1}^{N} g(x_n)\,\Delta x$$

$$\text{where} \qquad \Delta x \equiv \frac{1}{N}(x - x_0) \qquad \text{and} \qquad x_n = n\,\Delta x.$$

Why do I put this nice graphical description in a footnote? Because we can understand most of the Physics applications of integrals by thinking of them as "antiderivatives" and because when we go to *solve* an integral we almost always do it by asking the question, "What function is this the derivative of?" which means *thinking* of integrals as antiderivatives. This is not a complete description of the mathematics, but it is sufficient for the purposes of this course. [See? We really *do* "deemphasize mathematics!"]

can be turned into the *integral* equation

$$f(x) = \int_{x_0}^{x} g(x)\, dx \tag{2}$$

which reads, "$f(x)$ is the integral of $g(x)$ with respect to x from x_0 to x." We have used the rule that the *integral of the differential* of f [or any other quantity] is just the quantity itself,[3] in this case f:

$$\int df = f \tag{3}$$

Our example then reads

$$\int_{x_0}^{x} k\, x\, dx = k \int_{x_0}^{x} x\, dx = \frac{1}{2}\, k\, x^2 \;-\; \frac{1}{2}\, k\, x_0^2$$

where we have used the feature that any constant (like k) can be brought "outside the integral" — *i.e.* to the left of the integral sign $\int$.

Now let's use these new tools to transform Newton's SECOND LAW into something more comfortable.

11.2 Impulse and Momentum

Multiplying a *scalar* times a vector is easy, it just changes its dimensions and length — *i.e.* it is transformed into a new *kind* of vector with new units but which is still in the same direction. For instance, when we multiply the vector *velocity* $\vec{\boldsymbol{v}}$ by the scalar *mass* m we get the vector *momentum* $\vec{\boldsymbol{p}} \equiv m\,\vec{\boldsymbol{v}}$. Let's play a little game with differentials and the SECOND LAW:

$$\vec{\boldsymbol{F}} = \frac{d\vec{\boldsymbol{p}}}{dt}.$$

Multiplying both sides by dt and integrating gives

$$\vec{\boldsymbol{F}}\, dt = d\vec{\boldsymbol{p}} \quad \Rightarrow \quad \int_{t_0}^{t} \vec{\boldsymbol{F}}\, dt = \int_{\vec{\boldsymbol{p}}_0}^{\vec{\boldsymbol{p}}} d\vec{\boldsymbol{p}} = \vec{\boldsymbol{p}} - \vec{\boldsymbol{p}}_0. \tag{4}$$

[3] This also holds for the integrals of differentials of *vectors*.

The left hand side of the final equation is the time integral of the net externally applied force $\vec{\boldsymbol{F}}$. This quantity is encountered so often in Mechanics problems [especially when $\vec{\boldsymbol{F}}$ is known to be an explicit function of time, $\vec{\boldsymbol{F}}(t)$] that we give it a *name*:

$$\int_{t_0}^{t} \vec{\boldsymbol{F}}(t)\, dt \;\equiv\; \text{IMPULSE due to applied force } \vec{\boldsymbol{F}} \tag{5}$$

Our equation can then be read as a sentence:

> "The *impulse* created by the net external force applied to a system is equal to the *momentum change* of the system."

11.2.1 Conservation of Momentum

The IMPULSE AND MOMENTUM law is certainly a rather simple transformation of Newton's SECOND LAW; in fact one may be tempted to think of it as a trivial restatement of the same thing. However, it is much simpler to *use* in many circumstances. The most useful application, surprisingly enough, is when there is *no* external force applied to the system and therefore *no* impulse and *no* change in momentum! In such cases the total momentum of the system does not change. We call this the LAW OF CONSERVATION OF MOMENTUM and use it much the same as Descartes and Huygens did in the days before Newton.[4]

Momentum conservation goes beyond Newton's FIRST LAW, though it may appear to be the same idea. Suppose our "system" [trick word, that!] consists not of *one* object but of *several*. Then the "net" [another one!] momentum of the system is the vector *sum* of the momenta of its components. This is where the power of momentum conservation becomes apparent. As long as there are no *external* forces, there can

[4] It should be remembered that René Descartes and Christian Huygens formulated the LAW OF CONSERVATION OF MOMENTUM *before* Newton's work on Mechanics. They probably deserve to be remembered as the First Modern Conservationists!

be as many forces as we like *between the component parts* of the system without having the slightest effect on their combined momentum. Thus, to take a macabre but traditional example, if we lob a hand grenade through the air, just after it explodes (before any of the fragments hit anything) all its pieces *taken together* still have the same net momentum as before the explosion.

The Law of Conservation of Momentum is particularly important in analyzing the collisions of elementary particles. Since such collisions are the only means we have for performing experiments on the forces between such particles, you can bet that every particle physicist is very happy to have such a powerful (and simple-to-use!) tool.

Example: Volkwagen-Cadillac Scattering

Let's do a simple example in one dimension [thus avoiding the complications of adding and subtracting vectors] based on an apocryphal but possibly true story: A Texas Cadillac dealer once ran a TV ad showing a Cadillac running head-on into a parked Volkswagen Bug at 100 km/h. Needless to say, the Bug was squashed flat. Figs. 11.1 and 11.2 show a simplified sketch of this event, using the "before-and-after" technique with which our new paradigm works best. Figure 11.1 shows an *elastic* collision, in which the cars *bounce* off each other; Figure 11.2 shows a *plastic* collision in which they stick together. For quantitative simplicity we assume that the Cadillac has exactly twice the mass of the Bug ($M = 2m$). In both cases the net initial momentum of the "Caddy-Bug system" is $MV_i = 200m$, where I have omitted the "km/h" units of V_i, the initial velocity of the Caddy. Therefore, since all the forces act *between* the components of the system, the total momentum of the system is conserved and the net momentum *after* the collision must also be $200m$.

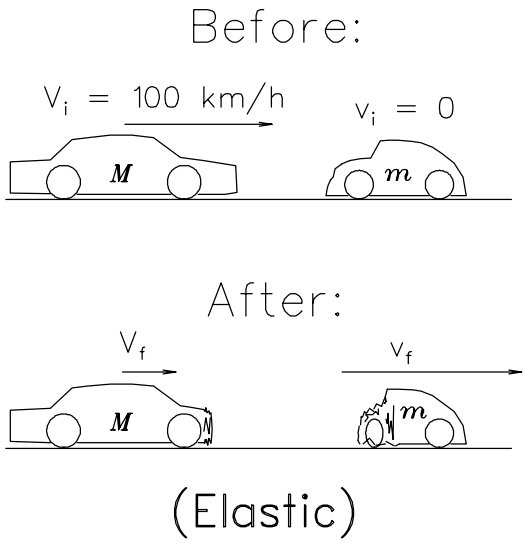


Figure 11.1 Sketch of a perfectly *elastic* collision between a Cadillac initially moving at 100 km/h and a parked Volkswagen Bug. For an elastic collision, the magnitude of the *relative* velocity between the two cars is the same before and after the collision. [The fact that the cars look "crunched" in the sketch reflects the fact that no actual collision between cars could ever be perfectly elastic; however, we will use this limiting case for purposes of illustration.]

In the *elastic* collision, the final *relative* velocity of the two cars must be the same as before the collision [this is one way of defining such a collision]. Thus if we assume (as on the drawing) that both cars move to the right after the collision, with velocities V_f for the Caddy and v_f for the Bug, then

$$v_f - V_f = 100 \qquad \text{or} \qquad v_f = V_f + 100.$$

Meanwhile the total momentum must be the same as initially:

$$MV_f \; + \; mv_f \; = \; 200m \qquad \text{or}$$

$$2mV_f \; + \; m(V_f + 100) \; = \; 200m$$

$$\text{or} \qquad 3mV_f = 100m$$

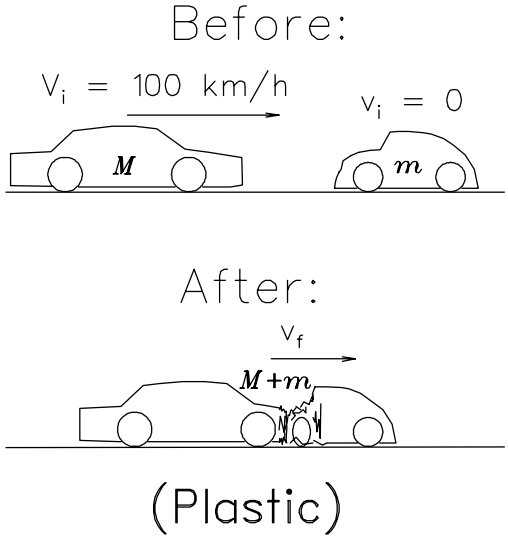


Figure 11.2 A perfectly inelastic or *plastic* collision in which the cars stick together and move as a unit after the collision.

---

giving the final velocities

$$V_f = 33\frac{1}{3} \text{ km/h} \qquad \text{and} \qquad v_f = 133\frac{1}{3} \text{ km/h}.$$

In the *plastic* collision, the final system consists of both cars stuck together and moving to the right at a common velocity v_f. Again the total momentum must be the same as initially:

$$(M+m)v_f = 200m \qquad \text{or}$$

$$3mv_f = 200m \qquad \text{or}$$

$$v_f = 66\frac{2}{3} \text{ km/h}.$$

Several features are worth noting: first, the final velocity of the Bug after the *elastic* collision is actually *faster* than the Caddy was going when it hit! If the Bug then runs into a brick wall, well.... For anyone unfortunate enough to be inside one of the vehicles the severity of the consequences would be worst for the largest sudden change in the velocity of that vehicle — *i.e.* for the largest *instantaneous acceleration* of the passenger. This quantity is far larger *for both cars* in the case of the *elastic* collision. This is why "collapsibility" is an important safety feature in modern automotive design. You want your car to be completely demolished in a severe collision, with only the passenger compartment left intact, in order to minimize the recoil velocity. This may be annoyingly expensive, but it is nice to be around to enjoy the luxury of being annoyed!

Back to our story: The Cadillac dealer was, of course, trying to convince prospective VW buyers that they would be a lot safer in a Cadillac — which is undeniable, except insofar as the Bug's greater maneuverability and smaller "cross-section" [the size of the "target" it presents to other vehicles] helps to *avoid* accidents. However, the local VW dealer took exception to the Cadillac dealer's stated editorial opinion that Bugs should not be allowed on the road. To illustrate *his* point, he ran a TV ad showing a Mack truck running into a parked Cadillac at 100 km/h. The Cadillac was quite satisfactorily squashed and the VW dealer suggested sarcastically that perhaps everyone should be required by law to drive Mack trucks to enhance road safety. His point was well taken.

11.2.2 Centre of Mass Velocity

If we calculate the total momentum of a composite system and then divide by the total mass, we obtain the velocity of the system-as-a-whole, which we call the *velocity of the centre of mass*. If we imagine "running alongside" the system at this velocity we will be "in a reference frame moving with the centre of mass," where everything moves together and bounces apart [or whatever] with a very satisfying symmetry. Regardless of the internal forces of collisions, *etc.*, the centre of mass [CM] will be motionless in this reference frame. This has many convenient features, especially for calculations, and has the advantage that the inifinite

number of other possible reference frames can all agree upon a common description in terms of the CM. Where exactly *is* the CM of a system? Well, wait a bit until we have defined *torques* and *rigid bodies*, and then it will be easy to show how to find the CM.

11.3 Work and Energy

We have seen how much fun it is to multiply the SECOND LAW by a scalar (dt) and integrate the result. What if we try multiplying through by a *vector*? As we have seen in the chapter on VECTORS, there are two ways to do this: the scalar or "dot" product $\vec{\boldsymbol{A}} \cdot \vec{\mathbf{B}}$, so named for the symbol $\cdot$ between the two vectors, which yields a *scalar* result, and the vector or "cross" product $\vec{\boldsymbol{A}} \times \vec{\mathbf{B}}$, whose name also reflects the appearance of the symbol $\times$ between the two vectors, which yields a *vector* result. The former is easier, so let's try it first.

In anticipation of situations where the applied force $\vec{\boldsymbol{F}}$ is an explicit function of the *position*[5] $\vec{\boldsymbol{x}}$ — *i.e.* $\vec{\boldsymbol{F}}(\vec{\boldsymbol{x}})$ — let's try using a differential change in $\vec{\boldsymbol{x}}$ as our multiplier:

$$\begin{aligned}
\vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} &= m\vec{\boldsymbol{a}} \cdot d\vec{\boldsymbol{x}} \\
&= m\frac{d\vec{\boldsymbol{v}}}{dt} \cdot d\vec{\boldsymbol{x}} \\
&= md\vec{\boldsymbol{v}} \cdot \frac{d\vec{\boldsymbol{x}}}{dt} \\
&= md\vec{\boldsymbol{v}} \cdot \vec{\boldsymbol{v}} \\
&= m\vec{\boldsymbol{v}} \cdot d\vec{\boldsymbol{v}}
\end{aligned}$$

where we have used the definitions of $\vec{\boldsymbol{a}}$ and $\vec{\boldsymbol{v}}$ with a little shifting about of the differential dt and a reordering of the dot product [which we may always do] to get the right-hand side [RHS] of the equation in the desired form. A delightful consequence of this form is that it allows us to convert the RHS into an explicitly *scalar* form: $\vec{\boldsymbol{v}} \cdot d\vec{\boldsymbol{v}}$ is zero if $d\vec{\boldsymbol{v}} \perp \vec{\boldsymbol{v}}$ — *i.e.* if the change in velocity is *perpendicular* to the velocity itself, so that the *magnitude* of the velocity does not change, only the *direction*. [Recall the case of circular motion!] If, on the other hand, $d\vec{\boldsymbol{v}} \parallel \vec{\boldsymbol{v}}$, then the whole effect of $d\vec{\boldsymbol{v}}$ is to change the *magnitude* of $\vec{\boldsymbol{v}}$, not its direction. Thus $\vec{\boldsymbol{v}} \cdot d\vec{\boldsymbol{v}}$ is precisely a measure of the *speed* v times the differential *change* in speed, dv:

$$\vec{\boldsymbol{v}} \cdot d\vec{\boldsymbol{v}} = v\,dv \tag{6}$$

so that our equation can now be written

$$\vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} = m\,v\,dv$$

and therefore

$$\int_{\vec{\boldsymbol{x}}_0}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} = m\int_{v_0}^{v} v\,dv = m\left(\frac{1}{2}v^2 - \frac{1}{2}v_0^2\right) \tag{7}$$

(Recall the earlier discussion of an equivalent *antiderivative*.)

Just to establish the connection to the mathematical identity $a\,dx = v\,dv$, we multiply that equation through by m and get $ma\,dx = mv\,dv$. Now, in *one dimension* (no vectors needed) we know to set $ma = F$ which gives us $F\,dx = mv\,dv$ or, integrating both sides,

$$\int_{x_0}^{x} F\,dx = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2$$

which is the same equation in one dimension.

OK, so what? Well, again this formula kept showing up over and over when people set out to solve certain types of Mechanics problems, and again they finally decided to recast the LAW in this form, giving new names to the left and right sides of the equation. We call

[5] In the section on CIRCULAR MOTION we chose $\vec{\boldsymbol{r}}$ to denote the vector position of a particle in a circular orbit, using the centre of the circle as the origin for the $\vec{\boldsymbol{r}}$ vector. Here we are switching to $\vec{\boldsymbol{x}}$ to emphasize that the current description works equally well for any type of motion, circular or otherwise. The two notations are interchangeable, but we tend to prefer $\vec{\boldsymbol{x}}$ when we are talking mainly about *rectilinear* (straight-line) motion and $\vec{\boldsymbol{r}}$ when we are referring our coordinates to some *centre* or *axis*.

$\vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}}$ the **work** dW done by exerting a force $\vec{\boldsymbol{F}}$ through a distance $d\vec{\boldsymbol{x}}$ [work is something we *do*] and we call $\frac{1}{2}mv^2$ the **kinetic energy** T. [kinetic energy is an *attribute* of a moving mass] Let's emphasize these definitions:

$$\int_{\vec{\boldsymbol{x}}_0}^{\vec{\boldsymbol{x}}} \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{x}} \equiv \Delta W \,, \tag{8}$$

the WORK done by $\vec{\boldsymbol{F}}(\vec{\boldsymbol{x}})$ over a path from $\vec{\boldsymbol{x}}_0$ to $\vec{\boldsymbol{x}}$, and

$$\frac{1}{2}mv^2 \equiv T \,, \tag{9}$$

the KINETIC ENERGY of mass m at speed v.

Our equation can then be read as a sentence:

> "When a force acts on a body, the *kinetic energy* of the body *changes* by an amount equal to the *work* done by the force exerted through a distance."

One nice thing about this "paradigm transformation" is that we have replaced a *vector* equation $\vec{\boldsymbol{F}} = m\,\vec{\boldsymbol{a}}$ by a *scalar* equation $\Delta W = \Delta T$. There are many situations in which the work done is easily calculated and the direction of the final velocity is obvious; one can then obtain the complete "final state" from the "initial state" in one quick step *without having to go through the details of what happens in between*. Another class of "before & after" problems solved!

11.3.1 Example: The Hill

Probably the most classic example of how the WORK AND ENERGY law can be used is the case of *a ball sliding*[6] *down a frictionless hill*, pictured schematically in Fig. 11.3. Now, Galileo was fond of this example and could have given us a calculation of the final speed of the ball *for the case of a straight-line path* (*i.e.* the inclined plane); but he would have thrown up his hands at the picture shown in Fig. 11.3! Consider one spot on the downward slope, say position **2**: the FBD of the ball is drawn in the expanded view, showing the two forces $\vec{\boldsymbol{N}}$ and $\vec{\boldsymbol{W}}$ acting on the mass m of the ball.[7] Now, the ball does not jump off the surface or burrow into it, so the motion is strictly tangential to the hill at every point.[8] Meanwhile, a *frictionless* surface cannot, by definition, exert any force *parallel* to the surface; this is why the *normal force* $\vec{\boldsymbol{N}}$ is called a "normal" force — it is always normal [perpendicular] to the surface.

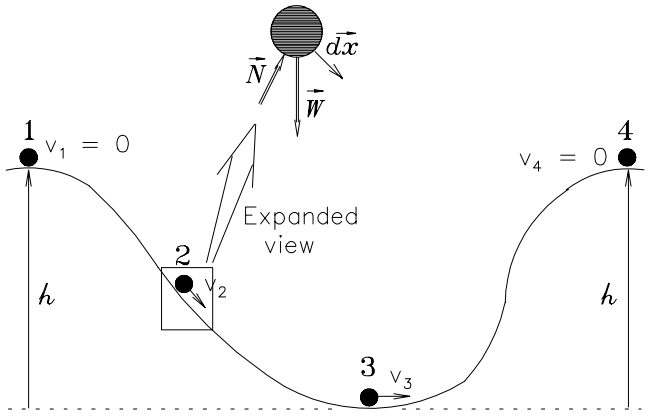


Figure 11.3 Sketch of a ball sliding down a frictionless hill. In position **1**, the ball is at rest. It is then given an infinitesimal nudge and starts to slide down the hill, passing position **2** on the way. At the bottom of the hill [position **3**] it has its maximum speed v_3 , which is then dissipated in sliding up the other side of the hill to position **4**. Assuming that it stops on a slight slope at both ends, the ball will keep sliding back and forth forever.

[6] We could have the ball *roll* up and down the hill instead of *sliding*, but that would involve *rotational kinetic energy*, and we're not there yet.

[7] It is unfortunate that the conventional symbol for the *weight*, $\vec{\boldsymbol{W}}$, uses the same letter as the conventional symbol for the *work*, W. I will try to keep this straight by referring to the weight always and only in its *vector* form and reserving the *scalar* W for the work. But this sort of difficulty is eventually inevitable.

[8] For now, I specifically *exclude* cases where the ball gets going so fast that it *does* get airborne at some places.

So $\vec{N} \perp d\vec{x}$ which means that $\vec{N} \cdot d\vec{x} = 0$ and *the normal force does no work!* This is an important general rule. Only the gravitational force $\vec{W}$ does any work on the mass m, and since $\vec{W} = -m\,g\,\hat{y}$ is a constant downward vector [where we define the unit vector $\hat{y}$ as "up"], *it is only the downward component* of $d\vec{x}$ that produces any work at all. That is, $\vec{W} \cdot d\vec{x} = -m\,g\,dy$, where dy is the component of $d\vec{x}$ directed *upward*.[9] That is, no matter what angle the hill makes with the vertical at any position, at that position the work done by gravity in *raising* the ball a differential height dy is given by $dW = -m\,g\,dy$ [notice that gravity does *negative* work going *up*hill and *positive* work going *down*hill] and the net work done in raising the ball a total distance Δy is given by a rather easy integral:

$$\Delta W = -m\,g \int dy = -m\,g\,\Delta y$$

where Δy is the height that the ball is *raised* in the process. By our LAW, this must be equal to the change in the *kinetic energy* $T \equiv \frac{1}{2}mv^2$ so that

$$\frac{1}{2}mv^2 - \frac{1}{2}mv_0^2 \;=\; -\,m\,g\,\Delta y. \qquad (10)$$

This formula governs both uphill slides, in which Δy is positive and the ball slows down, and downhill slides in which Δy is negative and the ball speeds up. For the example shown in Fig. 11.3 we start at the top with $v_0 = v_1 = 0$ and slide down to position **3**, dropping the height by an amount h in the process, so that the maximum speed (at position **3**) is given by

$$\frac{1}{2}mv_3^2 = mgh \qquad \text{or} \qquad v_3 = \sqrt{2gh}.$$

On the way up the other side the process exactly reverses itself [though the *details* may be completely different!] in that the altitude once again increases and the velocity drops back to zero.

The most pleasant consequence of this paradigm is that as long as the surface is truly frictionless, *we never have to know any of the details about the descent* to calculate the velocity at the bottom! The ball can drop straight down, it can slide up and down any number of little hills [as long as none of them are higher than its original position] or it can even slide through a tunnel or "black box" whose interior is hidden and unknown — and as long as I guarantee a frictionless surface you can be confident that it will come out the other end at the same speed as if it had just fallen the same vertical distance straight down. The *direction* of motion at the bottom will of course always be tangential to the surface.

For me it seems impossible to imagine the ball sliding up and down the hill without starting to think in terms of kinetic energy being *stored up* somehow and then automatically re-emerging from that storage as fresh kinetic energy. But I have already been indoctrinated into this way of thinking, so it is hard to know if this is really a compelling metaphor or just an extremely successful one. You be the judge. I will force myself to hold off talking about *potential energy* until I have covered the second prototypical example of the interplay between work and energy.

11.3.2 Captain Hooke

The *spring* embodies one of Physics' premiere paradigms, the *linear restoring force*. That is, a force which disappears when the system in question is in its "equilibrium position" x_0 [which we will define as the $x = 0$ position $(x_0 \equiv 0)$ to make the calculations easier] but increases as x moves away from equilibrium, in such a way that the *magnitude* of the force F is proportional to the displacement

[9] Alas, another unfortunate juxtaposition of symbols! We are using $d\vec{x}$ to describe the differential *vector* position change and dy to describe the *vertical component* of $d\vec{x}$. Fortunately we have no cause to talk about the horizontal component in this context, or we might wish we had used $d\vec{r}$ after all!

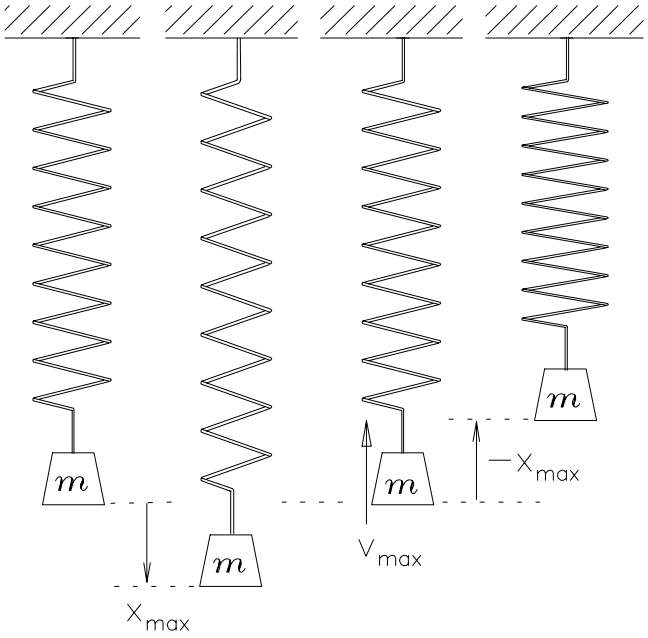


Figure 11.4 Sketch of a mass on a spring. In the leftmost frame the mass m is at rest and the spring is in its equilibrium position (*i.e.* neither stretched nor compressed). [If gravity is pulling the mass down, then in the equilibrium position the spring is stretched just enough to counteract the force of gravity. The equilibrium position can still be taken to define the $x = 0$ position.] In the second frame, the spring has been gradually pulled down a distance $x_{\max}$ and the mass is once again at rest. Then the mass is released and accelerates upward under the influence of the spring until it reaches the equilibrium position again [third frame]. This time, however, it is moving at its maximum velocity $v_{\max}$ as it crosses the centre position; as soon as it goes higher, it *compresses* the spring and begins to be *decelerated* by a linear restoring force in the opposite direction. Eventually, when $x = -x_{\max}$, all the kinetic energy has been been stored back up in the compression of the spring and the mass is once again instantaneously at rest [fourth frame]. It immediately starts moving downward again at maximum acceleration and heads back toward its starting point. In the absence of friction, this cycle will repeat forever.

from equilibrium [F is *linear* in x] and the *direction* of F is such as to try to *restore* x to the original position. The constant of proportionality is called the *spring constant*, always written k. Thus (using vector notation to account for the directionality)

$$\vec{\boldsymbol{F}} = -k\,\vec{\boldsymbol{x}} \tag{11}$$

which is the mathematical expression of the concept of a *linear restoring force*. This is known as HOOKE'S LAW. It is a lot more general than it looks, as we shall see later.

Keeping in mind that the $\vec{\boldsymbol{F}}$ given above is the force exerted *by the spring* against anyone or anything trying to stretch or compress it. If you are that stretcher/compressor, the force *you* exert is $-\vec{\boldsymbol{F}}$. If you *do work on the spring*[10] by stretching or compressing it[11] by a differential displacement $d\vec{\boldsymbol{x}}$ from equilibrium, the differential amount of work done is given by

$$dW \;=\; -\vec{\boldsymbol{F}}\cdot d\vec{\boldsymbol{x}} \;=\; k\,\vec{\boldsymbol{x}}\cdot d\vec{\boldsymbol{x}} \;=\; k\,x\,dx$$

which we can integrate from $x = 0$ (the equilibrium position) to x (the final position) to get the net work W:

$$W = k\int_0^x x\,dx = \frac{1}{2}k\,x^2 \tag{12}$$

Once you let go, *the spring will do the same amount of work back* against the only thing trying to impede it — namely, the inertia of the mass m attached to it. This can be used with the WORK AND ENERGY Law to calculate the speed $v_{\max}$ in the third frame of Fig. 11.4: since $v_0 = 0$,

$$\frac{1}{2}m\,v_{\max}^2 = \frac{1}{2}k\,x_{\max}^2 \qquad \text{or} \qquad v_{\max}^2 = \frac{k}{m}\,x_{\max}^2$$

[10] It is important to keep careful track of *who* is doing work on *whom*, especially in this case, because if you are careless the minus signs start jumping around and multiplying like cockroaches!

[11] It doesn't matter which — if you stretch it out you have to *pull* in the same direction as it moves, while if you compress it you have to *push* in the direction of motion, so either way the force and the displacement are in the same direction and you do *positive work* on the spring.

$$\text{or} \qquad v_{\max} = \sqrt{\frac{k}{m}}\ |x_{\max}|$$

where $|x_{\max}|$ denotes the *absolute value* of $x_{\max}$ (*i.e.* its magnitude, always positive). Note that this is a relationship between the *maximum* values of v and x, which occur *at different times* during the process.

Love as a Spring

Few other paradigms in Physics are so easy to translate into "normal life" terms as the *linear restoring force*. As a whimsical example, consider an intimate relationship between two lovers. In this case x can represent "emotional distance" — a difficult thing to quantify but an easy one to imagine. There is some *equilibrium distance* x_0 where at least one of the lovers is most comfortable[12] — this time, just to show how it works, we will not choose x_0 to be the zero position of x but leave it in the equations explicitly. When circumstances (usually *work*) *force* a greater emotional *distance* for a while, the lover experiences a sort of *tension* that *pulls* him or her back closer to the beloved. This is a perfect analogy to the linear restoring force:

$$F = -k\,(x - x_0)$$

What few people seem to recognize is that this "force," like any linear restoring force, is symmetric: it works the same in both directions, too far apart and too close. When circumstances permit a return to greater closeness, the lover rushes back to the beloved (figuratively — we are talking about emotional distance x here!) and very often "overshoots" the equilibrium position x_0 to get temporarily *closer* than is comfortable. The natural *repulsion* that then occurs is no cause for dismay — you can't really have an attraction without it — but some people seem surprised to discover that the *attraction* that binds them to their beloved does not just keep acting no matter how close they get; they are very upset that x cannot just keep getting closer and closer without limit.[13] In later chapters I will have much more to say about the *oscillatory* pattern that gets going [see Fig. 11.4] when the *overshoot* is allowed to occur without any *friction* to dissipate the energy stored in the stretched spring [a process known as *damping*]. But first I really must pick up another essential paradigm that has been begging to be introduced.

11.4 Potential Energy

Imagine yourself on skis, poised motionless at the top of a snow-covered hill: one way or another, you are deeply aware of the *potential* of the hill to increase your speed. In Physics we like to think of this obvious capacity as the *potential* for *gravity* to increase your *kinetic energy*. We can be quantitative about it by going back to the bottom of the hill and recalling the long trudge *uphill* that it took to get to the top: this took a lot of *work*, and we know the formula for how much: in raising your elevation by a height h you did an amount of work $W = mgh$ "against gravity" [where m is your mass, of course]. That work is now somehow "stored up" because if you slip over the edge it will all come back to you in the form of kinetic energy! What could be more natural than to think of that "stored up work" as *gravitational potential energy*

$$V_g = m\,g\,h \tag{13}$$

which will all turn into kinetic energy if we allow h to go back down to zero?[14]

[12] Sadly, x_0 is not always the same for both partners in the relationship; this is a leading cause of *tension* in such cases. [Doesn't this metaphor extend gracefully?]

[13] I suspect that such foolishness is merely an example of single-valued logic [closer = better] obsessively misapplied, rather than some more insidious psychopathology. But I could be wrong!

[14] The choice of a *zero point* for V_g is arbitrary, of course, just like our choice of where $h = 0$. This is not a problem if

We can then picture a skier in a bowl-shaped valley zipping down the slope to the bottom $[V_g \to T]$ and then coasting back up to stop at the original height $[T \to V_g]$ and (after a skillful flip-turn) heading back downhill again $[V_g \to T]$. *In the absence of friction*, this could go on forever: $V_g \to T \to V_g \to T \to V_g \to T \to \ldots$.

The case of the spring is even more compelling, in its way: if you push in the spring a distance x, you have done some work $W = \frac{1}{2}k\,x^2$ "against the spring." If you let go, this work "comes back at you" and will accelerate a mass until all the stored energy has turned into kinetic energy. Again, it is irresistible to call that "stored spring energy" the *potential energy* of the spring,

$$V_s = \frac{1}{2}k\,x^2 \tag{14}$$

and again the scenario after the spring is released can be described as a perpetual cycle of $V_s \to T \to V_s \to T \to V_s \to T \to \ldots$.

11.4.1 Conservative Forces

Physicists so love their ENERGY paradigm that it has been elevated to a higher status than the original SECOND LAW from which it was derived! In orer to make this switch, of course, we had to invent a way of making the *reverse derivation — i.e.* obtaining the vector *force* $\vec{\boldsymbol{F}}$ exerted "spontaneously" by the system in question from the scalar *potential energy* V of the system. Here's how: in one dimension we can forget the vector stuff and just juggle the differentials in $dW_{\rm me} = F_{\rm me}\,dx$, where the $W_{\rm me}$ is the work I do in exerting a force $F_{\rm me}$ "against the system" through a distance dx. Assuming that *all* the work I do against the system is *conserved* by the system in the form of its *potential energy* V, then $dV = dW_{\rm me}$. On the other hand, the force F exerted *by* the system [*e.g.* the force exerted by the spring] is the equal and opposite reaction force to the force I exert: $F = -F_{\rm me}$. The law for *conservative forces in one dimension* is then

$$F = -\frac{dV}{dx} \tag{15}$$

That is, the force of (*e.g.*) the spring is *minus* the *rate of change of the potential energy with distance*.

In three dimensions this has a little more complicated form, since $V(\vec{\boldsymbol{x}})$ could in principle vary with all three components of $\vec{\boldsymbol{x}}$: x, y and z. We can talk about the three components independently,

$$F_x = -\frac{\partial V}{\partial x}\,, \quad F_y = -\frac{\partial V}{\partial y} \quad \text{and} \quad F_z = -\frac{\partial V}{\partial z}$$

where the notation ∂ is used to indicate derivatives with respect to *one* variable of a function of *several* variables [here $V(x, y, z)$] *with the other variables held fixed*. We call $\partial V/\partial x$ the *partial derivative* of V with respect to x. In the same spirit that moved us to invent vector notation in the first place [*i.e.* making the notation more compact], we use the *gradient operator*

$$\vec{\boldsymbol{\nabla}} \;\equiv\; \hat{x}\,\frac{\partial}{\partial x} \;+\; \hat{y}\,\frac{\partial}{\partial y} \;+\; \hat{z}\,\frac{\partial}{\partial z} \tag{16}$$

to express the three equations above in one compact form:

$$\vec{\boldsymbol{F}} \;=\; -\,\vec{\boldsymbol{\nabla}} V \tag{17}$$

The *gradient* is easy to visualize in *two* dimensions: suppose you are standing on a *real* hill. Since your *height* $h \equiv z$ is actually proportional to your gravitational potential energy V_g, it is perfectly consistent to view the actual hill as a *graph* of the function $V_g(x, y)$ of East-West coordinate x and North-South coordinate y. In this picture, looking down on the

we allow *negative* potential energies [which we do!] since it is only the *change* in potential energy that appears in any actual mechanics problem.

hill from above, the direction of the gradient $\vec{\boldsymbol{\nabla}}V_g$ is *uphill*, and the magnitude of the gradient is the *slope* of the hill at the position where the gradient is evaluated. The nice feature is that $\vec{\boldsymbol{\nabla}}V_g$ will automatically point "straight up the hill" — *i.e.* in the steepest direction. Thus $-\vec{\boldsymbol{\nabla}}V_g$ points "straight downhill" — *i.e.* in the direction a marble will roll if it is released at that spot! There are lots of neat tricks we can play with the gradient operator, but for now I'll leave it to digest.

11.4.2 Friction

What about not-so-conservative forces? In the real world a lot of energy gets *dissipated* through what is loosely known as *friction*. Nowhere will you find an entirely satisfactory definition of precisely what friction is, so I won't feel guilty about using the cop-out and saying that it is the cause of all work that does *not* "get stored up as potential energy." That is, when I do work against frictional forces, it will not reappear as kinetic energy when I "let go."

Where does it go? We have already started getting used to the notion that energy is *conserved*, so it is disturbing to find some work just being *lost*. Well, relax. The energy dissipated by work against friction is still around in the form of *heat*, which is something like *disordered potential and kinetic energy*.[15] We will talk more about heat a few chapters later.

11.5 Torque & Angular Momentum

Finally we come to the formally trickiest transformation of the SECOND LAW, the one involving the *vector product* (or "cross product") of $\vec{\boldsymbol{F}}$ with the distance $\vec{\boldsymbol{r}}$ away from some *origin*[16] "O." Here goes:

$$\vec{\boldsymbol{r}} \times \left[d\frac{\vec{\boldsymbol{p}}}{dt} = \vec{\boldsymbol{F}} \right] \qquad \text{gives} \qquad \vec{\boldsymbol{r}} \times \frac{d\vec{\boldsymbol{p}}}{dt} = \vec{\boldsymbol{r}} \times \vec{\boldsymbol{F}}$$

Now, the distributive law for derivatives applies to cross products, so

$$\frac{d}{dt}\left[\vec{\boldsymbol{r}} \times \vec{\boldsymbol{p}}\right] = \frac{d\vec{\boldsymbol{r}}}{dt} \times \vec{\boldsymbol{p}} + \vec{\boldsymbol{r}} \times \frac{d\vec{\boldsymbol{p}}}{dt}$$

but

$$\frac{d\vec{\boldsymbol{r}}}{dt} \equiv \vec{\boldsymbol{v}} \qquad \text{and} \qquad \vec{\boldsymbol{p}} \equiv m\,\vec{\boldsymbol{v}}$$

$$\text{so} \qquad \frac{d\vec{\boldsymbol{r}}}{dt} \times \vec{\boldsymbol{p}} = m\,(\vec{\boldsymbol{v}} \times \vec{\boldsymbol{v}}) = 0$$

because the cross product of any vector with *itself* is zero.[17] Therefore

$$\frac{d}{dt}\left[\vec{\boldsymbol{r}} \times \vec{\boldsymbol{p}}\right] = \vec{\boldsymbol{r}} \times \vec{\boldsymbol{F}}.$$

If we define two new entities,

$$\vec{\boldsymbol{r}} \times \vec{\boldsymbol{p}} \equiv \vec{\boldsymbol{L}}_O, \tag{18}$$

the *Angular Momentum* about O

and

$$\vec{\boldsymbol{r}} \times \vec{\boldsymbol{F}} \equiv \vec{\boldsymbol{\tau}}_O, \tag{19}$$

the *Torque* generated by $\vec{\boldsymbol{F}}$ about O,

then we can write the above result in the form

$$\frac{d\vec{\boldsymbol{L}}_O}{dt} = \vec{\boldsymbol{\tau}}_O \tag{20}$$

This equation looks remarkably similar to the SECOND LAW. In fact, it is the *rotational analogue* of the SECOND LAW. It says that

[15] [Not quite, but you can visualize lots of little atoms wiggling and jiggling seemingly at random — that's heat, sort of.]

[16] Note that everything we discuss in this case will be *with reference to the chosen origin* O, which may be chosen arbitrarily but must then be carefully remembered!

[17] Remember from the chapter on VECTORS that only the *perpendicular* parts of two vectors contribute to the cross product. *Any* two *parallel* vectors have *zero cross product*. A vector crossed with itself is the simplest example.

> "The rate of change of the *angular momentum* of a body *about the origin* O is equal to the *torque* generated by forces acting about O."

So what? Well, *if we choose the origin cleverly* this "new" Law gives us some very nice generalizations. Consider for instance an example which occurs *very* often in physics: the *central force*.

11.5.1 Central Forces

Many [maybe even most] forces in nature are directed toward [or away from] some "source" of the force. An obvious example is Newton's Universal Law of Gravitation, but there are many others evident, especially in elementary particle physics.[18] We call these forces "central" because if we regard the point toward [or away from] which the force points as the *centre* (or *origin* O) of our coordinate system, from which the position vector $\vec{\boldsymbol{r}}$ is drawn, the cross product between $\vec{\boldsymbol{r}}$ and $\vec{\boldsymbol{F}}$ (which is along $\hat{r}$) is always zero. That is,

> "A *central force* produces *no torque* about the centre; therefore the *angular momentum* about the centre *remains constant* under a central force."

This is the famous Law of Conservation of Angular Momentum. Note the limitation on its applicability.

The Figure Skater

Again, so what? Well, there are numerous examples of central forces in which angular momentum conservation is used to make sense of otherwise counterintuitive phenomena. For instance, consider the classic image of the *figure skater* doing a pirouette: she starts spinning with hands and feet as far extended as possible, then pulls them in as close to her body. As a result, even though no *torques* were applied, she spins much faster. Why? I can't draw a good figure skater, so I will resort to a cruder example [shown in Fig. 11.5] that has the same qualitative features: imagine a ball (mass m) on the end of a string that emerges through a hole in an axle which is held rigidly fixed. The ball is swinging around in a circle in the end of the string. For an initial radius r and an initial velocity $v = r\omega$, the initial momentum is $mr\omega$ and the angular momentum about O is $L_O = mvr = mr^2\omega$. Now suppose we pull in the string until $r' = \frac{1}{2}r$. To keep the same L_O the momentum (and therefore the velocity) must increase by a factor of 2, which means that the *angular* velocity $\omega' = 4\omega$ since the ball is now moving at twice the speed but has only half as far to go around the circumference

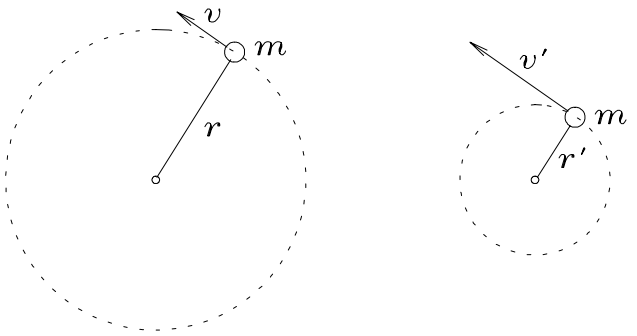


Figure 11.5 A contrived central-force problem. The ball swings around (without friction, of course) on the end of a string fixed at the origin O. The central force in the string cannot generate any torque about O, so the angular momentum $L_O = mvr$ about O must remain constant. As the string is pulled in slowly, the radius r gets shorter so the momentum $p = mv = mr\omega$ has to increase to compensate.

[18] For instance, the *electrostatic* force between two point *charges* obeys exactly the same "inverse square law" as gravitation, except with a much stronger constant of proportionality and the inclusion of both positive and negative charges. We will have lots more to do with that later on!

of the circle. The *period* of the "orbit" has thus *decreased* by a factor of four!

Returning to our more æsthetic example of the figure skater, if she is able to pull in all her mass a factor of 2 closer to her centre (on average) then she will spin 4 times more rapidly in the sense of revolutions per second or "Hertz" (Hz).

Kepler Again

A more formal example of the importance of the Law of Conservation of Angular Momentum under Central Forces is in its application to Celestial Mechanics, where the gravitational attraction of the Sun is certainly a classic central force. If we always use the Sun as our origin O, neglecting the influence of other planets and moons, the orbits of the planets must obey Conservation of Angular Momentum about the Sun. Suppose we draw a radius r from the Sun to the planet in question, as in Fig. 11.6. The

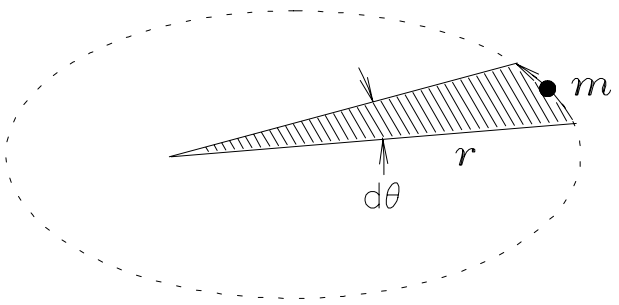


Figure 11.6 A diagram illustrating the *areal velocity* of an orbit. A planet (mass m) orbits the Sun at a distance r. the shaded area is equal to $\frac{1}{2}r \times r\,d\theta$ in the limit of infinitesimal intervals [*i.e.* as $d\theta \to 0$]. The *areal velocity* [*rate* at which this area is swept out] is thus $\frac{1}{2}r^2 d\theta/dt = \frac{1}{2}r^2\omega$.

rate at which this radius vector "sweeps out area" as the planet moves is $\frac{1}{2}r^2\omega$, whereas the *angular momentum about the Sun* is $mr^2\omega$. The two quantities differ only by the constants $\frac{1}{2}$ and m; therefore Kepler's *empirical* observation that the planetary orbits have constant "areal velocity" is *equivalent* to the requirement that the angular momentum about the Sun be a conserved quantity.

11.5.2 Rigid Bodies

Despite the fact that all Earthly matter is composed mostly of empty space sprinkled lightly with tiny bits of mass called atomic nuclei and even tinier bits called electrons, the forces between these bits are often so enormous that they hold the bits rigidly locked in a regular array called a *solid*. Within certain limits these arrays behave as if they were inseperable and perfectly rigid. It is therefore of some practical importance to develop a body of understanding of the behaviour of such *rigid bodies* under the influence of external forces. This is where the equations governing *rotation* come in.

A Moment of Inertia, Please!

Just as in the *translational* [straight-line motion] part of Mechanics there is an inertial factor m which determines how much p you get for a given $v \equiv dx/dt \equiv \dot{x}$ and how much $a \equiv dv/dt \equiv \dot{v} \equiv d^2x/dt^2 \equiv \ddot{x}$ you get for a given F,[19] so in *rotational* Mechanics there is an *angular analogue* of the inertial factor that determines how much L_O you get for a given $\omega \equiv \dot{\theta}$ and how much $\alpha \equiv \dot{\omega}$ you get for a given τ_O. This angular inertial factor is called the *moment of inertia* about O [we must always specify the origin about which we are defining torques and angular momentum] and is written I_O with the prescription

$$I_O = \int r_\perp^2\, dm \qquad (21)$$

[19] Here I have used the "dot notation" for *time derivatives*, defined in Ch. 6. This is merely a compact way of writing our favourite type of derivative — favourite because we consider knowing anything *as a function of time* as equivalent to complete knowledge of all its behaviour, since we can then find "everything else" by simply taking time derivatives.

where the integral represents a *summation* over all little "bits" of mass dm [we call these "mass elements"] which are distances $r_\perp$ away from an *axis* through the point O. Here we discover a slight complication: $r_\perp$ is measured from the *axis*, not from O itself. Thus a mass element dm that is a long way from O but right on the axis will contribute nothing to I_O. This continues to get more complicated until we have a complete description of Rotational Mechanics with I_O as a *tensor* of inertia and lots of other stuff I will never use again in this course. I believe I will stop here and leave the finer points of Rotational Mechanics for later Physics courses!

11.5.3 Rotational Analogies

It is, however, worth remembering that all the now-familiar [?] paradigms and equations of Mechanics come in "rotational analogues:"

Linear Version	**Angular Version**	**Name**
x	θ	angle
$\dot{x} \equiv v$	$\dot{\theta} \equiv \omega$	angular velocity
$\ddot{x} \equiv \dot{v} \equiv a$	$\ddot{\theta} \equiv \dot{\omega} \equiv \alpha$	angular acceleration
m	I_O	moment of inertia
$p = m\,v$	$L_O = I_O\omega$	angular momentum
F	τ_O	torque
$\dot{p} = F$	$\dot{L}_O = \tau_O$	Second Law
$T = \frac{1}{2}mv^2$	$T = \frac{1}{2}I_O\omega^2$	rotational kinetic energy
$dW = F\,dx$	$dW = \tau d\theta$	rotational work
$F = -kx$	$\tau = -\kappa\theta$	torsional spring law
$V_s = \frac{1}{2}k\,x^2$	$V_s = \frac{1}{2}\kappa\,\theta^2$	torsional potential energy

11.6 Statics

The enormous technology of *Mechanical Engineering* can be in some naïve sense be reduced to the two equations

$$\dot{\vec{\boldsymbol{p}}} = \vec{\boldsymbol{F}} \qquad \text{and} \qquad \dot{\vec{\boldsymbol{L}}}_O = \vec{\boldsymbol{\tau}}_O.$$

Whole courses are taught on what amounts to these two equations and the various tricks for solving them in different types of situations.

Fortunately, this isn't one of them! Just to give a flavour, however, I will mention the basic problem-solving technique of *Statics*, the science of things that are sitting still![20] That means $\dot{\vec{\boldsymbol{p}}} = 0$ and $\dot{\vec{\boldsymbol{L}}}_O = 0$ so that the relevant equations are now

$$\sum \vec{\boldsymbol{F}} \;=\; 0 \qquad \text{and} \qquad \sum \vec{\boldsymbol{\tau}}_O \;=\; 0$$

where the $\sum$ [summation] symbols emphasize that there is never just *one* force or *one* torque acting on a rigid body in equilibrium; if there were, it (the force or torque) would be unbalanced and acceleration would inevitably result!

To solve complex three-dimensional Statics problems it is often useful to back away from our nice tidy vector formalism and explicitly write out the "equations of equilibrium" in terms of the components of the forces along the $\hat{x}, \hat{y}$ and $\hat{z}$ directions as well as the torques about the x, y and z axes [which meet at the origin O]:

$$\sum F_x \;=\; 0 \qquad\qquad \sum \tau_x \;=\; 0 \quad (22)$$

$$\sum F_y \;=\; 0 \qquad\qquad \sum \tau_y \;=\; 0 \quad (23)$$

$$\sum F_z \;=\; 0 \qquad\qquad \sum \tau_z \;=\; 0 \quad (24)$$

If you have some civil engineering to do, you can work it out with these equations. Or hire an Engineer. I suggest the latter.

11.7 Physics as Poetry

This has been a long chapter; it needs some summary remarks. All I have set out to do here is to introduce the paradigms that emerged from Newton's Second Law through mathematical identity transformations. This process of *emergence* seems almost miraculous sometimes because by a simple [?] rearrangement of previously defined concepts we are able to create new *meaning* that wasn't there before! This is one of the ways Physics bears a family resemblance to Poetry and the other Arts. The Poet also juxtaposes familiar images in a new way and creates meaning that no one has ever seen before; this is the finest product of the human mind and one of the greatest inspirations to the human spirit.

In Physics, of course, the process is more sluggish, because we insist on working out all the ramifications of every new paradigm shift and evaluating its elegance and utility in some detail before we decide to "go with it." This explains why it is so easy to describe just how the concepts introduced in this chapter *emerged* from Newton's Mechanics, but not so easy to tidily describe the consequences (or even the *nature*) of more recent paradigm shifts whose implications are still being discovered. There is a lot of technical overhead to creativity in Physics.

A Physics paradigm shift is a profound alteration of the way Physicists see the world; but what do the rest of us care? It can be argued that such shifts have effects on our Reality even if we choose to exclude Physics from our immediate awareness. Examples of this are plentiful even in Classical Mechanics, but the first dramatic social revolution that can be clearly seen to have arisen largely from the practical consequences of breakthroughs in Physics was the Industrial Revolution, the origins of which will be discussed in the chapter on Thermal Physics.

[20] This is pretty boring from a Physicist's point of view, but even Physicists are grateful when bridges do not collapse.

Chapter 12

Equations of Motion

In the previous chapter we explored the process of *emergence* of new paradigms in Mechanics, using various mathematical identities to transform Newton's SECOND LAW into new equations whose left- and right-hand sides were given names of their own, like *impulse, momentum, work, energy, torque* and *angular momentum*. Eighteenth-Century physicists then learned to manipulate these "new" concepts in ways that greatly clarified the behaviour of objects in the material universe. As a result, previously mysterious or counterintuitive phenomena began to make sense in terms of *simple, easy-to-use models*, rather than long involved calculations. This is the essence of what Physics is all about. We work hard to make todays's difficult tasks easier, so that we will have more free time and energy tomorrow to work hard to make tomorrow's difficult tasks easier, so that....

Meanwhile, these new words made their way into day-to-day language and introduced new paradigms into society, whose evolution in "The Age of Reason" might have followed other paths were it not for Newton's work.[1] The effects of a more versatile and effective science of Mechanics were also felt in blunt practical terms: combined with the new science of Thermodynamics (to be discussed in a later chapter), Mechanics made possible an unprecedented growth of Mankind's ability to push Nature around by brute force, a profitable enterprise (in the short term) that led to the Industrial Revolution. Suddenly people no longer had to accept what Nature dealt, which enhanced their health and wealth considerably — but in taking new cards they found they also had a new dealer who was more merciless than Nature had ever been: Greed.

Here arises a perennial question: are the evils of "technology abuse," from pollution to exploitation to weapons of war, the "fault" of scientists who create the conceptual tools that make technology possible?[2] My own opinion is that we scientists have a responsibility for our creations in much the same way that parents have a responsibility for their children: we try to provide a wholesome and enlightened atmosphere in which they can grow and fulfill all their potential, offering our guidance and advice whenever it will be accepted, and setting the best example we can; but in the end ideas are like people — they will determine their own destiny. The best scientists can do to guide the impact of their ideas on society is to make sure the individual members of society have the opportunity to learn about those ideas. Whether

[1] Then again, maybe subtle sociological evolution had already made these changes inevitable and Newton was just the vehicle through which the emergent paradigms of the day infiltrated the world of science. Let's do the Seventeenth and Eighteenth Centuries over again without Newton and see how it comes out!

[2] I presume that I do not need to point out the distinction between *Science* and *Technology*. Even though politicians seem to be fond of the word "scienceandtechnology," I feel sure my readers are intelligent enough to find such a juxtaposition humourous.

anyone takes advantage of that opportunity or not is out of our control. Whether irresponsible or malign individuals make evil use of our ideas is also out of our control, though we can do our best to dissuade them.[3]

12.1 "Solving" the Motion

Getting back to the subject of Mechanics...

One of the reasons the paradigms in the previous chapter emerged was that physicists were always trying to "solve" certain types of "problems" using Newton's SECOND LAW,[4]

$$F = m\,\ddot{x}$$

This equation can be written

$$\ddot{x} = \frac{1}{m}\,F \tag{1}$$

to emphasize that it described a relationship between the *acceleration* $\ddot{x}$, the inertial coefficient m [usually constant] and the force F. It is conventional to call an equation in this form the "**equation of motion**" governing the problem at hand. When F is *constant* [as for "local" gravity] the "solution" to the equation of motion is the well-known set of equations governing *constant acceleration*, covered in the chapter on FALLING BODIES. Things are not always that simple, though.

Sometimes the problem is posed in such a way that the force F is explicitly a function of *time*, $F(t)$. This is not hard to work with, at least in principle, since the equation of motion (1) is then in the form

$$\ddot{x} = \frac{1}{m}\,F(t) \tag{2}$$

which can be straightforwardly *integrated* [assuming one knows a function whose time derivative is $F(t)$] using the formal operation

$$v(t) \;\equiv\; \dot{x} \;\equiv\; \int_0^t \ddot{x}\,dt \;=\; \frac{1}{m}\int_0^t F(t)\,dt \tag{3}$$

— which, when multiplied on both sides by m, leads to the paradigm of *Impulse and Momentum*.

In other cases the problem may be posed in such a way that the force F is explicitly a function of *position*, $F(x)$. Then the equation of motion has the form

$$\ddot{x} = \frac{1}{m}\,F(x) \tag{4}$$

which can be converted without too much trouble [using the identity $a\,dx = v\,dv$] into the paradigm of *Work and Energy*.

12.1.1 Timing is Everything!

If the equation of motion is the "question," what constitutes an "answer"? Surely the most *convenient* thing to know about any given problem is the *explicit time dependence of the position*, $x(t)$, because if we want the *velocity* $v(t) \equiv \dot{x}$, all we have to do is take the first time derivative — which may not be entirely trivial but is usually much easier than integrating! And if we want the *acceleration* $a(t) \equiv \dot{v} \equiv \ddot{x}$, all we have to do is take the time derivative again. Once you have found the acceleration, of course, you also know the net *force* on the object, by NEWTON'S SECOND LAW. A problem of this sort is therefore considered "solved" when we have discovered the explicit function $x(t)$ that "satisfies" the equation of motion.

For example, suppose we know that

$$x(t) = x_0\cos(\omega t), \tag{5}$$

where ω is some constant with units of *radians/unit time*, so that ωt is an *angle*. The time derivative of this is the velocity

$$\dot{x} \equiv v(t) = -\omega\,x_0\sin(\omega t)$$

[3] Some people feel that we should be prevented from *having* new ideas until those ideas have been "cleared" as innocuous. This would be hilarious if it weren't so dangerous.

[4] Let's limit our attention to *one dimensional* problems for the duration of this chapter, to keep things simple and avoid the necessity of using vector notation.

[look it up if needed] and the time derivative of *that* is the acceleration

$$\ddot{x} \equiv \dot{v} \equiv a(t) = -\omega^2 \, x_0 \cos(\omega t).$$

Note that the right-hand side of the last equation is just $-\omega^2$ times our original formula for $x(t)$, so we can also write

$$\ddot{x} = -\omega^2 \, x. \tag{6}$$

Multiplying through both sides by the mass m of the object in motion gives

$$ma = F = -m\omega^2 \, x,$$

which ought to look familiar to you: it is just HOOKE'S LAW with a force constant $k = m\omega^2$. Rearranging this a little gives

$$\omega = \sqrt{k/m},$$

which may also look familiar.... More on this later. Note, however, that we can very easily deduce what is going on in this situation, including the type of force being applied, just from knowing $x(t)$. That's why we think of it as "the solution."

12.1.2 Canonical Variables

Let's write the equation of motion in a *generalized* form,

$$\ddot{q} \; = \; \frac{1}{m} \, F \tag{7}$$

where I have used "q" as the "canonical coordinate" whose second derivative ($\ddot{q}$) is the "canonical acceleration." Normally q will be the spatial position x [measured in units of length like metres or feet], but you have already seen one case (rotational kinematics) in which "q" is the *angle* θ [measured in radians], "m" is the *moment of inertia* I_O and "F" is the *torque* Γ_O; then a completely analogous set of equations pertains. This turns out to be a quite common situation. Can we describe simply how to go about formulating the equations of motion for "systems" that might even be completely different from the standard objects of Classical Mechanics?

In general there can be any number of canonical coordinates q_i in a given "system" whose behaviour we want to describe. As long as we have an explicit formula for the *potential energy* V in terms of one or more q_i, we can define the *generalized force*

$$Q_i = -\frac{\partial V}{\partial q_i} \tag{8}$$

If we then generalize the "inertial coefficient" $m \to \mu$, we can write out i^{th} equation of motion in the form

$$\ddot{q}_i \; = \; \frac{Q_i}{\mu} \tag{9}$$

which in most cases will produce a valid and workable solution. There is an even more general and elegant formulation of the canonical equations of motion which we will discuss toward the end of this chapter.

I am not really sure how the term *canonical* came to be fashionable for referring to this abstraction/generalization, but Physicists are all so fond of it by now that you are apt to hear them using it in all their conversations to mean something like *archetypal*: "It was the canonical Government coverup..." or "This is a canonical cocktail party conversation...."

12.1.3 Differential Equations

What we are doing when we "solve the equation of motion" is looking for a "solution" [in the sense defined above] to the *differential equation* defined by Eq. (7). You may have heard horror stories about the difficulty of "solving differential equations," but it's really no big deal; like long division, basically you can only use a trial-and-error method: does this function have the right derivative? No? How about this one?

And so on. Obviously, you can quickly learn to recognize certain functions by their derivatives; more complicated ones are harder, and it doesn't take much to stump even a seasoned veteran. The point of all this is that "solving differential equations" is a difficult and arcane art only if you want to be able to solve *any* differential equation; solving the few *simple* ones that occur over and over in physics is no more tedious than remembering multiplication tables. Some of the other commonly-occuring examples have already been mentioned.

12.1.4 Exponential Functions

You have seen the procedure by which a new function, the *exponential* function $q(t) = q_0 \exp(kt)$, was constructed from a power series just to provide a solution to the differential equation $\dot{q} = k\,q$. (There are, of course, other ways of "inventing" this delightful function, but I like my story.) You may suspect that this sort of procedure will take place again and again, as we seek compact notation for the functions that "solve" other important differential equations. Indeed it does! We have Legendre polynomials, various Bessel functions, spherical harmonics and many other "named functions" for just this purpose. But — pleasant surprise! — we can get by with *just the ones we have so far* for almost all of Newtonian Mechanics, provided we allow just one more little "extension" of the *exponential* function....

Frequency = Imaginary Rate?

Suppose we have

$$q(t) = q_0 e^{\lambda t}.$$

It is easy to take the n^{th} time derivative of this function — we just "pull out a factor λ" n times. For $n = 2$ we get $\ddot{q} = \lambda^2 q_0 e^{\lambda t}$ or just

$$\ddot{q} = \lambda^2 q. \tag{10}$$

Now go back to the example "solution" in Eq. (5), which turned out to be equivalent to HOOKES'S LAW [Eq. (6)]: $\ddot{x} = -\omega^2\, x$, where $\omega = \sqrt{k/m}$ and k and m are the "spring constant" and the mass, respectively.

Equations (10) and (6) would be the *same equation* if only we could let $q \equiv x$ and $\lambda^2 = -\omega^2$. Unfortunately, there is no real number whose square is negative. Too bad. It would be *awfully nice* if we could just re-use that familiar exponential function to solve mass-on-a-spring problems too.... If we just use a little *imagination*, maybe we *can* find a λ whose square is negative. This would require having a number whose square is -1, which takes so much imagination that we might as well call it i. If there were such a number, then we could just write

$$\lambda = i\omega. \tag{11}$$

That is, the *rate* λ in the *exponential* formula would have to be an "*imaginary*" version of the *frequency* ω in the *oscillatory* version, which would mean (if the solution is to be *unique*) that

$$e^{i\omega t} = \cos\omega t.$$

It's not.

Oh well, maybe later....

12.2 Mind Your p's and q's!

Earlier we introduced the notion of *canonical coordinates* q_i and the *generalized forces* Q_i defined by the partial derivatives of the *potential energy* V with respect to q_i. I promised then that I would describe a more general prescription later. Well, here it comes!

If we may assume that *both* the *potential energy* $V(q_i, \dot{q}_i)$ and the *kinetic energy* $T(q_i, \dot{q}_i)$ are known as explicit functions of the canonical coordinates q_i and the associated "canonical velocities" $\dot{q}_i$, then it is useful to define the **Lagrangian** function

$$\mathcal{L}(q_i, \dot{q}_i) \;\equiv\; T(q_i, \dot{q}_i) \;- V(q_i, \dot{q}_i) \tag{12}$$

in terms of which we can then define the *canonical momenta*

$$p_i \equiv \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \tag{13}$$

These canonical momenta are then guaranteed to "act like" the conventional momentum $m\dot{x}$ in all respects, though they may *be* something entirely different.

How do we obtain the equations of motion in this new "all-canonical" formulation? Well, HAMILTON'S PRINCIPLE declares that the motion of the system will follow the *path* $q_i(t)$ for which the "path integral" of $\mathcal{L}$ from initial time t_1 to final time t_2,

$$\mathcal{I} = \int_{t_1}^{t_2} \mathcal{L}\, dt \tag{14}$$

is an *extremum* [either a maximum or a minimum]. There is a very powerful branch of mathematics called the *calculus of variations* that allows this principle to be used[5] to derive the LAGRANGIAN EQUATIONS OF MOTION,

$$\dot{p}_i = \frac{\partial \mathcal{L}}{\partial q_i} \tag{15}$$

Because the "q" and "p" notation is always used in advanced Classical Mechanics courses to introduce the ideas of canonical equations of motion, almost every Physicist attaches special meaning to the phrase, "Mind your p's and q's." Now you know this bit of jargon and can impress Physicist friends at cocktail parties. More importantly, you have an explicit prescription for determining the equations of motion of *any* system for which you are able to formulate analogues of the potential energy V and the kinetic energy T.

There is one last twist to this canonical business that bears upon greater things to come. That is the procedure by which the description is recast in a form which depends explicitly upon q_i and p_i, rather than upon q_i and $\dot{q}_i$. It turns out that if we define the **Hamiltonian** function

$$\mathcal{H}(q_i, p_i) \equiv \sum_i \dot{q}_i\, p_i - \mathcal{L}(q_i, \dot{q}_i) \tag{16}$$

then it is usually true that

$$\mathcal{H} = T + V \tag{17}$$

– that is, the Hamiltonian is equal to the *total energy* of the system! In this case the equations of motion take the form

$$\dot{q}_i = \frac{\partial \mathcal{H}}{\partial p_i} \qquad \text{and} \qquad \dot{p}_i = -\frac{\partial \mathcal{H}}{\partial q_i} \tag{18}$$

So what? Well, we aren't going to crank out any examples, but the Lagrangian and/or Hamiltonian formulations of Classical Mechanics are very elegant (and convenient!) generalizations that let us generate equations of motion for problems in which they are by no means self-evident. This is especially useful in solving complicated problems involving the rotation of rigid bodies or other problems where the motion is partially *constrained* by some mechanism [usually an actual machine of some sort]. It should also be useful to *you*, should you ever decide to apply the paradigms of Classical Mechanics to some "totally inappropriate" phenomenon like economics or psychology. First, however, you must invent analogues of *kinetic energy* V and *potential energy* T and give formulae for how they depend upon your canonical coordinates and velocities or momenta.

Note the dramatic paradigm shift from the *force* and *mass* of Newton's SECOND LAW to a complete derivation in terms of *energy* in "modern" Classical Mechanics. It turns out that this shift transfers smoothly into the not-so-classical realm of QUANTUM MECHANICS, where the HAMILTONIAN $\mathcal{H}$ takes on a whole new meaning.

[5] Relax, we aren't going to do it here.

Chapter 13

Simple Harmonic Motion

In the previous chapter we found several new classes of equations of motion. We now add one last paradigm to our repertoire — one so powerful and ubiquitous in Physics that it deserves a chapter all to itself.

13.1 Periodic Behaviour

Nature shows us many "systems" which return periodically to the same initial state, passing through the same sequence of intermediate states every period. Life is so full of periodic experiences, from night and day to the rise and fall of the tides to the phases of the moon to the annual cycle of the seasons, that we all come well equipped with "common sense" tailored to this paradigm.[1] It has even been suggested that the concept of *time* itself is rooted in the *cyclic* phenomena of Nature.

In Physics, of course, we insist on narrowing the definition just enough to allow precision. For instance, many phenomena are *cyclic* without being *periodic* in the strict sense of the word.[2] Here *cyclic* means that the same general pattern keeps repeating; *periodic* means that the system passes through the same "phase" at *exactly* the same time in every cycle and that all the cycles are *exactly* the same length. Thus if we know all the details of *one full cycle* of true periodic behaviour, then we know the subsequent state of the system at *all* times, future and past. Naturally, this is an idealization; but its utility is obvious.

Of course, there is an infinite variety of possible *periodic* cycles. Assuming that we can reduce the "state" of the system to a single variable "q" and its time derivatives, the graph of $q(t)$ can have any shape as long as it *repeats* after one full period. Fig. 13.1 illustrates a few examples. In (a) and (b) the "displacement" of q away from its "equilibrium" position [dashed line] is not symmetric, yet the phases repeat

[1] Many people are so taken with this paradigm that they apply it to all experience. The *I Ching*, for instance, is said to be based on the ancient equivalent of "tuning in" to the "vibrations" of Life and the World so that one's awareness resonates with the universe. By New Age reckoning, cultivating such resonances is supposed to be the fast track to enlightenment. Actually, Physics relies very heavily on the same paradigm and in fact supports the notion that many apparently random phenomena are actually superpositions of regular cycles; however, it offers little encouragement for expecting "answers" to emerge effortlessly from such a tuning of one's mind's resonances. Too bad. But I'm getting ahead of myself here.

[2] Examples of *cyclic* but not necessarily *periodic* phenomena are the mass extinctions of species on Earth that seem to have occurred roughly every 24 million years, the "seven-year cycle" of sunspot activity, the return of salmon to the river of their origin and recurring droughts in Africa. In some cases the basic reason for the cycle is understood and it is obvious why it only repeats approximately; in other cases we have no idea of the root cause; and in still others there is not even a consensus that the phenomenon is truly cyclic — as opposed to just a random fluctuation that just happens to mimic cyclic behaviour over a short time. Obviously the resolution of these uncertainties demands "more data," *i.e.* watching to see if the cycle continues; with the mass extinction "cycle," this requires considerable patience. When "periodicity debates" rage on in the absence of additional data, it is usually a sign that the combatants have some other axe to grind.

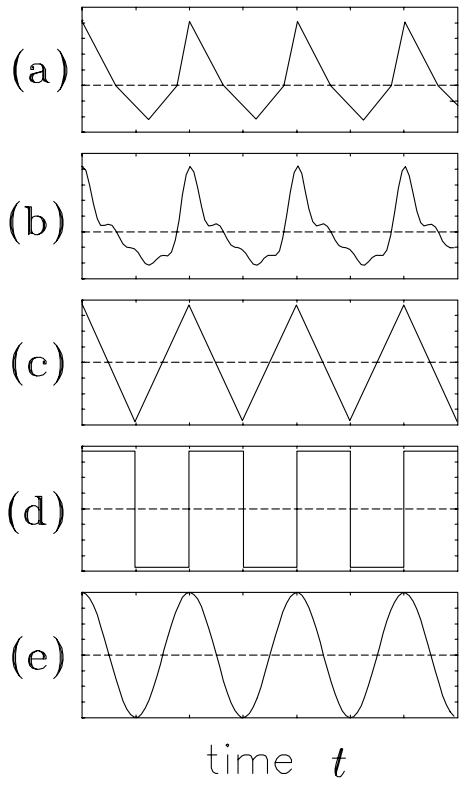


Figure 13.1 Some periodic functions.

every cycle. In (c) and (d) the cycle is symmetric with the same "amplitude" above and below the equilibrium axis, but at certain points the slope of the curve changes "discontinuously." Only in (e) is the cycle everywhere smooth and symmetric.

13.2 Sinusoidal Motion

There is one sort of periodic behaviour that is mathematically the *simplest possible* kind. This is the "sinusoidal" motion shown in Fig. 13.1(e), so called because one realization is the sine function, $\sin(x)$. It is easiest to see this by means of a crude mechanical example.

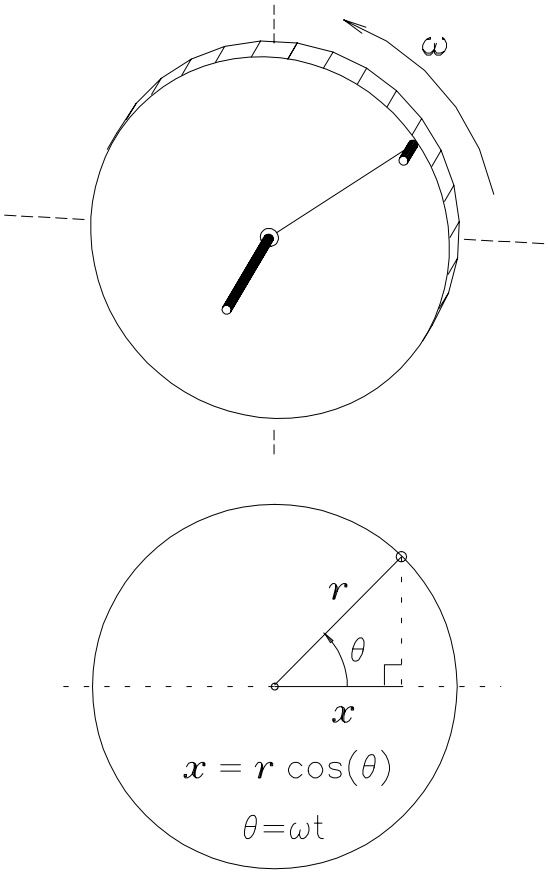


Figure 13.2 Projected motion of a point on the rim of a wheel.

13.2.1 Projecting the Wheel

Imagine a rigid wheel rotating at constant angular velocity about a fixed central axle. A bolt screwed into the rim of the wheel executes *uniform circular motion* about the centre of the axle.[3] For reference we scribe a line on the wheel from the centre straight out to the bolt and call this line the *radius vector*. Imagine now that we take this apparatus outside at high noon and watch the motion of the *shadow* of the bolt on the ground. This is (naturally enough) called the *projection* of the circular motion onto the horizontal axis. At some instant the radius vector makes an angle $\theta = \omega t + \phi$ with the hor-

[3] Note the frequency with which we periodically recycle our paradigms!

izontal, where ω is the angular frequency of the wheel (2π times the number of full revolutions per unit time) and ϕ is the initial angle (at $t = 0$) between the radius vector and the horizontal.[4] From a side view of the wheel we can see that the distance x from the shadow of the central axle to the shadow of the bolt [*i.e.* the *projected horizontal displacement* of the bolt from the centre, where $x = 0$] will be given by trigonometry on the indicated right-angle triangle:

$$\cos(\theta) \equiv \frac{x}{r}$$

$$\Rightarrow \quad x = r\,\cos(\theta) = r\,\cos(\omega t + \phi) \quad (1)$$

The resultant *amplitude* of the displacement as a *function of time* is shown in Fig. 13.3.

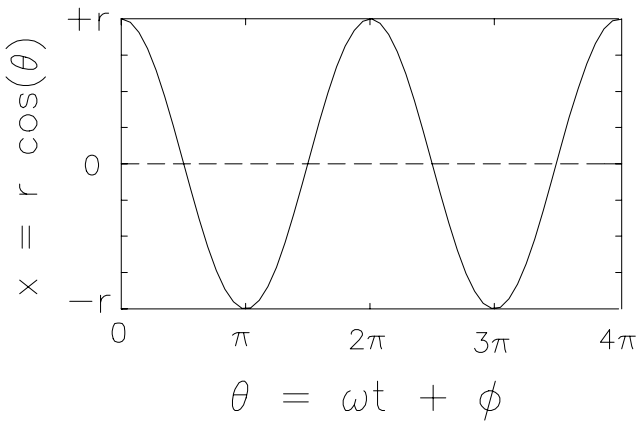


Figure 13.3 The cosine function.

The *horizontal velocity* v_x of the projected shadow of the bolt on the ground can also be obtained by trigonometry if we recall that the vector velocity $\vec{\boldsymbol{v}}$ is always perpendicular to the radius vector $\vec{\boldsymbol{r}}$. I will leave it as an exercise for the reader to show that the result is

$$v_x = -v\,\sin(\theta) = -r\,\omega\,\sin(\omega t + \phi) \quad (2)$$

where $v = r\omega$ is the constant speed of the bolt in its circular motion around the axle. It also follows (by the same sorts of arguments) that the *horizontal acceleration* a_x of the bolt's shadow is the projection onto the $\hat{\boldsymbol{x}}$ direction of $\vec{\boldsymbol{a}}$, which we know is back toward the centre of the wheel — *i.e.* in the $-\hat{\boldsymbol{x}}$ direction; its value at time t is given by

$$a_x = -a\,\cos(\theta) = -r\,\omega^2\,\cos(\omega t + \phi) \quad (3)$$

where $a = \dfrac{v^2}{r} = r\omega^2$ is the magnitude of the centripetal acceleration of the bolt.

13.3 Simple Harmonic Motion

The above mechanical example serves to introduce the idea of $\cos(\theta)$ and $\sin(\theta)$ as *functions* in the sense to which we have (I hope) now become accustomed. In particular, if we realize that (by definition) $v_x \equiv \dot{x}$ and $a_x \equiv \ddot{x}$, the formulae for $v_x(t)$ and $a_x(t)$ represent the *derivatives* of $x(t)$:

$$x = r\,\cos(\omega t + \phi) \quad (4)$$

$$\dot{x} = -r\,\omega\,\sin(\omega t + \phi) \quad (5)$$

$$\ddot{x} = -r\,\omega^2\,\cos(\omega t + \phi) \quad (6)$$

— which in turn tell us the derivatives of the sine and cosine functions:

$$\frac{d}{dt}\cos(\omega t + \phi) = -\omega\,\sin(\omega t + \phi) \quad (7)$$

$$\frac{d}{dt}\sin(\omega t + \phi) = \omega\,\cos(\omega t + \phi) \quad (8)$$

So if we want we can calculate the n^{th} derivative of a sine or cosine function almost as easily as we did for our "old" friend the exponential function. I will not go through the details this time, but this feature again allows us to express these functions as *series expansions*:

$$\begin{aligned}
\exp(z) &= 1 + z + \tfrac{1}{2}z^2 + \tfrac{1}{3!}z^3 + \tfrac{1}{4!}z^4 + \cdots \\
\cos(z) &= 1 - \tfrac{1}{2}z^2 + \tfrac{1}{4!}z^4 - \cdots \\
\sin(z) &= z - \tfrac{1}{3!}z^3 + \cdots
\end{aligned} \quad (9)$$

[4] The inclusion of the "initial phase" ϕ makes this description completely general.

where I have shown the exponential function along with the sine and cosine for reasons that will soon be apparent.

It is definitely worth remembering the SMALL ANGLE APPROXIMATIONS

$$\text{For} \quad \theta \ll 1, \quad \cos(\theta) \approx 1 - \tfrac{1}{2}\theta^2 \quad \text{and} \quad \sin(\theta) \approx \theta. \tag{10}$$

13.3.1 The Spring Pendulum

Another mecahnical example will serve to establish the paradigm of SIMPLE HARMONIC MOTION (*SHM*) as a solution to a particular type of *equation of motion*.[5]

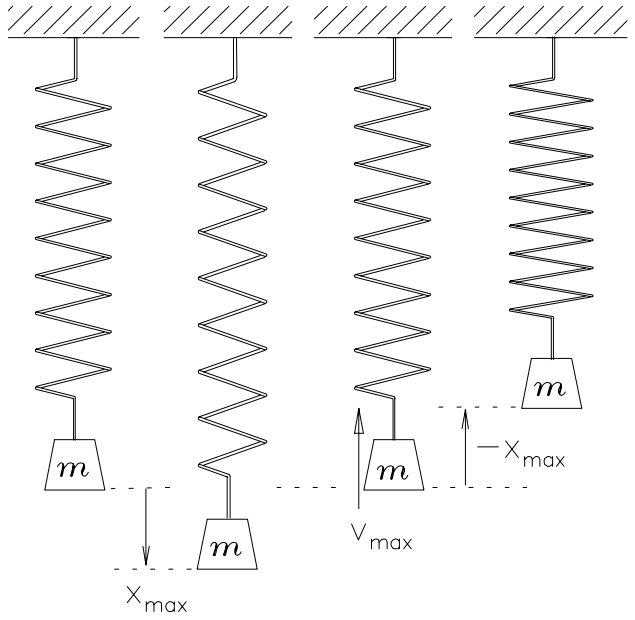


Figure 13.4 Successive "snapshots" of a mass bouncing up and down on a spring.

[5]Although we have become conditioned to expect such *mathematical* formulations of relationships to be more removed from our intuitive understanding than easily visualized *concrete* examples like the projection of circular motion, this is a case where the mathematics allows us to draw a sweeping conclusion about the detailed behaviour of *any* system that exhibits certain simple properties. Furthermore, these conditions *are actually satisfied* by an incredible variety of *real* systems, from the atoms that make up any solid object to the interpersonal "distance" in an intimate relationship. Just wait!

As discussed in a previous chapter, the *spring* embodies one of Physics' premiere paradigms, the *linear restoring force*. That is, a force which disappears when the system in question is in its "equilibrium position" x_0 [which we will define as the $x = 0$ position $(x_0 \equiv 0)$ to make the calculations easier] but increases as x moves away from equilibrium, in such a way that the *magnitude* of the force F is proportional to the displacement from equilibrium [F is *linear* in x] and the *direction* of F is such as to try to *restore* x to the original position. The constant of proportionality is called the *spring constant*, always written k. Thus $F = -kx$ and the resultant equation of motion is

$$\ddot{x} = -\left(\frac{k}{m}\right) x \tag{11}$$

Note that the *mass* plays a rôle just as essential as the *linear restoring force* in this paradigm. If $m \to 0$ in this equation, then the acceleration becomes infinite and in principle the spring would just return instantaneously to its equilibrium length and stay there!

In the leftmost frame of Fig. 13.4 the mass m is at rest and the spring is in its equilibrium position (*i.e.* neither stretched nor compressed) defined as $x = 0$. In the second frame, the spring has been gradually pulled down a distance $x_{\max}$ and the mass is once again at rest. Then the mass is released and accelerates upward under the influence of the spring until it reaches the equilibrium position again [third frame]. This time, however, it is moving at its maximum velocity $v_{\max}$ as it crosses the centre position; as soon as it goes higher, it *compresses* the spring and begins to be *decelerated* by a linear restoring force in the opposite direction. Eventually, when $x = -x_{\max}$, all the kinetic energy has been been stored back up in the compression of the spring and the mass is once again instantaneously at rest [fourth frame]. It immediately starts moving downward again at maximum acceleration and heads back toward its starting

point. In the absence of friction, this cycle will repeat forever.

I now want to call your attention to the acute similarity between the above differential equation and the one we solved for exponential decay:

$$\dot{x} = -\kappa\, x \tag{12}$$

and, by extension,

$$\ddot{x} = \kappa^2\, x \tag{13}$$

the solution to which equation of motion (*i.e.* the function which "satisfies" the differential equation) was

$$x(t) = x_0\, e^{-\kappa\, t} \tag{14}$$

Now, if only we could equate κ^2 with $-k/m$, these equations of motion (and therefore their solutions) would be exactly the same! The problem is, of course, that both k and m are intrinsically positive constants, so it is tough to find a real constant κ which gives a negative number when squared.

Imaginary Exponents

Mathematics, of course, provides a simple solution to this problem: just have κ be an *imaginary* number, say

$$\kappa \;\equiv\; i\,\omega \qquad \text{where} \qquad i \;\equiv\; \sqrt{-1}$$

and ω is a positive real constant. Let's see if this trial solution "works" (*i.e.* take its second derivative and see if we get back our equation of motion):

$$\begin{aligned} x(t) &= x_0\, e^{i\,\omega\, t} && (15)\\ \dot{x} &= i\,\omega\, x_0\, e^{i\,\omega\, t} && (16)\\ \ddot{x} &= -\omega^2\, x_0\, e^{i\,\omega\, t} && (17)\\ \text{or} \qquad \ddot{x} &= -\omega^2\, x && (18)\\ \text{so} \qquad \omega &\equiv \sqrt{\frac{k}{m}} && (19) \end{aligned}$$

OK, it works. But what does it *describe*? For this we go back to our series expansions for the exponential, sine and cosine functions and note that *if we let* $z \equiv i\theta$, the following *mathematical identity* holds:[6]

$$e^{i\,\theta} \;=\; \cos(\theta) \;+\; i\, \sin(\theta) \tag{20}$$

Thus, for the case at hand, if $\theta \equiv \omega\, t$ [you probably knew this was coming] then

$$x_0\, e^{i\,\omega\, t} \;=\; x_0\, \cos(\omega t) \;+\; i\, x_0\, \sin(\omega t)$$

— *i.e.* the formula for the projection of uniform circular motion, with an imaginary part "tacked on." (I have set the initial phase ϕ to zero just to keep things simple.) What does *this* mean?

I don't know.

What! How can I say, "I don't know," about the premiere paradigm of Mechanics? We're supposed to know *everything* about Mechanics! Let me put it this way: we have happened upon a nice tidy mathematical representation that *works* — *i.e.* if we use certain rules to manipulate the mathematics, it will faithfully give correct answers to our questions about how this thing will behave. The rules, by the way, are as follows:

> Keep the imaginary components through all your calculations until the final "answer," and then *throw away* any remaining *imaginary parts* of any actual *measurable quantity*.

The point is, there is a difference between understanding how something *works* and knowing what it *means*. Meaning is something we put

[6] You may find this unremarkable, but I have never gotten over my astonishment that functions so ostensibly unrelated as the *exponential* and the *sinusoidal* functions could be so intimately connected! And for once the mathematical oddity has profound *practical applications*.

into our world by act of will, though not always conscious will. How it works is there before us and after we are gone. No one asks the "meaning" of a screwdriver or a carburetor or a copy machine; some of the conceptual tools of Physics are in this class, though of course there is nothing to prevent anyone from *putting* meaning into them.[7]

13.4 Damped Harmonic Motion

Let's take stock. In the previous chapter we found that

$$x(t) \;=\; [\text{constant}] \;-\; \frac{v_0}{\kappa}\, e^{-\kappa t}$$

satisfies the basic differential equation

$$\ddot{x} = -\kappa \dot{x} \qquad \text{or} \qquad a = -\kappa v$$

defining *damped* motion (*e.g.* motion under the influence of a frictional force proportional to the velocity). We now have a solution to the equation of motion defining *SHM*,

$$\ddot{x} = -\omega^2\, x \quad \Rightarrow \quad x(t) = x_0\, e^{i\,\omega\, t},$$

where

$$\omega = \sqrt{\frac{k}{m}}$$

and I have set the initial phase ϕ to zero just to keep things simple. Can we put these together to "solve" the more general (and realistic) problem of *damped harmonic motion*? The differential equation would then read

$$\ddot{x} \;=\; -\omega^2\, x \;-\kappa\, \dot{x} \tag{21}$$

which is beginning to look a little hard. Still, we can sort it out: the first term on the *RHS* says that there is a linear restoring force and an inertial factor. The second term says that there is a damping force proportional to the velocity. So the differential equation itself is not all that fearsome. How can we "solve" it?

As always, by trial and error. Since we have found the *exponential* function to be so useful, let's try one here: *Suppose* that

$$x(t) = x_0\, e^{K\, t} \tag{22}$$

where x_0 and K are unspecified constants. Now plug this into the differential equation and see what we get:

$$\dot{x} \;=\; K\, x_0\, e^{K\, t} \;=\; K\, x$$

and

$$\ddot{x} \;=\; K^2\, x_0\, e^{K\, t} \;=\; K^2\, x$$

The whole thing then reads

$$K^2\, x \;=\; -\omega^2\, x \;-\kappa\, K\, x$$

which can be true "for all x" only if

$$K^2 = -\omega^2 \,-\kappa\, K \qquad \text{or} \qquad K^2 + \kappa K + \omega^2 = 0$$

which is in the standard form of a general quadratic equation for K, to which there are two solutions:

$$K = \frac{-\kappa \pm \sqrt{\kappa^2 - 4\omega^2}}{2} \tag{23}$$

Either of the two solutions given by substituting Eq. (23) into Eq. (22) will satisfy Eq. (21) describing *damped SHM*. In fact, generally any *linear combination* of the two solutions will also be a solution. This can get complicated, but we *have* found the answer to a rather broad question.

13.4.1 Limiting Cases

Let's consider a couple of "limiting cases" of such solutions. First, suppose that the linear restoring force is extremely weak compared to

[7] I am reminded of a passage in one of Kurt Vonnegut's books, perhaps *Sirens of Titan*, in which the story of creation is told something like this: God creates the world; then he creates Man, who sits up, looks around and says, "What's the meaning of all this?" God answers, "What, there has to be a meaning?" Man: "Of course." God: "Well then, I leave it to you to think of one."

the "drag" force — *i.e.*[8] $\kappa \gg \omega = \sqrt{\frac{k}{m}}$. Then $\sqrt{\kappa^2 - 4\omega^2} \approx \kappa$ and the solutions are $K \approx 0$ [*i.e.* $x \approx$ constant, plausible only if $x = 0$] and $K \approx -\kappa$, which gives the same sort of damped behaviour as if there were no restoring force, which is what we expected.

Now consider the case where the linear restoring force is very strong and the "drag" force extremely weak — *i.e.* $\kappa \ll \omega = \sqrt{\frac{k}{m}}$. Then $\sqrt{\kappa^2 - 4\omega^2} \approx 2\,i\,\omega$ and the solutions are $K \approx -\frac{1}{2}\kappa \pm i\,\omega$, or[9]

$$x(t) = x_0\, e^K \tag{24}$$
$$\approx x_0\, \exp(\pm i\,\omega t\ - \gamma t) \tag{25}$$
$$= x_0\, e^{\pm i\,\omega t} \cdot e^{-\gamma t} \tag{26}$$

where $\gamma \equiv \frac{1}{2}\kappa$. We may then think of $[i\,K]$ as a *complex frequency*[10] whose real part is $\pm\omega$ and whose imaginary part is γ. What sort of situation does this describe? It describes a *weakly damped harmonic motion* in which the usual sinusoidal pattern damps away within an "envelope" whose shape is that of an exponential decay. A typical case is shown in Fig. 13.5.

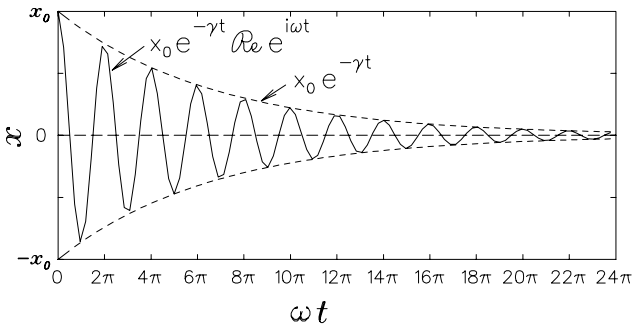


Figure 13.5 Weakly damped harmonic motion. The initial amplitude of x (whatever x is) is x_0, the angular frequency is ω and the damping rate is γ. The cosine-like oscillation, equivalent to the *real* part of $x_0\, e^{i\,\omega\,t}$, decays within the *envelope function* $x_0\, e^{-\gamma\,t}$.

13.5 Generalization of $\mathcal{SHM}$

As for all the other types of *equations of motion*, *SHM* need not have anything to do with masses, springs or even Physics. Even within Physics, however, there are so many different kinds of examples of *SHM* that we go out of our way to generalize the results: using "q" to represent the "coordinate" whose displacement from the equilibrium "position" (always taken as $q = 0$) engenders some sort of restoring "force" $Q = -k\,q$ and "μ" to represent an "inertial factor" that plays the rôle of the mass, we have

$$\ddot{q} = -\left(\frac{k}{\mu}\right) q \tag{27}$$

for which the solution is the real part of

$$q(t) = q_0\, e^{i\,\omega\,t} \qquad \text{where} \qquad \omega = \sqrt{\frac{k}{\mu}} \tag{28}$$

When some form of "drag" acts on the system, we expect to see the qualitative behaviour pictured in Fig. 13.5 and described by Eqs. (22) and (23). Although one might expect virtually every real example to have some sort of frictional damping term, in fact there are numerous physical examples with no damping what-

[8] The "$\gg$" symbol means "... is *much* greater than..." — there is an analogous "$\ll$" symbol that means "... is much *less* than...."

[9] There is a general rule about *exponents* that says, "A number raised to the sum of two powers is equal to the product of the same number raised to each power separately," or

$$a^{b+c} = a^b \cdot a^c.$$

[10] The word "complex" has, like "real" and "imaginary," been ripped off by Mathematicians and given a very explicit meaning that is not entirely compatible with its ordinary dictionary definition. While a *complex* number in Mathematics may indeed be complex — *i.e.* complicated and difficult to understand — it is *defined* only by virtue of its having both a *real* part and an *imaginary* part, such as $z = a + i\,b$, where a and b are both *real*. I hope that makes everything crystal clear....

soever, mostly from the microscopic world of solids.

13.6 The Universality of $\mathcal{SHM}$

If two systems satisfy the same equation of motion, their behaviour is the same. Therefore the motion of the mass on the spring is *in every respect identical* to the horizontal component of the motion of the peg in the rotating wheel, even though these two systems are physically quite distinct. In fact, *any* system exhibiting both a LINEAR RESTORING "FORCE" and an INERTIAL FACTOR analogous to MASS will exhibit *SHM*.[11] Moreover, since these arguments may be used equally well in reverse, the horizontal component of the *force* acting on the peg in the wheel must obey $F_x = -kx$, where k is an "effective spring constant."

Why is *SHM* characteristic of such an enormous variety of phenomena? Because *for sufficiently small* displacements from equilibrium, *every* system with an equilibrium configuration satisfies the first condition for *SHM*: the linear restoring force. Here is the simple argument: a linear restoring force is equivalent to a potential energy of the form $U(q) = \frac{1}{2}k\,q^2$ — *i.e.* a "quadratic minimum" of the potential energy at the equilibrium configuration $q = 0$. But if we "blow up" a graph of $U(q)$ near $q = 0$, *every* minimum looks quadratic under sufficient magnification! That means *any* system that *has* an equilibrium configuration also has some analogue of a "potential energy" which is a minimum there; if it also has some form of *inertia* so that it tends to stay at rest (or in motion) until acted upon by the analogue of a *force*, then it will automatically exhibit *SHM* for small-amplitude displacements. This makes *SHM* an extremely powerful paradigm.

13.6.1 Equivalent Paradigms

We have established previously that a LINEAR RESTORING FORCE $F = -kx$ is completely equivalent to a QUADRATIC MINIMUM IN POTENTIAL ENERGY $U = \frac{1}{2}kx^2$. We now find that, with the inclusion of an INERTIAL FACTOR (usually just the MASS m), either of these *physical paradigms* will guarantee the *mathematical paradigm* of *SHM* — *i.e.* the displacement x from equilibrium will satisfy the equation of motion

$$x(t) \;=\; x_{\max}\,\cos(\omega t + \phi) \tag{29}$$

where $x_{\max}$ is the *amplitude* of the oscillation. Any $x(t)$ of this form automatically satisfies the definitive equation of motion of *SHM*, namely

$$\frac{d^2x}{dt^2} \;=\; -\omega^2\,x \tag{30}$$

and *vice versa* — whenever x satisfies Eq. (30), the explicit time dependence of x will be given by Eq. (29).

13.7 Resonance

No description of *SHM* would be complete without some discussion of the general phenomenon of *resonance*, which has many practical consequences that often seem very counterintuitive.[12] I will, however, overcome my zeal for demonstrating the versatility of Mathematics and stick to a simple qualitative description of resonance. Just this once.

The basic idea is like this: suppose some system exhibits all the requisite properties for *SHM*,

[11] Examples are plentiful, especially in view of the fact that *any* potential energy minimum is approximately quadratic for small enough displacements from equilibrium. A prime example from outside Mechanics is the *electrical circuit*, in which the charge Q on a capacitor plays the rôle of the displacement variable x and the inertial factor is provided by an inductance, which resists changes in the current $I = dQ/dt$.

[12] It is, after all, one of the main purposes of this book to dismantle your intuition and rebuild it with the faulty parts left out and some shiny new paradigms added.

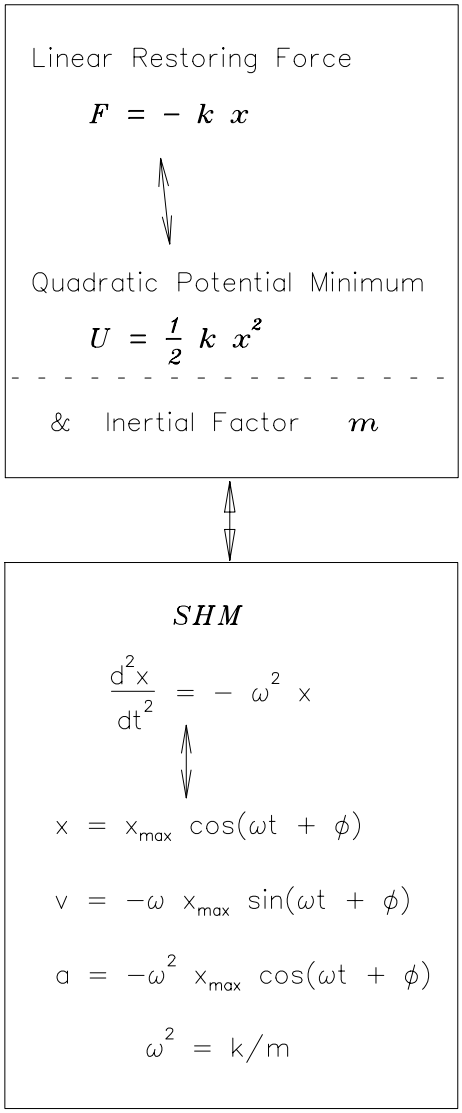


Figure 13.6 Equivalent paradigms of *SHM*.

namely a linear restoring "force" $Q = -k\,q$ and an inertial factor μ. Then *if once set in motion* it will oscillate forever at its "resonant frequency" $\omega = \sqrt{\frac{k}{\mu}}$, unless of course there is a "damping force" $D = -\kappa\mu q$ to dissipate the energy stored in the oscillation. As long as the damping is weak $[\kappa \ll \sqrt{\frac{k}{m}}]$, any oscillations will persist for many periods. Now suppose the system is initially at rest, in equilibrium, ho hum. What does it take to "get it going?"

The *hard* way is to give it a great whack to start it off with lots of kinetic energy, or a great tug to stretch the "spring" out until it has lots of potential energy, and then let nature take its course. The *easy* way is to give a tiny push to start up a small oscillation, then wait exactly one full period and give another tiny push to increase the amplitude a little, and so on. This works because *the frequency ω is independent of the amplitude q_0*. So if we "drive" the system *at its natural "resonant" frequency* ω, no matter how small the individual "pushes" are, we will slowly build up an *arbitrarily large oscillation*.[13]

Such resonances often have dramatic results. A vivid example is the famous movie of the collapse of the Tacoma Narrows bridge, which had a torsional [twisting] resonance[14] that was excited by a steady breeze blowing past the bridge. The engineer in charge anticipated all the other more familiar resonances [of which there are many] and incorporated devices specifically designed to safely damp their oscillations, but forgot this one. As a result, the bridge developed huge twisting oscillations [mistakes like this are usually painfully obvious when it is too late to correct them] and tore itself apart.

A less spectacular example is the trick of getting yourself going on a playground swing by leaning back and forth with arms and legs in synchrony with the natural frequency of oscillation of the swing [a sort of pendulum]. If your kinesthetic memory is good enough you may recall that it is important to have the "driving" push exactly $\frac{\pi}{2}$ radians [a quarter cycle] "out of phase" with your velocity — *i.e.* you *pull* when you reach the *motionless* position at the top of

[13] Of course, this assumes $\kappa = 0$. If damping occurs at the same time, we must put at least as much energy *in* with our driving force as friction takes *out* through the damping in order to build up the amplitude. Almost every system has some limiting amplitude beyond which the restoring force is no longer linear and/or some sort of losses set in.

[14] (something like you get from a blade of grass held between the thumbs to create a loud noise when you blow past it)

your swing, if you want to achieve the maximum result. This has an elegant mathematical explanation, but I promised....

Chapter 14

Waves

In a purely mathematical approach to the phenomenology of *waves*, we might choose to start with the WAVE EQUATION, a differential equation describing the qualitative features of wave propagation in the same way that *SHM* is characterized by $\ddot{x} = -\omega^2 x$. The advantage of such an approach is that one gains confidence that any phenomenon that can be shown to obey the WAVE EQUATION will *automatically* exhibit *all* the characteristic properties of *wave motion*. This is a very economical way of looking at things.

Unfortunately, the phenomenology of wave motion is not very familiar to most beginners — at least not in the mathematical form we will need here; so in this instance I will adopt the approach used in most first year Physics textbooks for almost everything: I will *start with the answer* (the simplest *solution* to the WAVE EQUATION) and explore its *properties* before proceeding to show that it is indeed a solution of the WAVE EQUATION — or, for that matter, before explaining what the WAVE EQUATION *is*.

14.1 Wave Phenomena

We can visualize a vivid example for the sake of illustration: suppose the "amplitude" A is the height of the water's surface in the ocean (measured from $A = 0$ at "sea level") and x is the distance toward the East, in which direction waves are moving across the ocean's surface.[1] Now imagine that we stand on a skinny piling and watch what happens to the water level on its sides as the wave passes: it goes up and down at a regular frequency, executing *SHM* as a function of time. Next we stand at a big picture window in the port side of a submarine pointed East, partly submerged so that the wave is at the same level as the window; we take a flash photograph of the wave at a given instant and analyze the result: the wave looks instantaneously just like the graph

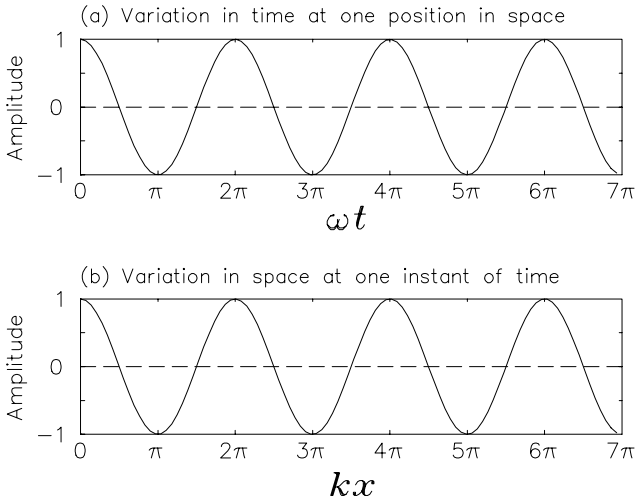


Figure 14.1 Two views of a wave.

[1] Technically speaking, I couldn't have picked a *worse* example, since water waves do *not* behave like our idealized example — a cork in the water does *not* move straight up and down as a wave passes, but rather in a vertical *circle*. Nevertheless I will use the example for illustration because it is the most familiar sort of easily visualized wave for most people and you have to watch closely to notice the difference anyway!

of *SHM* except the horizontal axis is *distance* instead of *time*. These two images are displayed in Fig. 14.1.

14.1.1 Traveling Waves

How do we represent this behaviour mathematically? Well, A is a function of position $\vec{r}$ and time t: $A(\vec{r},t)$. *At any fixed position* $\vec{r}$, A *oscillates in time* at a frequency ω. We can describe this statement mathematically by saying that the entire *time dependence* of A is contained in [the real part of] a factor $e^{-i\omega t}$ (that is, the amplitude *at any fixed position* obeys *SHM*).[2]

The oscillation with respect to *position* $\vec{r}$ *at any instant of time* t is given by the analogous factor $e^{i\vec{k}\cdot\vec{r}}$ where $\vec{k}$ is the *wave vector*;[3] it points in the *direction of propagation* of the wave and has a magnitude (called the "wavenumber") k given by

$$k \;=\; \frac{2\pi}{\lambda} \tag{1}$$

where λ is the *wavelength*. Note the analogy between k and

$$\omega \;=\; \frac{2\pi}{T} \tag{2}$$

where T is the *period* of the oscillation in time at a given point. You should think of λ as the "period in space."

We may simplify the above description by *choosing our coordinate system* so that the x axis is *in the direction of* $\vec{k}$, so that[4] $\vec{k}\cdot\vec{r} = k\,x$. Then the amplitude A no longer depends on y or z, only on x and t.

We are now ready to give a full description of the function describing this wave:

$$A(x,t) \;=\; A_0\; e^{ikx}\cdot e^{-i\omega t}$$

or, recalling the multiplicative property of the exponential function, $e^a\cdot e^b = e^{(a+b)}$,

$$A(x,t) \;=\; A_0\; e^{i(kx-\omega t)}. \tag{3}$$

To achieve complete generality we can restore the vector version:

$$\boxed{A(x,t) \;=\; A_0\; e^{i\left(\vec{k}\cdot\vec{r}-\omega t\right)}} \tag{4}$$

This is the preferred form for a general description of a PLANE WAVE, but for present purposes the scalar version (3) suffices. Using Eqs. (1) and (2) we can also write the plane wave function in the form

$$A(x,t) \;=\; A_0\;\exp\left[2\pi i\left(\frac{x}{\lambda}-\frac{t}{T}\right)\right] \tag{5}$$

but you should strive to become completely comfortable with k and ω — we will be seeing a lot of them in Physics!

14.1.2 Speed of Propagation

Neither of the images in Fig. 14.1 captures the most important qualitative feature of the wave: namely, that it *propagates* — *i.e.* moves steadily along in the direction of $\vec{k}$. If we were to let the *snapshot* in Fig. 14.1b become a *movie*, so that the time dependence could be seen vividly, what we would see would be the same wave pattern *sliding along the graph to the right* at a steady rate. *What* rate? Well, the answer is most easily given in simple qualitative terms:

The wave has a distance λ (one *wavelength*) between "crests." Every *period* T, one full

[2] Note that $e^{+i\omega t}$ would have worked just as well, since the real part is the same as for $e^{-i\omega t}$. The choice of sign does matter, however, when we write down the *combined* time and space dependence in Eq. (4), which see.

[3] The name "wave vector" is both apt and inadequate — apt because the term *vector* explicitly reminds us that its direction defines the direction of propagation of the wave; inadequate because the essential inverse relationship between k and the *wavelength* λ [see Eq. (1)] is not suggested by the name. Too bad. It is at least a little more descriptive than the name given to the *magnitude* k of $\vec{k}$, namely the "wavenumber."

[4] In general $\vec{k}\cdot\vec{r} = xk_x + yk_y + zk_z$. If $\vec{k} = k\,\hat{\imath}$ then $k_x = k$ and $k_y = k_z = 0$, giving $\vec{k}\cdot\vec{r} = k\,x$.

wavelength passes a fixed position. Therefore a given crest travels a distance λ in a time T so the *velocity of propagation* of the wave is just

$$c \;=\; \frac{\lambda}{T} \qquad \text{or} \qquad c \;=\; \frac{\omega}{k} \tag{6}$$

where I have used c as the symbol for the propagation velocity even though this is a completely *general* relationship between the frequency ω, the wave vector magnitude k and the propagation velocity c of *any* sort of wave, not just electromagnetic waves (for which c has its most familiar meaning, namely the speed of light).

This result can be obtained more easily by noting that A is a function *only* of the *phase* θ of the oscillation,

$$\theta \;\equiv\; kx - \omega t \tag{7}$$

and that the criterion for "seeing the same waveform" is θ = constant or $d\theta = 0$. If we take the differential of Eq. (7) and set it equal to zero, we get

$$d\theta \;=\; k\,dx \;-\; \omega\,dt \;=\; 0 \quad \text{or} \quad k\,dx \;=\; \omega\,dt$$

$$\text{or} \quad \frac{dx}{dt} = \frac{\omega}{k}.$$

But $dx/dt = c$, the propagation velocity of the waveform. Thus we reproduce Eq. (6). This treatment also shows why we chose $e^{-i\omega t}$ for the time dependence so that Eq. (7) would describe the phase: if we used $e^{+i\omega t}$ then the phase would be $\theta \;\equiv\; kx + \omega t$ which gives $dx/dt \;=\; -c$, — *i.e.* a waveform propagating in the negative x direction (to the *left* as drawn).

If we use the relationship (6) to write $(kx - \omega t) = k(x - ct)$, so that Eq. (4) becomes

$$A(x,t) \;=\; A_0\, e^{ik(x-ct)},$$

we can extend the above argument to waveforms that are not of the ideal sinusoidal shape shown in Fig. 14.1; in fact it is more vivid if one imagines some special shape like (for instance) a *pulse* propagating down a string at velocity c. As long as $A(x,t)$ *is a function only of* $x' = x - ct$, *no matter what its shape*, it will be *static in time* when viewed by an observer traveling along with the wave[5] at velocity c. This doesn't require any elaborate derivation; x' is just the position measured in such an observer's reference frame!

14.2 The Wave Equation

This is a bogus "derivation" in that we start with a *solution* to the Wave Equation and then show what sort of differential equation it satisfies. Of course, once we have the equation we can work in the other direction, so this is not *so* bad....

Suppose we know that we have a *traveling wave* $A(x,t) \;=\; A_0\,\cos(kx - \omega t)$.

At a *fixed position* (x = const) we see *SHM* in time:

$$\left(\frac{\partial^2 A}{\partial t^2}\right)_x \;=\; -\,\omega^2\,A \tag{8}$$

(Read: "The second partial derivative of A with respect to time [*i.e.* the *acceleration* of A] with x held fixed is equal to $-\omega^2$ times A itself.") *I.e.* we must have a *linear restoring force*.

Similarly, if we take a "snapshot" (hold t fixed) and look at the *spatial* variation of A, we find the oscillatory behaviour analogous to *SHM*,

$$\left(\frac{\partial^2 A}{\partial x^2}\right)_t \;=\; -\,k^2\,A \tag{9}$$

(Read: "The second partial derivative of A with respect to position [*i.e.* the *curvature* of A] with t held fixed is equal to $-k^2$ times A itself.")

Thus

$$A \;=\; -\frac{1}{\omega^2}\left(\frac{\partial^2 A}{\partial t^2}\right)_x \;=\; -\frac{1}{k^2}\left(\frac{\partial^2 A}{\partial x^2}\right)_t.$$

[5] Don't try this with an electromagnetic wave! The argument shown here is explicitly *nonrelativistic*, although a more mathematical proof reaches the same conclusion without such restrictions.

If we multiply both sides by $-k^2$, we get

$$\frac{k^2}{\omega^2}\left(\frac{\partial^2 A}{\partial t^2}\right)_x = \left(\frac{\partial^2 A}{\partial x^2}\right)_t .$$

But $\omega = ck$ so $\dfrac{k^2}{\omega^2} = \dfrac{1}{c^2}$, giving the WAVE EQUATION:

$$\boxed{\frac{\partial^2 A}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 A}{\partial t^2} = 0} \qquad (10)$$

In words, the *curvature* of A is equal to $1/c^2$ times the *acceleration* of A at any (x,t) point (what we call an *event* in *spacetime*).

Whenever you see this differential equation governing some quantity A, *i.e.* where the acceleration of A is proportional to its curvature, you know that $A(x,t)$ will exhibit wave motion!

14.3 Wavy Strings

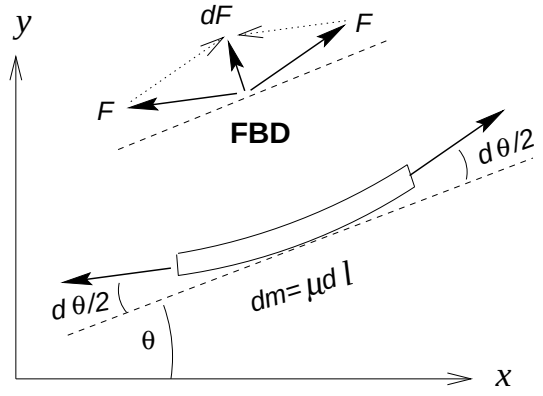


Figure 14.2 A small segment of a taut string.

One system that exhibits wave motion is the *taut string*. Picture a string with a uniform mass per unit length μ under tension F. Ignoring any effects of gravity, the undisturbed string will of course follow a straight line which we label the x axis. There are actually two ways we can "perturb" the quiescent string: with a "longitudinal" compression/stretch displacement (basically a sound wave in the string) or with a "transverse" displacement in a direction perpendicular to the x axis, which we will label the y direction.

The sketch in Fig. 14.2 shows a small string segment of length $d\ell$ and mass $dm = \mu\, d\ell$ which makes an average angle θ with respect to the x axis. The angle actually changes from $\theta - d\theta/2$ at the left end of the segment to $\theta + d\theta/2$ at the right end. For small displacements $\theta \ll 1$ [the large θ shown in the sketch is just for visual clarity] and we can use the SMALL-ANGLE APPROXIMATIONS

$$dx = d\ell\, \cos\theta \approx d\ell$$

$$dy = d\ell\, \sin\theta \approx \theta\, d\ell$$

$$\frac{dy}{dx} = \tan\theta \approx \theta \qquad (11)$$

Furthermore, for small θ the net force

$$dF = 2F\, \sin(d\theta/2) \approx 2F\,(d\theta/2) = F\, d\theta \quad (12)$$

acting on the string segment is essentially in the y direction, so we can use Newton's SECOND LAW on the segment at a fixed x location on the string:

$$dF \approx dm\, a_y = \ddot{y}\, dm \qquad \text{or}$$

$$F\, d\theta \approx \ddot{y}\, \mu\, d\ell \qquad \text{or}$$

$$\left(\frac{\partial^2 y}{\partial t^2}\right)_x \approx \frac{F\, d\theta}{\mu\, d\ell} \approx \frac{F}{\mu}\left(\frac{d\theta}{dx}\right). \qquad (13)$$

Referring now back to Eq. (11) we can use $\theta \approx dy/dx$ to set

$$\left(\frac{d\theta}{dx}\right) \approx \left(\frac{\partial^2 y}{dx^2}\right)_t \qquad (14)$$

— *i.e.* the *curvature* of the string at time t. Plugging Eq. (14) back into Eq. (13) gives

$$\left(\frac{\partial^2 y}{\partial t^2}\right)_x - \frac{F}{\mu}\left(\frac{\partial^2 y}{dx^2}\right)_t \approx 0 \qquad (15)$$

which is the WAVE EQUATION with

$$\frac{1}{c^2} = \frac{\mu}{F} \quad \text{or} \quad c = \sqrt{\frac{F}{\mu}}. \tag{16}$$

We may therefore jump right to the conclusion that *waves* will propagate down a taut string at this velocity.

14.3.1 Polarization

One nice feature of waves in a taut string is that they explicitly illustrate the phenomenon of *polarization*: if we change our notation slightly to label the string's equilibrium direction (and therefore the direction of propagation of a wave in the string) as z, then there are two orthogonal choices of "transverse" direction: x or y. We can set the string "wiggling" in either transverse direction, which we call the two orthogonal *polarization* directions.

Of course, one can choose an infinite number of transverse polarization directions, but these correspond to simple *superpositions* of x- and y-polarized waves with the same phase.

One can also superimpose x- and y-polarized waves of the same frequency and wavelength but with phases differing by $\pm\pi/2$. This gives left- and right-*circularly polarized* waves; I will leave the mathematical description of such waves (and the mulling over of its physical meaning) as an "exercise for the student. . . ."

14.4 Linear Superposition

The above derivation relied heavily on the SMALL-ANGLE APPROXIMATIONS which are valid only for *small displacements* of the string from its equilibrium position ($y = 0$ for all x). This almost always true: the simple description of a wave given here is only strictly valid in the limit of small displacements from equilibrium; for large displacements we usually pick up "anharmonic" terms corresponding to *nonlinear restoring forces*. But as long as the restoring force stays linear we have an important consequence: *several different waves can propagate independently through the same medium.* (*E.g.* down the same string.) The displacement at any given time and place is just the *linear sum* of the displacements due to each of the simultaneously propagating waves. This is known as the PRINCIPLE OF LINEAR SUPERPOSITION, and it is essential to our understanding of wave phenomena.

In general the overall displacement $A(x,t)$ resulting from the linear superposition of two waves $A_1 e^{i(k_1 x - \omega_1 t)}$ and $A_2 e^{i(k_2 x - \omega_2 t)}$ is given by

$$A(x,t) = A_1 e^{i(k_1 x - \omega_1 t)} + A_2 e^{i(k_2 x - \omega_2 t)}. \tag{17}$$

Let's look at a few simple examples.

14.4.1 Standing Waves

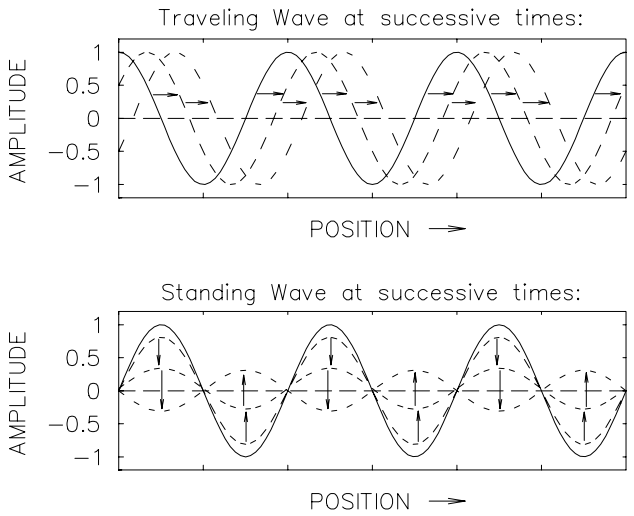


Figure 14.3 Traveling *vs.* standing waves.

A particularly interesting example of superposition is provided by the case where $A_1 = A_2 = A_0$, $k_1 = k_2 = k$ and $\omega_1 = -\omega_2 = \omega$. That is, two otherwise identical waves *propagating in*

opposite directions. The algebra is simple:

$$\begin{aligned} A(x,t) &= A_0\left[e^{i(kx-\omega t)} + e^{i(kx+\omega t)}\right] \\ &= A_0 e^{ikx}\left[e^{-i\omega t} + e^{+i\omega t}\right] \\ &= A_0 e^{ikx}[\cos(\omega t) - i\sin(\omega t) \\ &\qquad + \cos(\omega t) + i\sin(\omega t)] \\ &= 2A_0\cos(\omega t)e^{ikx}. \end{aligned} \tag{18}$$

The real part of this (which is all we ever actually use) describes a sinusoidal waveform of wavelength $\lambda = 2\pi/k$ whose *amplitude* $2A_0\cos(\omega t)$ oscillates in time but which does *not* propagate in the x direction — *i.e.* the lower half of Fig. 14.3. Standing waves are very common, especially in situations where a traveling wave is *reflected* from a boundary, since this automatically creates a second wave of similar amplitude and wavelength propagating back in the opposite direction — the very condition assumed at the beginning of this discussion.

14.4.2 Classical Quantization

None of the foregoing discussion allows us to *uniquely specify* any wavelike solution to the WAVE EQUATION, because nowhere have we given any BOUNDARY CONDITIONS forcing the wave to have any particular behaviour at any particular point. This is not a problem for the general phenomenology discussed so far, but if you want to actually describe one particular wave you have to know this stuff.

Boundary conditions are probably easiest to illustrate with the system of a taut string of length L with fixed ends, as shown in Fig. 14.4.[6] Fixing the ends forces the wave function $A(x,t)$ to have *nodes* (positions where the amplitude is always zero) at those

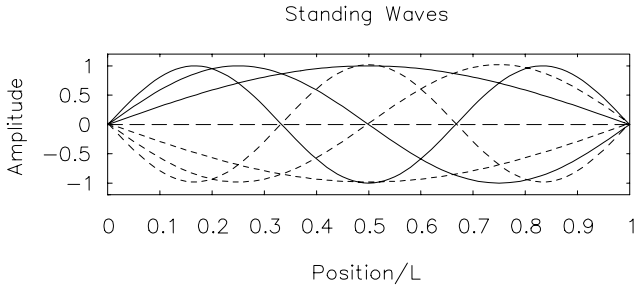


Figure 14.4 The first three allowed standing waves in a "closed box" (*e.g.* on a string with fixed ends).

positions. This immediately rules out traveling waves and restricts the simple sinusoidal "modes" to *standing waves* for which L is an integer number of half-wavelengths:[7]

$$\lambda_n = \frac{2L}{n}, \quad n = 1, 2, 3, \cdots \tag{19}$$

Assuming that $c = \omega/k = \lambda\nu = \text{const}$, the frequency ν [in *cycles per second* or *Hertz* (*Hz*)] of the n^{th} mode is given by $\nu_n = c/\lambda_n$ or

$$\nu_n = n\frac{c}{2L}, \quad n = 1, 2, 3, \cdots \tag{20}$$

For a string of linear mass density μ under tension F we can use Eq. (16 to write what one might frivolously describe as THE GUITAR TUNER'S EQUATION:

$$\nu_n = \frac{n}{2L}\sqrt{\frac{F}{\mu}}, \quad n = 1, 2, 3, \cdots \tag{21}$$

Note that a given string of a given length L under a given tension F has in principle an infinite number of modes (resonant frequencies); the guitarist can choose which modes to excite by plucking the string at the position of an *antinode* (position of *maximum* amplitude) for the desired mode(s). For the first few modes these antinodes are at quite different places,

[6] The Figure could also describe standing *sound waves* in an *organ pipe* closed at both ends, or the electric field strength in a resonant cavity, or the probability amplitude of an electron confined to a one-dimensional "box" of length L.

[7] Note that the n^{th} *mode* has $(n-1)$ *nodes* in addition to the two at the ends.

as evident from Fig. 14.4. As another "exercise for the student" try deducing the relationship between modes with a *common antinode* — these will all be excited as "harmonics" when the string is plucked at that position.

Exactly the same formulae apply to *sound waves* in *organ pipes* if they are *closed at both ends*. An organ pipe *open* at one end must however have an *antinode* at that end; this leads to a slightly different scheme for enumerating modes, but one that you can easily deduce by a similar sequence of logic.

This sort of restriction of the allowed modes of a system to a discrete set of values is known as QUANTIZATION. However, most people are not accustomed to using that term to describe macroscopic classical systems like taut strings; we have a tendency to think of quantization as something that only happens in QUANTUM MECHANICS. In reality, quantization is an ubiquitous phenomenon wherever *wave motion* runs up against *fixed boundary conditions*.

14.5 Energy Density

Consider again our little element of string at position x. We have shown that (for fixed x) the mass element will execute *SHM* as a function of time t. Therefore there is an effective LINEAR RESTORING FORCE in the y direction acting on the mass element $dm = \mu dx$: $dF = F\,d\theta = F\,(\partial^2 y/\partial x^2)\,dx$. But for a simple traveling wave we have[8] $y(x,t) = y_0\cos(kx-\omega t)$ so $(\partial^2 y/\partial x^2) = -k^2 y$, giving $dF = -\left[k^2\,F\,dx\right]\,y$. In other words, the *effective spring constant* for an element of string dx long is $\kappa_{\rm eff} = k^2\,F\,dx$ where I have used the unconventional notation κ for the effective spring constant to avoid confusing it with the *wavenumber* k, which is something completely different. Applying our knowledge of the potential energy stored in a stretched spring, $dU = \frac{1}{2}\kappa_{\rm eff}\,y^2$, we have the *elastic potential energy stored in the string per unit length*, $dU/dx = \frac{1}{2}k^2\,F\,y^2$ or, plugging in $y(x,t)$,

$$\frac{dU}{dx} \;=\; \frac{1}{2}\,k^2\,F\,y_0^2\,\cos^2(kx-\omega t) \qquad (22)$$

— that is, *the potential energy density is proportional to the amplitude squared*.

What about *kinetic* energy? From *SHM* we expect the energy to be shared between potential and kinetic energy as each mass element oscillates through its period. Well, the kinetic energy dK of our little element of string is just $dK = \frac{1}{2}dm\,v_y^2$. Again $dm = \mu\,dx$ and now we must evaluate v_y. Working from $y(x,t) = y_0\cos(kx-\omega t)$ we have $v_y = -\omega\,y_0\sin(kx-\omega t)$, from which we can write

$$\frac{dK}{dx} \;=\; \frac{1}{2}\,\mu\,\omega^2\,y_0^2\,\sin^2(kx-\omega t). \qquad (23)$$

The total energy density is of course the sum of these two:

$$\frac{dE}{dx} \;=\; \frac{dU}{dx} \;+\; \frac{dK}{dx} \qquad \text{or}$$

$$\frac{dE}{dx} \;=\; \frac{1}{2}\,y_0^2\,\left[k^2\,F\,\cos^2\theta \;+\; \mu\,\omega^2\,\sin^2\theta\right]$$

where $\theta \equiv kx-\omega t$. Using $c = \omega/k = \sqrt{F/\mu}$ we can write this as

$$\frac{dE}{dx} \;=\; \frac{1}{2}\,y_0^2\,\left[\mu\,\omega^2\,\cos^2\theta \;+\; \mu\,\omega^2\,\sin^2\theta\right] \qquad \text{or}$$

$$\frac{dE}{dx} \;=\; \frac{1}{2}\,\mu\,\omega^2\,y_0^2. \qquad (24)$$

You can use $F\,k^2$ in place of $\mu\,\omega^2$ if you like, since they are equal. [Exercise for the student.]

Note that the net energy density (potential plus kinetic) is constant in time and space for such a uniform traveling wave. It just switches back and forth between potential and kinetic energy

[8] I have avoided complex exponentials here to avoid confusion when I get around to calculating the transverse *speed* of the string element, v_y. The *acceleration* is the same as for the complex version.

twice every cycle. Since the *average* of either $\cos^2\theta$ or $\sin^2\theta$ is $1/2$, the energy density is *on average* shared equally between kinetic and potential energy.

If we want to know the energy per unit time (*power* P) transported past a certain point x by the wave, we just multiply dE/dx by $c = dx/dt$ to get

$$P \equiv \frac{dE}{dt} = \frac{1}{2}\mu\omega^2 c\, y_0^2. \qquad (25)$$

Again, you can play around with the constants; instead of $\mu\omega^2 c$ you can use $\omega^2\sqrt{F\mu}$ and so on.

Note that while the wave does not transport any *mass* down the string (all physical motion is *transverse*) it does transport *energy*. This is an ubiquitous property of waves, lucky for us!

14.6 Water Waves

Although all sorts of waves are ubiquitous in our lives,[9] our most familiar "wave experiences" are probably with *water waves*, which are unfortunately one of the *least simple* types of waves. Therefore, although water waves are routinely used for illustration, they are rarely discussed in great depth (heh, heh) in introductory Physics texts. They do, however, serve to illustrate one important feature of waves, namely that *not all waves obey* the simple relationship $c = \omega/k$ for their *propagation velocity* c.

Let's restrict ourselves to *deep ocean* waves, where the "restoring force" is simply *gravity*. (When a wave reaches shallow water, the bottom provides an immobile boundary that complicates matters severely, as anyone knows who has ever watched surf breaking on a beach!) The motion of an "element" of water in such a wave is *not* simply "up and down" as we pretended at the beginning of this chapter, but a *superposition* of "up and down" with "back and forth" in the direction of wave propagation. A cork floating on the surface of such a wave executes *circular* motion, or so I am told. (It is actually quite difficult to confirm this assertion experimentally since it requires a fixed reference that is *not* moving with the water — a hard thing to arrange in practice without disturbing the wave itself.) More importantly, the *propagation velocity* of such waves is *higher* for *longer wavelength*.

[9] Indeed, we are *made* of waves, as QUANTUM MECHANICS has taught us!

14.6.1 Phase *vs.* Group Velocity

The precise relationship between angular frequency ω and wavenumber k for deep-water waves is

$$\omega = \sqrt{\frac{g\,k}{2}} \qquad (26)$$

where g has its usual meaning. Such a functional relationship $\omega(k)$ between frequency and wavenumber is known as the DISPERSION RELATION for waves in the medium in question, for reasons that will be clear shortly.

If we have a simple traveling plane wave $A(x,t) = A_0\exp[i(kx - \omega t)]$, with no beginning and no end, the rate of propagation of a point of constant phase (known as the PHASE VELOCITY $v_{\rm ph}$) is still given by Eq. (6):

$$\boxed{v_{\rm ph} \equiv \frac{\omega}{k}} \qquad (27)$$

However, by combining Eq. (27) with Eq. (26) we find that the phase velocity is *higher* for *smaller* k (longer λ):

$$v_{\rm ph} = \sqrt{\frac{g}{2k}}. \qquad (28)$$

Moreover, such a wave *carries no information*. It has been passing by forever and will continue to do so forever; it is the same amplitude everywhere; and so on. Obviously our PLANE WAVE is a bit of an oversimplification. If we want to

send a *signal* with a wave, we have to turn it on and off in some pattern; we have to make wave *pulses* (or, anticipating the terminology of QUANTUM MECHANICS, "WAVE PACKETS"). And when we do that with water waves, we notice something odd: *the wave packets propagate slower than the "wavelets" in them!*

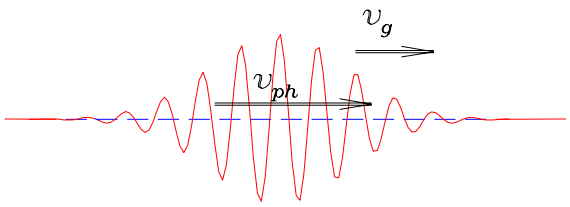


Figure 14.5 A WAVE PACKET moving at $v_{\rm g}$ with "wavelets" moving through it at $v_{\rm ph}$.

Such a packet is a superposition of waves with different wavelengths; the k-dependence of $v_{\rm ph}$ causes a phenomenon known as DISPERSION, in which waves of different wavelength, initially moving together in phase, will drift apart as the packet propagates, making it "broader" in both space and time. (Obviously such a DISPERSIVE MEDIUM is undesirable for the transmission of information!) But how do we determine the effective speed of transmission of said information — *i.e.* the propagation velocity *of the packet itself*, called the GROUP VELOCITY $v_{\rm g}$?

Allow me to defer an explanation of the following result until a later section. The *general* definition of the group velocity (the speed of transmission of information and/or energy in a wave packet) is

$$\boxed{v_{\rm g} \equiv \frac{\partial\omega}{\partial k}} \,. \tag{29}$$

For the particular case of deep-water waves, Eq. (29) combined with Eq. (26) gives

$$v_{\rm g} = \frac{1}{2}\sqrt{\frac{g}{2k}}. \tag{30}$$

That is, the *packet* propagates at *half* the speed of the "wavelets" within it. This behaviour can actually be observed in the wake of a large vessel on the ocean, seen from high above (*e.g.* from an airliner).

Such exotic-seeming wave phenomena are ubiquitous in all dispersive media, which are anything but rare. However, in the following chapters we will restrict ourselves to waves propagating through simple non-dispersive media, for which the DISPERSION RELATION is just $\omega = c\,k$ with c constant, for which $v_{\rm ph} = v_{\rm g} = c$.

14.7 Sound Waves

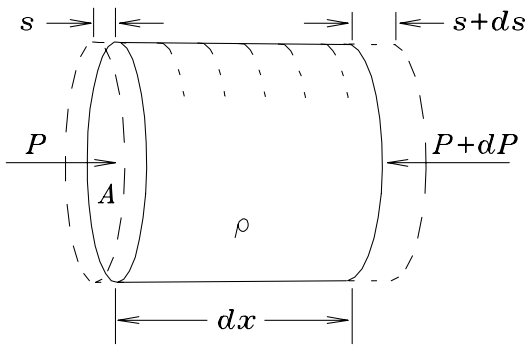


Figure 14.6 Cylindrical element of a compressible medium.

Picture a "snapshot" (holding time t fixed) of a small cylindrical section of an elastic medium, shown in Fig. 14.6: the cross-sectional area is A and the length is dx. An excess pressure P (over and above the ambient pressure existing in the medium at equilibrium) is exerted on the left side and a slightly different pressure $P + dP$ on the right. The resulting volume element $dV = A\,dx$ has a mass $dm = \rho\,dV = \rho\,A\,dx$, where ρ is the mass density of the medium. If we choose the positive x direction to the right, the net force acting on dm in the x direction is $dF_x = PA - (P+dP)A = -A\,dP$.

Now let s denote the *displacement* of particles of the medium from their equilibrium positions.

(I didn't use A here because I am using that symbol for the *area*. This may also differ between one end of the cylindrical element and the other: s on the left *vs.* $s + ds$ on the right. We assume the displacements to be in the x direction but *very small* compared to dx, which is itself no great shakes.[10]

The *fractional change in volume* dV/V of the cylinder due to the *difference* between the displacements at the two ends is

$$\frac{dV}{V} = \frac{(s+ds)\,A - s\,A}{A\,dx} = \frac{ds}{dx} = \left(\frac{\partial s}{\partial x}\right)_t \qquad (31)$$

where the rightmost expression reminds us explicitly that this description is being constructed around a "snapshot" with t held fixed.

Now, any elastic medium is by definition compressible but "fights back" when compressed ($dV < 0$) by exerting a pressure in the direction of increasing volume. The BULK MODULUS B is a constant characterizing how hard the medium fights back — a sort of 3-dimensional analogue of the SPRING CONSTANT. It is defined by

$$P = -B\,\frac{dV}{V}. \qquad (32)$$

Combining Eqs. (31) and (32) gives

$$P = -B\left(\frac{\partial s}{\partial x}\right)_t \qquad (33)$$

so that the *difference* in pressure between the two ends is

$$dP = \left(\frac{\partial P}{\partial x}\right)_t dx = -B\left(\frac{\partial^2 s}{\partial x^2}\right)_t dx. \qquad (34)$$

We now use $\sum F_x = m\,a_x$ on the mass element, giving

$$-A\,dP = AB\left(\frac{\partial^2 s}{\partial x^2}\right)_t dx = dm\,a_x = \rho\,A\,dx\left(\frac{\partial^2 s}{\partial t^2}\right)_x \qquad (35)$$

where we have noted that the acceleration of all the particles in the volume element (assuming $ds \ll s$) is just $a_x \equiv (\partial^2 s/\partial t^2)_x$.

If we cancel $A\,dx$ out of Eq. (35), divide through by B and collect terms, we get

$$\left(\frac{\partial^2 s}{\partial x^2}\right)_t - \frac{\rho}{B}\left(\frac{\partial^2 s}{\partial t^2}\right)_x = 0 \quad \text{or}$$

$$\left(\frac{\partial^2 s}{\partial x^2}\right)_t - \frac{1}{c^2}\left(\frac{\partial^2 s}{\partial t^2}\right)_x = 0 \qquad (36)$$

which the acute reader will recognize as the WAVE EQUATION in one dimension (x), provided

$$c = \sqrt{\frac{B}{\rho}} \qquad (37)$$

is the velocity of propagation.

The fact that disturbances in an elastic medium obey the WAVE EQUATION *guarantees* that such disturbances will propagate as simple *waves* with phase velocity c given by Eq. (37).

We have now progressed from the strictly one-dimensional propagation of a wave in a taut string to the two-dimensional propagation of waves on the surface of water to the three-dimensional propagation of pressure waves in an elastic medium (*i.e.* sound waves); yet we have continued to pretend that the only *simple* type of traveling wave is a *plane wave* with constant $\vec{\boldsymbol{k}}$. This will never do; we will need to treat all sorts of wave phenomena, and although in general we can treat most types of waves as *local approximations to plane waves* (in the same way that we treat the Earth's surface as a flat plane in most mechanics problems), it is important to recognize the most important features of at least one other common idealization — the SPHERICAL WAVE.

[10] Note also that any of s, ds, P or dP can be either positive or negative; we merely illustrate the math using an example in which they are all positive.

14.8 Spherical Waves

The utility of thinking of $\vec{\boldsymbol{k}}$ as a "ray" becomes even more obvious when we get away from plane waves and start thinking of waves with *curved* wavefronts. The simplest such wave is the type that is emitted when a pebble is tossed into a still pool — an example of the "point source" that radiates waves isotropically in all directions. The wavefronts are then *circles* in two dimensions (the surface of the pool) or *spheres* in three dimensions (as for sound waves) separated by one wavelength λ and heading outward from the source at the propagation velocity c. In this case the "rays" k point along the radius vector $\hat{\boldsymbol{r}}$ from the source at any position and we can once again write down a rather simple formula for the "wave function" (displacement A as a function of position) that depends only on the time t and the *scalar* distance r from the source.

A plausible first guess would be just $A(x,t) = A_0\ e^{i(kr-\omega t)}$, but this cannot be right! Why not? Because it violates energy conservation. The energy density stored in a wave is proportional to the square of its amplitude; in the trial solution above, the amplitude of the outgoing spherical wavefront is constant as a function or r, but the *area* of that wavefront increases as r^2. Thus the energy in the wavefront increases as r^2? I think not. We can get rid of this effect by just dividing the amplitude by r (which divides the energy density by r^2). Thus a trial solution is

$$\boxed{A(x,t) \;=\; A_0\ \frac{e^{i(kr-\omega t)}}{r}.} \qquad (38)$$

which is, as usual, correct.[11] The factor of $1/r$ accounts for the conservation of energy in the outgoing wave: since the spherical "wave front" distributes the wave's energy over a surface area $4\pi r^2$ and the flux of energy per unit area through a spherical surface of radius r is proportional to the *square* of the wave amplitude at that radius, the integral of $|f|^2$ over the entire sphere (*i.e.* the total outgoing *power*) is independent of r, as it must be.

We won't use this equation for anything right now, but it is interesting to know that it does accurately describe an outgoing[12] spherical wave.

The perceptive reader will have noticed by now that Eq. (38) is not a solution to the WAVE EQUATION as represented in one dimension by Eq. (10). That is hardly surprising, since the spherical wave solution is an intrinsically 3-dimensional beast; what happened to y and z? The correct *vector* form of the WAVE EQUATION is

$$\boxed{\nabla^2 A - \frac{1}{c^2}\,\frac{\partial^2 A}{\partial t^2} \;=\; 0} \qquad (39)$$

where the LAPLACIAN operator ∇^2 can be expressed in Cartesian[13] coordinates (x,y,z) as[14]

$$\nabla^2 \;=\; \frac{\partial^2}{\partial x^2} \;+\; \frac{\partial^2}{\partial y^2} \;+\; \frac{\partial^2}{\partial z^2}. \qquad (40)$$

[11] I should probably show you a few wrong guesses first, just to avoid giving the false impression that we always guess right the first time in Physics; but it would use up a lot of space for little purpose; and besides, "knowing the answer" is always the most powerful problem-solving technique!

[12] One can also have "incoming" spherical waves, for which Eq. (38) becomes

$$A(x,t) \;=\; A_0\ \frac{e^{i(kr+\omega t)}}{r}.$$

[13] The LAPLACIAN operator can also be represented in other coordinate systems such as spherical (r,θ,ϕ) or cylindrical (ρ,θ,z) coordinates, but I won't get carried away here.

[14] The LAPLACIAN operator can also be thought of as the inner (scalar or "dot") product of the GRADIENT operator $\vec{\boldsymbol{\nabla}}$ with itself: $\nabla^2 = \vec{\boldsymbol{\nabla}}\cdot\vec{\boldsymbol{\nabla}}$, where

$$\vec{\boldsymbol{\nabla}} \;=\; \hat{\boldsymbol{\imath}}\frac{\partial}{\partial x} \;+\; \hat{\boldsymbol{\jmath}}\frac{\partial}{\partial y} \;+\; \hat{\boldsymbol{k}}\frac{\partial}{\partial z}$$

in Cartesian coordinates. This VECTOR CALCULUS stuff is really elegant — you should check it out sometime — but it is usually regarded to be beyond the scope of an introductory presentation like this.

With a little patient effort you can show that Eq. (38) does indeed satisfy Eq. (39), if you remember that $r = \sqrt{x^2 + y^2 + z^2}$. Or you can just take my word for it....

14.9 Electromagnetic Waves

We have some difficulty visualizing a *wave* consisting only of electric and magnetic *fields*. However, if we plot the strength of $\vec{\boldsymbol{E}}$ along one axis and the strength of $\vec{\boldsymbol{B}}$ along another (perpendicular) axis, as in Fig. 14.7, then the direction of propagation $\hat{\boldsymbol{k}}$ will be perpendicular to both $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$, as shown.

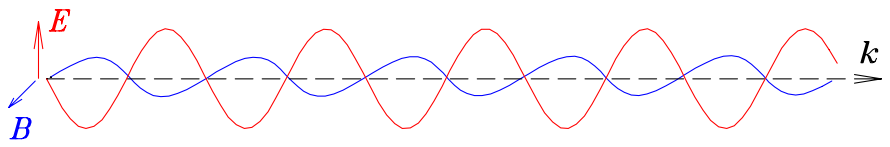


Figure 14.7 A linearly polarized electromagnetic wave. The electric and magnetic fields $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ are mutually perpendicular and both are perpendicular to the direction of propagation $\hat{\boldsymbol{k}}$ ($\vec{\boldsymbol{k}}$ is the *wave vector*).

14.9.1 Polarization

The case shown in Fig. 14.7 is *linearly polarized*, which means simply that the $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ fields are in specific fixed directions. Of course, the directions of $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ could be interchanged, giving the "opposite" polarization. Polaroid sunglasses transmit the light waves with $\vec{\boldsymbol{E}}$ vertical (which are not reflected efficiently off horizontal surfaces) and absorb the light waves with $\vec{\boldsymbol{E}}$ horizontal (which are), thus reducing "glare" (reflected light from horizontal surfaces) without blocking out all light.

There is another possibility, namely that the two linear polarizations be *superimposed* so that both the $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ vectors *rotate* around the direction of propagation $\hat{\boldsymbol{k}}$, remaining always perpendicular to $\hat{\boldsymbol{k}}$ and to each other. This is known as *circular polarization*. It too comes in two versions, *right* circular polarization and *left* circular polarization, referring to the hand whose fingers curl in the direction of the rotation if the thumb points along $\hat{\boldsymbol{k}}$.

14.9.2 The Electromagnetic Spectrum

We have special names for electromagnetic (EM) waves of different wavelengths and frequencies.[15] We call EM waves with $\lambda \gtrsim 1$ m "radio waves," which are subdivided into various ranges or "bands" like "short wave" (same thing as high frequency), VHF (very high frequency), UHF (ultra high frequency) and so on.[16] The dividing line between "radar" and "microwave" bands (for example) is determined by arbitrary convention, if at all, but the rule of thumb is that if the wavelength fits inside a *very* small appliance it is "microwave." Somewhere towards the short end of the microwave spectrum is the beginning of "far infrared," which of course becomes "near infrared" as the wavelength gets still shorter. The name "infrared" is meant to suggest frequencies *below* those of the *red* end of the *visible light* spectrum of EM waves, which extends (depending on the

[15] If the wavelength λ *in*creases (so that the wavenumber $k = 2\pi/\lambda$ *de*creases), then the frequency ω must *de*crease to match, since the ratio ω/k must always be equal to the same propagation velocity c.

[16] One can detect a history of proponents of different bands claiming ever higher (and therefore presumably "better") frequency ranges....

individual eye) from a wavelength of roughly 500 nm (5000 Å) for red light through orange, yellow, green and blue to roughly 200 nm (2000 Å) for violet light. Beyond that we lost sight of the shorter wavelengths (so to speak) and the next range is called "near ultraviolet," the etymology of which is obvious. Next comes "far ultraviolet" which fades into "soft x-rays" and in turn "hard x-rays" and finally "gamma rays" as the frequency increases and the wavelength gets shorter. Note all the different kinds of "rays" that are all just other forms of *light* — *i.e.* *EM* waves — with different wavelengths!

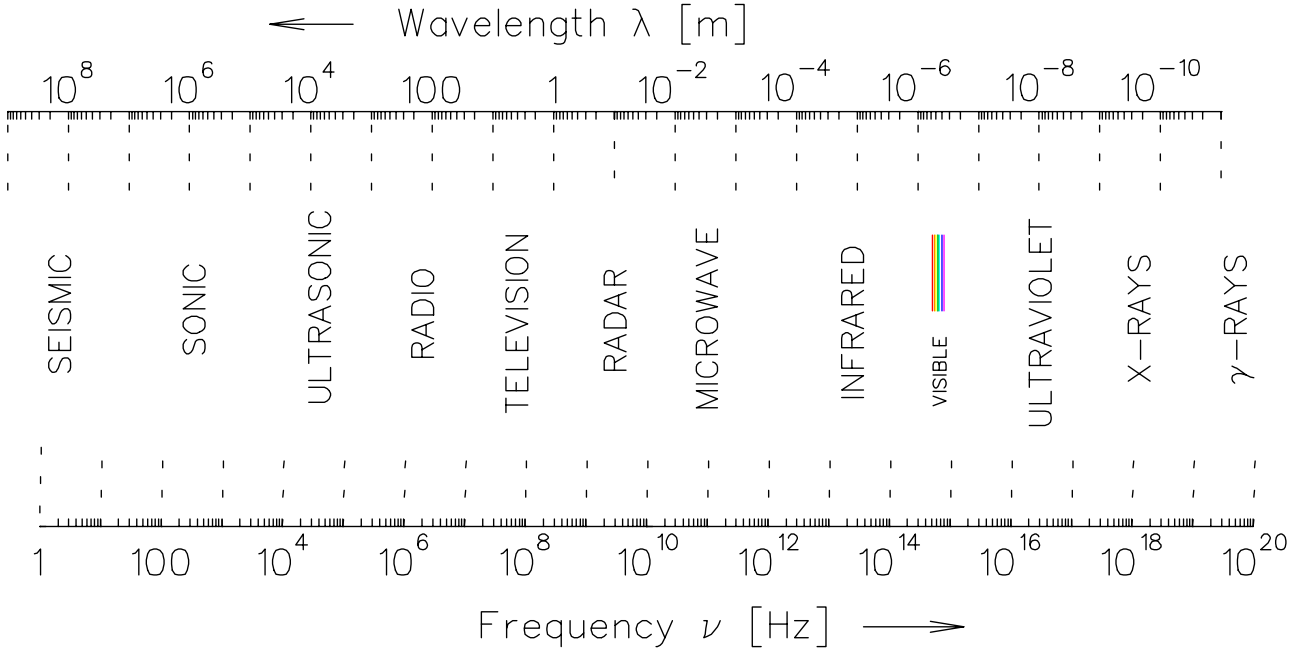


Figure 14.8 The electromagnetic spectrum. Note logarithmic wavelength and frequency scales.

14.10 Reflection

The simplest thing waves do is to REFLECT off flat surfaces. Since billiard balls do the same thing quite nicely, this is not a particularly distinctive behaviour of waves — which was probably one of the reasons why Newton was convinced that light consisted of *particles*.[17] The reflection of waves looks something like Fig. 14.9.

The incoming wave vector $\vec{\boldsymbol{k}}$ makes the same angle with the surface (or, equivalently, with the direction *normal* to the surface) as the outgoing wavevector $\vec{\boldsymbol{k}}'$:

$$\theta \;=\; \theta' \tag{41}$$

This is the most important property of reflection, and it can be stated in words thus:

The *incident* [incoming] angle is equal to the *reflected* [outgoing] angle.

[17] He was actually correct, but it is equally true that light consists of *waves*. If you are hoping that these apparently contradictory statements will be reconciled with common sense by the Chapter on QUANTUM MECHANICS, you are in for a disappointment. Common sense will have to be beaten into submission by the utterly implausible facts.

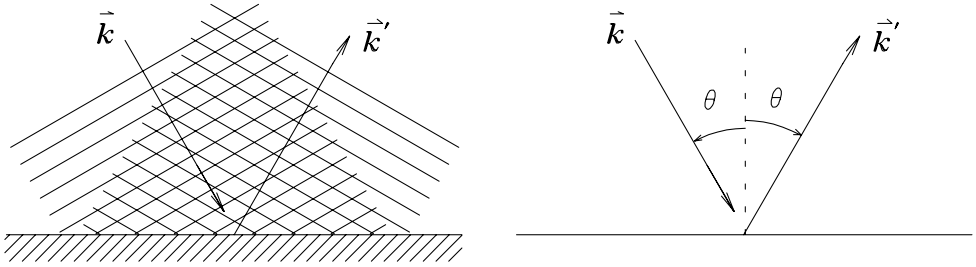


Figure 14.9 Reflection of a wave from a flat surface.

14.11 Refraction

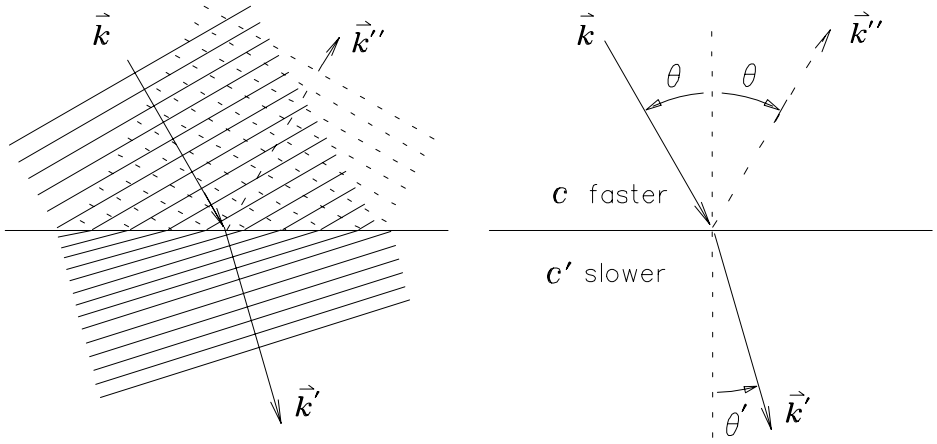


Figure 14.10 Refraction of a wave at a boundary between two media where the propagation velocity (c) of the wave in the first medium is *greater* than that (c') in the second medium. The diagram on the left shows the *wavefronts* ("crests" of the waves) and the corresponding perpendicular wavevectors $\vec{k}$ (incoming wave), $\vec{k}'$ (transmitted wave) and $\vec{k}$" (reflected wave). The diagram on the right shows the angles between the wavevectors and the normal to the interface.

When a wave crosses a boundary between two regions in which its velocity of propagation has different values, it "bends" toward the region with the *slower* propagation velocity. The following mnemonic image can help you remember the qualitative sense of this phenomenon, which is known as REFRACTION: picture the wave front approaching the boundary as a *yardstick* moving through some *fluid* in a direction perpendicular to its length. If one end runs into a *thicker* fluid first, it will "drag" that end a little so that the trailing end gets ahead of it, changing the direction of motion gradually until the whole meter stick is in the thicker fluid where it will move more slowly.[18]

Conversely, if one end emerges first into a *thinner* fluid (where it can move faster) it will pick up speed and the trailing end will fall behind. This picture also explains why there is no "bending" if the wave hits the interface *normally* (at right angles). The details are revealed mathematically (of

[18] Boy, is this ever Aristotelian!

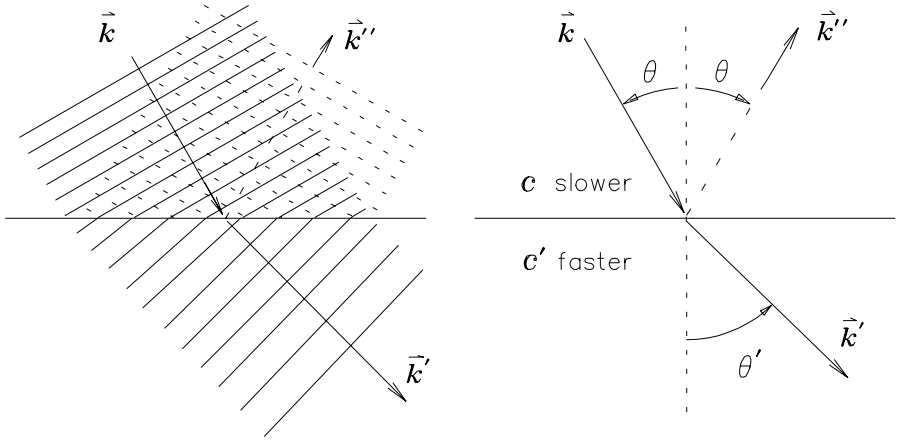


Figure 14.11 Refraction of a wave at a boundary between two media where the propagation velocity (c) of the wave in the first medium is *less* than that (c') in the second medium.

course) in SNELL'S LAW:[19]

$$\frac{\sin(\theta)}{\sin(\theta')} = \frac{c}{c'} \tag{42}$$

where θ is the angle of incidence of the incoming wave (the angle that $\vec{\boldsymbol{k}}$ makes with the normal to the interface), θ' is the angle that the refracted wavevector $\vec{\boldsymbol{k}}'$ makes with the same normal, c is the propagation velocity of the wave in the first medium and c' is the propagation velocity of the wave in the second medium.

[19] SNELL'S LAW is normally expressed in terms of the INDEX OF REFRACTION n in each medium:

$$n\ \sin(\theta)\ =\ n'\ \sin(\theta'),$$

where (we now know) the INDEX OF REFRACTION is the ratio of the speed of light in vacuum to the speed of light in the medium:

$$n\ \equiv\ \frac{c_0}{c}.$$

The reason for inventing such a semicircular definition was that when Willebrord Snell discovered this empirical relationship in 1621 he had no idea what n was, only that every medium had its own special value of n. (This is typical of anything that gets the name "index.") I see no pedagogical reason to even define the dumb thing.

Another semi-obvious consequence of the fact that the "crests" of the waves remain continuous[20] is that the wavelength gets shorter as the wave enters the "thicker" medium or longer as it enters a "thinner" medium. Another way of putting this is that *the frequency stays the same* (and therefore so does the *period* T) as the wave crosses the boundary. Since $c = \lambda/T$ this means that if the velocity decreases, so does the wavelength. One can follow this argument a bit further to *derive* SNELL'S LAW from a combination of geometry and logic. I haven't done this, but you might want to....

There is also always a *reflected* wave at any interface, though it may be weak. The reflected wave is shown as dotted lines in Figs. 14.10 and 14.11, where its wavevector is denoted "$\vec{\boldsymbol{k}}$". This phenomenon is familiar as a source of annoyance to anyone who has tried to watch television in a room with a sunny window facing the TV screen. However, it does have some redeeming features, as can be deduced from a thoughtful analysis of Eq. (42). For instance, if the wave is emerging from a "thick" medium into a "thin" medium as in Fig. 14.11 (like light emerging from glass into air), then there is some incoming angle θ_c, called the CRITICAL ANGLE, for which the *refracted* wave will actually be *parallel to the interface — i.e.* $\theta' = \pi/2$ (90°). This implies $\sin(\theta') = 1$ so that SNELL'S LAW reads

$$\sin(\theta_c) \;=\; \frac{c}{c'} \qquad (43)$$

which has a solution only if $c' > c$ — *i.e.* for emergence into a "thinner" medium with a higher wave propagation velocity, as specified earlier.

What happens, qualitatively, is that as θ gets larger and larger (closer and closer to "grazing incidence") the *amplitude* (strength) of the transmitted wave gets weaker and weaker, while the amplitude of the *reflected* wave gets stronger and stronger, until for incoming angles $\theta \geq \theta_c$ there *is no* transmitted wave and the wave is *entirely reflected*. This phenomenon is known as TOTAL INTERNAL REFLECTION and has quite a few practical consequences.

Because of total internal reflection, a fish cannot see out of the water except for a limited "cone" of vision overhead bounded by the critical angle for water, which is about $\sin^{-1}(1/1.33)$ or 49°. Lest this lend reckless abandon to fishermen, it should be kept in mind that the light "rays" which appear to come from just under 49° from the vertical are actually coming from just across the water's surface, so the fish has a pretty good view of the surrounding environment — it just looks a bit distorted. To observe this phenomenon with your own eyes, put on a good diving mask, carefully slip into a still pool and hold your breath until the surface is perfectly calm again. Looking up at the surface, you will see the world from the fish's perspective (except that the fish is probably a good deal less anoxic) — inside a cone of about 49° from the vertical, you can see out of the water; but outside that cone, the surface forms a perfect mirror!

How total *is* total internal reflection? Total! If the surface has no scratches *etc.*, the light is *perfectly* reflected back into the denser medium. This is how "light pipes" work — light put into one end of a long Lucite rod will follow the rod through bends and twists (as long as they are "gentle" so that the light never hits the surface at less than the critical angle) and emerge at the other end attenuated only by the absorption in the Lucite itself. Even better transmission is achieved in FIBER OPTICS, where fine threads of special glass are prepared with extremely low absorption for the wavelengths of light that are used to send signals down them. A faint pulse of light sent into one end of a fiber optic transmission line will emerge many kilometers down the line with nothing "leaking out" in between. (This feature is especially attractive to those who don't want their

[20] A "crest" doesn't turn into a "trough" just because the propagation velocity changes!

conversations bugged, or so I am told.) Another application was invented by Lorne Whitehead while he was a UBC Physics graduate student: by an ingenious trick he was able to make a large-diameter *hollow* LIGHT PIPE [trademark] which avoids even the small losses in the Lucite itself! Using this trick he is able to "pipe" large amounts of light from single (efficient) light sources [including rooftop solar collectors] into other areas [like the interiors of office buildings] using strictly passive components that do not wear out. He founded a company called TIR — see if you can guess what the acronym stand for!

14.12 Huygens' Principle

At the beginning of this chapter we pictured only PLANE WAVES, in which the wavefronts ("crests" of the waves) form long straight lines (or, in space, flat planes) moving along together in parallel (separated by one wavelength λ) in a common direction $\hat{\boldsymbol{k}}$. One good reason for sticking to this description for as long as possible (and returning to it every chance we get) is that it is so *simple* — we can write down an explicit formula for the amplitude of a plane wave as a function of time and space whose qualitative features are readily apparent (with a little effort). Another good reason has to do with the fact that *all* waves *look pretty much like plane waves* when they are *far from their origin*.[21] We will come back to this shortly. A final reason for our love of plane waves is that they are so easily related to the idea of "RAYS."

In GEOMETRICAL OPTICS it is convenient to picture the wavevector $\vec{\boldsymbol{k}}$ as a "ray" of light (though we can adopt the same notion for any kind of wave) that propagates along a straight line like a billiard ball. In fact, the analogy between $\vec{\boldsymbol{k}}$ and the *momentum* $\vec{\boldsymbol{p}}$ of a *particle* is more than just a metaphor, as we shall see later. However, for now it will suffice to borrow this imagery from Newton and company, who used it very effectively in describing the *corpuscular* theory of light.[22]

However, *near* any *localized source* of waves the outgoing wavefronts are nothing like plane waves; if the dimensions of the source are *small compared to the wavelength* then the outgoing waves look pretty much like SPHERICAL WAVES. For sources similar in size to λ, things can get very complicated.

Christian Huygens (1629-1695) invented the following gimmick for constructing actual wavefronts from spherical waves:

HUYGENS' PRINCIPLE:

> "All points on a wavefront can be considered as *point sources* for the production of *spherical secondary wavelets*. At a later time, the new position of the wavefront will be the *surface of tangency* to these secondary wavelets."

This may be seen to make some sense (try it yourself) but its profound importance to our qualitative understanding of the behaviour of light was really brought home by Fresnel (1788-1827), who used it to explain the phenomenon of *diffraction*, which we will discuss shortly. But first, let's familiarize ourselves with the simpler phenomena of *interference*.

14.13 Interference

To get more quantitative about this "addition of amplitudes," we make the following assumption, which is crucial for the arguments to follow and is even *valid* for the most important

[21] This is sort of like the mathematical assertion that all lines look straight if we look at them through a powerful enough microscope.

[22] "Corpuscles" are hypothetical *particles* of light that follow trajectories Newton called "rays," thus starting a long tradition of naming every new form or radiation a "ray."

kinds of waves, namely *EM* waves, under all but the most extreme conditions:

Linear Superposition of Waves:

> As several waves pass the same point in space, the total *amplitude* at that point at any instant is simply the *sum* of the amplitudes of the individual waves.

For water waves this is not perfectly true (water waves are very peculiar in many ways) but to a moderately good approximation the amplitude (height) of the surface disturbance at a given position and time is just the sum of the heights of all the different waves passing that point at any instant. This has some alarming implications for sailors! If you are sailing along a coastline with steep cliffs, the incoming swells are apt to be *reflected* back out to sea with some efficiency; if the reflected waves from many parts of the shoreline happen to interfere constructively with the incoming swells at the position of your boat, you can encounter "freak waves" many times higher than the mean swell height. Experienced sailors stay well out from the coastline to avoid such unpredictable interference maxima.

14.13.1 Interference in Time

Suppose we add together two *equal amplitude* waves with slightly different *frequencies*

$$\omega_1 = \bar{\omega} + \delta/2 \quad \text{and} \quad \omega_2 = \bar{\omega} - \delta/2 \qquad (44)$$

where $\bar{\omega}$ is the average frequency and δ is the difference between the two frequencies. If we measure the combined amplitude at a fixed point in space, a little algebra reveals the phenomenon of BEATS. This is usually done with sin or cos functions and a lot of trigonometric identities; let's use the complex notation instead — I find it more self-evident, at least algebraically:

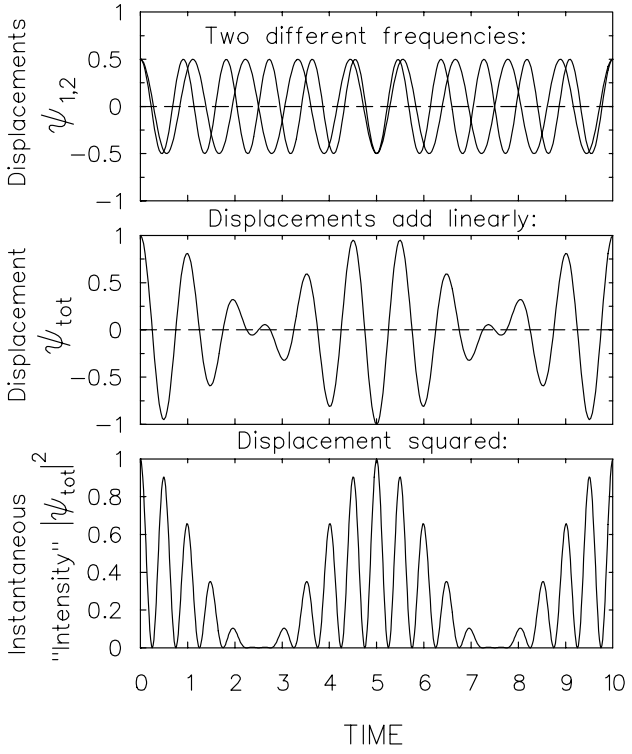


Figure 14.12 Beats.

$$\begin{aligned}
\psi(z,t) &= \psi_0 \left[e^{i\omega_1 t} + e^{i\omega_2 t}\right] \\
&= \psi_0 \left[e^{i(\bar{\omega}+\delta/2)t} + e^{i(\bar{\omega}-\delta/2)t}\right] \\
&= \psi_0\, e^{i\bar{\omega}t} \left[e^{+i(\delta/2)t} + e^{-i(\delta/2)t}\right] \\
&= 2\,\psi_0\, e^{i\bar{\omega}t}\, \cos[(\delta/2)t] \qquad (45)
\end{aligned}$$

That is, the combined signal consists of an oscillation at the *average* frequency, *modulated* by an oscillation at one-half the *difference* frequency. This phenomenon of "BEATS" is familiar to any musician, automotive mechanic or pilot of a twin engine aircraft.

One seemingly counterintuitive feature of BEATS is that the "envelope function" $\cos[(\delta/2)t]$ has only half the angular frequency of the difference between the two original frequencies. What we *hear* when two frequencies interfere is the variation of the

sound INTENSITY with time; and the *intensity* is proportional to the *square* of the displacement.[23] Squaring the envelope effectively doubles its frequency (see Fig. 14.12) and so the detected BEAT FREQUENCY is the full frequency difference $\delta = \omega_1 - \omega_2$.

This is a universal feature of waves and interference: the detected signal is the *average intensity*, which is proportional to the *square* of the *amplitude* of the displacement oscillations; and it is the *displacements* themselves that add linearly to form the interference pattern. Be sure to keep this straight.

14.13.2 Interference in Space

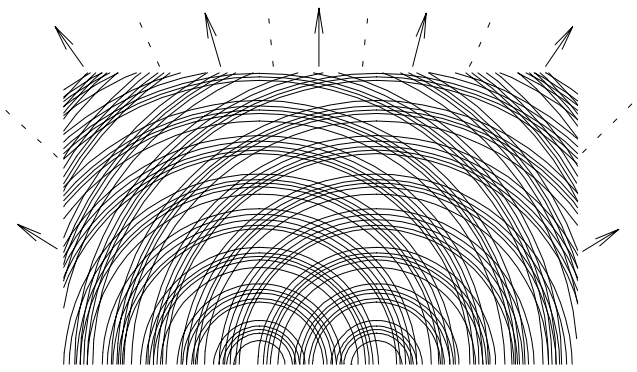


Figure 14.13 A replica of Thomas Young's original drawing (1803) showing the interference pattern created by two similar waves being emitted "in phase" (going up and down simultaneously) from two sources separated by a small distance. The arrows point along lines of constructive interference (crests on top of crests and troughs underneath troughs) and the dotted lines indicate "lines of nodes" where the crests and troughs cancel.

Suppose spherical waves emanate from two *point sources* oscillating *in phase* (one goes "up" at the same time as the other goes "up") at the same frequency, so that the two wave-generators are like synchronized swimmers in water ballet.[24] Each will produce outgoing *spherical waves* that will *interfere* wherever they meet.

The qualitative situation is pictured in Fig.14.13, which shows a "snapshot" of two outgoing spherical[25] waves and the "rays" ($\vec{\boldsymbol{k}}$ directions) along which their peaks and valleys (or crests and troughs, whatever) coincide, giving *constructive interference*. This diagram accompanied an experimental observation by Young of "interference fringes"" (a pattern of intensity maxima and minima on a screen some distance from the two sources) that is generally regarded as the final proof of the wave nature of light.[26]

If we want to precisely locate the angles at which constructive interference occurs ("interference maxima") then it is most convenient to think in terms of "rays" ($\vec{\boldsymbol{k}}$ vectors) as pictured in Fig. 14.14.

The mathematical criterion for constructive in-

[23] Actually the INTENSITY is defined in terms of the *average* of the square of the displacement over times long compared with the average frequency $\bar{\omega}$. This makes sense as long as the beat frequency $\delta \ll \bar{\omega}$; but if ω_1 and ω_2 differ by an amount $\delta \sim \bar{\omega}$ then it is hard to define what is meant by a "time average". We will just duck this issue.

[24] This notion of being "in phase" or "out of phase" is one of the most archetypal metaphors in Physics. It is so compelling that most Physicists incorporate it into their thinking about virtually everything. A Physicist at a cocktail party may be heard to say, "Yeah, we were 90° out of phase on everything. Eventually we called it quits." This is slightly more subtle than, "... we were 180° out of phase..." meaning diametrically opposed, opposite, cancelling each other, *destructively interfering*. To be "90° out of phase" means to be moving at top speed when the other is sitting still (in SHM, this would mean to have all your energy in *kinetic* energy when the other has it all in *potential* energy) and *vice versa*. The $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ fields in a linearly polarized EM wave are 90° out of phase, as are the "push" and the "swing" when a *resonance* is being driven (like pushing a kid on a swing) at maximum effect, so in the right circumstances "90° out of phase" can be productive.... Just remember, "in phase" at the point of interest means *constructive interference* (maximum amplitude) and "180° out of phase" at the point of interest means *destructive interference* (minimum amplitude — zero, in fact, if the two waves have equal amplitude).

[25] OK, they are *circular* waves, not spherical waves. *You* try drawing a picture of spherical waves!

[26] Young's classic experiment is in fact the archetype for all subsequent demonstrations of wave properties, as shall be seen in the Chapter(s) on QUANTUM MECHANICS.

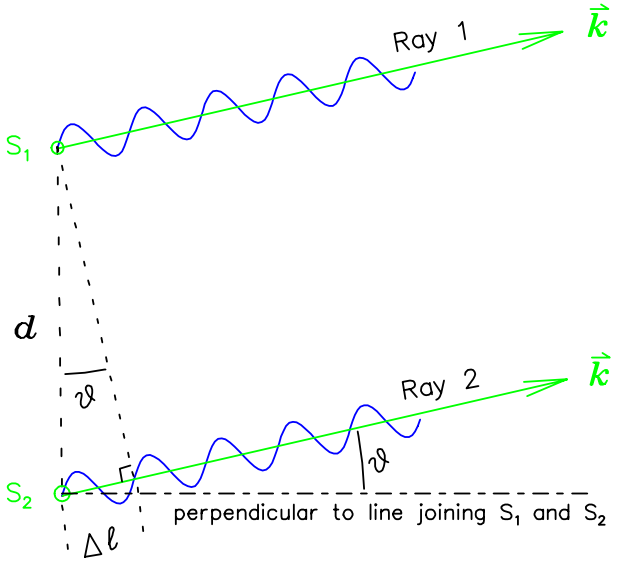


Figure 14.14 Diagram showing the condition for *constructive interference* of two "rays" of the same frequency and wavelength λ emitted in phase from two sources separated by a distance d. At angles for which the difference in path length $\Delta\ell$ is an integer number (m) of wavelengths, $m\lambda$, the two rays arrive at a distant detector in phase so that their amplitudes add constructively, maximizing the intensity. The case shown is for $m = 1$.

terference is simply a statement that the difference in path length, $\Delta\ell = d\,\sin\vartheta_m$, for the two "rays" is an integer number m of wavelengths λ, where the $_m$ subscript on ϑ_m is a reminder that this will be a different angle for each value of m:

$$\boxed{d\,\sin\vartheta_m \;=\; m\,\lambda}\;. \qquad (46)$$

(criterion for CONSTRUCTIVE INTERFERENCE)

Conversely, if the path length difference is a *half-integer* number of wavelengths, the two waves will arrive at the distant detector exactly *out of phase* and cancel each other out. The angles at which this happens are given by

$$\boxed{d\,\sin\vartheta_m^{\rm destr} \;=\; \left(m+\frac{1}{2}\right)\,\lambda}\;. \qquad (47)$$

(criterion for DESTRUCTIVE INTERFERENCE)

Phasors

What happens when coherent light comes through more than two slits, all equally spaced a distance d apart, in a line parallel to the incoming wave fronts? The same criterion (46) still holds for completely *con*structive interference (what we will now refer to as the PRINCIPAL MAXIMA) but (47) is no longer a reliable criterion for *de*structive interference: each successive slit's contribution cancels out that of the adjacent slit, but if there are an *odd number of slits*, there is still one left over and the combined amplitude is not zero.

Does this mean there are *no* angles where the intensity goes to zero? Not at all; but it is not quite so simple to locate them. One way of making this calculation easier to visualize (albeit in a rather abstract way) is with the geometrical aid of PHASORS: A single wave

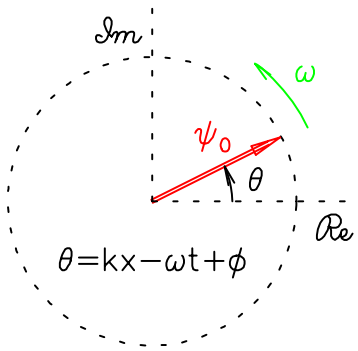


Figure 14.15 A single "PHASOR" of length ψ_0 (the wave amplitude) precessing at a frequency ω in the complex plane.

can be expressed as $\psi(x,t) = \psi_0 e^{i\theta}$ where $\theta = kx - \omega t + \phi$ is the *phase* of the wave at a fixed position x at a given time t. (As usual, ϕ is the "initial" phase at $x = 0$ and $t = 0$. At this stage it is usually ignored; I just retained it one last time for completeness.) If we focus our attention on one particular location in space, this single wave's "displacement" ψ at that location can be represented geometrically as a vector of length ψ_0 (the wave amplitude) in

the complex plane called a "PHASOR" As time passes, the "direction" of the phasor rotates at an angular frequency ω in that abstract plane.

There is not much advantage to this geometrical description for a *single* wave (except perhaps that it engages the right hemisphere of the brain a little more than the algebraic expression) but when one goes to "add together" two or more waves with *different phases*, it helps a lot! For example, two waves of equal amplitude

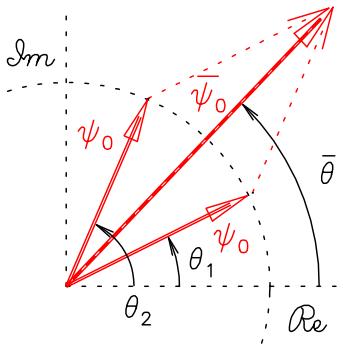


Figure 14.16 Two waves of equal amplitude ψ_0 but different phases θ_1 and θ_2 are represented as PHASORS in the complex plane. Their vector sum has the resultant amplitude $\bar{\psi}_0$ and the average phase $\bar{\theta}$.

but different phases can be added together algebraically as in Eq. (45)

$$\begin{aligned} \bar{\psi} &= \psi_0 \left[e^{i\theta_1} + e^{i\theta_2} \right] \\ &= 2\,\psi_0\, e^{i\bar{\theta}} \cos(\delta/2) \\ &= \bar{\psi}_0\, e^{i\bar{\theta}} \end{aligned} \tag{48}$$

where

$$\begin{aligned} \bar{\psi}_0 &= 2\,\psi_0 \cos(\delta/2) \\ \bar{\theta} &\equiv \tfrac{1}{2}(\theta_1 + \theta_2) \\ \delta &\equiv \theta_2 - \theta_1 \,. \end{aligned} \tag{49}$$

That is, the combined amplitude $\bar{\psi}_0$ can be obtained by adding the phasors "tip-to-tail" like ordinary vectors. Like the original components, the whole thing continues to precess in the complex plane at the common frequency ω.

We are now ready to use PHASORS to find the amplitude of an arbitrary number of waves of arbitrary amplitudes and phases but a common frequency and wavelength interfering at a given position. This is illustrated in Fig. 14.17 for 5 phasors. In practice, we rarely attempt such an

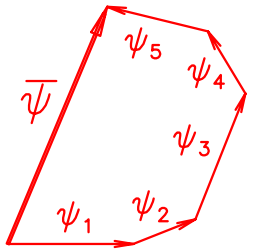


Figure 14.17 The net amplitude of a wave produced by the interference of an arbitrary number of other waves of the same frequency of arbitrary amplitudes ψ_j and phases θ_j can in principle be calculated geometrically by "tip-to-tail" vector addition of the individual PHASORS in the complex plane.

arbitrary calculation, since it cannot be simplified algebraically.

Instead, we concentrate on simple combinations of waves of equal amplitude with well defined phase differences, such as those produced by a regular array of parallel slits with an equal spacing between adjacent slits. Figure 14.18 shows an example using 6 identical slits with a spacing $d = 100\lambda$. The angular width of the interference pattern from such widely spaced slits is quite narrow, only 10 mrad (10^{-2} radians) between principal maxima where all 6 rays are in phase. In between the principal maxima there are 5 minima and 4 secondary maxima; this can be generalized:

The interference pattern for N equally spaced slits exhibits $(N-1)$ *minima* and $(N-2)$ *secondary maxima* between each pair of principal maxima.

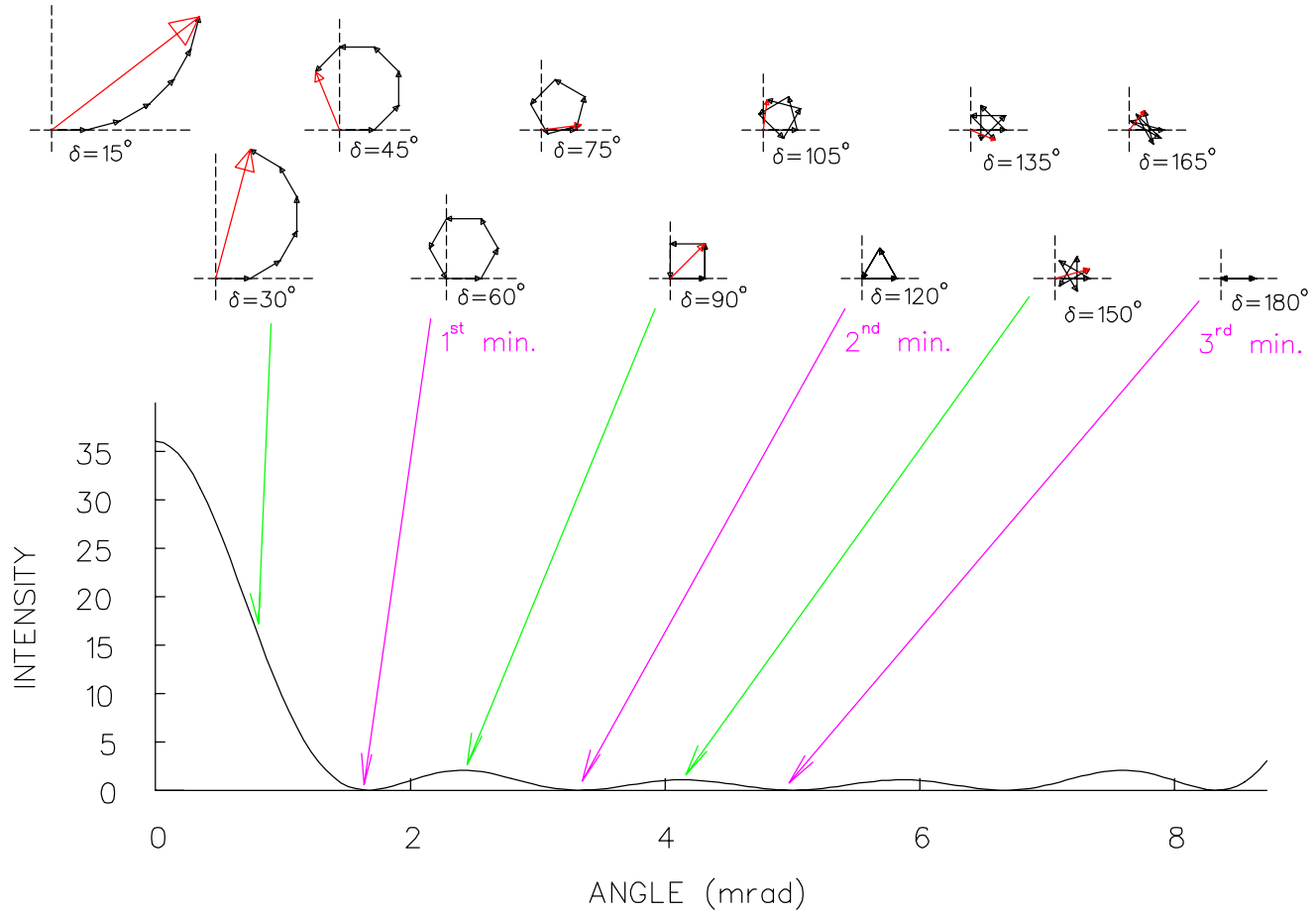


Figure 14.18 The intensity pattern produced by the interference of coherent light passing through six parallel slits 100 wavelengths apart. PHASOR DIAGRAMS are shown for selected angles. Note that, while the *phase* angle difference δ between rays from adjacent slits is a monotonically increasing function of the angle ϑ (plotted horizontally) that the rays make with the "forward" direction, the latter is a real geometrical angle in space while the former is a pure abstraction in "phase space". The exact relationship is $\delta/2\pi = (d/\lambda)\,\sin\vartheta \approx (d/\lambda)\,\vartheta$ for very small ϑ. Note the symmetry about the $3^{\rm rd}$ minimum at $\vartheta \approx 5$ mrad. At $\vartheta \approx 10$ mrad the intensity is back up to the same value it had in the central maximum at $\vartheta = 0$; this is called the first PRINCIPAL MAXIMUM. Then the whole pattern repeats....

It may be conceptually helpful to show the geometrical explanation of the 6-slit interference pattern in Fig. 14.18 in terms of phasor diagrams, but clearly the smooth curve shown there is not the result of an infinte number of geometrical constructions. It comes from an algebraic formula that we can derive for an arbitrary angle ϑ and a corresponding phase difference $\delta = (2\pi d/\lambda)\sin\vartheta$ between rays from adjacent slits. The formula itself is obtained by analysis of a geometrical construction like that illustrated in Fig. 14.19 for 7 slits, each of which contributes a wave of amplitude a, with a phase difference of δ between adjacent slits.

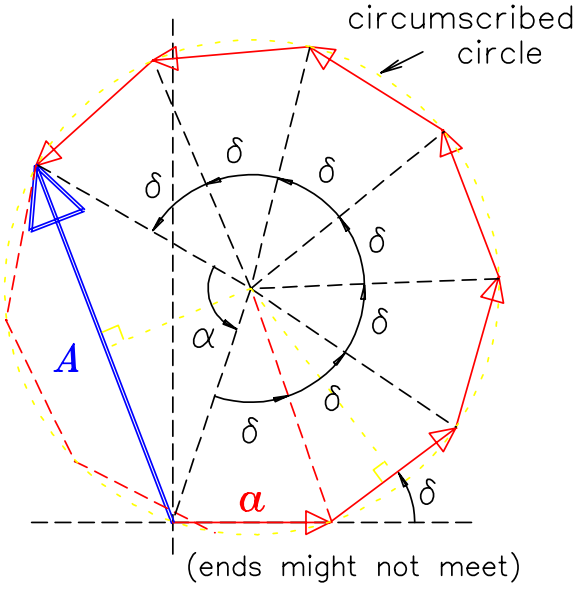


Figure 14.19 PHASOR DIAGRAM for calculating the intensity pattern produced by the interference of coherent light passing through 7 parallel, equally spaced slits.

After adding all 7 equal-length phasors in Fig. 14.19 "tip-to-tail", we can draw a vector from the starting point to the tip of the final phasor. This vector has a length A (the net amplitude) and makes a chord of the circumscribed circle, intercepting an angle

$$\alpha = 2\pi - N\,\delta\,, \tag{50}$$

where in this case $N = 7$. The radius r of the circumscribed circle is given by

$$\frac{a}{2} = r \sin\left(\frac{\delta}{2}\right)\,, \tag{51}$$

as can be seen from the blowup in Fig. 14.20; this can be combined with the analogous

$$\frac{A}{2} = r \sin\left(\frac{\alpha}{2}\right) \tag{52}$$

to give the net amplitude

$$A = a\left[\frac{\sin\left(\frac{\alpha}{2}\right)}{\sin\left(\frac{\delta}{2}\right)}\right]\,. \tag{53}$$

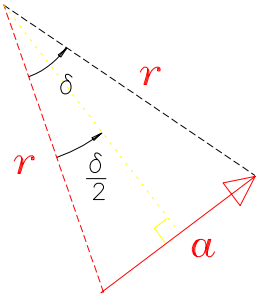


Figure 14.20 Blowup of one of the isosceles triangles formed by a single phasor and two radii from the center of the circumscribed circle to the tip and tail of the phasor.

From Eq. (50) we know that $\alpha/2 = \pi - N\,\delta/2$, and in general $\sin(\pi - \theta) = \sin\theta$, so

$$\boxed{A = a\left[\frac{\sin\left(N\frac{\delta}{2}\right)}{\sin\left(\frac{\delta}{2}\right)}\right]} \tag{54}$$

where

$$\boxed{\delta = 2\pi\left(\frac{d}{\lambda}\right)\sin\vartheta} \tag{55}$$

Although the drawing shows $N = 7$ phasors, this result is valid for an arbitrary number N of equally spaced and evenly illuminated slits.

Chapter 15

Thermal Physics

NOTE.[1]

> *"A theory is the more impressive the greater the simplicity of its premises, the more different kinds of things it relates, and the more extended its area of applicability. Therefore the deep impression that classical thermodynamics made upon me. It is the only physical theory of universal content which I am convinced will never be overthrown, within the framework of applicability of its basic concepts."* — A. Einstein

> *"But although, as a matter of history, statistical mechanics owes its origin to investigations in thermodynamics, it seems eminently worthy of an independent development, both on account of the elegance and simplicity of its principles, and because it yields new results and places old truths in a new light in departments quite outside of thermodynamics."*
>
> — J.W. Gibbs

We have seen how a few simple laws (in particular NEWTON'S SECOND LAW) can be combined with not-too-sophisticated mathematics to solve a great variety of problems — problems which eventually are perceived to fall into a number of reasonably well-defined categories by virtue of the mathematical manipulations appropriate to each — and that those distinct classes of mathematical manipulations eventually become familiar enough to acquire familiar names of their own, such as "conservation of impulse and momentum" or "conservation of work and energy" or "conservation of torque and angular momentum." This *emergence* of new tacit paradigms was the great conceptual gift of the Newtonian revolution. But the most profound *practical* impact of the new sciences on society came in the form of the Industrial Revolution, which was made possible only when the science of mechanics was combined with an understanding of how to extract usable mechanical **work** from that most mysterious of all forms of energy, **heat**.

Historically, heat was recognized as a form of energy and *temperature* was understood in terms of its qualitative properties long before anyone truly understood what either of these terms actually meant in any rigorous microscopic model of matter. The link between Newton's mechanics and the thermodynamics of Joule and Kelvin was forged by Boltzmann long after steam power had changed the world, and a simple understanding of many of the finer points of Boltzmann's *statistical mechanics* had

[1] I have "borrowed" the notation, general approach, basic derivations and most of the quotations shown here from the excellent textbook of the same name by Kittel & Kroemer, who therefore deserve all the credit (and none of the blame) for the abbreviated version displayed before you.

to wait even longer until Quantum Mechanics provided a natural explanation for the requisite fact that the number of possible states of any system, while huge, is not infinite, and that small, simple systems are in fact restricted to a countable number of discrete "stationary states." In this drama Albert Einstein was to play a rather important role.

The following conceptual outline of Statistical Mechanics is designed to make the subject as clear as possible, not to be historically accurate or even fair. Having made this choice, however, I hope to be able to display the essence of the most astonishing product of human Science without undue rigamarole, and perhaps to convey the wonder that arises from a deeper and more fundamental understanding.

15.1 Random Chance

With so many miracles to choose from, where do I get off declaring Statistical Mechanics to be "the most astonishing product of human Science?" This is of course a personal opinion, but it is one shared by many physicists — perhaps even a majority. The astonishment is a result of the incredible precision with which one can predict the outcome of experiments on very complicated systems (the more complicated, the more precise!) based on the FUNDAMENTAL ASSUMPTION of STATISTICAL MECHANICS:

> *A system in thermal equilibrium is a priori equally likely to be found in any one of the fully-specified states accessible to it.*

This seemingly trivial statement contains a couple of ringers: the word "accessible" means, for instance, that the total "internal" energy of the system — which is always written U — *i.e.* the sum of the kinetic and potential energies of all the little particles and waves that make up the big system — is fixed. There are many ways to divide up that energy, giving more to one particle and less to another, and the FUNDAMENTAL ASSUMPTION says that they are all equally likely; but in every case the energy must add up to the same U. This can obviously be very confusing, but fortunately we rarely attempt to count up the possibilities on our fingers!

It is the assumption itself that is so amazing. How can anything but total ignorance result from the assumption that we know *nothing at all* about the minute biases a real system might have for one state over another? More emphatically, how can such an outrageous assumption lead to anything but wrong predictions? It amounts to a pronouncement that Nature runs a perfectly honest casino, in which every possible combination of the roll of the dice is *actually* equally likely! And yet every prediction derived from this assumption has been demonstrated to be accurate to the best precision our measurements can provide. And the consequences are numerous indeed!

15.2 Counting the Ways

If we accept the FUNDAMENTAL ASSUMPTION at face value, then it is easy to calculate the *probability* of finding the equilibrated system in any given fully specified state: if the state is *not accessible* [*e.g.* if it takes more energy U than we have at our disposal] then the probability is *zero*; if it *is* accessible, then its probability is just $\frac{1}{\Omega}$, where Ω is the *total number of accessible states*. The first step is therefore to calculate Ω. In general this can get difficult, but we can choose a few simple examples to illustrate how the calculation goes.

15.2.1 Conditional Multiplicity

Suppose we have a jar full of pennies, say N pennies, all of which have had unique numbers painted on them so that they can be easily dis-

tinguished from each other. Now suppose we shake it thoroughly and dump it out on a nice flat table; each penny falls either "heads" or "tails" with equal *a priori* probability. The probability of penny #1 being "heads" is $\frac{1}{2}$. The probability of penny #1 being "heads" *and* penny #2 being "tails" is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. The probability of penny #1 being "heads" *and* penny #2 being "tails" *and* penny #3 being "tails" is $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$. And so on. If the pennies are all "statistically independent" (*i.e.* how one penny falls has no influence on the other pennies), the probability of *any specific arrangement* of specific pennies falling specific ways [what we call a *fully specified state* of the system] is

$$\left(\frac{1}{2}\right)^N = \frac{1}{2^N}$$

where N is the total number of pennies.

Unfortunately, *this is not what we want to know*. We don't care *which* pennies fall which way,[2] only *how many* of each. This is what we call a *partially specified* or *partially constrained* state of the system. What we really want to know is the *number of ways* we can get n heads and $(N-n)$ tails.[3]

Suppose we *specify* that n pennies are "heads" and the remaining $(N-n)$ are "tails." The *number of ways we can do this* is what we call $\Omega(n, N)$, the *multiplicity function* for the *partially constrained* state specified only by n and N. Here's how we calculate $\Omega(n, N)$: the number of different ways we can rearrange all N coins is

$$N! \equiv N \cdot (N-1) \cdot (N-2) \cdots 3 \cdot 2 \cdot 1$$

because we have N choices of which coin will be first, then we have $(N-1)$ choices of which coin will be second, then we have $(N-2)$ choices of which coin will be third, and so on. The total number of choices is the *product* of the numbers of choices at each step. However, we have *overcounted* by the number of different ways the *heads* can be rearranged *among themselves*, which by the same argument is $n!$, *and* by the number $(N-n)!$ of ways the *tails* can be rearranged among *them*selves. Therefore the total number of *distinguishable* combinations that all give n heads and $(N-n)$ tails is

$$\Omega(n, N) = \frac{N!}{n! \cdot (N-n)!} \qquad (1)$$

Another example would be a parking lot with N spaces in which n cars are parked. The number of different ways we can label the spaces is $N!$ but the n occupied spaces can be rearranged amongst themselves $n!$ different ways and the $(N-n)$ empty spaces can be rearranged $(N-n)!$ different ways without altering the partial constraint [namely, that only n of the spaces are filled].[4] Then Eq. (1) describes the number of different ways the cars can be parked without changing the total number of parked cars.

The Binomial Distribution

To generalize, we talk about a *system of* N *particles*,[5] each of which can only be in one

[2] In the present case, we have a *choice* of whether to treat the pennies as "indistinguishable" or not. No two pennies are *really* indistinguishable, of course; even without our painted-on numbers, each one has unique scratches on its surface and was crystallized from the molten state in a unique microscopic pattern. We *could* tell one from another; we just don't *care*, for circumstantial reasons. In QUANTUM MECHANICS, however, you will encounter the concept of *elementary particles* [*e.g.* electrons] which are so uncomplicated that they truly *are* indistinguishable [*i.e. perfectly* identical]; moreover, STATISTICAL MECHANICS provides a means of actually *testing* to see whether they are *really absolutely indistinguishable* or just very similar!

[3] It might be that we get to keep all the pennies that come up heads, but for every penny that comes up tails we have to chip in another penny of our own. In that case our *profit* would be $n - (N-n) = 2n - N$ cents.

[4] If you were the parking lot owner and were charging \$1 per space [cheap!], your profit would be $\$n$. I keep coming back to monetary examples — I guess *cash* is the social analogue of *energy* in this context.

[5] The term "particle" is [in this usage] meant to be as vague as possible, just like "system:" the *particles* are "really simple things that are all very much alike" and the *system* is "a bunch of particles taken together."

of two possible *single-particle states*. A *fully specified N-particle state* of the system would have the single-particle state of *each individual particle* specified, and is not very interesting. The *partially specified N-particle state* with n of the particles in the first single-particle state and the remaining $(N-n)$ particles in the other single-particle state can be realized in $\Omega(n,N)$ different ways, with $\Omega(n,N)$ given by Eq. (1). Because there are only *two* possible single-particle states, this case of Ω is called the *binomial* distribution. It is plotted[6] in Fig. 15.1 for several values of N.

Note what happens to $\Omega(n,N)$ as N gets bigger: the peak value, which always occurs at $n_{\rm peak} = \frac{1}{2}N$, gets very large [in the plots it is compensated by dividing by 2^N, which is a big number for large N] and the *width* of the distribution grows steadily *narrower* — *i.e.* values of $\frac{n}{N}$ far away from the peak get less and less likely as N increases. The width is in fact the *standard deviation*[7] of a hypothetical random sample of n, and is proportional to $\sqrt{N}$. The *fractional* width (expressed as a fraction of the total range of n, namely N) is therefore proportional to $\frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}}$:

$$\text{Fractional Width} \propto \frac{1}{\sqrt{N}} \qquad (2)$$

which means that for *really* large N, like $N = 10^{20}$, the binomial distribution will get *really* narrow, like a part in 10^{10}, in terms of the *fraction of the average*.

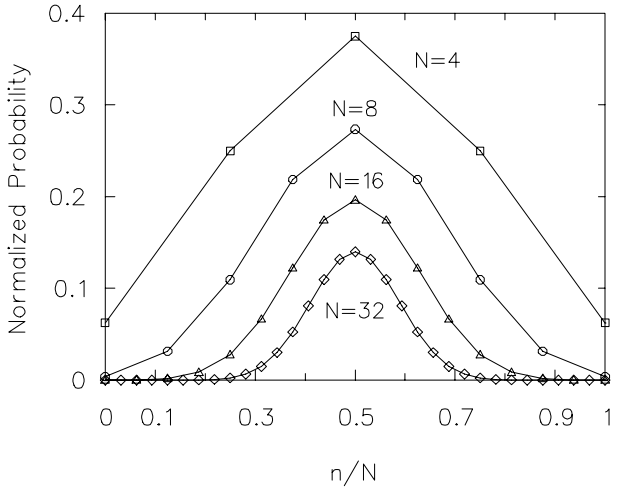


Figure 15.1 The *normalized binomial distribution* for several values of N. In order to put several cases on a single graph, the horizontal axis shows n divided by its maximum possible value N [giving the *fraction* of the total range] and the binomial coefficient $\Omega(n,N)$ given by Eq. (1) has been divided by the total number of possible fully specified N-particle states, 2^N, to give the "normalized" probability — *i.e.* if we add up the values of $\Omega(n,N)/2^N$ for all possible n from 0 to N, the total probability must be 1. [This is eminently sensible; the probability of n having *some* value is surely equal to unity!]

15.2.2 Entropy

> *"If we wish to find in rational mechanics an a priori foundation for the principles of thermodynamics, we must seek mechanical definitions of temperature and entropy."* — J.W. Gibbs

The function $\Omega(n,N)$ is called the MULTIPLICITY FUNCTION for the partially specified system. If N and n get to be large numbers (which is usually the case when we are talking about things like the numbers of electrons in a crystal), $\Omega(n,N)$ can get *really huge*.[8] It is so huge, in fact, that it becomes very diffi-

[6] Actually what is plotted in Fig. 15.1 is the *probability* function

$$\mathcal{P}(n) \equiv \frac{1}{2^N}\cdot\Omega(n,N) = \frac{1}{2^N}\cdot\frac{N!}{n!\,(N-n)!}$$

vs. $\frac{n}{N}$, as explained in the caption. Otherwise it would be difficult to put more than one plot on the same graph, as the peak value of $\Omega(n,N)$ gets very large very fast as N increases!

[7] Recall your Physics Lab training on MEASUREMENT!

[8] A good estimate of the size of $N!$ for large N is given

cult to cope with, and we do what one usually does with ungainly huge numbers to make them manageable: we take its *logarithm*. We define the [natural] logarithm of Ω to be the ENTROPY σ:

$$\sigma \equiv \ln \Omega \tag{3}$$

Let's say that again: the ENTROPY σ is the *natural logarithm* of the MULTIPLICITY FUNCTION Ω — *i.e.* of the *number of different ways we can get the partially specified conditions* in this case defined by n.

Is this all there is to the most fearsome, the most arcane, the most incomprehensible quantity of THERMODYNAMICS? Yep. Sorry to disappoint. That's it. Of course, we haven't played around with σ yet to see what it might be good for — this can get very interesting — nor have I told this story in an historically accurate sequence; the concept of ENTROPY preceded this definition in terms of "statistical mechanics" by many years, during which all of its properties were elucidated and armies of thermal physicists and engineers built the machines that powered the Industrial Revolution. But understanding THERMODYNAMICS the old-fashioned way is *hard*! So we are taking the easy route — sort of like riding a helicopter to the top of Mt. Everest.

15.3 Statistical Mechanics

Before we go on, I need to move away from our examples of binomial distributions and cast the general problem in terms more appropriate to Mechanics. We can always go back and generalize the paradigm[9] but I will develop it along traditional lines.

The owner of the parking lot described earlier is only interested in the total number of cars parked because that number will determine his or her profit. In Mechanics the "coin of the realm" is *energy*, which we have already said is always written U in thermal physics. The abstract problem in STATISTICAL MECHANICS involves a complex system with many possible states, each of which has a certain total energy U. This energy may be in the form of the sum of the kinetic energies of all the atoms of a gas confined in a box of a certain volume, or it may be the sum of all the vibrational energies of a crystal; there is no end of variety in the physical examples. But we are always talking about the *random, disordered energy* of the system, the so-called *internal* energy, when we talk about U.

Now, *given a certain amount of internal energy* U, the *number of different fully-specified states of the system* whose total internal energy is U [our *partial constraint*] is the conditional MULTIPLICITY FUNCTION $\Omega(U)$. Taking the binomial distribution as our example again, we could substitute *crystal lattice sites* for "parking places" and *defects* for "cars" [a defect could be an atom out of place, for instance]. If it takes an energy ε to create one defect, then the total internal energy stored in n defects would be $U = n\,\varepsilon$. Lots of other examples can be imagined, but this one has the energy U proportional to the number n of defects, so that you can see how the U-dependence of Ω in this case is just like the n-dependence of Ω before.

So what?

Well, things start to get interesting when you put *two* such systems *in contact* so that U can flow freely between them through random statistical fluctuations.

15.3.1 Ensembles

One of the more esoteric notions in STATISTICAL MECHANICS is the concept of an *ensem-*

by *Stirling's approximation*:

$$N! \approx \sqrt{2\pi N} \cdot N^N \cdot e^{-N}$$

[9] Count on it!

ble. This has nothing to do with music; it goes back to the original meaning of the French word *ensemble*, which is a collection or gathering of things — much more general and abstract than the small band of musicians we tend to visualize. Anyway, the Statistical Mechanical "ENSEMBLE" is a collection of *all the possible fully specified states* of some *system*.

Of course, there are different *kinds* of ENSEMBLES depending upon what global *constraints* are in effect. For instance, the set of all possible states of an *isolated* system $\mathcal{S}$ consisting of a fixed number N of "particles"[10] with a well defined total energy U is called a MICROCANONICAL ENSEMBLE. This is what we have been discussing so far.

The set of all possible states of a system $\mathcal{S}$ consisting of a fixed number N of particles but in "thermal contact" with a *much, much larger* system $\mathcal{R}$ (called a "heat reservoir") so that the energy U of $\mathcal{S}$ can flow in or out of $\mathcal{R}$ at random is called a CANONICAL ENSEMBLE.

And the set of all possible states of a system $\mathcal{S}$ in contact with a reservoir $\mathcal{R}$ with which it can exchange *both* energy (U) *and* particles (N) is called a GRAND CANONICAL ENSEMBLE.

If the utility of these concepts is less than obvious to you, join the club. I won't need to use them to derive the good stuff below, but now you will be able to scoff at pedants that pretend you can't understand "Stat Mech" unless you know what the various types of Ensembles are.

15.4 Temperature

> *"The general connection between energy and temperature may only be established by probability considerations.* [Two systems] *are in statistical equilibrium when a transfer of energy does not increase the probability."*
>
> — M. Planck

When we put *two* systems $\mathcal{S}_1$ and $\mathcal{S}_2$ (with N_1 and N_2 particles, respectively) into "thermal contact" so that the (constant) total energy $U = U_1 + U_2$ can redistribute itself randomly between $\mathcal{S}_1$ and $\mathcal{S}_2$, the *combined* system $\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2$ will, we postulate, obey the FUNDAMENTAL PRINCIPLE — it is equally likely to be found in any one of its accessible states. The *number* of accessible states of $\mathcal{S}$ (partially constrained by the requirement that N_1, N_2 and $U = U_1 + U_2$ remain constant) is given by

$$\Omega \;=\; \Omega_1(U_1) \;\times\; \Omega_2(U_2) \tag{4}$$

where Ω_1 and Ω_2 are the MULTIPLICITY FUNCTIONS for $\mathcal{S}_1$ and $\mathcal{S}_2$ taken separately [both depend upon their internal energies U_1 and U_2] and the overall multiplicity function is the *product* of the two individual multiplicity functions because the rearrangements within one system are statistically independent of the rearrangements within the other.[11] Since the ENTROPY is the log of the MULTIPLICITY and the log of a product is the sum of the logs, Eq. (4) can also be written

$$\sigma \;=\; \sigma_1(U_1) \;+\; \sigma_2(U_2) \tag{5}$$

— *i.e.* the entropy of the combined system is the *sum* of the entropies of its two subsystems.

15.4.1 The Most Probable

So what? Well, here's the thing: we know that all accessible states of the system are *a priori* equally likely; however, the *number* Ω of accessible states will depend upon the division of

[10] Remember, a "particle" is meant to be an abstract concept in this context!

[11] If I flip my coin once and you flip your coin twice, there are $2^1 = 2$ ways my flip can go [h, t] and $2^2 = 4$ ways your 2 flips can go [HH, HT, TH, TT]; the total number of ways the *combination* of your flips and mine can go [hHH, hHT, hTH, hTT, tHH, tHT, tTH, tTT] is $2 \times 4 = 8$. And so on.

the total energy U between U_1 and U_2. Moreover, for a certain value of U_1 (and therefore of $U_2 = U - U_1$), Ω will be a *maximum* — *i.e.* that value of U_1 will make possible the largest variety of equally likely random states of the system and consequently we will be more likely, on average, to find the system in states with that value of U_1 than in other states[12] with different values of U_1.

This special value of U_1 is called (reasonably enough) the "most probable value" and is given the symbolic representation $\hat{U}_1$.

15.4.2 Criterion for Equilibrium

If our two systems are initially prepared separately with energies U_1 and U_2 *other than* the most probable, *what will happen* when we bring them into contact so that U can flow between them? The correct answer is, of course, "Everything that possibly *can* happen." But there is a *bigger variety* of possibilities for certain gross distributions of energy than for others, and this makes those gross distributions *more likely* than others. The overall *entropy* is thus a *measure* of this likelihood. It seems inevitable that one will eventually feel compelled to anthropomorphize this behaviour and express it as follows:[13]

> All random systems "like" *variety* and will "seek" arrangements that maximize it.

In any case, the tendency of energy to flow from one system to the other will *not* be governed by equalization of either energy or entropy themselves, but by equalization of the *rate of change* of entropy with energy, $\frac{\partial\sigma}{\partial U}$. To see why, suppose (for now) that more energy always gives more entropy. Then suppose that the entropy σ_1 of system $\mathcal{S}_1$ depends only *weakly* on its energy U_1, while the entropy σ_2 of system $\mathcal{S}_2$ depends *strongly* on its energy U_2. In mathematical terms, this reads

$$\text{Suppose} \qquad \frac{\partial\sigma_1}{\partial U_1} < \frac{\partial\sigma_2}{\partial U_2}$$

Then *removal* of a small amount of energy dU from $\mathcal{S}_1$ will *decrease* its entropy σ_1, but *not by as much* as the *addition* of that same energy dU to $\mathcal{S}_2$ will *increase its* entropy σ_2. Thus the *net* entropy $\sigma_1 + \sigma_2$ will be *increased* by the transfer of dU from $\mathcal{S}_1$ to $\mathcal{S}_2$. This argument is as convoluted as it sounds, but it contains the irreducible essence of the definition of temperature, so don't let it slip by!

The converse also holds, so we can combine this idea with our previous statements about the system's "preference" for higher entropy and make the following claim:

> Energy U will flow spontaneously *from* a system with *smaller* $\dfrac{\partial\sigma}{\partial U}$ *to* a system with *larger* $\dfrac{\partial\sigma}{\partial U}$.

If the rate of increase of entropy with energy $\left(\frac{\partial\sigma}{\partial U}\right)$ is the *same* for $\mathcal{S}_1$ and $\mathcal{S}_2$, then the combined system will be "happy," the energy will stay where it is (on average) and a state of "thermal equilibrium" will prevail.

Mathematical Derivation

Is there any way to *derive* a formal (mathematical) criterion for the condition of thermal equilibrium, starting from a hypothetical knowledge of Ω_1 as a function of U_1 and Ω_2 as a function of $U_2 = U - U_1$? Of course! Why else would I be doing this? The thing about a *maximum* of a function (or a minimum, for that matter;

[12] Nothing precludes finding the system in states with other values of U_1, of course. In fact we *must* do so sometimes! Just less often.

[13] Perhaps the converse is actually true: human "wants" are actually manifestations of random processes whose variety is greater in the direction of perceived desire. I find this speculation disturbing.

either type of *extremum* obeys the same rule) is that the *slope* of the function must be *zero* at the extremum. [Otherwise it would still have further up or down to go!] Since the slope is given by the *derivative*, this reads

$$\text{Criterion for an extremum:} \qquad \frac{\partial \Omega}{\partial U_1} \;=\; 0 \tag{6}$$

In this case, since $\Omega = \Omega_1 \cdot \Omega_2$, the PRODUCT RULE for derivatives gives

$$\frac{\partial \Omega}{\partial U_1} \;=\; \frac{\partial \Omega_1}{\partial U_1} \cdot \Omega_2 \;+\; \Omega_1 \cdot \frac{\partial \Omega_2}{\partial U_1} \;=\; 0 \tag{7}$$

Now, Ω_2 is a function of U_2, not U_1; but we can get around that by using the CHAIN RULE,

$$\frac{\partial \Omega_2}{\partial U_1} \;=\; \frac{\partial \Omega_2}{\partial U_2} \cdot \frac{\partial U_2}{\partial U_1}.$$

where $U_2 = U - U_1$ and U is a constant, so

$$\frac{\partial U_2}{\partial U_1} \;=\; -1$$

We can therefore substitute $-\dfrac{\partial \Omega_2}{\partial U_2}$ for $\dfrac{\partial \Omega_2}{\partial U_1}$ in Eq. (7):

$$\frac{\partial \Omega_1}{\partial U_1} \cdot \Omega_2 \;-\; \Omega_1 \cdot \frac{\partial \Omega_2}{\partial U_2} \;=\; 0$$

or

$$\frac{\partial \Omega_1}{\partial U_1} \cdot \Omega_2 \;=\; \Omega_1 \cdot \frac{\partial \Omega_2}{\partial U_2}$$

If we now divide both sides by the product $\Omega_1 \cdot \Omega_2$, we get

$$\frac{1}{\Omega_1} \cdot \frac{\partial \Omega_1}{\partial U_1} \;=\; \frac{1}{\Omega_2} \cdot \frac{\partial \Omega_2}{\partial U_2}. \tag{8}$$

Now we need to recall the property of the *natural logarithm* that was so endearing when we first encountered it: $\ln(x)$ is the function *whose derivative is the inverse*,

$$\frac{d}{dx} \ln(x) = \frac{1}{x}$$

and, by the CHAIN RULE,

$$\frac{d}{dx} \ln(y) = \frac{1}{y} \cdot \frac{dy}{dx}$$

In this case "y" is Ω and "x" is U, so we have

$$\frac{\partial}{\partial U} \ln(\Omega) = \frac{1}{\Omega} \cdot \frac{\partial \Omega}{\partial U}$$

which means that Eq. (8) can be written

$$\frac{\partial}{\partial U_1} \ln(\Omega_1) \;=\; \frac{\partial}{\partial U_2} \ln(\Omega_2)$$

But the logarithm of the MULTIPLICITY FUNCTION Ω is the definition of the ENTROPY σ, so the equation can be simplified further to read

$$\frac{\partial \sigma_1}{\partial U_1} \;=\; \frac{\partial \sigma_2}{\partial U_2} \tag{9}$$

where of course we are assuming that all the other parameters (like N_1 and N_2) are held constant.

Note that we have recovered, by strict mathematical methods, the same criterion dictated by common sense earlier. The only advantage of the formal derivation is that it is rigourous, general and involves no questionable assumptions.[14]

15.4.3 Thermal Equilibrium

Eq. (9) establishes the criterion for the MOST PROBABLE CONFIGURATION — *i.e.* the value of $\hat{U}_1$ for which the combined systems have the maximum total entropy, the maximum total number of accessible states and the highest probability. This also defines the condition of THERMAL EQUILIBRIUM between the two systems — that is, if $U_1 = \hat{U}_1$, any flow of energy from $\mathcal{S}_1$ to $\mathcal{S}_2$ or back will *lower* the number of accessible states and will therefore be *less likely*

[14] Or, at least, none that are readily apparent....

than the configuration[15] with $U_1 = \hat{U}_1$. Therefore if we leave the systems alone and come back later, we will be *most likely* to find them in the "configuration" with $\hat{U}_1$ in system $\mathcal{S}_1$ and $(U - \hat{U}_1)$ in system $\mathcal{S}_2$.

This seems like a pretty weak statement. Nothing certain, just a *bias* in favour of $\hat{U}_1$ over other possible values of U_1 all the way from zero to U. That is true. STATISTICAL MECHANICS has nothing whatever to say about what *will* happen, only about what is *likely* to happen — and *how* likely! However, when the numbers of particles involved become very large (and in Physics they do become *very* large), the fractional width of the binomial distribution [Eq. (2)] becomes *very* narrow, which translates into a probability distribution that is *incredibly sharply peaked* at $\hat{U}_1$. As long as energy conservation is not violated, there is *nothing but luck* to prevent all the air molecules in this room from vacating the region around my head until I expire from asphyxiation. However, I trust my luck in this. A quotation from Boltzmann confirms that I am in distinguished company:

> *"One should not imagine that two gases in a 0.1 liter container, initially unmixed, will mix, then again after a few days separate, then mix again, and so forth. On the contrary, one finds ... that not until a time enormously long compared to $10^{10^{10}}$ years will there be any noticeable unmixing of the gases. One may recognize that this is practically equivalent to never...."*
>
> — L. Boltzmann

[15] Note the distinction between the words *configuration* and *state*. The latter implies we specify *everything* about the system — all the positions and velocities of all its particles, *etc.* — whereas the former refers only to some gross *overall macroscopic* specification like the total energy or how it is split up between two subsystems. A *state* is *completely specified* while a *configuration* is only *partly specified*.

15.4.4 Inverse Temperature

What do we expect to happen if the systems are *out of equilibrium*? For instance, suppose system $\mathcal{S}_1$ has an energy $U_1 < \hat{U}_1$. What will random chance "do" to the two systems? Well, a while later it would be more likely to find system $\mathcal{S}_1$ with the energy $\hat{U}_1$ again. That is, energy would tend to "spontaneously flow" from system $\mathcal{S}_2$ into system $\mathcal{S}_1$ to *maximize the total entropy*.[16] Now stop and think: is there any *familiar, everyday property* of physical objects that governs whether or not internal energy (HEAT) will spontaneously flow from one to another? Of course! Every child who has touched a hot stove knows that heat flows spontaneously from a *hotter* object [like a stove] to a *cooler* object [like a finger]. We even have a *name* for the quantitative measure of "hotness" — we call it TEMPERATURE.

Going back to Eq. (9), we have a mathematical expression for the criterion for THERMAL EQUILIBRIUM, whose familiar everyday-life equivalent is to say that *the two systems have the same temperature*. Therefore we have a compelling motivation to associate the quantity $\frac{\partial\sigma}{\partial U}$ for a given system with the TEMPERATURE of that system; then the equation reads the same as our intuition. The only problem is that we expect heat to flow *from* a system at *high* temperature *to* a system at *low* temperature; let's check to see what is predicted by the mathematics.[17] Let's suppose that for some initial value of $U_1 < \hat{U}_1$ we have

$$\frac{\partial\sigma_1}{\partial U_1} > \frac{\partial\sigma_2}{\partial U_2}.$$

Then adding a little extra energy dU to $\mathcal{S}_1$ will increase σ_1 by *more* than we decrease σ_2 by subtracting the same dU from $\mathcal{S}_2$ [which

[16] This is the same as maximizing the probability, but from now on I want to use the terminology "maximizing the entropy."

[17] We have already done this once, but it bears repeating! To avoid complete redundancy, this time we will reverse the order of hot and cold.

we must do, because the total energy is conserved]. So the *total* entropy will *increase* if we move a little energy *from* the system with a *smaller* $\frac{\partial\sigma}{\partial U}$ *to* the system with a *larger* $\frac{\partial\sigma}{\partial U}$. The region of *smaller* $\frac{\partial\sigma}{\partial U}$ must therefore be *hotter* and the region of *larger* $\frac{\partial\sigma}{\partial U}$ must be *cooler*. This is the *opposite* of what we expect of TEMPERATURE, so we do the obvious: we define $\frac{\partial\sigma}{\partial U}$ to be the INVERSE TEMPERATURE of a system:

$$\frac{\partial\sigma}{\partial U} \equiv \frac{1}{\tau} \tag{10}$$

where (at last) τ is the TEMPERATURE of the system in question. We can now express Eq. (9) in the form that agrees with our intuition:

Condition of THERMAL EQUILIBRIUM:

$$\tau_1 = \tau_2 \tag{11}$$

— *i.e.* if the *temperatures* of the two systems are the same, then they will be in *thermal equilibrium* and everything will be most likely to stay pretty much as it is.

As you can see, TEMPERATURE is not quite such a simple or obvious concept as we may have been led to believe! But now we have a universal, rigorous and valid *definition* of temperature. Let's see what use we can make of it.

15.4.5 Units & Dimensions

I have borrowed from several authors the convention of expressing the ENTROPY σ in explicitly *dimensionless* form [the logarithm of a pure number is another pure number]. By the same token, the simple definition of TEMPERATURE τ given by Eq. (10) automatically gives τ dimensions of *energy*, just like U. Thus τ can be measured in joules or ergs or other more esoteric units like electron-volts; but we are accustomed to measuring TEMPERATURE in other, less "physical" units called *degrees*. What gives?

The story of how temperature units got invented is fascinating and sometimes hilarious; suffice it (for now) to say that these units were invented *before anyone knew what temperature really was*![18] There are two types of "degrees" in common use: Fahrenheit degrees[19] and Celsius degrees (written °C) which are moderately sensible in that the interval between the freezing point of water (0°C) and the boiling point of water (100°C) is divided up into 100 equal "degrees" [hence the alternate name "Centigrade"]. However, in Physics there are only one kind of "degrees" in which we measure temperature: degrees *absolute* or "Kelvin"[20] which are written just "K" without any ° symbol. One K is the same size as one °C, but the *zero* of the Kelvin scale is at *absolute zero*, the coldest temperature *possible*, which is itself an interesting concept. The freezing temperature of water is at 273.15 K, so to convert °C into K you just add 273.15 degrees. Temperature measured in K is always written T.

What relationship does τ bear to T? The latter had been invented long before the development of Statistical Mechanics and the explanation of what temperature really *was*; but these clumsy units never go away once people have

[18] Well, to be fair, people had a pretty good working knowledge of the *properties* of temperature; they just didn't have a *definition* of temperature in terms of nuts-and-bolts mechanics, like Eq. (10).

[19] These silly units were invented by an instrument maker called Fahrenheit [1686-1736] who was selling thermometers to meteorologists. He picked *body temperature* [a handy reference, constant to the precision of his measurements] for one "fiducial" point and for the other he picked the *freezing point of saturated salt water*, presumably from the North Sea. Why not pure water? Well, he didn't like negative temperatures [neither do we, but he didn't go far enough!] so he picked a temperature that was, for a meteorologist, as cold as was worth measuring. [Below that, presumably, it was just "damn cold!"] Then he (sensibly) divided up the interval between these two fiducials into $96 = 64 + 32$ equal "degrees" [can you see why this is a pragmatic choice for the number of divisions?] and *voilá!* he had the Fahrenheit temperature scale, on which pure water freezes at 32°F and boils at 212°F. A good system to forget, if you can.

[20] Named after Thomson, Lord Kelvin [1852], a pioneer of thermodynamics.

gotten used to them. The two types of units must, of course, differ by some constant conversion factor. The factor in this case is $k_{\rm B}$, BOLTZMANN'S CONSTANT:

$$\tau \equiv k_{\rm B} T \quad \text{where}$$

$$k_{\rm B} \equiv 1.38066 \times 10^{-23}\ \text{J/K} \tag{12}$$

By the same token, the "conventional entropy" S defined by the relationship

$$\frac{1}{T} = \frac{\partial S}{\partial U} \tag{13}$$

must differ from our dimensionless version σ by the same conversion factor:

$$S \equiv k_{\rm B}\, \sigma \tag{14}$$

This equivalence completes the definition of the mysterious entities of classical thermodynamics in terms of the simple "mechanical" paradigms of Statistical Mechanics. I will continue to use σ and τ here.

15.4.6 A Model System

Some of the more peculiar properties of temperature can be illustrated by a simple example:

Certain particles such as electrons have "spin $\frac{1}{2}$" which (it turns out) prevents their spins from having any orientation in a magnetic field $\vec{\boldsymbol{B}}$ other than parallel to the field ("spin up") or antiparallel to it ("spin down"). Because each electron has a magnetic moment $\vec{\mu}$ (sort of like a tiny compass needle) lined up along its spin direction, there is an energy $\varepsilon = -\vec{\mu} \cdot \vec{\boldsymbol{B}}$ associated with its orientation in the field.[21] For a "spin up" electron the energy is $\varepsilon_\uparrow = +\mu B$ and for a "spin down" electron the energy is $\varepsilon_\downarrow = -\mu B$.

[21] The rate of change of this energy with the angle between the field and the compass needle is in fact the torque which tries to align the compass in the Earth's magnetic field, an effect of considerable practical value.

Consider a *system* consisting of N electrons in a magnetic field and neglect all other interactions, so that the total energy U of the system is given by

$$U = (N_\uparrow - N_\downarrow)\, \mu B$$

where $N_\uparrow$ is the number of electrons with spin up and $N_\downarrow$ is the number of electrons with spin down. Since $N_\downarrow = N - N_\uparrow$, this means

$$U = (2N_\uparrow - N)\, \mu B \quad \text{or}$$

$$N_\uparrow = \frac{N}{2} + \frac{U}{2\mu B} \tag{15}$$

— that is, $N_\uparrow$ and U are basically the same thing except for a couple of simple constants. As $N_\uparrow$ goes from 0 to N, U goes from $-N\mu B$ to $+N\mu B$.

This system is another example of the *binomial distribution* whose multiplicity function was given by Eq. (1), with $N_\uparrow$ in place of n. This can be easily converted to $\Omega(U)$. The *entropy* $\sigma(U)$ is then just the logarithm of $\Omega(U)$, as usual. The result is plotted in the top frame of Fig. 15.2 as a function of energy. Note that the entropy has a *maximum* value for equal numbers of spins up and down — *i.e.* for zero energy. There must be some such peak in $\sigma(U)$ whenever the energy is *bounded above* — *i.e.* whenever there is a *maximum possible energy* that can be stored in the system. Such situations do occur [this is a "real" example!] but they are rare; usually the system will hold as much energy as you want.

Negative Temperature

The "boundedness" of U and the consequent "peakedness" of $\sigma(U)$ have some interesting consequences: the *slope* of $\sigma(U)$ [which, by Eq. (10), defines the *inverse temperature*] decreases steadily and smoothly over the entire range of U from $-N\mu B$ to $+N\mu B$, going through zero at $U = 0$ and becoming negative

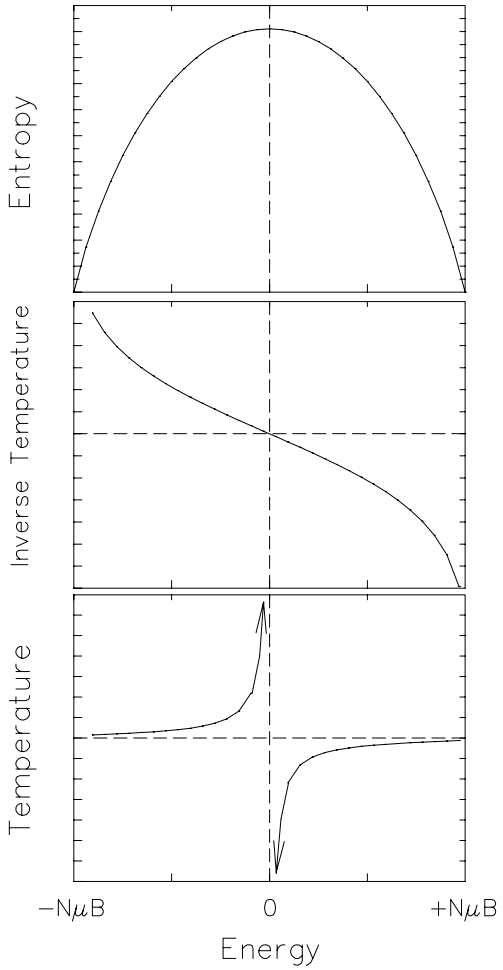


Figure 15.2 Entropy, inverse temperature and temperature of a system consisting of $N = 32$ spin-$\frac{1}{2}$ particles (with magnetic moments μ) in a magnetic field B.

for positive energies. This causes the *temperature* itself to diverge toward $+\infty$ as $U \to 0$ from the left and toward $-\infty$ as $U \to 0$ from the right. Such discontinuous behaviour is disconcerting, but it is only the result of our insistence upon thinking of τ as "fundamental" when in fact it is $1/\tau$ that most sensibly defines how systems behave. Unfortunately, it is too late to get thermometers calibrated in inverse temperature and get used to thinking of objects with *lower* inverse temperature as being *hotter*. So we have to live with some pretty odd properties of "temperature."

Consider, for instance, the whole notion of *negative temperature*, which is actually exhibited by this system and can actually be prepared in the laboratory.[22] What is the behaviour of a system with a negative temperature? Our physical intuition, which in this case is trustworthy, declares that one system is *hotter* than another if, when the two are placed in thermal contact, heat energy spontaneously flows *out of* the first *into* the second. I will leave it as an exercise for the reader to decide which is most hot — infinite positive temperature or finite negative temperature.

15.5 Time & Temperature

Let's do the following *Gedankenexperiment*: Suppose I show you a movie of a swimming pool full of waves and splashes; suddenly (in the movie) all the waves come together and squirt a diver out of the pool. She flies gracefully through the air to land on the diving board while the pool's surface has miraculously returned to mirror smoothness. What is wrong with this picture? Wait! Before you answer, you also get the following movie: A box full of 100 black and 100 white marbles sits on a table; the marbles are arranged randomly. An anonymous assistant picks up the box, closes the lid, shakes the box for a while, puts it down and opens the lid. All the white marbles are now on the left side and all the black marbles are on the right side. Why do you keep thinking there is a problem? Try this: The same box, the same assistant, the same story; except this time there are only 4 marbles, two of each. Not so sure, hmmm? How about 2 marbles, one black and one white? Now we can't tell a thing about whether the movie is being shown forward or backward, right? What is going on here?

Our concept of the "arrow of time" is somehow bound up with statistical mechanics and

[22] [by reversing the direction of the magnetic field before the spins have a chance to react]

is alarmingly *fragile* — we can lose our bearings completely just by confining our attention to *too small* a system! As we will see later, the "fundamental" laws governing the microscopic interactions of matter will be no help at all in clarifying this mystery.

15.6 Boltzmann's Distribution

In defining the concept of *temperature*, we have examined the behaviour of systems in thermal contact (*i.e.* able to exchange energy back and forth) when the total energy U is fixed. In the real world, however, it is not often that we *know* the total energy of an arbitrary system; there is no "energometer" that we can stick into a system and read off its energy! What we often *do* know about a system it its *temperature*. To find this out, all we have to do is stick a calibrated thermometer into the system and wait until equilibrium is established between the thermometer and the system. Then we read its temperature off the thermometer. So what can we say about a *small* system[23] $\mathcal{S}$ (like a single molecule) in thermal equilibrium with a *large* system (which we usually call a "heat reservoir" $\mathcal{R}$) at temperature $\tau = k_{\rm B}T$?

Well, the small system can be in any one of a large number of fully-specified states. It is convenient to be invent an abstract *label* for a given fully-specified state so that we can talk about its properties and probability. Let's call such a state $|\alpha\rangle$ where α is a "full label" — *i.e.* α includes *all the information there is* about the state of $\mathcal{S}$. It is like a complete list of which car is parked in which space, or exactly which coins came up heads or tails in which order, or whatever. For something simple like a single particle's spin, α may only specify whether the spin is up or down. Now consider some particular fully-specified state $|\alpha\rangle$ whose *energy* is ε_α. As long as $\mathcal{R}$ is *very big* and $\mathcal{S}$ is *very small*, $\mathcal{S}$ can — and sometimes will — absorb from $\mathcal{R}$ the energy ε_α required to be in the state $|\alpha\rangle$, no matter how large ε_α may be. However, you might expect that states with really *big* ε_α would be excited somewhat less often than states with *small* ε_α, because the extra energy has to come from $\mathcal{R}$, and every time we take energy out of $\mathcal{R}$ we decrease its entropy and make the resultant configuration that much less probable. You would be right. Can we be quantitative about this?

Well, the *combined system* $\{\mathcal{S}+\mathcal{R}\}$ has a multiplicity function Ω which is the *product* of the multiplicity function $\Omega_\mathcal{S} = 1$ for $\mathcal{S}$ [which equals 1 because we have already postulated that $\mathcal{S}$ is in a specific *fully specified* state $|\alpha\rangle$] and the multiplicity function $\Omega_\mathcal{R} = e^{\sigma_\mathcal{R}}$ for $\mathcal{R}$:

$$\Omega \;=\; \Omega_\mathcal{S} \times \Omega_\mathcal{R} \;=\; 1 \times e^{\sigma_\mathcal{R}}$$

Moreover, the *probability* $\mathcal{P}_\alpha$ of finding $\mathcal{S}$ in state $|\alpha\rangle$ with energy ε_α will be proportional to this net multiplicity:

$$\mathcal{P}_\alpha \;\propto\; e^{\sigma_\mathcal{R}}$$

We must now take into account the effect on this probability of removing the energy ε_α from $\mathcal{R}$ to excite the state $|\alpha\rangle$.

The energy of the reservoir $\mathcal{R}$ before we brought $\mathcal{S}$ into contact with it was U. We don't need to know the value of U, only that it was a fixed starting point. The entropy of $\mathcal{R}$ was then $\sigma_\mathcal{R}(U)$. Once contact is made and an energy ε_α has been "drained off" into $\mathcal{S}$, the energy of $\mathcal{R}$ is $(U - \varepsilon_\alpha)$ and its entropy is $\sigma_\mathcal{R}(U - \varepsilon_\alpha)$.

Because ε_α is *so tiny* compared to U, we can treat it as a "differential" of U (like "dU") and estimate the resultant *change* in $\sigma_\mathcal{R}$ [relative to its old value $\sigma_\mathcal{R}(U)$] in terms of the *derivative* of $\sigma_\mathcal{R}$ with respect to energy:

$$\sigma_\mathcal{R}(U + dU) \;=\; \sigma_\mathcal{R}(U) \;+\; \left(\frac{\partial \sigma_\mathcal{R}}{\partial U}\right) \cdot dU$$

[23] A "small system" can even be a "particle," since both terms are intentionally vague and abstract enough to mean anything we want!

or in this case (with $dU \equiv -\varepsilon_\alpha$)

$$\sigma_{\mathcal{R}}(U - \varepsilon_\alpha) \;=\; \sigma_{\mathcal{R}}(U) \;-\; \left(\frac{\partial \sigma_{\mathcal{R}}}{\partial U}\right) \cdot \varepsilon_\alpha$$

But this derivative is by definition the *inverse temperature* of $\mathcal{R}$: $\dfrac{\partial \sigma_{\mathcal{R}}}{\partial U} \equiv \dfrac{1}{\tau}$. Thus

$$\sigma_{\mathcal{R}}(U - \varepsilon_\alpha) \;=\; \sigma_{\mathcal{R}}(U) \;-\; \frac{\varepsilon_\alpha}{\tau}$$

and thus the probability of finding $\mathcal{S}$ in the state $|\alpha\rangle$ obeys

$$\mathcal{P}_\alpha \;\propto\; e^{\sigma_{\mathcal{R}}(U - \varepsilon_\alpha)} \;=\; \exp\left[\sigma_{\mathcal{R}}(U) - \frac{\varepsilon_\alpha}{\tau}\right]$$

$$\text{or} \quad \mathcal{P}_\alpha \;\propto\; e^{\sigma_{\mathcal{R}}(U)} \cdot \exp\left(-\frac{\varepsilon_\alpha}{\tau}\right)$$

Since $e^{\sigma_{\mathcal{R}}(U)}$ is a constant independent of either ε_α or τ, that term will be the same for any state $|\alpha\rangle$ so we may ignore it and write simply

$$\mathcal{P}_\alpha \;\propto\; \exp\left(-\frac{\varepsilon_\alpha}{\tau}\right) \tag{16}$$

This is the famous BOLTZMANN FACTOR that describes exactly how to calculate the *relative probabilities* of different states $|\alpha\rangle$ of a system in thermal contact with a heat reservoir at temperature τ. It is probably the single most *useful* rule of thumb in all of thermal physics.

15.6.1 The Isothermal Atmosphere

The gravitational potential energy of a gas molecule of mass m at an altitude h above sea level is given approximately by $\varepsilon = mgh$, where $g = 9.81$ m/s^2. Here we neglect the decrease of g with altitude, which is a good approximation over a few dozen miles. Next we pretend that the *temperature* of the atmosphere does not vary with altitude, which is untrue, but perhaps relative to 0 K it is not all that silly, since the difference between the freezing (273.15 K) and boiling (373.15 K) points of water is less than 1/3 of their average. For convenience we will assume that the whole atmosphere has a temperature $T = 300$ K (a slightly warm "room temperature").

In this approximation, the probability $\mathcal{P}(h)$ of finding a given molecule of mass m at height h will drop off exponentially with h:

$$\mathcal{P}(h) \;=\; \mathcal{P}(0)\, \exp\left(-\frac{mgh}{\tau}\right)$$

Thus the *density* of such molecules per unit volume and the *partial pressure* p_m of that species of molecule will drop off exponentially with altitude h:

$$p_m(h) \;=\; p_m(0)\, \exp\left(-\frac{h}{h_0}\right)$$

where h_0 is the altitude at which the partial pressure has dropped to $1/e$ of its value $p_m(0)$ at sea level. We may call h_0 the "mean height of the atmosphere" for that species of molecule. A quick comparison and a bit of algebra shows that

$$h_0 = \frac{\tau}{mg}$$

For *oxygen molecules* (the ones we usually care about most) $h_0 \approx 8$ km. For *helium atoms* $h_0 \approx 64$ km and in fact He atoms rise to the "top" of the atmosphere and disappear into interplanetary space. This is one reason why we try not to lose any helium from superconducting magnets *etc.* — helium is a non-renewable resource!

15.6.2 How Big are Atoms?

Wait a minute! How did I calculate h_0? I had to know m for the different molecules, and that requires some knowledge of the *sizes* of atoms — information that has not yet been set forth in this book! In fact, empirical observations about how fast the pressure of the atmosphere *does* drop off with altitude could give a pretty good idea of his big atoms are; this isn't how it was done historically, but let's give it a try anyway:

Suppose that, by climbing mountains and measuring the density of oxygen molecules (O_2) as a function of altitude, we have determined *empirically* that h_0 for O_2 is about 8,000 m. Then, according to this simple model, it must be true that the mass m of an O_2 molecule is about

$$m \approx \frac{\tau}{h_0\, g} = \frac{300 \times 1.38 \times 10^{-23}}{8 \times 10^3 \times 9.81} \text{ kg}$$

$$\text{or} \quad m \approx 5.3 \times 10^{-26} \text{ kg}$$

This is a mighty small mass!

Now to mix in just a pinch of actual history: Long ago, chemists discovered (again empirically) that different pure substances combined with other pure substances in fixed ratios of small integers times a certain characteristic mass (characteristic for each pure substance) called its *molecular weight* A. People had a pretty good idea even then that these pure substances were made up of large numbers of identical units called "atoms,"[24] but no one had the faintest idea how *big* atoms were — except of course that they must be pretty small, since we never could see any directly. The number N_0 of molecules in one *molecular weight* of a pure substance was (correctly) presumed to be the same, to explain why chemical reactions obeyed this rule. This number came to be called a "mole" of the substance. For oxygen (O_2), the molecular weight is roughly 32 grams or 0.032 kg.

If we now combine this conventional definition of a *mole* of O_2 with our previous estimate of the mass of one O_2 molecule, we can estimate

$$N_0 \approx \frac{0.032}{5.3 \times 10^{-26}} \approx 6 \times 10^{23}$$

The exact number, obtained by quite different means, is

$$N_0 \equiv 6.02205 \times 10^{23} \tag{17}$$

[24] I will cover the history of "Atomism" in a bit more detail later on!

molecules per mole. This is known as AVOGADRO'S NUMBER.

Turning the argument around, the mass of a molecule can be obtained from its molecular weight A as follows: One *mole* of any substance is defined as a mass $A \times 1$ gram, and contains N_0 molecules (or atoms, in the case of monatomic molecules) of the substance. Thus helium, with $A = 4$, weighs 4 gm (or 0.004 kg) per mole containing N_0 atoms, so one He atom weighs $(0.004/N_0)$ kg or 6.6×10^{-27} kg.

15.7 Ideal Gases

We have argued on an abstract basis that the state of highest entropy (and hence the most probable state) for any complicated system is the one whose macroscopic properties can be obtained in the largest possible number of different ways; if the model systems we have considered are any indication, a good rule of thumb for how to do this is to let each "degree of freedom" of the system contain (on average) an equal fraction of the total energy U. We can justify this argument by treating that degree of freedom as a "system" in its own right (almost anything can be a "system") and applying Boltzmann's logic to show that the probability of that microsystem having an energy ε while in thermal equilibrium at temperature τ decays exponentially as $\exp(-\varepsilon/\tau)$. This implies a mean ε on the order of τ, if we don't quibble over factors comparable to 1.

The Equipartition Theorem, which is more rigourously valid than the above hand-waving would suggest,[25] specifies the factor to be exactly 1/2:

[25] If you want the details, here they are: Suppose that p_i is the CANONICAL MOMENTUM characterizing the i^{th} degree of freedom of a system and that $\varepsilon(p_i) = bp_i^2$ is the energy associated with a given value of p_i. Assume further that p_i can have a *continuous* distribution of values from $-\infty$ to $+\infty$. Then the probability of p_i having a given value is proportional to $\exp(-bp_i^2/\tau)$ and therefore the average

> *A system in thermal equilibrium with a heat reservoir at temperature τ will have a* **mean energy** *of $\frac{1}{2}\tau$ per degree of freedom.*

In an ideal monatomic gas of N atoms at temperature τ each atom has three degrees of freedom: left–right (x), back–forth (y) and up–down (z). Thus the average internal energy of our monatomic ideal gas is

$$U \;=\; \frac{3}{2}\,N\,\tau \tag{18}$$

In spite of the simplicity of the above argument[26] this is a profound and useful result. It tells us, for instance, that the energy U of an ideal gas *does not depend upon its pressure*[27] p! This is not strictly true, of course; interactions between the atoms of a gas make its *potential energy* different when the atoms are (on average) close together or far apart. But for most gases at (human) room temperature and (Earth) atmospheric pressure, the ideal-gas approximation is extremely accurate!

It also means that if we change the temperature of a container of gas, the *rate of change* of the internal energy U *with temperature*, which is the definition of the HEAT CAPACITY

$$C \;\equiv\; \frac{\partial U}{\partial T}, \tag{19}$$

is extremely simple: since $\tau \equiv k_{\rm B}T$ and $U = \frac{3}{2}N\tau$, $U = \frac{3}{2}Nk_{\rm B}T$ and so the heat capacity of an ideal gas is *constant*:

$$C\;[\text{ideal gas}] \;=\; \frac{3}{2}\,N\,k_{\rm B} \tag{20}$$

Now let's examine our gas from a more microscopic, "mechanical" point of view: picture *one atom* bouncing around inside a cubical container which is a length L on a side. In the "ideal" approximation, atoms never hit each other, but only bounce off the walls, so our consideration of a *single* atom should be independent of whether or not there are other atoms in there with it. Suppose the atom in question has a velocity $\vec{\boldsymbol{v}}$ with components v_x, v_y and v_z along the three axes of the cube.

Thinking only of the wall at the $+x$ end of the box, our atom will bounce off this wall at a rate $1/t$ where t is the time taken to travel a distance $2L$ (to the far wall and back again) at a speed v_x: $t = 2L/v_x$. We assume *perfectly elastic* collisions — *i.e.* the magnitude of v_x does not change when the particle bounces, it just changes sign. Each time our atom bounces off the wall in question, it imparts an *impulse* of $2mv_x$ to that wall. The average *impulse per unit time* (*force*) exerted on said wall by said atom is thus $F_1 = 2mv_x/t$ or $F_1 = mv_x^2/L$. This force is (on average) spread out all over the wall, an area $A = L^2$, so that the *force per unit area* (or *pressure*) due to that one particle is given by $p_1 = F_1/A = mv_x^2/L^3$. Since $L^3 = V$, the volume of the container, we can write $p_1 = mv_x^2/V$ or

$$p_1\,V \;=\; m\,v_x^2$$

The average pressure p exerted by *all* N *atoms together* is just N times the mean value

---

energy associated with that degree of freedom is given by

$$\langle\varepsilon(p_i)\rangle = \frac{\int_{-\infty}^{+\infty} bp_i^2 e^{-bp_i^2/\tau}dp_i}{\int_{-\infty}^{+\infty} e^{-bp_i^2/\tau}dp_i}$$

These definite integrals have "well known" solutions:

$$\int_{-\infty}^{+\infty} x^2 e^{-ax^2}dx = \frac{1}{2}\sqrt{\frac{\pi}{a^3}}, \qquad \int_{-\infty}^{+\infty} e^{-ax^2}dx = \sqrt{\frac{\pi}{a}},$$

where in this case $a = b/\tau$ and $x = p_i$, giving

$$\langle\varepsilon(p_i)\rangle = \frac{\tau}{2}. \qquad \mathcal{QED}$$

[26] We can, of course, make the explanation more elaborate, thus satisfying both the demands of rigourous logic and the Puritan conviction that nothing of real value can be obtained without hard work. I will leave this as an exercise for other instructors.

[27] Unfortunately, we use the same notation (p) for both *momentum* and *pressure*. Worse yet, the notation for *number density* (number of atoms per unit volume) is n. Sorry, I didn't set up the conventions.

of p_1: $p = N\langle p_1\rangle$, where the "$\langle\cdots\rangle$" notation means *the average of* the quantity within the angle brackets. Thus

$$p\,V \;=\; N\,m\,\langle v_x^2\rangle \tag{21}$$

Now, the kinetic energy of our original atom is explicitly given by

$$\frac{1}{2}mv^2 = \frac{1}{2}m(v_x^2 + v_y^2 + v_z^2)$$

since $\vec{\boldsymbol{v}}$ is the vector velocity. We expect each of the *mean square* velocity components $\langle v_x^2\rangle$, $\langle v_y^2\rangle$ and $\langle v_z^2\rangle$ to average about the same in a random gas, so each one has an average value of $\frac{1}{3}$ of their sum.[28] Thus $\langle v_x^2\rangle = \langle v_y^2\rangle = \langle v_z^2\rangle = \frac{1}{3}\langle v^2\rangle$ and the mean kinetic energy of a *single* particle is $U_1 = \frac{3}{2}\,m\langle v_x^2\rangle$. The kinetic energy of all N atoms is just $U = NU_1$, or

$$U \;=\; \frac{3}{2}\,N\,m\,\langle v_x^2\rangle \tag{22}$$

But according to Eq. (18) we have $U = \frac{3}{2}\,N\tau$; so we may write[29]

$$m\,\langle v_x^2\rangle \;=\; \tau \tag{23}$$

Combining Eqs. (21) and (23), we obtain the famous IDEAL GAS LAW:

$$p\,V \;=\; N\,\tau \tag{24}$$

Despite the flimsiness of the foregoing arguments, the IDEAL GAS LAW is a quantum mechanically correct description of the interrelationship between the pressure p, the volume V and the temperature $\tau \equiv k_{\rm B}T$ of an ideal gas of N particles, as long as the only way to store energy in the gas is in the form of the kinetic energy of individual particles (usually atoms or molecules). Real gases can also store some energy in the form of rotation or vibration of larger molecules made of several atoms or in the form of potential energies of interaction (attraction or repulsion) between the particles themselves. It is the latter interaction that causes gases to spontaneously condense, below a certain *boiling point* T_b, into *liquids* and, at a still lower temperature T_m (called the *melting point*), into *solids*. However, in the gaseous phase even carbon [vaporized diamond] will behave very much like an ideal gas at sufficiently high temperature and low pressure. It is a pretty good Law!

15.8 Things I Left Out

As you can tell by the length of this chapter, I find it hard to stop talking about this wonderful subject. Thermal Physics is like an old but vibrantly modern city with a long, fabulous and meticulously preserved history: around every corner there is a host of fascinating shops, theatres, galleries and restaurants offering the latest goodies from a cosmopolitan state of the art, intermixed with libraries and museums that tell stories of heroic acts and world-changing events. "Shop till you drop!" Still, I have to stop somewhere.

The foregoing has been a rather unusual *introduction* to Thermal Physics. I have completely left out THE LAWS OF THERMODYNAMICS — the traditional *starting point* for the subject — in favour of a strictly conceptual (though often painfully formal, I know) explanation of the meaning of entropy and temperature, in the conviction that these notions can be *generalized* to provide tools for quantitative analysis of random statistical processes in realms where no one ever dreamed of applying the paradigms of Physics. In my zeal to convey this conviction,

[28] We may say that the average kinetic energy "stored in the x *degree of freedom*" of an atom is $\frac{1}{2}m\langle v_x^2\rangle$.

[29] This is equivalent to saying that the average energy stored in the x degree of freedom of one atom [or, for that matter, in any other degree of freedom] is $\frac{1}{2}\tau$ — which is just what we originally claimed in the EQUIPARTITION THEOREM. We could have just jumped to this result, but I thought it might be illuminating to show an explicit argument for the equality of the mean energies stored in several different degrees of freedom.

I have also omitted any discussion of the profound *practical applications* of Thermodynamics, like ENGINES and REFRIGERATORS. Worst of all, I have not told any stories of the bizarre spontaneous behaviour of large numbers of similar atoms under different conditions of temperature and pressure — the so-called EQUATIONS OF STATE and PHASE DIAGRAMS of gases, liquids and solids, from FERMI GASES to SUPERFLUIDS and SUPERCONDUCTORS. Part of the reason for this is that you need a bit more introduction to the phenomenology of Physics — QUANTUM MECHANICS in particular — before you can fully appreciate (or even, in some cases, *describe*) much of the above-mentioned behaviour. All I can hope to have done in this *HyperReference* is to have unlocked the door (and perhaps opened it a crack) to a world of wonder and magic where analytical thinking and mathematics play the role of spells and incantations. I urge you to continue this adventure beyond the limits (and end) of this *HyperReference*!

Chapter 16

Weird Science

The English word "weird" is self-descriptive, violating for no apparent reason the grammatical rule, "*i* before *e* except after *c*." No doubt there is some interesting etymological reason for this particular exception, but to students of English as a Second Language it must seem a completely arbitrary booby-trap set for hapless victims.

The numerous breakdowns of the "Laws of Physics" discovered in the early part of the Twentieth Century must have elicited similar reactions in students of Physics as a Second Language [which is, of course, what we are all trying to learn].

There is a story [which may even be historically accurate, but for my purposes it doesn't matter] about a distinguished physicist around the end of the 19$^{\text{th}}$ Century who advised his bright student to go into some other more promising field [today it would be Computer Science or Microbiology] because "Physics is just about wrapped up — all that remains is to tie up some loose ends and work out a lot of engineering details." Imagine the consternation of that student when, a decade or two later, it became clear that the basic classical "Laws" of Physics were *all wrong* and that the world behaves *essentially differently* from our "common sense" expectations! The success of Classical Physics [before Relativity and Quantum Mechanics] was just a lucky accident: in the world we perceive — naturally enough, a world of objects of roughly our own size — the *true* qualitative behaviour of matter and energy is obscured by the enormous *size* of objects we can handle and the miniscule *speeds* we can achieve with our own huge, puny bodies; in this anthropocentric limit [virtually infinite size relative to atoms and virtually zero velocity relative to light] Newton's "Laws" turn out to be an excellent *approximation* to the truth, so we can still make good use of them. But they are wrong in an absolute qualitative sense. Of course, the "Laws" of Relativity and Quantum Mechanics are almost certainly wrong in an absolute qualitative sense, too. In fact, ever since their "discovery" (if that is the right word), their "truth" has been challenged continuously, often no more aggressively than by those who formulated them in the first place. Einstein in particular was convinced that Quantum Mechanics was merely a provisional calculational technology, that "God does not play dice." And he was surely right; sooner or later we are bound to find where these new descriptions break down [*e.g.* in the description of gravity...] and there we will doubtless find the more "true" theory of which they are merely limiting cases under restricted conditions. [Ain't it always the way?] But it is no criticism of any theory to predict that it is ultimately wrong in an absolute sense; and in any case I am getting much too far ahead of myself here.

16.1 Maxwell's Demon

One hint that there is more to physics than meets the Classical eye can be obtained by the following *Gedankenexperiment* credited to J.C. Maxwell [whom we shall meet again soon]:

We know that a system prepared initially in a highly *ordered* state — *i.e.* one whose gross macroscopic properties can only be achieved by a very small subset of all the possible fully specified microscopic states (*e.g.* a box full of marbles with all the white ones on one side and all the black ones on the other side) — is sure to drift toward more probable, less ordered (more random) states (*e.g.* all the marbles mixed up) as time goes on, if some "jiggling" is provided by the world around it. This intuitively obvious conclusion is translated by Physicists into the Second Law of Thermodynamics, which states that *entropy will always increase* in any spontaneous process involving a highly complex system.[1] When examined critically, this conclusion can be seen to contain virtually everything we know about the "arrow of time" — *i.e.* the only practical way to tell whether a movie of some process is being shown forward or backward. So it is a pretty basic idea.

Now suppose that we build a modern, microminiaturized robot[2] that sits by a hole in a divider between the left and right sides of the box of marbles and opens the door only for white marbles heading toward the right side and for black marbles heading toward the left side. This action can presumably take far less energy than the marbles' kinetic energy; we simply substitute "will" (in this case, the programmer's will as translated into action by the robot) for "brute force" and avoid any "waste" of energy. *Is it possible to reverse* the Second Law of Thermodynamics using a "Maxwell's Demon?"

The answer is not obvious. One can see why by examining the analogous example of keeping one's office or bedroom tidy: in this case a simple application of *will* should suffice to maintain Order (keeping Entropy at bay) by simply putting every article in its proper place every time the opportunity arises; however one is apt to notice some dissipation of energy as such good habits are put into practice. With the possible exception of a few "Saints of Order," we all think of "tidying up" as *work*; and the human machine is fuelled by a form of internal combustion which entails a massive increase of "global" entropy as food is consumed and digested. Therefore we may be able to suppress the Second Law of Thermodynamics *locally* (*e.g.* in our office or bedroom), but only at the expense of a far greater increase in the entropy of our surroundings.[3]

Can we, however, *beat* this "entropy backlash" by building a much more *efficient* machine into which we program our will? Can we build a housekeeping robot that will keep our office/bedroom tidy without consuming more than a fraction of the energy it saves? Or, driving the analogy back to the microscopic level, can we build a "Maxwell's Demon" robot that will let only *fast* air molecules into our house and let only *slow* ones out, so that the average kinetic energy increases (*i.e.* the air warms up) and we can stop paying our heating bill? One problem is the cost (in energy or entropy increase) of *building* such a Demon-robot; but this can be disregarded if the robot is so well-constructed that it never wears out, since any such system that gains on the Second Law will *eventually* gain back any finite initial outlay.[4] If such a device is possible, then we can

[1] There are, of course, many other ways of stating the Second Law, but this suffices for my purposes.

[2] Maxwell specified a "demon," but as A.C. Clarke says, "Any sufficiently advanced technology is indistinguishable from magic," so there is no practical difference.

[3] An awareness of such consequences is perhaps a first step toward an enlightened form of "environmentalism."

[4] Another lesson for the wise consumer: always consider the long term energy-economics of a prospective appliance purchase. For example, a fluorescent light takes as little as 1/4 as much power as an incandescent bulb to generate the

make as many of them as we please and use them to store up energy which we can use in even our less efficient machines to push back the tide of Entropy on all fronts. We can even picture *self-replicating* Maxwell's Demons that get sent out into the Universe to reverse the SECOND LAW everywhere — the ultimate Conservationist scheme! Never mind whether this sounds like a good idea; *could it work?*

The answer is still not obvious. We will have to come back to this question after we have a working knowledge of Quantum Mechanics — and even then it will probably not be obvious, but at least we may be able to find an answer.

16.2 Action at a Distance

Another perplexing problem for turn-of-the-Century scientists was the issue of whether two objects had to "touch" in order to exert forces on each other. The car's wheels touch the road, the crane lifts the concrete block by a cable attached to it and the arrow's flight is slowed by air molecules rubbing against it; so how exactly is the Earth's gravitational force *transmitted* to the cannonball?[5]

Physicists might have been willing to live with the idea that "gravity is weird," were it not for the fact that other types of forces also appeared to act "at a distance" without any strings attached (as it were) — namely, the *electrical* and *magnetic* forces whose simplest properties had been know for millenia but whose detailed behaviour was only beginning to be understood empirically in the late 19th Century. An amber rod rubbed with rabbit fur attracts or repels bits of lint or paper even when separated by hard vacuum; a lodestone's alignment will seek magnetic North wherever it is carried [an important practical property!] except at the North Pole, where we seldom need to go. How does the North Pole "touch" the magnetic compass needle? What is going on here? How can things act on each other without touching? Weird.

There are other examples of "weird science" that kept cropping up around the turn of the Century; I will append some more to this Chapter as we go on, but for now it's time to get on with ELECTRICITY AND MAGNETISM.

same amount of light; on the other hand, turning the fluorescent light on and off may shorten its lifetime even more dramatically than for the equivalent incandescent bulb, and the *replacement* fluorescent light costs *far* more (in energy) to make! So one should strive to use fluorescent light in applications where the light stays on essentially all the time, but in on-and-off applications it is not so clear.

[5]This question has still not been answered in an intuitively satisfactory way; the General Theory of Relativity [coming up!] nicely avoids the issue by making gravitational acceleration equivalent to warped space-time — and thus replies, "the question is meaningless." Maybe *all* "forces" will eventually be shown to be false constructs, misleading paradigms conjured up to satisfy foolish prejudices and ill-posed questions; it wouldn't surprise me a bit. But for the time being we still cling to the image of two "things" acting on each other and have managed to reconcile this image (sort of) with Quantum Mechanics and Relativity in all cases except Gravity, where even stretching the metaphor to the breaking point has not sufficed. More on this later.

Chapter 17

Electromagnetism

As suggested in the previous Chapter, Electricity and Magnetism (or $\mathcal{E}\&\mathcal{M}$, as they are known in the trade) are "weird" phenomena because the palpable *forces* they generate on objects seem to come from nowhere — nothing is "touching" the objects and yet they are moved. The related fact that we are unable to wilfully exert significant electrical or magnetic forces directly on objects around us using any combination of muscles or mechanical devices removes $\mathcal{E}\&\mathcal{M}$ still further from our personal sensory experience and thus makes them seem "weirder." Even the most seasoned $\mathcal{E}\&\mathcal{M}$ veteran still experiences a sense of primitive wonder when a magnet on top of the table moves "by magic" under the influence of another magnet underneath the table.

On the one hand, this makes $\mathcal{E}\&\mathcal{M}$ a fun subject to study. On the other hand, it makes $\mathcal{E}\&\mathcal{M}$ hard to teach, because it will never make "common sense" like nuts-and-bolts Mechanics. *C'est la vie.* As our first foray into "Weird Science" it is only fitting that $\mathcal{E}\&\mathcal{M}$ should be something we know is there but that we will just have to get used to instead of ever hoping to rectify it with our common sense. It is, of course, "common sense" itself that is defective. . . .

17.1 "Direct" Force Laws

There are two fundamental kinds of forces in $\mathcal{E}\&\mathcal{M}$: the *electrostatic* force between two *charges* and the *magnetic* force between two *currents*. Let's start with the easy one.

17.1.1 The Electrostatic Force

First, what is a *charge*? We don't know! But then, we don't know what a *mass* is, either, except in terms of its behaviour: a mass resists acceleration by forces and attracts other masses with a gravitational force. The analogy is apt, in the sense that electrical charges exert forces on each other in almost exactly the same way as masses do, except for two minor differences, which I will come to shortly. Recall Newton's Universal Law of Gravitation in its most democratic form: the force $\vec{\boldsymbol{F}}^G_{12}$ acting *on* body #2 (mass m_2) *due to* body #1 (mass m_1) is

$$\vec{\boldsymbol{F}}^G_{12} \;=\; -G\,\frac{m_1 m_2}{r_{12}^2}\,\hat{r}_{12}$$

where G is the Universal Gravitational Constant, r_{12} is the distance between the two masses and $\hat{r}_{12}$ is the unit vector pointing *from* #1 *to* #2. The *electrostatic* force $\vec{\boldsymbol{F}}^E_{12}$ between two *charges* q_1 and q_2 is of exactly the same form:

$$\vec{\boldsymbol{F}}^E_{12} \;=\; k_E\,\frac{q_1 q_2}{r_{12}^2}\,\hat{r}_{12} \qquad (1)$$

where k_E is some constant to make all the units come out right [allow me to sidestep this can of worms for now!]. Simple, eh? This force law, also known as the Coulomb force,[1] has al-

[1] The Coulomb force law, like the "coulomb" unit for electric charge (to be discussed later), is named after a guy

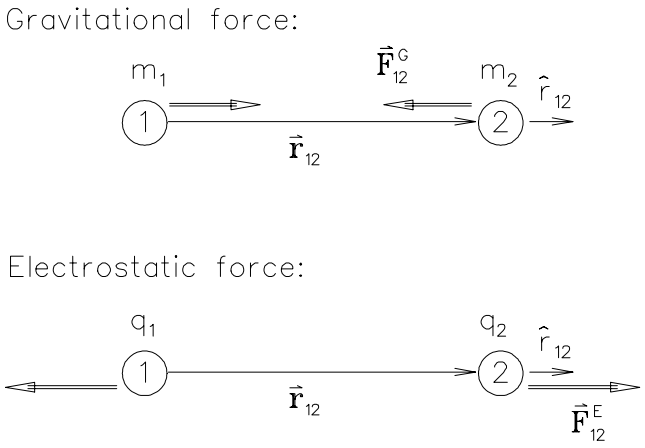


Figure 17.1 Comparison of the gravitational force $\vec{\boldsymbol{F}}_{12}^{G}$ on mass m_2 due to mass m_1 and the electrostatic (Coulomb) force $\vec{\boldsymbol{F}}_{12}^{E}$ on electric charge q_2 due to charge q_1.

most the same qualitative earmarks as the force of gravity: the force is "central" — *i.e.* it acts along the line joining the centres of the charges — and drops off as the inverse square of the distance between them; it is also proportional to each of the charges involved. [We could think of *mass* as a sort of "gravitational charge" in this context.]

So what are the "minor differences?" Well, the first one is in the *sign*. Both "coupling constants" (G and k_E) are defined to be *positive*; therefore the $-$ sign in the first equation tells us that the gravitational force $\vec{\boldsymbol{F}}_{12}^{G}$ on mass #2 is in the *opposite direction* from the unit vector $\hat{r}_{12}$ pointing form #1 to #2 — *i.e.* the force is *attractive*, back toward the other body. All masses attract all other masses gravitationally; there are (so far as we know) no repulsive forces in gravity. Another way of putting it would be to say that "there are no negative masses." By contrast, electric charges come in both positive (+) and negative (−) varieties; moreover, Eq. (1) tells us that the electrical force $\vec{\boldsymbol{F}}_{12}^{E}$ on charge #2 is in the *same* direction as $\hat{r}_{12}$ as long as the product $q_1 q_2$ is positive — *i.e.*

charges of *like* sign [*both* + or *both* −] *repel*

whereas *unlike* charges *attract*.

This means that a positive charge and a negative charge of equal magnitude will get pulled together until their net charge is zero, whereupon they "neutralize" each other and cease interacting with all *other* charges. To a good approximation, this is just what happens! Most macroscopic matter is electrically *neutral*, meaning that it has the positive and negative charges pretty much piled on top of each other.[2]

The second "minor difference" between electrical and gravitational forces is in their *magnitudes*. Of course, each depends on the size of the "coupling constant" [G for gravity *vs.* k_E for electrostatics] as well as the sizes of the "sources" [m_1 and m_2 for gravity *vs.* q_1 and q_2 for electrostatics] so any discussion of magnitude has to be in reference to "typical" examples. Let's choose the heaviest stable elementary particle that has both charge and mass: the *proton*, which constitutes the nucleus of a hydrogen atom.[3] A proton has a positive charge of

$$e = 1.60217733(49) \times 10^{-19} \text{ C (coulomb)} \quad (2)$$

[Don't worry about what a coulomb is just yet.] and a mass of

$$m_p = 1.6726231(10) \times 10^{-27} \text{ kg} \quad (3)$$

called Coulomb; $\mathcal{E}\&\mathcal{M}$ units are littered with the names of the people who invented them or discovered related phenomena. Generally I find this sort of un-mnemonic naming scheme counterdidactic, but since we have no experiential referents in $\mathcal{E}\&\mathcal{M}$ it's as good a scheme as any.

[2] On a *microscopic* scale there are serious problems with this picture. As the two charges get closer together, the force grows bigger and bigger and the *work* required to pull them apart grows without limit; in principle, according to Classical Electrodynamics, an infinite amount of work can be performed by two opposite charges that are allowed to "fall into" each other, providing we can set up a tiny system of levers and pulleys. Worse yet, the "self energy" of a *single* charge of vanishingly small size becomes infinite in the classical limit. But I am getting ahead of myself again....

[3] Now I am *'way* ahead of myself; but we do need something for an example here!

For *any* separation distance r, two protons *attract* each other (gravitationally) with a force whose magnitude F_G is $\frac{G\,m_p^2}{k_E\,e^2}$ times the magnitude F_E of the (electrostatic) force with which they *repel* each other. This ratio has an astonishing value of 0.80915×10^{-36} — the gravitational attraction between the two protons is roughly a *trillion trillion trillion* times *weaker* than the electrostatic repulsion. The electrical force wins, hands down. However, *in spite of its phenomenal puniness, gravity can overcome all other forces if enough mass gets piled up in one place.* This feature will be discussed at length later on, but for now it is time to discuss the basic *magnetic* force.

17.1.2 The Magnetic Force

As we shall see later, the "Laws" of $\mathcal{E}\&\mathcal{M}$ are so symmetric between electrical and magnetic phenomena that most Physicists are extremely frustrated by the fact that no one has ever been able to conclusively demonstrate the existence (other than theoretical) of a "magnetic charge" (also known as a *magnetic monopole*). If there were magnetic charges, the magnetic force equation would look just like the gravitational and electrostatic force laws above and this part of the description would be nice and simple. Alas, this is not the case. Static (constant in time) magnetic phenomena are generated instead by the steady *motion* of electric charges, represented by a *current* I (the charge passing some fixed point per unit time) in some direction $\vec{\boldsymbol{\ell}}$. Usually (at least at the outset) we talk about currents flowing in a *conductor* (*e.g.* a wire) through which the charges are free to move with minimal resistance. Then $\vec{\boldsymbol{\ell}}$ is a vector length pointing along the wire, or (if the wire is curved) $d\vec{\boldsymbol{\ell}}$ is an infinitesimal *element* of the wire at some point. We may then think in terms of a "current element" $I d\vec{\boldsymbol{\ell}}$.

One such current element $I_1 d\vec{\boldsymbol{\ell}}_1$ exerts a magnetic force $\mathrm{d}\vec{\boldsymbol{F}}_{12}^{\,M}$ on a second current element $I_2 d\vec{\boldsymbol{\ell}}_2$ at a distance $\vec{\boldsymbol{r}}_{12}$ (the vector from #1 to #2) given by

$$d\vec{\boldsymbol{F}}_{12}^{\,M} \;=\; k_M\,\frac{I_1 I_2}{r_{12}^2}\,d\vec{\boldsymbol{\ell}}_2 \times (d\vec{\boldsymbol{\ell}}_1 \times \hat{r}_{12}) \qquad (4)$$

where k_M is yet another unspecified constant to make all the units come out right [just wait!] and again $\hat{r}_{12}$ is the *unit vector* defining the direction *from* #1 *to* #2.

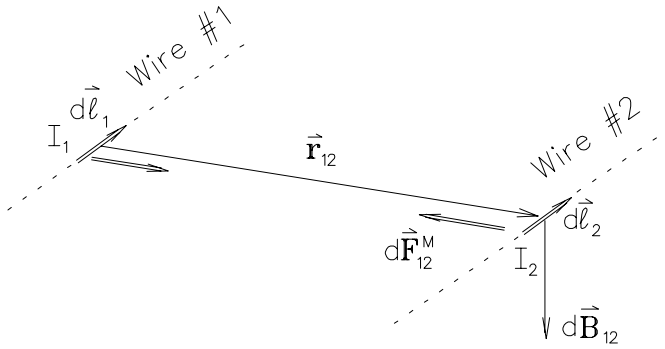


Figure 17.2 The magnetic force $d\vec{\boldsymbol{F}}_{12}^{\,M}$ on current element $I_2 d\vec{\boldsymbol{\ell}}_2$ due to current element $I_1 d\vec{\boldsymbol{\ell}}_1$.

This ugly equation (4) does give us some important qualitative hints about the force between two current-carrying wires: the force between any two *elements* of wire drops off as the inverse square of the distance between them, just like the gravitational and electrostatic forces [although this isn't much use in guessing the force between real current-carrying wires, which don't come in infinitesimal lengths] and the force is in a direction perpendicular to both wires. In fact, if we are patient we can see which way the magnetic forces will act between two *parallel* wires: we can visualize a distance vector $\vec{\boldsymbol{r}}$ from the first wire (#1) over to the second wire (#2); let it be perpendicular to both for convenience. The "RIGHT-HAND RULE" will then tell us the *direction* of $(d\vec{\boldsymbol{\ell}}_1 \times \hat{r}_{12})$: if we "turn the screw" in the sense of cranking through the angle *from* $d\vec{\boldsymbol{\ell}}_1$ *to* $\hat{r}_{12}$), a right-handed screw [the conventional kind] would move in the direction labelled $d\vec{\boldsymbol{B}}_{12}$ in

Fig. 17.2. This is the direction of $(d\vec{\boldsymbol{\ell}}_1 \times \hat{r}_{12})$. Now if we crank $d\vec{\boldsymbol{\ell}}_2$ into $d\vec{\boldsymbol{B}}_{12}$, the turn of the screw will cause it to head back toward the first wire! Simple, eh?

Seriously, *no one* is particularly enthused over this equation! All anyone really retains from this intricate exercise is the following pair of useful rules:

1. Two parallel wires with electrical currents flowing in the *same* direction will *attract* each other.

2. Two parallel wires with electrical currents flowing in *opposite* directions will *repel* each other.

Nevertheless, electrical engineers and designers of electric motors and generators need to know just what sorts of forces are exerted by one complicated arrangement of current-carrying wires on another; moreover, once it had been discovered that moving charges create this weird sort of action-at-a-distance, no one wanted to just give up in disgust and walk away from it. What can we possibly do to make magnetic calculations manageable? Better yet, is there any way to make this seem more *simple*?

17.2 Fields

In Classical Mechanics we found several conceptual aids that not only made calculations easier by skipping over inessential details but also made it possible to carry around the bare essence of Mechanics in our heads in a small number of compact "Laws." This is generally regarded as a good thing, although of course we pay a price for every entrenched paradigm — we may lose the ability (if we ever had it!) to "see things as they are" without filtering our experience through models. I will leave that debate to the philosophers, psychologists and mystics; it is true even in Physics, however, that the more successful the paradigm the bigger the blind spot it creates for alternative descriptions of the same phenomena. This bothers most Physicists, too, but there doesn't seem to be a practical alternative; so we content ourselves with maintaining an awareness of our own systematic prejudices.

Perhaps the best example of this from the days of "Classical" Physics [*i.e.* before Relativity and Quantum Mechanics rained confusion down on all of us] is the invention of the Electric and Magnetic Fields, written $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$, respectively. The idea of fields is to break down the nasty problems described in the previous Section into two easier parts:

1. First, calculate the field due to the *source* charge or current.

2. Then calculate the *force* on a *test* charge or current *due to* that field.

This also makes it a lot easier to organize our calculations in cases where the *sources* are complicated arrays of charges and/or currents. Here's how it works:

17.2.1 The Electric Field

The electric field $\vec{\boldsymbol{E}}$ at any point in space is defined to be the *force per unit test charge* due to all the other charges in the universe. That is, there is probably no "test charge" q there to experience any force, but *if there were* it would experience a force

$$\vec{\boldsymbol{F}}_E \;=\; q\,\vec{\boldsymbol{E}} \tag{5}$$

Note that since the force is a vector, $\vec{\boldsymbol{E}}$ is a *vector field*.

Since by definition $\vec{\boldsymbol{E}}$ is there even if there *isn't* any test charge present, it follows that there is an electric field at every point in space, all the

time! [It might be pretty close to zero, but it's still there!][4]

Is the ELECTRIC FIELD real? No. Yes. You decide.[5] This paradigm makes everything so much easier that most Physicists can't imagine thinking about $\mathcal{E}\&\mathcal{M}$ any other way. Does this blind us to other possibilities? Undoubtedly.

A single isolated electric "source" charge Q [I am labelling it differently from my "test" charge q just to avoid confusion. Probably that won't work.] generates a *spherically symmetric* electric field

$$\vec{\boldsymbol{E}} \;=\; k_E\,\frac{Q}{r^2}\,\hat{r} \tag{6}$$

at any point in space specified by the vector distance $\vec{\boldsymbol{r}}$ from Q to that point. That is, the field $\vec{\boldsymbol{E}}$ is *radial* [in the direction of the radius vector] and has the same *magnitude* E at all points on an imaginary spherical surface a distance r from Q.

It might be helpful to picture the *acceleration of gravity* as a similar *vector field*:

$$\vec{\boldsymbol{g}} \;=\; -G\,\frac{M_E}{r^2}\,\hat{r} \tag{7}$$

— *i.e.* $\vec{\boldsymbol{g}}$ always points back toward the centre of the Earth (mass M_E) and drops off as the inverse square of the distance r from the centre of the Earth.

17.2.2 The Magnetic Field

Any current element $Id\vec{\boldsymbol{\ell}}$ contributes $d\vec{\boldsymbol{B}}$ to the magnetic field $\vec{\boldsymbol{B}}$ at a given point in space:

$$d\vec{\boldsymbol{B}} \;=\; k_M\,\frac{I\,d\vec{\boldsymbol{\ell}}\times\hat{r}}{r^2} \tag{8}$$

where $\hat{r}$ is the unit vector in the direction of $\vec{\boldsymbol{r}}$, the vector distance from the current element to the point in space where the magnetic field is being evaluated. Eq. (8) is known as the LAW OF BIOT AND SAVART. It is still not perfectly transparent, I'm sure you will agree, but it beats Eq. (4)!

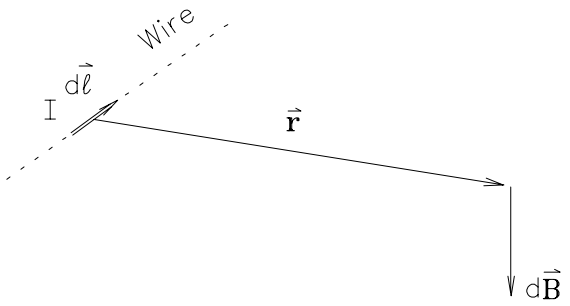


Figure 17.3 The magnetic field $d\vec{\boldsymbol{B}}$ at position $\vec{\boldsymbol{r}}$ due to a current element $Id\vec{\boldsymbol{\ell}}$ at the origin.

17.2.3 Superposition

While it may seem obvious, it bears saying that the electric fields due to several different "source" charges or the magnetic fields due to several different "source" current elements are just added together (vectorially, of course) to make the net $\vec{\boldsymbol{E}}$ or $\vec{\boldsymbol{B}}$ field. Horrible as it might seem, this might in principle *not* be true — we might have to "add up" such fields in some hopelessly more complicated way. But it didn't turn out that way in this universe. Lucky us!

17.2.4 The Lorentz Force

We can now put the second part of the procedure [calculating the *forces* on a test charge due to known FIELDS] into a very compact form combining both the electric and the magnetic forces into one equation. If a particle with charge q and mass m moves with velocity $\vec{\boldsymbol{v}}$ in the combination of a uniform electric field $\vec{\boldsymbol{E}}$ and a uniform magnetic field $\vec{\boldsymbol{B}}$, the net force acting on the particle is the LORENTZ FORCE,

[4] We often try to represent this graphically by drawing "lines of force" that show which way $\vec{\boldsymbol{E}}$ points at various positions; unfortunately it is difficult to draw in $\vec{\boldsymbol{E}}$ at *all* points in space. I will discuss this some more in a later Section.

[5] Define "real."

which can be written (in one set of units)

$$\vec{\boldsymbol{F}} \;=\; q\left(\vec{\boldsymbol{E}} \;+\; \frac{\vec{\boldsymbol{v}}}{c}\times\vec{\boldsymbol{B}}\right), \qquad (9)$$

where (for now) we can think of c as just some constant with units of velocity.

If $\vec{\boldsymbol{E}} = 0$ and $\vec{\boldsymbol{v}}$ is *perpendicular* to $\vec{\boldsymbol{B}}$, the Lorentz force is perpendicular to both $\vec{\boldsymbol{B}}$ and the momentum $\vec{\boldsymbol{p}} = m\vec{\boldsymbol{v}}$. The force will deflect the momentum sideways, changing its direction but not its magnitude.[6] As $\vec{\boldsymbol{p}}$ changes direction, $\vec{\boldsymbol{F}}$ changes with it to remain ever perpendicular to the velocity — this is an automatic property of the cross product — and eventually the *orbit* of the particle closes back on itself to form a circle. In this way the magnetic field produces UNIFORM CIRCULAR MOTION with the plane of the circle perpendicular to both $\vec{\boldsymbol{v}}$ and $\vec{\boldsymbol{B}}$.

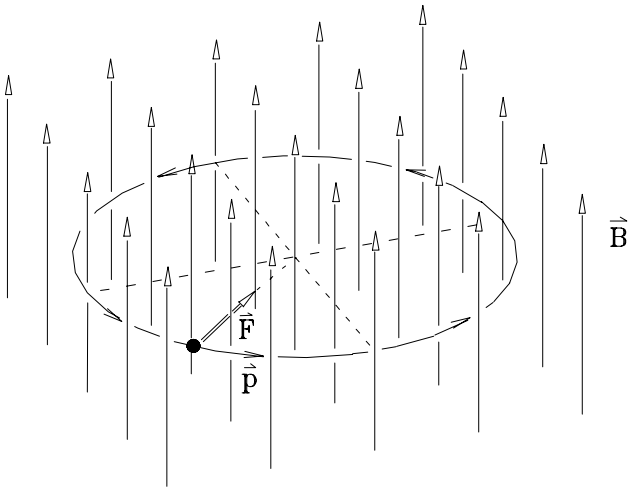


Figure 17.4 Path of a charged particle with momentum $\vec{\boldsymbol{p}}$ in a uniform, static magnetic field $\vec{\boldsymbol{B}}$ perpendicular to $\vec{\boldsymbol{p}}$.

Using Newton's SECOND LAW and a general knowledge of circular motion, one can derive a formula for the *radius* of the circle (r) in terms of the *momentum* of the particle ($p = mv$), its *charge* (q) and the magnitude of the *magnetic field* (B). In "Gaussian units" (grams, centimeters, Gauss) the formula reads[7]

$$r = \frac{pc}{qB}. \qquad (10)$$

It is also interesting to picture qualitatively what will happen to the particle if an *electric* field $\vec{\boldsymbol{E}}$ is then applied *parallel* to $\vec{\boldsymbol{B}}$: since $\vec{\boldsymbol{E}}$ accelerates the charge in the direction of $\vec{\boldsymbol{E}}$, which is also the direction of $\vec{\boldsymbol{B}}$, and since $\vec{\boldsymbol{B}}$ only produces a force when the particle moves *perpendicular* to $\vec{\boldsymbol{B}}$, in effect the "perpendicular part of the motion" is unchanged (circular motion) while the "parallel part" is unrestricted acceleration. The path in space followed by the particle will be a spiral with steadily increasing "pitch":

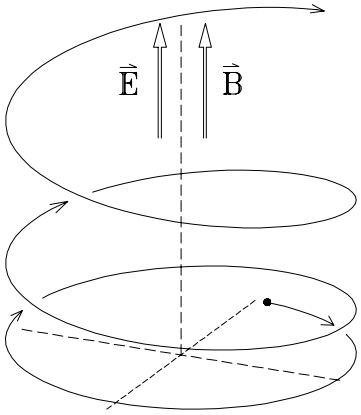


Figure 17.5 Path of a charged particle in *parallel* $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ fields.

[6] A force perpendicular to the motion does no work on the particle and so does not change its kinetic energy or speed — recall the general qualitative features of CIRCULAR MOTION under the influence of a CENTRAL FORCE.

[7] In "practical" units the formula reads

$$r\,[\mathrm{cm}] \;=\; \frac{p\,[\mathrm{MeV/c}]}{0.3\;B\,[\mathrm{kG}]\;q\,[\text{electron charges}]}$$

where cm are (as usual) centimeters, MeV/c are millions of "electron volts" divided by the speed of light (believe it or not, a unit of momentum!) and kG ("kilogauss") are thousands of Gauss. I only mention this now because I will use it later on and because it illustrates the madness of electromagnetic units — see next Section!

17.2.5 "Field Lines" and Flux

In Fig. 17.4 the uniform magnetic field is pictured as a forest of little parallel arrows of equal length, equally spaced. Something like this is always necessary if we want to make a visual representation of $\vec{\boldsymbol{B}}$, but it leaves a lot to be desired. For instance, a uniform magnetic field has the same magnitude and direction at *every* point in space, not just where the lines are drawn. Moreover, as we have seen, the magnetic *force*, if any, is *never* in the direction of the "lines of $\vec{\boldsymbol{B}}$" but rather perpendicular to them, as shown in Fig. 17.4.

Nevertheless, the visual appeal of such a graphical representation in terms of "field lines" is so compelling that a whole description of $\mathcal{E}\&\mathcal{M}$ has been developed in terms of them. In that description one speaks of "lines per unit area" as a measure of the *strength* of an electric or magnetic field. The analogy is with *hydrodynamics*, the flow of incompressible fluids, in which we may actually see "lines" of fluid flow if we drop packets of dye in the water.

In fluid dynamics there is actually "stuff" flowing, a transfer of mass that has momentum and density. In that context one naturally thinks of the FLUX of material through imaginary surfaces perpendicular to the flow[8] and indeed $\vec{\boldsymbol{B}}$ is sometimes referred to as the *magnetic flux per unit* (perpendicular) *area*.

By the same token, if "lines" of $\vec{\boldsymbol{B}}$ pass through a surface of area A normal (perpendicular) to $\vec{\boldsymbol{B}}$, then we can (and do) talk about the MAGNETIC FLUX Φ through the surface; Φ has units of magnetic field times area. If we want, we can turn this around and say that a magnetic field has units of *flux per unit area*.

Even though we rarely take this "lines of $\vec{\boldsymbol{B}}$" business literally, it makes such a good image that we make constant use of it in handwaving arguments. Moreover, the concept of MAGNETIC FLUX is well ensconced in modern $\mathcal{E}\&\mathcal{M}$ terminology.

[8] For instance, the flux of a river past a fixed point might be measured in gallons per minute per square meter of area perpendicular to the flow. A hydroelectric generator will intercept twice as many gallons per minute if it presents twice as large an area to the flow. And so on.

17.3 Potentials and Gradients

Recall from MECHANICS that if we move a particle a vector distance $d\vec{\boldsymbol{\ell}}$ under the influence of a force $\vec{\boldsymbol{F}}$, that force does $dW = \vec{\boldsymbol{F}} \cdot d\vec{\boldsymbol{\ell}}$ worth of *work* on the particle — which appears as *kinetic energy*. *Etc.* If the force is due to the action of an electric field $\vec{\boldsymbol{E}}$ on a charge q, the work done is $dW = q\vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}}$. This work gets "stored up" as *potential energy* V as usual: $dV = -dW$. Just as we defined $\vec{\boldsymbol{E}}$ as the *force per unit charge*, we now define the ELECTRIC POTENTIAL ϕ to be the *potential energy per unit charge*, *viz.*

$$dV \;=\; q\,d\phi \quad \text{where} \quad d\phi \;=\; -\vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} \qquad (11)$$

or, summing the contributions from all the infinitesimal elements $\vec{\boldsymbol{\ell}}$ of a finite path through space in the presence of electric fields,[9]

$$\phi \;\equiv\; -\int \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} \qquad (12)$$

When multiplied by q, ϕ gives the potential *energy* of the charge q in the electric field $\vec{\boldsymbol{E}}$.

Just as we quickly adapted our formulation of MECHANICS to use *energy* (potential and kinetic) as a starting point instead of *force*, in $\mathcal{E}\&\mathcal{M}$ we usually find it easier to start from $\phi(\vec{\boldsymbol{r}})$ as a function of position ($\vec{\boldsymbol{r}}$) and derive $\vec{\boldsymbol{E}}$ the same way we did in MECHANICS:

$$\vec{\boldsymbol{E}} \;\equiv\; -\vec{\boldsymbol{\nabla}}\phi \qquad (13)$$

[9] Note that, just as in the case of the mechanical potential energy V, the *zero* of ϕ is chosen arbitrarily at some point in space; we are really only sensitive to *differences* in potential. However, for a *point charge* it is conventional to choose an infinitely distant position as the zero of the electrostatic potential, so that $\phi(r)$ for a point charge Q is the work required to bring a unit test charge up to a distance r away from Q, starting at infinite distance.

where, as before,[10]

$$\vec{\boldsymbol{\nabla}} \equiv \hat{x}\,\frac{\partial}{\partial x} + \hat{y}\,\frac{\partial}{\partial y} + \hat{z}\,\frac{\partial}{\partial z} \qquad (14)$$

The most important example is, of course, the electric potential due to a single "point charge" Q at the origin:

$$\phi(\vec{\boldsymbol{r}}) = k_E\,\frac{Q}{r} \qquad (15)$$

Note that $\phi(r) \to 0$ as $r \to 0$, as discussed in the previous footnote. This is a convenient convention. I will leave it as an exercise for the enthusiastic reader to show that

$$\vec{\boldsymbol{\nabla}}\left(\frac{1}{r}\right) = -\frac{\hat{r}}{r^2}.$$

Electric potential is most commonly measured in *volts* (abbreviated V) after Count Volta, who made the first useful batteries. We often speak of the "voltage" of a battery or an appliance. [The latter does not ordinarily have any electric potential of its own, but it is designed to be *powered* by a certain "voltage." A light bulb would be a typical case in point.] The *volt* is actually such a familiar unit that eletric *field* is usually measured in the derivative unit, *volts per meter* (V/m). It really is time now to begin discussing *units* — what are those constants k_E and k_M, for instance? But first I have one last remark about *potentials*.

The electrostatic potential ϕ is often referred to as the Scalar Potential, which immediately suggests that there must be such a thing as a Vector Potential too. Just so. The Vector Potential $\vec{\boldsymbol{A}}$ is used to calculate the *magnetic* field $\vec{\boldsymbol{B}}$ but not quite as simply as we get $\vec{\boldsymbol{E}}$ from $\vec{\boldsymbol{\nabla}}\phi$. In this case we have to take the "curl" of $\vec{\boldsymbol{A}}$ to get $\vec{\boldsymbol{B}}$:

$$\vec{\boldsymbol{B}} = \vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{A}}. \qquad (16)$$

Never mind this now, but we will get back to it later.

[10] Remember the metaphor of $\vec{\boldsymbol{\nabla}}\phi$ as the "slope" of a "hill" whose height is given by $\phi(\vec{\boldsymbol{r}})$.

17.4 Units

When Physicists are working out problems "formally" (that is, trying to understand "how things behave") they are usually only concerned with deriving a formula which describes the behaviour, not so much with getting "numbers" out of the formula. This is why we can tolerate so much confusion in the details of the alternate electromagnetic unit systems. We never actually calculate any "answers" that an engineer could use to build devices with; we simply derive a formula for such calculations, preferably in a form as free of specific units as possible, and leave the practical details up to the engineer (who may be us, later).

So I have left the unspecified "coupling constants" k_E and k_M undefined while we talked about the *qualitative* behaviour of electric and magnetic fields. Now we finally have to assign some *units* to all these weird quantities.

The history of *units* in $\mathcal{E}\&\mathcal{M}$ is a long horror story. It isn't even very entertaining, at least to my taste. Numerous textbooks provide excellent summaries of the different systems of units used in $\mathcal{E}\&\mathcal{M}$ [there are at least three!] but even when one understands perfectly there is not much satisfaction in it. Therefore I will provide only enough information on $\mathcal{E}\&\mathcal{M}$ units to define the unavoidable units one encounters in everyday modern life and to allow me to go on to the next subject.

As long as electric and magnetic fields are not both involved in the same problem, one can usually stick to familiar units expressed in a reasonably clear fashion. Let's discuss them one at a time.

17.4.1 Electrical Units

I will give the old-fashioned version of this saga, in which one picks either Volts or Coulombs as the "fundamental" unit and derives the rest from that. Today the Ampere [A] is actu-

ally the most basic unit; it is *defined* to be the current required to flow in *both* of two "very long" parallel wires 1 *m* apart in order to give a *magnetic force per unit length* of exactly 2×10^{-7} *N/m* acting on each wire. No, I'm not kidding. Then the COULOMB [C] is defined as the electric charge that flows past any point in 1 *s* when a steady current of 1 A is maintained in a wire. *I.e.* we have 1 C = 1 A-*s*. Anyway, I will start with COULOMBS because it is more mnemonic.

Coulombs and Volts

As indicated in Eq. (2), electric *charge* is usually measured in COULOMBS (abbreviated C). If we take this as a fundamental unit, we can analyze the definition of the *volt* (V) by reference to Eq. (11): moving a charge of $q = 1$ C through an electric potential difference $\Delta\phi = 1$ V produces a potential energy difference of $\Delta V = 1$ J. Therefore

a VOLT is a *joule per coulomb*.

If we prefer to think of the *volt* as a more fundamental unit, we can turn this around and say that

a COULOMB is a *joule per volt*.

However, I think the former is a more comfortable definition.

Electron Volts

We can also take advantage of the fact that Nature supplies electric charges in integer multiples of a fixed quantity of charge[11] to define some more "natural" units. For instance, the electric charge of an electron is $-e$ [where e is the charge of a proton, defined in Eq. (2)]. An ELECTRON VOLT (eV) is the kinetic *energy* gained by an electron [or any other particle with the same size charge] when it is accelerated through a one volt (1 V) electric potential. Moving a charge of 1 C through a potential of 1 V takes 1 J of work (and will produce 1 J of kinetic energy), so we know immediately from Eq. (2) that

$$1 \text{ eV} \;=\; 1.60217733(49) \times 10^{-19} \text{ J} \qquad (17)$$

This is not much energy if you are a toaster, but for an electron (which is an *incredibly tiny* particle) it is enough to get it up to a velocity of 419.3828 km/s, which is 0.14% of the speed of light! Another way of looking at it is to recall that we can express *temperature* in energy units using Boltzmann's constant as a conversion factor. You can easily show for yourself that 1 eV is equivalent to a temperature of 11,604 degrees Kelvin or about 11,331°C. So in the microscopic world of electrons the eV is a pretty convenient (or "natural") unit. But not in the world of toasters and light bulbs. So let's get back to "conventional" units.

Amperes

Electric currents (the rate at which charges pass a fixed point in a wire, for instance) have dimensions of *charge per unit time*. If the COULOMB is our chosen unit for electric charge and we retain our fondness for *seconds* as a time unit, then *current* must be measured in *coulombs per second*. We call these units AMPERES or Amps [abbreviated A] after a Frenchman named Ampère. Thus

$$1\,\text{A [AMPERE]} \;\equiv\; 1\,\text{C/s [COULOMB per second]} \qquad (18)$$

I have a problem with Amps. It makes about as much sense to give the coulomb per second its own name as it would to make up a name for meters per second. No one frets over the complexity of expressing speed in m/s or kph or whatever — in fact it serves as a good re-

[11] This is what we mean when we say that charge is *quantized*.

minder that velocity is a rate of change of distance with time — but for some reason we feel obliged to give C/s their own name. Ah well, it is probably because all this electrical stuff is so weird.[12] Whatever the reason, we are stuck with them now!

The Coupling Constant

We are now ready to define our electrical "coupling constant" k_E. Referring to Eq.(15) we have

$$\phi\,[V] \;=\; k_E\,\frac{Q\,[C]}{r\,[m]}$$

which we can rearrange to read

$$k_E \;=\; \frac{\phi\,[V]\;\cdot\;r\,[m]}{Q\,[C]}$$

Thus k_E must have *dimensions* of {electric potential times distance per unit charge}; we can pick *units* of V-m/C to stick with convention. This still doesn't tell us the *value* of k_E. This must be *measured*. The result is

$$k_E \;=\; 8.98755\cdots\times 10^9\ \text{V-m/C} \qquad (19)$$

17.4.2 Magnetic Units

Gauss *vs.* Tesla

There are two "accepted" units for the magnetic field $\vec{\boldsymbol{B}}$: GAUSS [abbreviated G] and TESLA [abbreviated T]. Needless to say, both are named after great $\mathcal{E}\&\mathcal{M}$ researchers. The former is handy when describing *weak* magnetic fields — for instance, the Earth's magnetic field is on the order of 1 G — but the unit that goes best with our selected electrical units (because it is defined in terms of meters and coulombs and seconds) is the TESLA. Fortunately the conversion factor is simple:

$$1\ \text{T} \equiv 10{,}000\ \text{G}.$$

[12] And also, I suspect, because people were looking for a good way to honour the great Physicist Ampère and all the best units were already taken.

The TESLA is also defined in terms of the WEBER [W] (named after guess whom), a conventional unit of magnetic *flux*. The definition is

$$1\ \text{T} \equiv 1\ \text{W/m}^2 \qquad \text{or} \qquad 1\ \text{W} = 1\ \text{T}\times 1\ \text{m}^2$$

if you're interested. So referring back to Eq. (8), we have

$$B\,[\text{TESLA}] \;=\; k_M\,\frac{I\,[\text{A}]\;d\ell[\text{m}]}{(r[\text{m}])^2}$$

which we can rearrange to read

$$k_M \;=\frac{B\,[\text{TESLA}]\;(r\,[\text{m}])^2}{I\,[\text{A}]\;d\ell\,[\text{m}]}$$

so that k_M must have *dimensions* of magnetic flux [Webers (W)] per unit current [Amp (A)] per unit length or *units* of W/A-m. Its *value* is again determined by experiment:

$$k_M \;\equiv\; \frac{\mu_0}{4\pi} \;=\; 10^{-7}\ \text{W/A-m} \qquad (20)$$

where μ_0 is the PERMEABILITY OF FREE SPACE.

I will leave it as an exercise for the student to plug these coupling constants back into the equations where they appear and show that everything is, though weird, dimensionally consistent.

17.5 Exercises

17.5.1 Rod of Charge

As an exercise in the "brute force" integration of COULOMB'S LAW (unavoidable in many cases), here is one way to find the electric field due to a uniformly charged, skinny rod of finite length L:

If the total charge Q is uniformly distributed along the rod, then the charge per unit length is

$$\lambda = \frac{Q}{L}\ . \qquad (21)$$

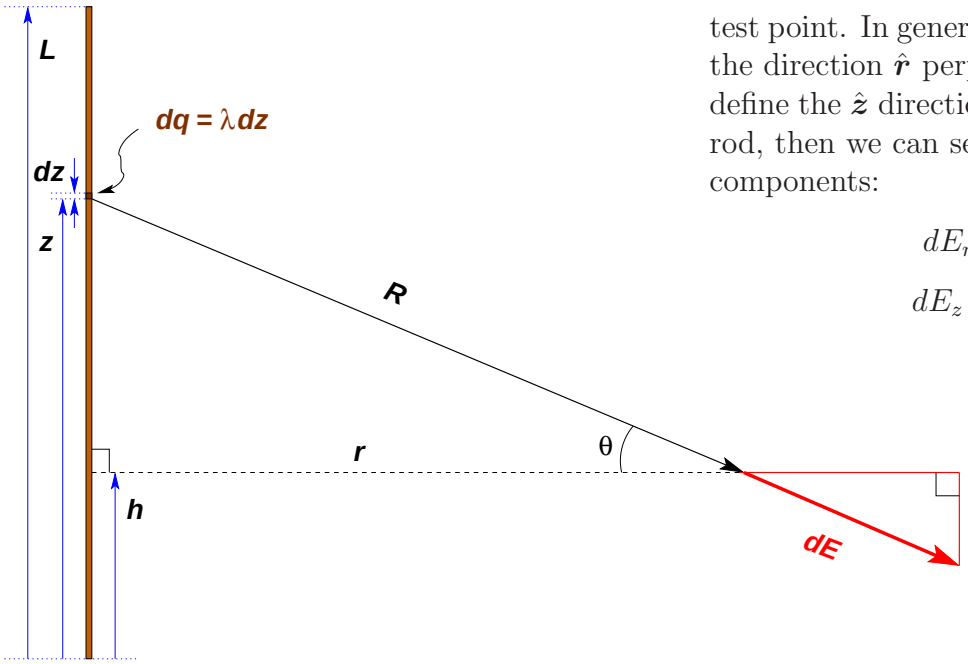


Figure 17.6 Electric field due to a uniformly charged rod of finite length.

We want to evaluate the electric field $\vec{\boldsymbol{E}}$ at an arbitrary "test point" in space. Such a point can be characterized completely by its perpendicular distance r from the rod and its distance h up from the bottom end of the rod, measured parallel to the rod, as shown. We choose to look at the rod and the point in their common plane.

Then we pick an arbitrarily position a distance z up from the bottom end of the rod, as shown. A small slice of the rod (width dz) at that position contains a "charge element" $dq = \lambda\, dz$ which contributes $d\vec{\boldsymbol{E}}$ to the electric field vector $\vec{\boldsymbol{E}}$ at the test point. Coulomb's Law says that $d\vec{\boldsymbol{E}}$ points away from the charge element (assuming positive charge) and has a magnitude

$$dE = \frac{k_E \lambda\, dz}{R^2} \tag{22}$$

where

$$R = \sqrt{r^2 + (z-h)^2} \tag{23}$$

is the distance from the charge element to the test point. In general $d\vec{\boldsymbol{E}}$ makes an angle θ with the direction $\hat{\boldsymbol{r}}$ perpendicular to the rod. If we define the $\hat{\boldsymbol{z}}$ direction to be "up" parallel to the rod, then we can separate $d\vec{\boldsymbol{E}}$ until its r and z components:

$$dE_r = dE\ \cos\theta \tag{24}$$

$$dE_z = -dE\ \sin\theta \tag{25}$$

To integrate these equations we need to convert all variables to match the differential (over which we integrate). We could use Eq. (41) to express R in terms of z (where r and h are constants) and use

$$\cos\theta = \frac{r}{R} \tag{26}$$

$$\sin\theta = \frac{(z-h)}{R} \tag{27}$$

but this would leave us with integrals that cannot be solved by inspection.

If we want to solve this problem without reference to external aids (like tables of integrals), it is better to convert into angles and trigonometric functions as follows:

Equation (42) can be rewritten

$$\frac{1}{R^2} = \frac{\cos^2\theta}{r^2} \tag{28}$$

and since

$$z - h = r\tan\theta\ , \tag{29}$$

giving

$$dz = r\sec^2\theta\ d\theta = \frac{r\, d\theta}{\cos^2\theta}\ , \tag{30}$$

we can write Eq. (22) as

$$dE = k_E\lambda \left(\frac{\cos^2\theta}{r^2}\right)\left(\frac{r\,d\theta}{\cos^2\theta}\right) = \frac{k_E\lambda}{r}\,d\theta \quad (31)$$

and from that, Eqs. (24) and (25), respectively, as

$$dE_r = \frac{k_E\lambda}{r}\,\cos\theta\;d\theta = \frac{k_E\lambda}{r}\,du \qquad (32)$$

where $u \equiv \sin\theta$, and

$$dE_z = -\frac{k_E\lambda}{r}\,\sin\theta\;d\theta = \frac{k_E\lambda}{r}\,dv \qquad (33)$$

where $v \equiv \cos\theta$.

Integrating these differentials is trivial; we are left with just the differences between u (or v) at the limits of integration (the top and bottom of the rod):

$$E_r = \frac{k_E\lambda}{r}\left[\frac{(L-h)}{\sqrt{r^2+(L-h)^2}} + \frac{h}{\sqrt{r^2+h^2}}\right] \qquad (34)$$

(note that u is negative at the bottom) and

$$E_z = k_E\lambda\left[\frac{1}{\sqrt{r^2+(L-h)^2}} - \frac{1}{\sqrt{r^2+h^2}}\right] \qquad (35)$$

These equations express a completely general solution to this problem.

Let's check to see what these give for the field directly out from the *midpoint* of the rod — *i.e.* for $h = L/2$:

$$\begin{aligned} E_r &= \frac{k_E\lambda}{r}\left[\frac{L/2}{\sqrt{r^2+L^2/4}} + \frac{L/2}{\sqrt{r^2+L^2/4}}\right] \\ &= \frac{k_E\lambda}{r}\,\frac{L}{\sqrt{r^2+L^2/4}} \end{aligned} \qquad (36)$$

and

$$\begin{aligned} E_z &= k_E\lambda\left[\frac{1}{\sqrt{r^2+L^2/4}} - \frac{1}{\sqrt{r^2+L^2/4}}\right] \\ &= 0\,. \end{aligned} \qquad (37)$$

Let's also check to see what we get for E_r (at the midpoint) very far from the rod ($r \gg L$):

$$E_r \underset{r\to\infty}{\longrightarrow} \frac{k_E\lambda L}{r^2} = \frac{k_E Q}{r^2} \qquad (38)$$

(*i.e.* Coulomb's Law) $^{\surd}$

and very close to the rod ($r \ll L$):

$$E_r \underset{r\to 0}{\longrightarrow} \frac{2k_E\lambda}{r}\,. \qquad (39)$$

The last result can be used as the field due to an *infinitely long* uniform line of charge. But there is a much easier way to obtain it....

17.5.2 Rod of Current

As an exercise in the "brute force" integration of Law of Biot & Savart (unavoidable in many cases), here is one way to find the magnetic field due to a current I flowing in a skinny rod of finite length L. (Of course the current cannot just start at one end and stop at the other, but the field due to the rest of the circuit can be added in separately.)

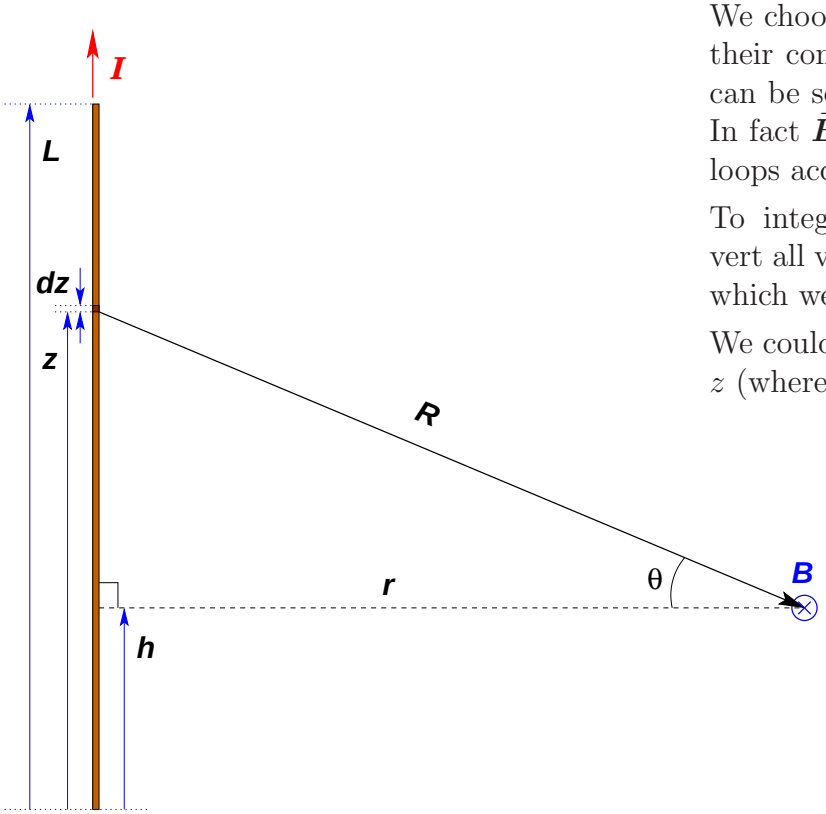


Figure 17.7 Magnetic field due to a current-carrying rod of finite length.

According to the Law of Biot & Savart, the current element $I\,dz$ shown in the Figure contributes

$$d\vec{\boldsymbol{B}} = \frac{\mu_0}{4\pi}\, I\, dz \left(\frac{\hat{k} \times \hat{\boldsymbol{R}}}{R^2} \right) \tag{40}$$

to the magnetic field $\vec{\boldsymbol{B}}$ at the "test point" shown, where

$$R = \sqrt{r^2 + (z-h)^2} \tag{41}$$

is the distance from the current element to the test point, r is its perpendicular distance from the rod and h is its distance up from the bottom end of the rod, measured parallel to the rod. The "test point" is characterized completely by r and h.

We choose to look at the rod and the point in their common plane. Thus the *direction* of $\vec{\boldsymbol{B}}$ can be seen by inspection to be into the page. In fact $\vec{\boldsymbol{B}}$ circulates around the rod in circular loops according to the Right Hand Rule.

To integrate equation (40) we need to convert all variables to match the differential (over which we integrate).

We could use Eq. (41) to express R in terms of z (where r and h are constants) and use

$$\cos\theta = \frac{r}{R} \tag{42}$$

$$\sin\theta = \frac{(z-h)}{R} \tag{43}$$

but this would leave us with integrals that cannot be solved by inspection.

If we want to solve this problem without reference to external aids (like tables of integrals), it is better to convert into angles and trigonometric functions as follows:

Equation (42) can be rewritten

$$\frac{1}{R^2} = \frac{\cos^2\theta}{r^2} \tag{44}$$

and since

$$z - h = r\tan\theta\ , \tag{45}$$

giving

$$dz = r \sec^2\theta \, d\theta = \frac{r \, d\theta}{\cos^2\theta} \,, \tag{46}$$

we can write Eq. (40) as

$$\begin{aligned} dB &= \frac{\mu_0 I}{4\pi} \cos\theta \left(\frac{\cos^2\theta}{r^2}\right)\left(\frac{r \, d\theta}{\cos^2\theta}\right) \\ &= \frac{\mu_0 I}{4\pi r} \cos\theta d\theta \end{aligned} \tag{47}$$

or

$$dB = \frac{\mu_0 I}{4\pi r} \, du \tag{48}$$

where $u \equiv \sin\theta$.

Integrating this differential is trivial; we are left with just the difference between u at the limits of integration (the top and bottom of the rod):

$$B = \frac{\mu_0 I}{4\pi r} \left[\frac{(L-h)}{\sqrt{r^2 + (L-h)^2}} + \frac{h}{\sqrt{r^2 + h^2}}\right] \tag{49}$$

(note that u is negative at the bottom).

This equation expresses a completely general solution to this problem.

Let's check to see what this gives for the field directly out from the *midpoint* of the rod — *i.e.* for $h = L/2$:

$$\begin{aligned} B &= \frac{\mu_0 I}{4\pi r} \left[\frac{L/2}{\sqrt{r^2 + L^2/4}} + \frac{L/2}{\sqrt{r^2 + L^2/4}}\right] \\ &= \frac{\mu_0 I}{4\pi r} \frac{L}{\sqrt{r^2 + L^2/4}} \end{aligned} \tag{50}$$

Let's also check to see what we get (at the midpoint) very far from the rod ($r \gg L$):

$$B \underset{r\to\infty}{\longrightarrow} \frac{\mu_0 I}{4\pi r} \frac{L}{r} = \frac{\mu_0 I L}{4\pi r^2} \tag{51}$$

and very close to the rod ($r \ll L$):

$$B \underset{r\to 0}{\longrightarrow} \frac{\mu_0 I}{4\pi r} \frac{L}{L/2} = \frac{\mu_0 I}{2\pi r} \,. \tag{52}$$

The last result can be used as the field due to an *infinitely long* current-carrying wire. But there is a much easier way to obtain it....

Note that the general formula (49) (and the right-hand rule to determine directions) can be used to find the net $\vec{\boldsymbol{B}}$ from a circuit composed of *any arrangement of straight wire segments* carrying a current I, by the principle of superposition. Note also, however, that the result is a *vector* sum.

Chapter 18

Gauss' Law

If you go on in Physics you will learn all about Gauss' Law along with vector calculus in your advanced course on Electricity and Magnetism, where it is used to calculate the electric field strength at various distances from highly symmetric distributions of electric charge. However, Gauss' Law can be applied to a huge variety of interesting situations having nothing to do with electricity except by analogy. Moreover, the rigourous statement of Gauss' Law in the mathematical language of vector calculus is not the only way to express this handy concept, which is one of the few powerful modern mathematical tools which can be accurately deduced from "common sense" and which really follows from a statement so simple and obvious as to seem trivial and uninteresting, to wit:

(Colloquial form of Gauss' Law)

"When something passes out of a region, it is no longer inside that region."

How, you may ask, can such a dumb tautology teach us anything we don't already know? The power of Gauss' Law rests in its combination with our knowledge of *geometry* (*e.g.* the surface area A of a sphere of radius r is $A = 4\pi r^2$) and our instinctive understanding of *symmetry* (*e.g.* there is no way for a *point* of zero size to define a favoured *direction*). When we put these two skills together with Gauss' Law we are able to easily derive some not-so-obvious *quantitative* properties of many commonly-occurring natural phenomena.

18.1 The Point Source

For example, consider a hypothetical "spherically symmetric" *sprinkler head* (perhaps meant to uniformly irrigate the inside surface of a hollow spherical space colony): located at the centre of the sphere, it "emits" (squirts out) dQ/dt gallons per second of water *in all directions equally*, which is what we mean by "spherically symmetric" or "*isotropic*."[1] Here Q is the "amount of stuff" — in this case measured in gallons. Obviously (beware of that word, but it's OK here), since water is *conserved* the total *flow* of water is conserved: once a "steady-state" (equilibrated) flow has been established, the rate at which water is deposited on the walls of the sphere is the same as the rate at which water is emitted from the sprinkler head at the centre. That is, if we add up (integrate) the "flux" $\vec{\boldsymbol{J}}$ of water per second per square meter of surface area at the sphere wall over the whole spherical surface, we must get dQ/dt. Mathematically, this is written

$$\oiint_{\mathcal{S}} \vec{\boldsymbol{J}} \cdot d\vec{\boldsymbol{A}} = \frac{dQ}{dt} \qquad (1)$$

[1] Note how our terminology of spherical coordinates stems from terrestrial navigation (Tropics of Cancer, Capricorn, *etc.*). Since the 16th Century, our most familiar spherical object (next to the cannonball) has been the Earth.

where the $\oiint_{\mathcal{S}}$ stands for an integral (sum of elements) over a closed surface $\mathcal{S}$. [This part is crucial, inasmuch as an open surface (like a hemisphere) does not account for all the flux and cannot be used with GAUSS' LAW]. Now, we must pay a little attention to the *vector* notation: the "flux" $\vec{\boldsymbol{J}}$ always has a *direction*, like the flux (current) of water flowing in a river or in this case the flux of water droplets passing through space.

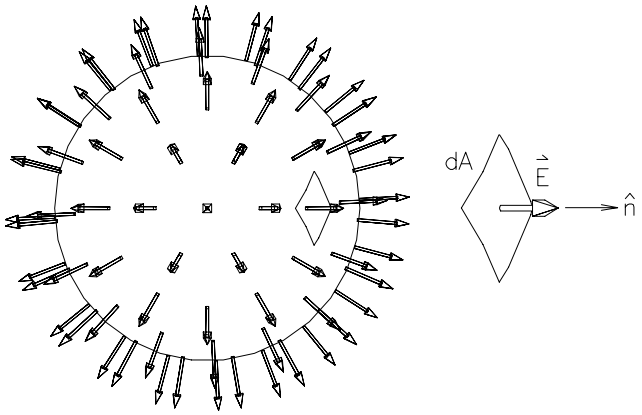


Figure 18.1 An isotropic source.

Each droplet has a (vector) velocity, and the velocity and the density of droplets combine to form the "flux" as described above. Not so trivial is the idea of a vector *area* element $d\vec{\boldsymbol{A}}$, but the sense of this is clear if we think of what happens to the scalar flux (in gallons/sec) through a hoop of wire of area $d\vec{\boldsymbol{A}}$ when we place it in a river: if the direction of the flow of the river is perpendicular ("normal") to the plane of the hoop, we get the maximum possible flux, namely the vector flux magnitude (the flow rate of the river) times the area of the hoop; if we reorient the hoop so that its area intercepts no flow (*i.e.* if the direction $\hat{n}$ "normal" to the plane of the hoop is perpendicular to the direction of flow of the river) then we get zero flux through the hoop. In general, the scalar rate of flow (here measured in gallons/sec) through a "surface element" $d\vec{\boldsymbol{A}}$ whose "normal" direction $\hat{n}$ is given by $(\vec{\boldsymbol{J}}\cdot\hat{n})dA$ or just $\vec{\boldsymbol{J}}\cdot d\vec{\boldsymbol{A}}$ where we have now defined the *vector* surface element $d\vec{\boldsymbol{A}} \equiv \hat{n}dA$. This is pictured in Fig. 18.1 above.

Returning now to our sprinkler-head example, we have a Law [Eq. (1)] which is a mathematical (and therefore quantitative) statement of the colloquial form, which in principle allows us to calculate something. However, it is still of only academic interest in general. Why? Because the integral described in Eq. (1) is so general that it may well be hopelessly difficult to solve, unless (!) there is something about the *symmetry* of the particular case under consideration that makes it easy, or even "trivial." Fortunately (though hardly by accident) in this case there is — namely, the isotropic nature of the sprinkler head's emission, plus the spherically symmetric (in fact, spherical) shape of the surface designated by "$\mathcal{S}$" in Eq. (1). These two features ensure that

1. the *magnitude* $J = |\vec{\boldsymbol{J}}|$ of the flux is the same everywhere on the surface $\mathcal{S}$; and

2. the *direction* of $\vec{\boldsymbol{J}}$ is *normal* to the surface everywhere it hits on $\mathcal{S}$.

In this case, $\vec{\boldsymbol{J}}\cdot d\vec{\boldsymbol{A}} = JdA$ and J is now a constant which can be taken outside the integral sign, leaving

$$J \oiint_{\mathcal{S}} dA = \dot{Q}$$

where $\dot{Q}$ is just a compact notation for dQ/dt. But $\oiint_{\mathcal{S}} dA$ is just the *area* of the sphere, $4\pi r^2$, where r is the *radius* of the sphere, so (1) becomes

$$4\pi r^2 J = \dot{Q}$$

or

$$J(r) = \frac{\dot{Q}}{4\pi r^2} \tag{2}$$

which states the *general* conclusion for *any* spherically symmetric emission of a conserved quantity, namely

> *The flux from an isotropic source points away from the centre and falls off proportional to the inverse square of the distance from the source.*

This holds in an amazing variety of situations. For instance, consider the "electric field lines" from a spherically symmetric electric charge distribution as measured at some point a distance r away from the centre. We visualize these electric field "lines" as streams of some mysterious "stuff" being "squirted out" by positive charges (or "sucked in" by negative charges). The idea of an electric field line is of course a pure construct; no one has ever seen or ever will see a "line" of the electric field $\vec{\boldsymbol{E}}$, but if we think of the *strength* of $\vec{\boldsymbol{E}}$ as the "number of field lines per unit area perpendicular to $\vec{\boldsymbol{E}}$" and treat these "lines of force" as if they were *conserved* in the same way as streams of water, we get a useful graphical picture as well as a model which, when translated into mathematics, gives correct answers. As suspicious as this may sound, it is really all one can ask of a physical model of something we cannot see. This is the sense of all sketches showing electric field lines. For every little bit ("element") of charge dq on one side of the symmetric distribution there is an equal charge element exactly opposite (relative to the radius vector joining the centre to the point at which we are evaluating $\vec{\boldsymbol{E}}$); the "transverse" contributions of such charge elements to $\vec{\boldsymbol{E}}$ all cancel out, and so the only possible direction for $\vec{\boldsymbol{E}}$ to point is along the radius vector — *i.e.* as described above. An even simpler argument is that *there is no way to pick a preferred direction* (other than the radial direction) if the charge distribution truly has spherical symmetry. This "symmetry argument" is implied in Fig. 18.1.

Now we must change our notation slightly from the general description of Eqs. (1) and (2) to the specific example of electric charge and field. Inasmuch as one's choice of a system of *units* in electromagnetism is rather flexible, and since each choice introduces a different set of constants of proportionality with odd units of their own, I will merely state that "J turns into E, $dQ/dt \to q$ now stands for *electric charge*, and there is a $1/\epsilon_0$ in front of the $dQ/dt \equiv q$ on the right-hand side of Eq. (1)" to give us the electrostatics version of (1):

$$\oiint_{\mathcal{S}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{A}} = \frac{q}{\epsilon_0} \tag{3}$$

which, when applied to the isotropic charge distribution, gives the result

$$E(r) = \frac{q}{4\pi\epsilon_0} \cdot \frac{1}{r^2} \tag{4}$$

The implication of Eq. (3) is then that, since the spherical shell contains the same amount of charge for all radii $r > R$, where R is the physical radius of the charge distribution itself, it cannot matter *how* the charge is distributed (as long as it is spherically symmetric); to the distant observer, the $\vec{\boldsymbol{E}}$ field it produces will always look just like the $\vec{\boldsymbol{E}}$ field due to a *point* charge q at the centre; *i.e.* Eq. (4).

18.1.1 Gravity

Another example is *gravity*, which differs from the electrostatic force only in its relative weakness and the innocuous-looking fact that it only comes in one sign, namely attractive, whereas the electric force can be either attractive (for unlike charges) or repulsive (for like charges). That is, "There are no negative masses." So all these equations hold equally well for gravity, except of course that we must again shuffle constants of proportionality around to make sure we are not setting apples equal to oranges. In this case we can use some symbol, say $\vec{\boldsymbol{g}}$, to represent the force *per unit mass* at some position, as we did for $\vec{\boldsymbol{E}}$ = force per unit *charge*, and talk about the "gravitational field" as if it were really there, rather than being what would

be there (a force) if we placed a mass there. (Note that $\vec{\boldsymbol{g}}$ will be measured in units of acceleration.) Then the role of "dQ/dt" in Eq. (1) is played by M, the total mass of the attracting body, and the constant of proportionality is $4\pi G$, where G is Newton's Universal Gravitational Constant:

$$\oiint_{\mathcal{S}} \vec{\boldsymbol{g}} \cdot d\vec{\boldsymbol{A}} = 4\pi GM \tag{5}$$

and

$$g(r) = \frac{GM}{r^2} \tag{6}$$

for any spherically symmetric mass distribution of total mass M. Note that we have "derived" this fundamental relationship from arguments about symmetry, geometry and common sense, plus the weird notion that "lines" of gravitational force are "emitted" by masses and are "conserved" in the sense of streams of water — a pretty kinky idea, but evidently one with powerful applications. Be sure you are satisfied that this is *not* a "circular argument;" we really have derived Eq. (6) without using it in the development at all! Now, besides being suggestive of deeper knowledge, this trick can be used to draw amusing conclusions about interesting physical situations.

The Spherical Shell

For instance, suppose that one day we assemble all the matter in the Solar System and build one gigantic spherical shell out of it. We arrange its radius so that the force of gravity at its surface (standing on the outside) is "Earth normal," *i.e.* 9.81 N/kg or $g = 9.81\ \mathrm{m/s^2}$. This is all simple so far, and Gauss' Law tells us that as long as we are *outside* of the spherical shell enclosing the whole spherically symmetric mass distribution, the gravitational field we will see is indistinguishable from that produced by the entire mass concentrated at a point at the centre. The amazing prediction is that if we merely step *inside* the shell, there is still spherical symmetry, but the spherical surface touching our new radius *does not enclose any mass* and therefore sees *no* gravitational field at all! This is actually correct: inside the sphere we are weightless, and travel opportunities to other parts of the shell (across the inside) become quite interesting. There are many more examples of entertaining properties of spherically symmetric charge or mass distributions, all of which you can easily deduce from similar arguments to dazzle your friends. Let us now ask, however, if any *less symmetric* situations can also be treated easily with this technique.

18.1.2 The Uniform Sphere

Another familiar example of spherical symmetry is the uniformly dense solid sphere of mass (if we are interested in gravity) or the solid sphere of insulating material carrying a uniform charge density ρ (if we want to do electrostatics). Let's pick the latter, just for variety. If we imagine a spherical "Gaussian surface" concentric with the sphere, with a radius r less than the sphere's radius R, the usual isotropic symmetry argument gives $\oiint_{\mathcal{S}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{A}} = 4\pi r^2 E$, where E is the (constant) radial electric field strength at radius $r < R$. The net charge enclosed within the Gaussian surface is $\frac{4}{3}\pi r^3 \rho$, giving $4\pi r^2 E = \frac{1}{\epsilon_0}\frac{4}{3}\pi r^3 \rho$, or

$$E(r < R) = \frac{\rho}{3\epsilon_0}\, r \tag{7}$$

for the electric field inside such a uniform spherical charge density.

A similar linear relationship holds for the gravitational field within a solid sphere of uniform mass density, of course, except in that case the force on a "test mass" is always back toward the centre of the sphere — *i.e.* a *linear restoring force* with all that implies....

18.2 The Line Source

A sphere, as we have seen, can be collapsed to a point without affecting the external field; and a point is essentially a "zero-dimensional object" — it has no properties that can help us to define a unique direction in space. The next higher-dimensional object would be *one*-dimensional, namely a *line*. What can we do with this?

In the spirit of the normal physics curriculum, we will now stick to the example of electrostatics, remembering that all the same arguments can be used on gravity or indeed on other situations not involving "force fields" at all. (Consider the sprinkler, or a source of "rays" of light.) Suppose that we have an "infinite line of charge," *i.e.* a straight wire with a charge λ *per unit length*. This is pictured in Fig. 18.2.

The same sort of symmetry arguments used in Fig. 18.1 tell us that for every element of charge a distance d *above* position x on the wire, there is an equal element of charge an equal distance d *below* position x, from which we can conclude that the "transverse" contributions to the $\vec{\boldsymbol{E}}$ field from the opposite charge elements cancel, leaving only the components pointing *directly away from the wire*; *i.e.* perpendicular to the wire. In what are referred to as "cylindrical coordinates," the perpendicular distance from the wire to our field point is called r, and the direction described above is the r direction. Thus $\vec{\boldsymbol{E}}$ points in the $\hat{r}$ direction. (Indeed, if it wanted to point in another direction, it would have to choose it arbitrarily, as there *is* no other direction that can be defined uniquely by reference to the wire's geometry!) Given the direction of $\vec{\boldsymbol{E}}$ and the "obvious" (but nevertheless correct) fact that it must have the same *strength* in all directions (*i.e.* it must be independent of the "azimuthal angle" ϕ — another descriptive term borrowed from celestial navigation), we can guess at a shape for the closed surface of Eq. (3) which will give us $\vec{\boldsymbol{E}}$ either parallel to the surface (no contribution to the outgoing flux) or normal to the surface and constant, which will let us take E outside the integral and just determine the total area perpendicular to $\vec{\boldsymbol{E}}$: we choose a cylindrical shaped "pillbox" centred on the wire. No flux escapes from the "end caps" because $\vec{\boldsymbol{E}}$ is parallel to the surface; $\vec{\boldsymbol{E}}$ is constant in magnitude over the curved outside surface and everywhere perpendicular (normal) to it. Thus

$$\oiint_{\mathcal{S}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{A}} = E \oiint_{\mathcal{S}} dA = (E)(2\pi r L)$$

where $2\pi r L$ is the curved surface area of a cylinder of radius r and height L.

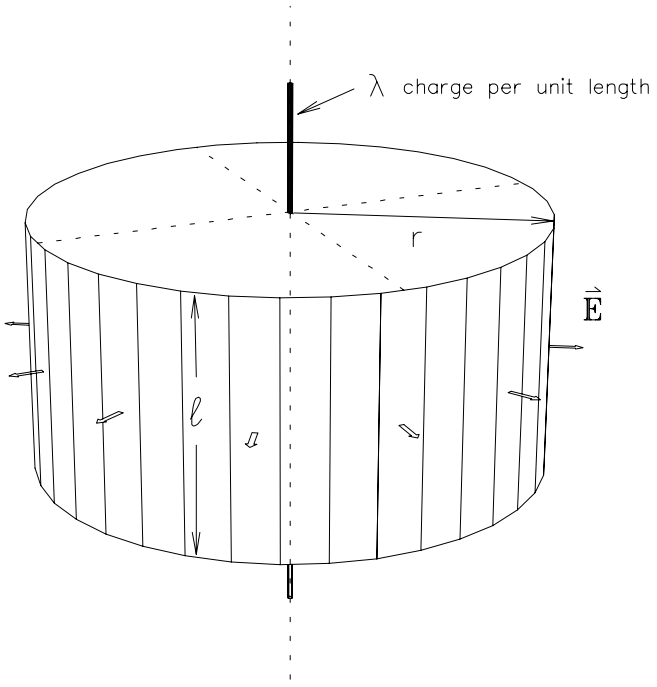


Figure 18.2 An infinite, uniform line of charge.

The same surface, clipping off a length L of wire, encloses a net charge $q = \lambda L$. Plugged into (3), this gives

$$2\pi r L E = \frac{\lambda L}{\epsilon_0}$$

or

$$E(r) = \frac{\lambda}{2\pi\epsilon_0} \cdot \frac{1}{r} \tag{8}$$

which states the *general* conclusion for *any cylindrically* symmetric charge distribution, namely that

> *The electric field from a cylindrically symmetric charge distribution points away from the central line and falls off proportional to the inverse of the distance from the centre.*

This also holds in an amazing variety of situations. Applications are left to the interested student.

18.3 The Plane Source

Note the interesting trend: a zero-dimensional distribution (a point) produces a field that drops off as r^{-2}, while a one-dimensional distribution (a line) produces a field that drops off as r^{-1}. We have to be tempted to see if a *two*-dimensional distribution (a *plane*) will give us a field that drops off as r^{0} — *i.e.* which does not drop off at all with the distance from the plane, but remains *constant* throughout space. This application of Gauss' Law is a straightforward analogy to the other two, and can be worked out easily by the reader. ;-)

Chapter 19

Faraday's Law

Like Gauss' Law, Faraday's Law is most elegantly expressed using vector calculus, but Faraday's Law can also be accurately (if informally) deduced from "common sense" combined with the Lorentz force:

$$\vec{\boldsymbol{F}} = q(\vec{\boldsymbol{E}} + \vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}} \qquad (1)$$

20.1 Handwaving Faraday

Figure 20.1 offers a hand-waving "derivation" of Faraday's Law from the Lorentz force: We start with the magnetic force $\vec{\boldsymbol{F}}_M$ on the charges in a conducting bar that moves through a uniform magnetic field $\vec{\boldsymbol{B}}$ at a speed $\vec{\boldsymbol{v}} \perp \vec{\boldsymbol{B}}$. A positive charge moving with the bar experiences an upward magnetic force and will try to move to the top of the bar while negative charges (*e.g.* electrons) will be forced downward until enough + charges pile up at the top (and − charges at the bottom) to create an electric field $\vec{\boldsymbol{E}}$ whose force on a charge *cancels* $\vec{\boldsymbol{F}}_M$. For a bar of length ℓ, the resulting voltage between the ends is $V = \ell E$ or $V = \ell v B$. But ℓv is the *area swept out by the bar* per unit time, $\ell v = dA/dt$.

Time to define magnetic flux Φ_M:

$$\Phi_M \equiv \iint \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{A}} \qquad (2)$$

is here the magnetic flux $\Phi_M = BA$ passing through the closed loop of which an area $A = \ell x$ between the bar and the edge of the magnetic

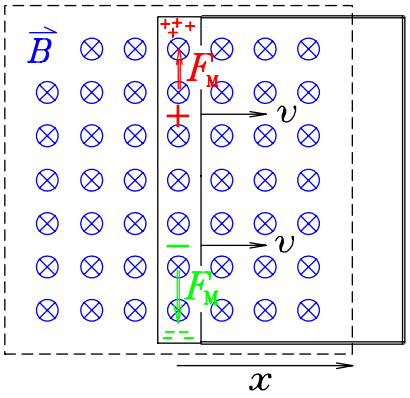


Figure 20.1 Sketch for "deriving" Faraday's law from the Lorentz force. Within the dashed line, a uniform magnetic field $\vec{\boldsymbol{B}}$ points into the page. Elsewhere there is no magnetic field. A metal bar moves to the right (perpendicular to its length), causing the positive charges in the bar to experience an upward magnetic force and negative charges (electrons) a downward one. Usually the positive charges can't move, but in any case negative charges quickly "pile up" on the bottom end of the bar, leaving an excess positive charge at the top end.

field region. The loop shown in black includes A and sticks out into a region where there is no magnetic field. This loop may be made of physical wires or it may only exist in our imagination; either way, the flux through it is *changing* at a rate given by $d\Phi_M/dt = \ell v B$. This is the same as the voltage across the bar. Since no other voltages act, it is also the voltage *around*

the loop:

$$\mathcal{E}_{\rm ind} = -\frac{d\Phi_M}{dt}, \qquad (3)$$

— *i.e.* FARADAY'S LAW!

(What does that − sign signify?)

20.1.1 Lenz's Law

Referring again to Fig. 20.1, if we imagine for the moment that the black path *actually is a wire*, then a current can flow around the loop due to $\mathcal{E}_{\rm ind}$. (We imagine the wire to have a small resistance, to avoid confusing aspects of superconductivity.) The current will flow *clockwise* here, so as to reunite the + and − charges. This current *makes its own magnetic field* into the page, and thus *adds* to the net flux through the loop in that direction, which was *decreasing* as the loop was pulled out of the field. So the induced EMF "tries" to make a current flow to *counteract the change in flux*.

This is LENZ'S LAW.

Reaction Force

LENZ' LAW predicts a current to the left through the moving wire in the scenario shown in Fig. 20.2. The *Lorentz force* on that current-carrying wire is therefore *fighing the motion*. *Work* must be performed by the person pulling the wire in the direction shown. That work goes into the energy stored in the circulating current. This is a crude version of a *dynamo*. Hydroelectric dams use the same principle to generate electrical power.

20.1.2 Magic!

We have "derived" FARADAY'S LAW and LENZ'S LAW for the particular scenario shown in Fig. 20.1. This may seem an artificial way of expressing what is obvious from the simple /sc Lorentz force law. What's so amazing about a simple change of terminology? The "magic" of this Law is that it applies (and works!) equally in situations that bear no resemblance to this example.

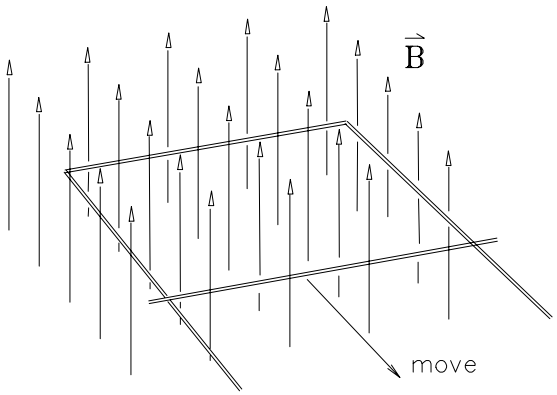


Figure 20.2 Viewing Fig. 20.1 from the side, reversing the direction of the magnetic field and replacing the moving bar on the left with a sliding wire on the right that completes a loop, we can see that for a constant magnetic field the motion shown increases the flux through the loop, so LENZ' LAW predicts a clockwise current through the wire to "fight the change in flux".

If there is *no physical motion at all*, but only a change in the *strength* of the magnetic field, we still get an induced EMF according to the same equation. This accounts for the enormous impact of "electric power" on the modern world.

It also leads to our understanding of the nature of light itself, as shown by Maxwell.

20.2 The Hall Effect

In the HALL EFFECT (discovered by Edwin Hall in 1879) the metal bar is at rest and the magnetic field stays constant, but a *current* flows through the conducting bar. If the current is "up" as shown in Fig. 20.3, the result depends on whether the charge carriers are positive or negative: if they are positively charged (like "holes" in a *p*-type semiconductor) then the *motion* of the actual individual charges is

"up", as shown in the top frame of Fig. 20.3. If the field is into the page, as shown, then the LORENTZ FORCE on the positive charges is to the right, and the positive charges "pile up" on the right side of the bar, leaving negative charges behind on the left side. This accumulated charge separation eventually produces an electric field big enough to counteract the magnetic force on the moving charges, and generates the *Hall voltage*, an electric potential difference across the bar.

If the charge carriers are *negatively* charged (like electrons), then the upward current corresponds to *downward* motion of the electrons; but the sign of the charge q also appears in Eq. (1), so that *the magnetic force is still to the right*. Thus the negative carriers also "pile up" on the right side and the resulting Hall voltage is in the opposite direction.

The HALL EFFECT it therefore an essential tool in determining the *sign of the charge carriers* in any conductor — which would be silly if they were always electrons; but they aren't!

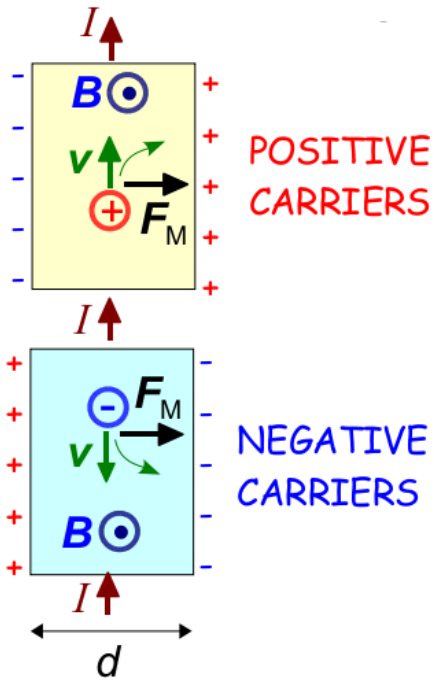


Figure 20.3 Sketch for deriving the HALL EFFECT from the LORENTZ FORCE. A uniform magnetic field $\vec{B}$ points out of the page. A current I flows upward; this can be a current of positive charge carriers moving upward (top drawing) or a current of negative charges (like electrons) movibng down (bottom drawing). In either case, the magnetic force on the carriers is to the right. Eventually the magnetic force will be cancelled by the electric force due to charge accumulations on left and right, creating a "Hall voltage" across the conductor; but this voltage is in opposite directions for positive carriers and negative carriers.

Chapter 21

Vector Calculus

So far I have conjured up somewhat bogus "derivations" of GAUSS' LAW and FARADAY'S LAW using the basic "force laws" for electricity and magnetism. I looked for a way to do this for AMPÈRE'S LAW too, but found that, while it is *possible* to derive AMPÈRE'S LAW from the LAW OF BIOT AND SAVART, the derivation involves very sophisticated mathematics; this sort of defeats the purpose, so I'll just pull AMPÈRE'S LAW out of a hat in the next chapter and ask you to trust that it has been fully checked out by experiment.[1] But before I can even *state* AMPÈRE'S LAW in a simple and elegant form, I need some better notation — namely, that of VECTOR CALCULUS. If you are mathematically inclined you will surely enjoy the elegance and economy of vector notation when applied to calculus; if nothing else this is an æsthetic treat — read it just for fun!

21.1 Functions of Several Variables

Suppose we go beyond $f(x)$ and talk about $F(x, y, z)$ — *e.g.* a function of the *exact position in space*. This is just an example, of course; the abstract idea of a function of several variables can have "several" be as many as you like and "variables" be anything you choose. Another place where we encounter lots of functions of "several" variables is in THERMODYNAMICS, but for the time being we will focus our attention on the three *spatial* variables x (left-right), y (back-forth) and z (up-down).

How can we tackle *derivatives* of this function?

21.1.1 Partial Derivatives

Well, we do the obvious: we say, "Hold all the *other* variables *fixed* except [for instance] x and then treat $F(x, y, z)$ as a function only of x, with y and z as fixed *parameters*." Then we know just how to define the derivative with respect to x. The short name for this derivative is the PARTIAL DERIVATIVE *with respect to* x, written symbolically

$$\frac{\partial F}{\partial x}$$

where the fact that there are other variables being held fixed is implied by the use of the symbol ∂ instead of just d.

Similarly for $\dfrac{\partial F}{\partial y}$ and $\dfrac{\partial F}{\partial z}$.

21.2 Operators

The foregoing description applies for *any* function of (x, y, z); the concept of "taking partial derivatives" is independent of what function we are taking the derivatives *of*. It is therefore practical to learn to think of

$$\frac{\partial}{\partial x} \quad \text{and} \quad \frac{\partial}{\partial y} \quad \text{and} \quad \frac{\partial}{\partial y}$$

[1] I hate doing this, but so far I haven't been able to think of an alternative.

as OPERATORS that can be applied to *any* function (like F). Put the operator on the left of a function, perform the operation and you get a partial derivative — a new function of (x, y, z). In general, such "operators" *change one function into another*. Physics is loaded with operators like these.

21.2.1 The GRADIENT Operator

The GRADIENT operator is a *vector* operator, written $\vec{\boldsymbol{\nabla}}$ and called "grad" or "del." It is defined (in Cartesian coordinates x, y, z) as[2]

$$\boxed{\vec{\boldsymbol{\nabla}} \equiv \hat{\boldsymbol{\imath}}\frac{\partial}{\partial x} + \hat{\boldsymbol{\jmath}}\frac{\partial}{\partial y} + \hat{\boldsymbol{k}}\frac{\partial}{\partial z}} \qquad (1)$$

and can be applied directly to any *scalar* function of (x, y, z) — say, $\phi(x, y, z)$ — to turn it into a *vector* function,

$$\vec{\boldsymbol{\nabla}}\phi = \hat{\boldsymbol{\imath}}\frac{\partial\phi}{\partial x} + \hat{\boldsymbol{\jmath}}\frac{\partial\phi}{\partial y} + \hat{\boldsymbol{k}}\frac{\partial\phi}{\partial z}\,.$$

21.3 GRADIENTS of Scalar Functions

It is instructive to work up to this "one dimension at a time." For simplicity we will stick to using ϕ as the symbol for the function of which we are taking derivatives.

21.3.1 GRADIENTS in 1 Dimension

Let the dimension be x. Then we have no "extra" variables to hold constant and the gradient of $\phi(x)$ is nothing but $\hat{\boldsymbol{\imath}}\frac{d\phi}{dx}$. We can illustrate the "meaning" of $\vec{\boldsymbol{\nabla}}\phi$ by an example: let $\phi(x)$ be the mass of an object times the acceleration of gravity times the height h of a hill at horizontal position x. That is, $\phi(x)$ is the *gravitational potential energy* of the object when it is at horizontal position x. Then

$$\vec{\boldsymbol{\nabla}}\phi = \hat{\boldsymbol{\imath}}\frac{d\phi}{dx} = \hat{\boldsymbol{\imath}}\frac{d}{dx}(mgh) = mg\left(\frac{dh}{dx}\right)\hat{\boldsymbol{\imath}}.$$

Note that $\frac{dh}{dx}$ is the *slope* of the hill and $-\vec{\boldsymbol{\nabla}}\phi$ is the *horizontal component of the net force* (gravity plus the normal force from the hill's surface) on the object. That is, $-\vec{\boldsymbol{\nabla}}\phi$ is the *downhill force*.

21.3.2 GRADIENTS in 2 Dimensions

In the previous example we disregarded the fact that most hills extend in *two* horizontal directions, say x = East and y = North. [If we stick to small distances we won't notice the curvature of the Earth's surface.] In this case there are two *components* to the slope: the Eastward slope $\frac{\partial h}{\partial x}$ and the Northward slope $\frac{\partial h}{\partial y}$. The former is a measure of how steep the hill will seem if you head due East and the latter is a measure of how steep it will seem if you head due North. If you put these together to form a *vector* "steepness" (gradient)

$$\vec{\boldsymbol{\nabla}}h = \hat{\boldsymbol{\imath}}\frac{\partial h}{\partial x} + \hat{\boldsymbol{\jmath}}\frac{\partial h}{\partial y}$$

then the vector $\vec{\boldsymbol{\nabla}}h$ points *uphill* — *i.e.* in the direction of the *steepest ascent*. Moreover, the gravitational potential energy $\phi = mgh$ as before [only now ϕ is a function of *2* variables, $\phi(x, y)$] so that $-\vec{\boldsymbol{\nabla}}\phi$ is once again the *downhill force* on the object.

21.3.3 GRADIENTS in 3 Dimensions

If the potential ϕ is a function of *3* variables, $\phi(x, y, z)$ [such as the three *spatial coordinates* x, y and z — in which case we can write it a little more compactly as $\phi(\vec{\boldsymbol{r}})$ where

$$\vec{\boldsymbol{r}} \equiv x\hat{\boldsymbol{\imath}} + y\hat{\boldsymbol{\jmath}} + z\hat{\boldsymbol{k}}\,,$$

[2] I am using the conventional notation for $\hat{\boldsymbol{\imath}}, \hat{\boldsymbol{\jmath}}, \hat{\boldsymbol{k}}$ as the UNIT VECTORS in the x, y, z directions, respectively.

the vector distance from the origin of our coordinate system to the point in space where ϕ is being evaluated], then it is a little more difficult to make up a "hill" analogy — try imagining a topographical map in the form of a 3-dimensional hologram where instead of *lines* of constant *altitude* the "equipotentials" are *surfaces* of constant ϕ. (This is just what Physicists do picture!) Fortunately the *math* extends easily to 3 dimensions (or any larger number, if that has any meaning in the context we choose).

In general, any time there is a *potential energy* function $\phi(\vec{\boldsymbol{r}})$ we can immediately write down the *force* $\vec{\boldsymbol{F}}$ associated with it as

$$\vec{\boldsymbol{F}} \;\equiv\; -\vec{\boldsymbol{\nabla}}\phi \tag{2}$$

A perfectly analogous expression holds for the *electric field* $\vec{\boldsymbol{E}}$ [force per unit charge] in terms of the *electrostatic potential* ϕ [potential energy per unit charge]:[3]

$$\vec{\boldsymbol{E}} \;\equiv\; -\vec{\boldsymbol{\nabla}}\phi \tag{3}$$

21.3.4 GRADIENTS in N Dimensions

Although we won't be needing to go beyond 3 dimensions very often in Physics, you might want to borrow this metaphor for application in other realms of human endeavour where there are more than 3 variables of which your scalar field is a function. You could have ϕ be a measure of *happiness*, for instance [though it is hard to take reliable measurements on such a subjective quantity]; then ϕ might be a function of lots of factors, such as x_1 = freedom from violence, x_2 = freedom from hunger, x_3 = freedom from poverty, x_4 = freedom from oppression, and so on.[4] Note that with an arbitrary number of variables we get away from thinking up different names for each one and just call the i^{th} variable "x_i."

Then we can define the GRADIENT in N dimensions as

$$\vec{\boldsymbol{\nabla}}\phi \;=\; \hat{\boldsymbol{\imath}}_1\,\frac{\partial\phi}{\partial x_1} \;+\; \hat{\boldsymbol{\imath}}_2\,\frac{\partial\phi}{\partial x_2} \;+\; \cdots \;+\; \hat{\boldsymbol{\imath}}_N\,\frac{\partial\phi}{\partial x_N}$$

$$\text{or}\quad \vec{\boldsymbol{\nabla}}\phi \;=\; \sum_{i=1}^{N} \hat{\boldsymbol{\imath}}_i\,\frac{\partial\phi}{\partial x_i}$$

where $\hat{\boldsymbol{\imath}}_i$ is a UNIT VECTOR in the x_i direction.

21.4 DIVERGENCE of a Field

If we form the scalar ("dot") product of $\vec{\boldsymbol{\nabla}}$ with a *vector* function $\vec{\boldsymbol{A}}(x,y,z)$ we get a *scalar* result called the DIVERGENCE of $\vec{\boldsymbol{A}}$:

$$\boxed{\text{div}\vec{\boldsymbol{A}} \equiv \vec{\boldsymbol{\nabla}}\cdot\vec{\boldsymbol{A}} \equiv \frac{\partial A_x}{\partial x} \;+\; \frac{\partial A_y}{\partial y} \;+\; \frac{\partial A_z}{\partial z}} \tag{4}$$

This name is actually quite mnemonic: the DIVERGENCE of a vector field is a *local* measure of its "outgoingness" — *i.e.* the extent to which there is more *exiting* an infinitesimal region of space than *entering* it. If the field is represented as "flux lines" of some indestructible "stuff" being emitted by "sources" and absorbed by "sinks," then a nonzero DIVERGENCE at some point means there must be a *source* or *sink* at that position. That is to say,

"What leaves a region is no longer in it."

For example, consider the divergence of the CURRENT DENSITY $\vec{\boldsymbol{J}}$, which describes the FLUX of a CONSERVED QUANTITY such as electric charge Q. (Mass, as in the current of a river, would do just as well.)

To make this as easy as possible, let's picture a *cubical* volume element $dV = dx\,dy\,dz$. In general, $\vec{\boldsymbol{J}}$ will (like any vector) have three components (J_x, J_y, J_z), each of which may be a function of position (x,y,z). If we take the lower

[3] I know, I know, I am using the ϕ symbol for two different things. Well, I *said* it was the preferred symbol for a scalar field, so you shouldn't be surprised to see it "recycled" many times. This won't be the last!

[4] These are rotten examples, of course — the first practical criterion for the variables of which any ϕ is a function is that they should be *linearly independent* [*i.e. orthogonal*] so that the dependence on one is not all mixed up with the dependence on another!

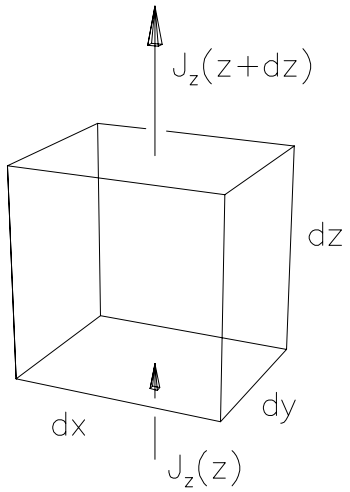


Figure 21.1 Flux into and out of a volume element $dV = dx\,dy\,dz$.

left front corner of the cube to have coordinates (x, y, z) then the upper right back corner has coordinates $(x + dx,\ y + dy,\ z + dz)$. Let's concentrate first on J_z and how it depends on z.

It may not depend on z at all, of course. In this case, the amount of Q coming into the cube through the bottom surface (per unit time) will be the same as the amount of Q going out through the top surface and there will be no net gain or loss of Q in the volume — at least not due to J_z.

If J_z is bigger at the top, however, there will be a net loss of Q within the volume dV due to the "divergence" of J_z. Let's see how much: the difference between $J_z(z)$ at the bottom and $J_z(z + dz)$ at the top is, by definition, $dJ_z = \left(\frac{\partial J_z}{\partial z}\right) dz$. The flux is over the same area at top and bottom, namely $dx\,dy$, so the total *rate of loss* of Q due to the z-dependence of J_z is given by

$$\dot{Q}_z \;=\; -dx\,dy\left(\frac{\partial J_z}{\partial z}\right) dz \;=\; -\left(\frac{\partial J_z}{\partial z}\right) dx\,dy\,dz$$

$$\text{or} \quad \dot{Q} \;=\; -\left(\frac{\partial J_z}{\partial z}\right) dV\,.$$

A perfectly analogous argument holds for the x-dependence if J_x and the y-dependence of J_y, giving a total rate of change of Q

$$\dot{Q} \;=\; -\left(\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \frac{\partial J_z}{\partial z}\right) dV$$

$$\text{or} \quad \dot{Q} \;=\; -\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{J}}\, dV$$

The total amount of Q *in* our volume element dV at a given instant is just $\rho\, dV$, of course, so the rate of change of the enclosed Q is just

$$\dot{Q} \;=\; \dot{\rho}\, dV$$

which means that we can write

$$\frac{\partial \rho}{\partial t}\, dV \;=\; -\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{J}}\, dV$$

or, just cancelling out the common factor dV on both sides of the equation,

$$\boxed{\frac{\partial \rho}{\partial t} \;=\; -\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{J}}} \tag{5}$$

which is the compact and elegant "differential form" of the EQUATION OF CONTINUITY.

This equation tells us that the "Q sourciness" of *each point* in space is given by the degree to which flux "lines" of $\vec{\boldsymbol{J}}$ tend to radiate *away* from that point more than they converge *toward* that point — namely, the DIVERGENCE of $\vec{\boldsymbol{J}}$ at the point in question. This esoteric-looking mathematical expression is, remember, just a formal way of expressing our original dumb tautology!

21.5 CURL of a Vector Field

If we form the vector ("cross") product of $\vec{\boldsymbol{\nabla}}$ with a *vector* function $\vec{\boldsymbol{A}}(x, y, z)$ we get a *vector* result called the **curl** of $\vec{\boldsymbol{A}}$:

$$\begin{aligned}\textbf{curl}\,\vec{\boldsymbol{A}} \equiv \vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{A}} \;\equiv\; & \hat{\boldsymbol{\imath}}\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right) \\ +\; & \hat{\boldsymbol{\jmath}}\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\end{aligned}$$

$$+ \quad \hat{\boldsymbol{k}}\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) \quad (6)$$

This is a lot harder to visualize than the DIVERGENCE, but not impossible. Suppose you are in a boat in a huge river (or Pass) where the current flows mainly in the x direction but where the speed of the current (flux of water) varies with y. Then if we call the current $\vec{\boldsymbol{J}}$, we have a nonzero value for the derivative $\frac{\partial J_x}{\partial y}$, which you will recognize as one of the terms in the formula for $\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{J}}$. What does this imply? Well, if you are sitting in the boat, moving with the current, it means the current on your port side moves faster — *i.e.* forward relative to the boat — and the current on your starboard side moves slower — *i.e.* backward relative to the boat — and this implies a *circulation* of the water around the boat — *i.e.* a *whirlpool!* So $\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{J}}$ is a measure of the local "swirliness" of the current $\vec{\boldsymbol{J}}$, which means "**curl**" is not a bad name after all!

21.6 STOKES' THEOREM

$$\boxed{\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \iint_A (\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{B}}) \cdot d\vec{\boldsymbol{S}}} \quad (7)$$

where the surface integral on the right is over a surface A bounded by the path C in the path integral on the left. This can be proven formally, but the proof is not trivial, so I am just going to state it and let you decide whether to look up the proof to satisfy your skepticism.

21.7 The LAPLACIAN Operator

If we form the scalar ("dot") product of $\vec{\boldsymbol{\nabla}}$ with *itself* we get a scalar *second derivative* operator called the LAPLACIAN:

$$\boxed{\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{\nabla}} \equiv \nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}} \quad (8)$$

What does the ∇^2 operator "mean?" It is the three-dimensional generalization of the one-dimensional CURVATURE operator d^2/dx^2. Consider the familiar one-dimensional function $h(x)$ where h is the height of a hill at horizontal position x. Then dh/dx is the *slope* of the hill and d^2h/dx^2 is its *curvature* (the rate of change of the slope with position). This property appears in every form of the WAVE EQUATION. In three dimensions, a nice visualization is harder (there is no extra dimension "into which to curve") but $\nabla^2\phi$ represents the equivalent property of a scalar function $\phi(x, y, z)$.

21.8 GAUSS' LAW

The EQUATION OF CONTINUITY [see Eq. (5)] describes the conservation of "actual physical stuff" entering or leaving an infinitesimal region of space dV. For example, $\vec{\boldsymbol{J}}$ may be the *current density* (charge flow per unit time per unit area normal to the direction of flow) in which case ρ is the *charge density* (charge per unit volume); in that example the conserved "stuff" is electric charge itself. Many other examples exist, such as FLUID DYNAMICS (in which *mass* is the conserved stuff) or HEAT FLOW (in which *energy* is the conserved quantity). In ELECTROMAGNETISM, however, we deal not only with the conservation of *charge* but also with the *continuity* of abstract *vector fields* like $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$. In order to visualize $\vec{\boldsymbol{E}}$, we have developed the notion of "electric field lines" that cannot be broken except where they originate (from positive charges) and terminate (on negative charges). [This description only holds for *static* electric fields; when things *move* or otherwise change with time, things get a lot more complicated ... and interesting!] Thus a positive charge is a "source of electric field lines" and a negative charge is a "sink" — the charges themselves stay put, but the lines of $\vec{\boldsymbol{E}}$ *diverge* out of or into them. You can probably see where this is heading.

GAUSS' LAW states that the net flux of electric field "lines" *out* of a closed surface $\mathcal{S}$ is proportional to the net electric charge enclosed *within* that surface. The constant of proportionality depends on which system of units one is using; in *SI* units it is $1/\epsilon_\circ$. In mathematical shorthand, this reads

$$\epsilon_\circ \oiint_{\mathcal{S}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{A}} \;=\; Q_{\text{encl}}\,.$$

Recalling our earlier discussion of DIVERGENCE, we can think of $\vec{\boldsymbol{E}}$ as being a sort of flux density of conserved "stuff" *emitted* by positive electric charges. Remember, in this case the charges themselves do not go anywhere; they simply emit (or absorb) the electric field "lines" which emerge from (or disappear into) the enclosed region. The rate of generation of this "stuff" is $Q_{\text{encl}}/\epsilon_\circ$. We can then apply GAUSS' LAW to an infinitesimal volume element using Fig. 21.1 with $\vec{\boldsymbol{D}} \equiv \epsilon_\circ \vec{\boldsymbol{E}}$ in place of $\vec{\boldsymbol{J}}$.[5] Except for the "fudge factor" $\epsilon_\circ$ and the replacement of $\dot{Q}$ by Q_{encl}, the same arguments used to derive the EQUATION OF CONTINUITY lead in this case to a formula relating the divergence of $\vec{\boldsymbol{D}}$ to the electric charge density ρ *at any point in space*, namely

$$\boxed{\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{D}} \;=\; \rho}\,. \tag{9}$$

This is the *differential form* of GAUSS' LAW.

21.9 Poisson and Laplace

Even in its differential form, GAUSS' LAW is a little tricky to solve analytically, since it is a *vector* differential equation. Generally we have an easier time solving *scalar* differential equations, even though they may involve higher order partial derivatives. Fortunately, we can convert the former into the latter: recall that the vector electric *field* can always be obtained from the scalar electrostatic *potential* using

$$\vec{\boldsymbol{E}} \;\equiv\; -\vec{\boldsymbol{\nabla}}\phi\,.$$

Thus $\text{div}\vec{\boldsymbol{E}} \equiv \vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{E}} = -\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{\nabla}}\phi$ or

$$\boxed{\nabla^2\phi \;=\; -\,\frac{1}{\epsilon_\circ}\,\rho}\,. \tag{10}$$

This relation is known as POISSON'S EQUATION. Its simplified cousin, LAPLACE'S EQUATION, applies in regions of space where there are *no free charges*:

$$\boxed{\nabla^2\phi \;=\; 0}\,. \tag{11}$$

Each of these equations finds much use in real electrostatics problems. Advanced students of electromagnetism learn many types of functions that satisfy LAPLACE'S EQUATION, with different *symmetries*; since a *conductor* is always an *equipotential* (every point in a given conductor must have the same ϕ, otherwise there would be an electric field in the conductor that would cause charges to move until they cancelled out the differences in ϕ), empty regions surrounded by conductors of certain shapes must have ϕ with a spatial dependence satisfying those BOUNDARY CONDITIONS as well as LAPLACE'S EQUATION. One can often write down a complicated-looking formula for ϕ almost by inspection, using this favourite method of Physicists and Mathematicians, namely ... KNOWING THE ANSWER.

21.10 Faraday Revisited

With these tools we can express FARADAY'S LAW more elegantly.

[5] Note how I cleverly slipped in the definition of the "ELECTRICAL DISPLACEMENT" field $\vec{\boldsymbol{D}}$ there... but I left out the possibility of these fields existing inside an *electrically polarizable medium* (like a dielectric) with polarization in which case we have $\vec{\boldsymbol{D}} = \epsilon\vec{\boldsymbol{E}} = \epsilon_\circ\vec{\boldsymbol{E}} + \vec{\boldsymbol{P}}$.

21.10.1 Integral Form

The induced voltage around a closed loop C can be expressed as a line integral:

$$\mathcal{E}_{\text{ind}} = \oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}}$$

and the magnetic flux through that closed loop can be expressed as a surface integral over any surface *A bounded by* the loop C:

$$\Phi_M = \iint_A \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{S}}$$

whose time derivative (all other things being constant) is given by

$$\frac{\partial \Phi_M}{\partial t} = \iint_A \frac{\partial \vec{\boldsymbol{B}}}{\partial t} \cdot d\vec{\boldsymbol{S}}$$

so we can write FARADAY'S LAW in the form

$$\boxed{\oint_C \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{\ell}} = -\iint_A \frac{\partial \vec{\boldsymbol{B}}}{\partial t} \cdot d\vec{\boldsymbol{S}}}\,. \qquad (12)$$

21.10.2 Differential Form

Using STOKES' THEOREM we can convert Eq. (12) into its differential form:

$$\boxed{\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{E}} = -\frac{\partial \vec{\boldsymbol{B}}}{\partial t}}\,, \qquad (13)$$

which relates the rate of change of the magnetic field to the "curliness" of the electric field *at every point in space*. **Cool!**

Chapter 22

Ampère's law

With VECTUR CALCULUS firmly under our belts (?) we are now ready to tackle AMPÈRE'S LAW, right?

22.1 Integral Form

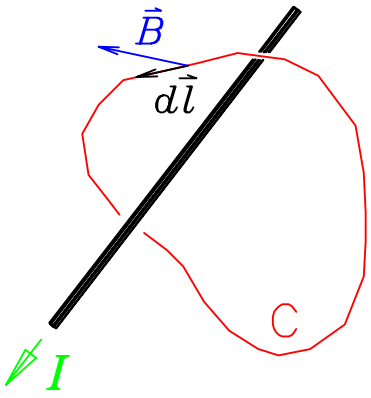


Figure 22.1 A wire carrying a current I passes through an arbitrary closed loop C, generating a magnetic field $\vec{\boldsymbol{B}}$ in the region around the wire. At every point on C there is a path element $d\vec{\boldsymbol{\ell}}$ in the direction around the loop corresponding to the direction the fingers of your right hand would point if you grabbed the wire with your thumb pointing along the current, and a magnetic field $\vec{\boldsymbol{B}}$ in some direction (not necessarily the same direction as $d\vec{\boldsymbol{\ell}}$).

If at each step $d\vec{\boldsymbol{\ell}}$ around the path C in Fig. 22.1 we find the component of the magnetic field $\vec{\boldsymbol{B}}$ in the direction of $d\vec{\boldsymbol{\ell}}$, multiply the two, and add up all the results for the whole loop, we get the *integral form* of AMPÈRE'S LAW:

$$\boxed{\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = \mu_0 I} \tag{1}$$

where $\mu_0 = 4\pi \times 10^{-7}$ Webers/(Amp·m) [or Newtons/Amp^2, or Henries/m, or Tesla·m/Amp, or Volt·s/(Amp·m)] is the PERMEABILITY OF FREE SPACE.[1]

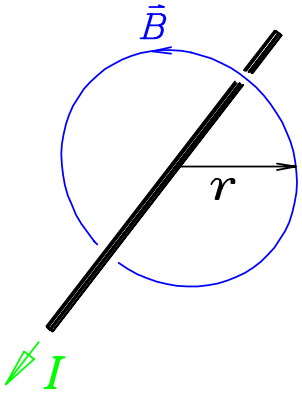


Figure 22.2 By symmetry, a wire carrying a current I generates a magnetic field $\vec{\boldsymbol{B}}$ that forms circular loops centered on the wire at every radius r.

In cases where the direction of $\vec{\boldsymbol{B}}$ at every point along the path C is not known, this form is pretty useless for practical calculations. But the LAW OF BIOT & SAVART tells us that the contribution to $\vec{\boldsymbol{B}}$ from each element of current is always perpendicular to the current and proportional to the inverse square of the distance

[1] What can I say? Electromagnetic units are *weird*!

from that current element; so SYMMETRY demands that a "line of $\vec{\boldsymbol{B}}$" forms a *circular* loop centered on the wire, as shown in Fig. 22.2, and that its magnitude is the same everywhere around that loop. So we simply pick such a loop of radius r as our path C, and the path integral on the left side of Eq. (1) becomes just

$$\oint_C \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{\ell}} = 2\pi r B$$

giving

$$B(r) = \frac{\mu_0 I}{2\pi r} \tag{2}$$

as we found in the earlier Exercise.

22.2 Differential Form

We can apply STOKES' THEOREM to the integral in Eq. (1) to get

$$\iint_A (\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{B}}) \cdot d\vec{\boldsymbol{S}} = \mu_0 I$$

and note that

$$I = \iint_A \vec{\boldsymbol{J}} \cdot d\vec{\boldsymbol{S}}$$

on the same surface A bounded by the path C in Fig. 22.1. Therefore the integrands of the two surface integrals must be equal:

$$\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{B}} = \mu_0 \vec{\boldsymbol{J}}$$

or

$$\boxed{\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{H}} = \vec{\boldsymbol{J}}} \tag{3}$$

where $\vec{\boldsymbol{J}}$ is the CURRENT DENSITY and we have defined

$$\vec{\boldsymbol{B}} = \mu_0 \vec{\boldsymbol{H}} \tag{4}$$

in free space. (In magnetic materials $\vec{\boldsymbol{B}} = \mu \vec{\boldsymbol{H}}$ where μ is the magnetic permeability of the material.) Equation (3) expresses the relationship between the CURRENT DENSITY and the curl of the magnetic field *at any point in space*. This is pretty cool too!

But we have left something out....

22.3 Displacement Current

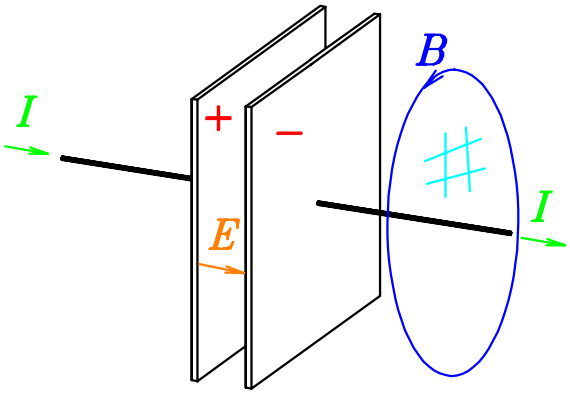


Figure 22.3 A capacitor consists of two adjacent plates of conductor separated by an insulator (*e.g.* air). The plates are initially uncharged. If a current begins flowing onto the left plate, it starts to accumulate a positive charge; this attracts negative charges on the light plate, which must come down the wire on the right (from "elsewhere"). negative charges flowing to the left constitutes a positive current to the right, so the current appears (at least initially) to *pass through* the capacitor, even though one plate is isolated from the other. The surface charges produce an electric field $\vec{\boldsymbol{E}}$ between the plates (and a voltage $V = Ed$ where d is the distance between the plates). Since E is proportional to the accumulated charge on the plate, $\partial E/\partial t \propto I$.

James Maxwell reasoned that an application of the integral form of AMPÈRE'S LAW to find the magnetic field encircling the wire far from the capacitor was supposed to work for **any** surface bounded by the path over which the line integral of $\vec{\boldsymbol{B}}$ is evaluated, it should give the same answer whether that surface is "punctured" by the current or not.

Visualize, if you will, a soap bubble across the blue loop shown in Fig. 22.3. The current I clearly "punctures" that surface. Now blow to the left through the blue loop and imagine that the right plate of the capacitor somehow fails

to pop the resultant bubble, so that the surface bounded by the blue loop now passes *between* the capacitor plates, where there are no moving charges. What gives?

Let's review the electric field between two capacitor plates: By GAUSS' LAW it's constant far from the edges, points from the + plate to the − plate, and has a magnitude $E = \sigma/\epsilon_0$, where $\sigma = Q/A$ (the charge on one plate divided by the area of the plate). Thus $\epsilon_0 E = D = Q/A$ and taking the time derivative gives

$$A \cdot \frac{\partial D}{\partial t} = \frac{\partial Q}{\partial t} \equiv I.$$

But since $\vec{\boldsymbol{D}}$ is constant over the area A and zero outside the capacitor, we can write this as

$$\iint_A \frac{\partial \vec{\boldsymbol{D}}}{\partial t} \cdot d\vec{\boldsymbol{S}} = I.$$

That is, a changing electric field is equivalent to an actual current.

Maxwell called this surface integral of the changing electric field a DISPLACEMENT CURRENT after the name of $\vec{\boldsymbol{D}}$ (the "electric displacement"). It turns out (with a little more rigorous derivation) to hold equally well for less simple geometries, giving us MAXWELL'S EXTENSION OF AMPÈRE'S LAW,

$$\boxed{\oint_C \vec{\boldsymbol{H}} \cdot d\vec{\boldsymbol{\ell}} = \iint_A \left(\vec{\boldsymbol{J}} + \frac{\partial \vec{\boldsymbol{D}}}{\partial t} \right) \cdot d\vec{\boldsymbol{S}}.} \qquad (5)$$

which is equivalent to the differential version,

$$\boxed{\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{H}} = \vec{\boldsymbol{J}} + \frac{\partial \vec{\boldsymbol{D}}}{\partial t}} \qquad (6)$$

Chapter 23

Maxwell's Equations

In 1860, while Americans were waging a bloody civil war, a "thorough old Scotch laird" (then only 29) named James Clerk Maxwell was assembling the known laws of electromagnetism into a compact and elegant form that was to lead, a year later, to the discovery that *light* is in fact a propagating disturbance in the electromagnetic fields. That discovery was later to overturn all the conceptual foundations of classical Physics and leave "common sense" in much the same condition as the United States after the Civil War. It was hard times all around, but *exciting*....

23.1 Gauss' Law

By now you are familiar with GAUSS' LAW in its integral form,

$$\epsilon_\circ \oiint_{\mathcal{S}} \vec{\boldsymbol{E}} \cdot d\vec{\boldsymbol{A}} = Q_{\text{encl}} \qquad (1)$$

where Q_{encl} is the electric charge enclosed within the closed surface $\mathcal{S}$. Except for the "fudge factor" $\epsilon_\circ$, which is just there to make the units come out right, GAUSS' LAW is just a simple statement that electric field "lines" are continuous except when they start or stop *on electric charges*. In the absence of "sources" (positive charges) or "sinks" (negative charges), electric field lines obey the simple rule, "What goes in must come out." This is what GAUSS' LAW says.

There is also a GAUSS' LAW for the *magnetic* field $\vec{\boldsymbol{B}}$; we can write it the same way,

$$(\text{some constant}) \oiint_{\mathcal{S}} \vec{\boldsymbol{B}} \cdot d\vec{\boldsymbol{A}} = Q_{\text{Magn}} \qquad (2)$$

where in this case Q_{Magn} refers to the enclosed *magnetic charges*, of which (so far) none have ever been found! So GAUSS' LAW FOR MAGNETISM is usually written with a zero on the right-hand side of the equation, even though no one is very happy with this lack of symmetry between the electric and magnetic versions.

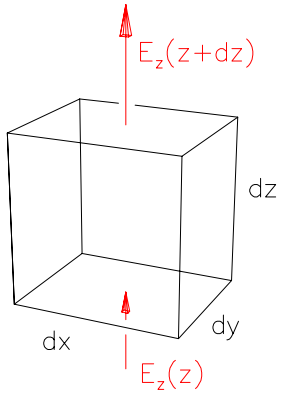


Figure 23.1 An infinitesimal volume of space.

Suppose now we apply GAUSS' LAW to a small rectangular region of space where the z axis is chosen to be in the direction of the electric field, as shown in Fig. 23.1.[1] The flux of electric field

[1] This Figure is very similar to the one used to derive the EQUATION OF CONTINUITY, which in fact expresses the same basic principles (conservation of some "stuff" produced locally), although it is generally used for different purposes.

into this volume at the bottom is $E_z(z)\,dx\,dy$. The flux *out* at the top is $E_z(z+dz)\,dx\,dy$; so the *net* flux *out* is just $[E_z(z+dz)-E_z(z)]\,dx\,dy$. The definition of the *derivative of E with respect to z* gives us $[E_z(z+dz)-E_z(z)] = (\partial E_z/\partial z)\,dz$ where the partial derivative is used in acknowledgement of the possibility that E_z may also vary with x and/or y. GAUSS' LAW then reads $\epsilon_\circ(\partial E_z/\partial z)\,dx\,dy\,dz = Q_{\rm encl}$. What is $Q_{\rm encl}$? Well, in such a small region there is some approximately constant *charge density* ρ (charge per unit volume) and the volume of this region is $dV = dx\,dy\,dz$, so GAUSS' LAW reads $\epsilon_\circ(\partial E_z/\partial z)\,dV = \rho\,dV$ or just $\epsilon_\circ\,\partial E_z/\partial z = \rho$. If we now allow for the possibility of electric flux entering and exiting through the other faces (*i.e.* $\vec{\boldsymbol{E}}$ may also have x and/or y components), perfectly analogous arguments hold for those components, with the resultant "outflow-ness" given by

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \vec{\boldsymbol{\nabla}}\cdot\vec{\boldsymbol{E}} \equiv \mathrm{div}\,\vec{\boldsymbol{E}}$$

where the GRADIENT operator $\vec{\boldsymbol{\nabla}}$ is shown in its cartesian representation (in rectangular coordinates x, y, z). It has completely equivalent representations in other coordinate systems such as spherical (r, θ, ϕ) or cylindrical coordinates, but for illustration purposes the cartesian coordinates are simplest.

We are now ready to write GAUSS' LAW in its compact *differential* form,

$$\boxed{\epsilon_\circ\,\vec{\boldsymbol{\nabla}}\cdot\vec{\boldsymbol{E}} = \rho} \qquad (3)$$

and for the magnetic field, assuming no magnetic charges (MONOPOLES),

$$\boxed{\vec{\boldsymbol{\nabla}}\cdot\vec{\boldsymbol{B}} = 0} \qquad (4)$$

These are the first two of MAXWELL'S EQUATIONS.

23.2 Faraday's Law

You should now be familiar with the long *integral* mathematical form of FARADAY'S LAW of MAGNETIC INDUCTION: in SI units,

$$\oint_{\mathcal{C}} \vec{\boldsymbol{E}}\cdot d\vec{\boldsymbol{\ell}} = -\frac{\partial}{\partial t}\iint_{\mathcal{S}} \vec{\boldsymbol{B}}\cdot d\vec{\boldsymbol{S}} \qquad (5)$$

where the line integral of $\vec{\boldsymbol{E}}$ around the closed loop $\mathcal{C}$ is (by definition) the *induced* $\mathcal{EMF}$ around the loop and the right hand side refers to the *rate of change* of the *magnetic flux* through the area $\mathcal{S}$ *bounded* by that closed loop.

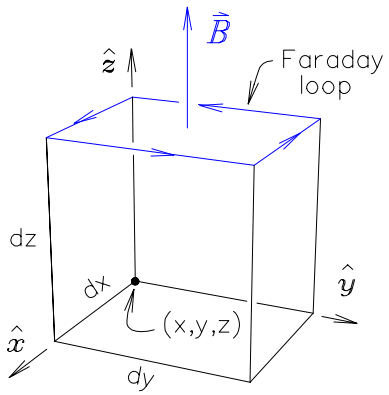


Figure 23.2 Another infinitesimal volume of space.

To make this easy to visualize, let's again draw an infinitesimal rectangular box with the z axis along the direction of the magnetic field, which can be considered more or less uniform over such a small region. Then the flux through the "Faraday loop" is just $B_z\,dx\,dy$ and the line integral of the electric field is

$$E_x(y)dx + E_y(x+dx)dy - E_x(y+dy)dx - E_y(x)dy.$$

(Yes it is. Study the diagram!) Here, as before, $E_y(x+dx)$ denotes the magnitude of the y component of $\vec{\boldsymbol{E}}$ along the front edge of the box, and so on. As before, we note that $[E_y(x+dx) - E_y(x)] = (\partial E_y/\partial x)\,dx$ and

$[E_x(y+dy) - E_x(y)] = (\partial E_x/\partial y)\,dy$ so that FARADAY'S LAW reads

$$\left(\frac{\partial E_y}{\partial x}dx\right)dy - \left(\frac{\partial E_x}{\partial y}dy\right)dx = -\left(\frac{\partial B_z}{\partial t}\right)dx\,dy$$

which reduces to the *local* relationship

$$\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) = -\left(\frac{\partial B_z}{\partial t}\right)$$

between the "swirlyness" of the spatial dependence of the electric field and the rate of change of the magnetic field with time.

If you have studied the definition of the CURL of a vector field, you may recognize the left-hand side of the last equation as the z component of

$$\begin{aligned}\mathbf{curl}\,\vec{\boldsymbol{E}} &\equiv \vec{\boldsymbol{\nabla}}\times\vec{\boldsymbol{E}}\\ &\equiv \hat{\boldsymbol{\imath}}\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\\ &+ \hat{\boldsymbol{\jmath}}\left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right)\\ &+ \hat{\boldsymbol{k}}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right).\end{aligned}$$

The x and y components of $\mathbf{curl}\,\vec{\boldsymbol{E}}$ are related to the corresponding components of $\partial\vec{\boldsymbol{B}}/\partial t$ in exactly the same way, allowing us to write FARADAY'S LAW in a *differential form* which describes part of the behaviour of electric and magnetic fields at every point in space:

$$\boxed{\vec{\boldsymbol{\nabla}}\times\vec{\boldsymbol{E}} = -\frac{\partial\vec{\boldsymbol{B}}}{\partial t}} \qquad (6)$$

This says, in essence, that *any change in the magnetic field with time induces an electric field perpendicular to the changing magnetic field.* Hold that thought.

23.3 Ampère's Law

You are probably also adept at using the trick developed by Henri Ampère for calculating the magnetic field ($\vec{\boldsymbol{H}} \equiv \vec{\boldsymbol{B}}/\mu$) due to various symmetrical arrangements of electric current (I). In its integral form and SI units, AMPÈRE'S LAW reads

$$\oint_{\mathcal{C}} \vec{\boldsymbol{H}}\cdot d\vec{\boldsymbol{\ell}} = I + \frac{\partial}{\partial t}\iint_{\mathcal{S}} \vec{\boldsymbol{D}}\cdot d\vec{\boldsymbol{S}} \qquad (7)$$

where Maxwell's "DISPLACEMENT CURRENT" associated with a time-varying electric displacement $\vec{\boldsymbol{D}} \equiv \epsilon\vec{\boldsymbol{E}}$ has been included. This equation says (sort of), "The *circulation* of the magnetic field around a *closed loop* is equal to a constant times the total electric current *linking* that loop, *except when there is a changing electric field* in the same region."

As you know, this "Law" is used with various symmetry arguments to "finesse" the evaluation of magnetic fields due to arrangements of electric currents, much as GAUSS' LAW was used to calculate electric fields due to different arrangements of electric charges. Skipping over the details, let me draw your attention to the formal similarity to FARADAY'S LAW and state (this time without showing the derivation) that there is an analogous *differential form* of AMPÈRE'S LAW describing the behaviour of the fields at any point in space:

$$\boxed{\vec{\boldsymbol{\nabla}}\times\vec{\boldsymbol{H}} = \vec{\boldsymbol{J}} + \frac{\partial\vec{\boldsymbol{D}}}{\partial t}} \qquad (8)$$

If we ignore the current density $\vec{\boldsymbol{J}}$ then this equation says (sort of), "*A changing electric field generates a magnetic field at right angles to it,*" which is rather reminiscent of what FARADAY'S LAW said.

Now we're getting somewhere.

23.4 Maxwell's Equations

In 1865, James Clerk Maxwell assembled all the known "Laws" of $\mathcal{E}\&\mathcal{M}$ in their most compact, elegant (differential) form, shown here in SI units:

GAUSS' LAW FOR ELECTROSTATICS:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{D}} \;=\; \rho \tag{9}$$

GAUSS' LAW FOR MAGNETOSTATICS:

$$\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{B}} \;=\; 0 \tag{10}$$

FARADAY'S LAW:

$$\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{E}} + \frac{\partial \vec{\boldsymbol{B}}}{\partial t} \;=\; 0 \tag{11}$$

AMPÈRE'S LAW:

$$\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{H}} - \frac{\partial \vec{\boldsymbol{D}}}{\partial t} \;=\; \vec{\boldsymbol{J}} \tag{12}$$

These four basic equations are known collectively as MAXWELL'S EQUATIONS; they are considered by most Physicists to be a beautifully concise summary of $\mathcal{E}\&\mathcal{M}$ phenomenology.

Well, actually, a *complete* description of $\mathcal{E}\&\mathcal{M}$ also requires two additional laws:

EQUATION OF CONTINUITY:

$$\frac{\partial \rho}{\partial t} \;=\; -\vec{\boldsymbol{\nabla}} \cdot \vec{\boldsymbol{J}} \tag{13}$$

LORENTZ FORCE:

$$\vec{\boldsymbol{F}} \;=\; q\left(\vec{\boldsymbol{E}} \;+\; \vec{\boldsymbol{v}} \times \vec{\boldsymbol{B}}\right). \tag{14}$$

23.5 The Wave Equation

The two "Laws" of ELECTRODYNAMICS — FARADAY'S LAW and AMPÈRE'S LAW — can be combined to produce a very important result.

First let's simplify matters by considering the behaviour of electromagnetic fields in *empty space*, where

$$\rho = 0, \quad \vec{\boldsymbol{J}} = 0, \quad \vec{\boldsymbol{D}} = \epsilon_\circ \vec{\boldsymbol{E}} \quad \text{and} \quad \vec{\boldsymbol{B}} = \mu_\circ \vec{\boldsymbol{H}}.$$

Our two equations then read

$$\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{E}} = -\frac{\partial \vec{\boldsymbol{B}}}{\partial t} \quad \text{and} \quad \frac{\vec{\boldsymbol{\nabla}} \times \vec{\boldsymbol{B}}}{\mu_\circ} = \epsilon_\circ \frac{\partial \vec{\boldsymbol{E}}}{\partial t}.$$

We can simplify further by assuming that the electric field is in the $\hat{y}$ direction and the magnetic field is in the $\hat{z}$ direction. In that case,

$$\frac{\partial E}{\partial x} = -\frac{\partial B}{\partial t} \quad \text{and} \quad \frac{\partial B}{\partial x} = -\epsilon_\circ \mu_\circ \frac{\partial E}{\partial t}$$

where the second equation has been multiplied through by $\mu_\circ$.

If we now take the derivative of the first equation with respect to x and derivative of the second equation with respect to t, we get

$$\frac{\partial^2 E}{\partial x^2} = -\frac{\partial^2 B}{\partial x \partial t} \quad \text{and} \quad \frac{\partial^2 B}{\partial t \partial x} = -\epsilon_\circ \mu_\circ \frac{\partial^2 E}{\partial^2 t}.$$

$$\text{Since} \qquad \frac{\partial^2 B}{\partial x \partial t} = \frac{\partial^2 B}{\partial t \partial x},$$

the combination of these two equations yields

$$\frac{\partial^2 E}{\partial x^2} = \epsilon_\circ \mu_\circ \frac{\partial^2 E}{\partial^2 t}$$

which the discerning reader will recognize as the one-dimensional WAVE EQUATION for E,

$$\boxed{\frac{\partial^2 E}{\partial x^2} - \frac{1}{c^2}\frac{\partial^2 E}{\partial^2 t} = 0} \tag{15}$$

where the propagation velocity is

$$\boxed{c = \frac{1}{\sqrt{\epsilon_\circ \mu_\circ}}.} \tag{16}$$

You can easily show that there is an identical equation for B.

A more general derivation yields the 3-dimensional version,

$$\boxed{\nabla^2 \vec{\boldsymbol{E}} \;=\; \frac{1}{c^2}\,\frac{\partial^2 \vec{\boldsymbol{E}}}{\partial t^2} \quad \text{or} \quad \Box^2 \vec{\boldsymbol{E}} = 0\,.} \tag{17}$$

In either form, this equation expresses the fact that, since a changing electric field generates a magnetic field but that change in the magnetic field generates, in turn, an electric field, and so on, we can conclude that *electromagnetic fields*

will propagate spontaneously through regions of total vacuum in the form of a *wave* of $\vec{\boldsymbol{E}}$ and $\vec{\boldsymbol{B}}$ fields working with/against each other.

This startling conclusion (in 1865) led to the revision of all the "classical" paradigms of Physics, even such fundamental concepts as space and time.

Chapter 24

A Short History of Atoms

An accurate historical account of the development of Atomic Physics is probably the most hopeless task in the History of Science discipline. The story began, almost certainly, before the dawn of recorded history. Written records from Western antiquity date from as early as 450 BC, when the Greek Leucippus proposed that all matter was composed of $\alpha\tau o\mu$s, *i.e.* miniscule indestructible subunits of which there are only a few basic species. This view was picked up by Leucippus' student, Democritos of Abdera, some 50 years later and popularized by Epicurus of Samos around 300 BC, who developed the "Atomist" philosophical system that was epitomized by the Roman philosopher and poet Titus Lucretius Carus in about 60 BC.

Meanwhile, in 335 BC Aristotle countered with the proposition that matter was not grainy (as would seem to be required by the Atomist view) but smoothly *continuous* and composed of four basic elements, also continuous: earth, air, fire and water. This picture gained popularity around 300 BC under Zeno of Cition, founder of the Stoics.

Thus the battle lines between a "bricks and mortar" view of matter and a "continuous" image of space, time and substance had been drawn well before the birth of Jesus; it took until the Twentieth Century to find the synthesis that allowed these two pictures (both of which, incidentally, are correct) to coexist in peace, though perhaps at the expense of what once passed for common sense.

Probably one key paradigm was Newton's calculus, which taught everyone to understand continuous mathematical behaviour in terms of discrete "differentials" whose intervals were allowed to go to zero. Thus by the Nineteenth Century all scientists and mathematicians were intimately familiar with this trick for making the *smooth* look *grainy* and *vice versa*. The psychological stage was set for a new *physical* paradigm that reconciled Democritus' Atomism with Aristotle's Elements.

There was also an enormous amount of work done in the Middle Ages on determining exactly which ordinary household materials were true elements and which were *combinations* of several elements — what we now call chemical compounds. This was the work of untold numbers of Alchemists, most of whose work was done in secret for fear of persecution by those who considered such matters to be none of Humanity's business. Nevertheless, by the turn of the Nineteenth Century, a great many true elements had been correctly identified and some regularities had begun to appear.

The next difficulty with the History of Atomic Physics is that a lot of it is Chemistry. Even after Alchemy became respectable under the new name of Chemistry, a certain mutual disdain was cherished between Physicists and Chemists — which unfortunately lives on to this day — and consequently the History of Atomism reads a little differently in the Chemistry textbooks from the Physics version. Both are equally le-

gitimate, of course, but since History is subject to politics and revisionism, one must always read any account with a certain healthy skepticism.

I will therefore make no claim that my account is fair, or even historically accurate; rather, my goal will be to show how the ideas *might* have developed in a perfectly logical sequence, using the powerful optics of hindsight. If you are stimulated by this "fake history" to go learn for yourself what *really* happened, then I will consider my goal achieved.

24.1 Modern Atomism

Most Physicists (and all Chemists) will probably agree that the crucial empirical observations that set modern science on the track of atoms (as we now know them) occurred around the transition between the Eighteenth and Nineteenth Centuries when a number of scientists including Antoine Laurent Lavoisier, Bryan and William Higgins, Joseph Louis Proust, John Dalton and Joseph Louis Gay-Lussac[1] discovered that certain chemical agents combined in simple integer ratios of their "MOLECULAR WEIGHTS" with other agents, a phenomenon most easily explained by assuming that these agents were the true chemical *elements* sought by the Alchemists[2] and furthermore that one MOLECULAR WEIGHT of *any* ELEMENT contained *the same number of* ATOMS of that element! This specific hypothesis is credited to Lorenzo Romano Amadeo Avogardo who in 1811 made a clear distinction between ATOMS (irreducible chemical units) and MOLECULES, which are clumps of atoms. For his trouble he got AVOGADRO'S NUMBER N_0 named after him. The actual *number* of atoms (or, for that matter, molecules) in one MOLECULAR WEIGHT (or MOLE) of the corresponding element is

$$N_0 \equiv 6.02205 \times 10^{23} \text{molecules per mole.} \quad (1)$$

You may recognize this number from the Chapter on THERMAL PHYSICS, in particular the Section on the KINETIC THEORY OF GASES, the qualitative assumptions of which dated back as far as Robert Boyle, Robert Hooke and Isaac Newton himself in the late Seventeenth Century. The work of Daniel Bernoulli in 1738 foreshadowed the use of kinetic theory by Joseph Loschmidt in 1865 to make the first determination of the *value* of N_0 from measurements of the actual behaviour of gases. STATISTICAL MECHANICS actually played a major rôle in the development of modern Atomic theory, but its rôle is often downplayed in historical accounts simply because its is harder to understand. I will probably do likewise — but at least I admit it!

24.2 What are Atoms Made of?

By the end of the Nineteenth Century [I am leaving out a lot here!] most scientists were convinced that *atoms* were "real" (as opposed to a mere calculational aid or a handy mnemonic paradigm) and were looking for ways to determine their true structure.

24.2.1 Thomson's Electron and e/m

It was found that negatively charged particles called "cathode rays" could be coaxed out of a hot metal filament by a large enough electric potential and accelerated to hit a screen covered with phosphorescent material where they made a bright spot [the forerunner of today's *c*athod *r*ay *t*ubes or CRT's], but until 1897 no one knew much about the properties of these particles. In that year Joseph John Thomson used magnetic

[1] As you might guess, the details of the history of these discoveries also tend to vary with the nationality of the Historian!

[2] The Alchemists were already pretty certain of many of these, of course; but they were accustomed to keeping their mouths shut.

deflection (the LORENTZ FORCE) to determine the charge-to-mass ratio q/m of the cathode rays.[3] He found an astonishingly large negative ratio: $q/m = -1.76 \times 10^{11}$ coulombs/kg, indicating that the ELECTRON (as the "cathode ray" particle soon came to be known) must be a very light particle (mass m_e) with a very large electric charge ($q = -e$) where the "electronic charge" e was thought until recently to be the QUANTUM of electric charge — *i.e.* the irreducible minimum nonzero *quantity* of electric charge, in integer multiples of which all larger charges must come.[4]

24.2.2 Milliken's Oil Drops and e

Of course, this result revealed nothing about either e or m_e, just their *ratio*. But the absolute magnitude of e was determined ten years later by Robert A. Millikan, who watched tiny droplets of mineral oil through a microscope: the spherical oil drops, created with an ordinary atomizer (no pun intended), *fell* through still air in the Earth's gravity at a terminal velocity determined by their weight and the frictional drag of the air, both of which can be calculated from their radius. Now, every once in a while one of the drops would pick up a stray electron and become charged. If the experiment was performed in a vertical electric field of adjustable strength, the charged droplets could be made to "hover" by applying just the right voltage to overcome the force of gravity. Then, knowing the electric field, Millikan was able to calculate the charge.[5] The result was $e \approx 1.6 \times 10^{-19}$ C, which meant that the *mass* of the electron must be *really* small, namely $m_e \approx 9.1 \times 10^{-31}$ kg.

24.2.3 "Plum Puddings" *vs.* Rutherford

The discovery of that the ELECTRON was such an incredibly *lightweight* particle with such a huge *charge* made it perfectly clear that an ATOM must be something like a "plum pudding" — a homogeneous, featureless matrix of positive charge (carrying most of the mass) with the electrons embedded in it like raisins. Otherwise the electrons were apt to be *moving*, and this was unthinkable! If they were in motion but stayed inside the atom, then they must be continually changing direction. That means they must be *accelerated*, and by that time everyone understood only too well that

accelerated charges radiate!

Specifically, an accelerated charge (especially one with such a *large* charge-to-mass ratio) must always radiate away energy in the form of electromagnetic waves — it is a sort of *antenna* — and so the normal quiescence of matter "proves" that the electrons must be *at rest* in their atoms; this can only be so if they are "stuck" in a "plum pudding" of positive charge.[6]

In about 1910 a new type of "radioactivity" was discovered: certain nuclei spontaneously emit "α rays" which were shown to have a q/m ratio nearly 4000 times smaller than electrons and where therefore much heavier particles. Soon afterwards, Ernest Rutherford set out to demonstrate the correctness of the "plum

[3] Such a device (for measuring the charge-to-mass ratio of electrically charged particles) is known as a MAGNETIC SPECTROMETER. Thomson's version was pretty crude by today's standards, but this is still the most accurate method for measuring the q/m ratio of particles (and hence, if we know their charge by some other means, their *mass*).

[4] This is really the original prototype example of a QUANTIZED property. Many others were to follow, as we shall see.

[5] Naturally, sometimes he got *two* or *three* electrons on a drop; but this was simple enough to take into account: sometimes he got a result of e, sometimes he got a result of $2e$, sometimes he got a result of $3e$, but he never got a result of $\frac{1}{2}e$, for instance, so it was clear which result was the true charge quantum.

[6] This is truly an unavoidable conclusion if we accept the theory of classical electrodynamics at face value; it was not just a misinterpretation. You may be sure that hordes of Physicists looked high and low for a way out of this and found none.

pudding" model of atoms by SCATTERING these α particles off gold atoms comprising a thin gold foil.

The picture is analogous to firing cannon balls at great slabs of gelatin in which are embedded many small marbles. The cannon balls will lose a lot of energy going through the gelatin walls, but they certainly won't change their direction of motion much.

To Rutherford's astonishment, most of the α particles passed right through the target foil without being deflected *or* losing much energy — indicating that what seemed to be "solid" metal was actually composed mainly of sheer vacuum. Even more alarmingly, *some* of the α particles *bounced backward* off the gold atoms — indicating that the mass of the gold atom was almost all concentrated in a tiny hard kernel of positive charge some 10,000 to 100,000 times smaller than the size of the atoms themselves!

As Rutherford himself put it, "*It is like firing shells at a piece of paper handkerchief and having them bounce back at you.*"

Scattering Cross Sections

Inasmuch as we are going to discuss modern ELEMENTARY PARTICLE PHYSICS later on, it is appropriate to stop for a moment and contemplate Rutherford's classic experiment, for the art of interpreting the distributions of SCATTERING ANGLES when a beam of one type of particle in a well-defined initial state is slammed into a target composed of other types of particles is essentially the entire experimental *repertoire* of the modern Particle Physicist.

Consider: the goal of the experimenter is to learn more about the *structure* of particles that are, individually, too small to be detected with a microscope. [If the particle is much smaller in size than the *wavelength* λ of the light used in the microscope, the best it can do is scatter the light into spherical outgoing wavefronts (HUYGENS' PRINCIPLE), from which we can learn nothing about the shape of the particle itself. The approved terminology for this limitation is that the RESOLUTION of the microscope can never be finer than the *wavelength* of the light it uses.] So how *can* we learn anything about the shape of the object particle? By SCATTERING other particles off it!

Imagine that there is an object hidden from sight behind a thin piece of paper; you have a BB gun which you can use to bounce BBs off the object. You get to to see which way the BBs bounce, and if you have a more fancy apparatus you may get to measure their velocities (momenta) before and after their collisions with the object; moreover, if any bits fly off the object as a result of a BB collision, you get to measure their directions and momenta as well. This is essentially the situation of the Particle Physicist. We may have a variety of PARTICLE BEAMS ranging from electrons to heavy nuclei, with energies ranging from a few eV to many GeV (billions of eV) or even TeV (trillions of eV) per particle — corresponding to peashooters, BB guns, rifles, howitzers and rail guns — but the only way we can use them is to shoot "blind" at our target particles and study the SCATTERING DISTRIBUTION.

You should try to imagine for yourself some qualitative phenomena you might look for to test various hypotheses about the target object — starting with Rutherford's test for "plum puddings" *vs.* hard-kernel ATOMIC NUCLEI. I will not attempt to develop the arcane terminology of scattering theory here, but I will mention the basic paradigm: the thing one can measure and describe most easily about a particle is the *area* it presents to an incoming beam; we call this the SCATTERING CROSS SECTION and measure it in area units such as BARNS [one BARN $\equiv (10^{-13}\ \mathrm{cm})^2$ or $10^{-30}\ \mathrm{m}^2$)] — about the size of an average nucleus.[7]

[7] This humourous name for the *size of a target* may have

24.2.4 A Short, Bright Life for Atoms

A new picture of the atom thus emerged, in which all the positive charge and virtually all the mass was concentrated in a tiny NUCLEUS at the centre of the atom and the light, negatively charged ELECTRONS orbited about it at rather large distances, much like the Earth and other planets about the Sun. This is a compelling and pretty image, and there is no problem calculating the orbital velocities of the electrons in the attractive central force of the nucleus.

The problem is, the *accelerations* of said electrons are *enormous*, causing them to *radiate* away their energy as electromagnetic waves (light) and spiral down into the nucleus. The lifetime of such an atom must be less than about 1 ns (or 10^{-9} seconds), during which time the atom gives off a bright pulse of light. Then, nothing.

This doesn't quite fit the data. Atoms are apparently quite stable and we are still here to talk about it, so there must be something wrong with this picture. Naturally, armies of Physicists went to work trying to find fault with the logic of classical electrodynamics, but there was no way out; the predictions were too simple to be mistaken. Something was seriously wrong.

---

marked the start of a trend toward "cute" nomenclature in Particle Physics, which manifested itself later in *strangeness*, *quarks* and (most recently) *truth* and *beauty* as particle properties — the latter pair now being retracted in favour of *top* and *bottom*, which I regard as a failure of nerve and will on the part of Particle Physicists. But that is yet another story. . . .

24.3 Timeline: "Modern" Physics

450-300 BC Greek Atomists: **Leucippus, Democritos, Epicurus** . . .

335 BC **Aristotle**: continuous elements (earth, air, fire, water)

300 BC **Zeno** of Cition (founder of Stoics) popularizes Aristotelian view.

60 BC Titus Lucretius **Carus** of Rome epitomizes "Atomist" philosophy.

.
.
.

1879 Josef **Stefan** [expt] power emitted as blackbody radiation $\boldsymbol{P = A\sigma T^4}$

1884 Ludwig **Boltzmann** [theor] explains Stefan's empirical law

1885 Johann Jakob **Balmer** [expt]
empirical description of **line spectra** emitted by **H atoms**

1890 Johannes Robert **Rydberg** [expt]

1893 Wilhelm **Wien** [expt] blackbody spectrum **displacement law**:
peak wavelength varies as $\boldsymbol{T^{-1}}$

1895 Wilhelm Conrad **Roentgen** [expt] discovers **X-rays**

1897 Joseph John **Thomson** [expt] measures boldmath q/m of the **electron**

1900 Max **Planck** [theor] derives correct **blackbody radiation spectrum**

1902 Philipp E.A. von **Lenard** [expt] measures **photoelectric effect**

1905 Albert **Einstein** [theor] explains **photoelectric effect**

1905 Albert **Einstein** [theor] publishes **Special Theory of Relativity (STR)**

1905 Albert **Einstein** [theor] explains **Brownian motion** (gives **mass of atoms**!)

1905 Ernest **Rutherford** [expt] performs first **alpha-scattering** experiments at McGill Univ. (Canada)

1907 Robert A. **Milliken** [expt] measures **electron charge** (now know both $\boldsymbol{q_e}$ and $\boldsymbol{m_e}$).

1912 William (H. & L.) **Bragg** [expt] shows that **X-rays** scatter off **crystal lattices**

1913 Hans **Geiger** & Ernest **Marsden** [expt] confirm **Rutherford scattering** results at Univ. of Manchester (U.K.)

1913 Niels Henrik David **Bohr** [theor] pictures **H atom** with **quantized angular momentum**

1916 Albert **Einstein** [theor] publishes **General Theory of Relativity (GTR)**

1916 Robert Andrews **Milliken** [expt] confirms **photoelectric effect** in detail

1922 Arthur Holly **Compton** [expt] scatters **X-rays** off **electrons**

1924 Louis Victor **de Broglie** [theor] hypothesizes "**matter waves**" with $\boldsymbol{\lambda = h/p}$

1925 Wolfgang **Pauli** [theor] formulates his **exclusion principle**

1925 Max **Born** & Werner **Heisenberg** [theor] introduce **quantum mechanics**

1926 Erwin **Schroedinger** [theor] develops a nonrelativistic **wave equation** for **quantum mechanics**

1927 Werner **Heisenberg** [theor] formulates his **uncertainty principle**

1928 Paul A.M. **Dirac** [theor] develops a **relativistic wave equation** for electrons and predicts **antimatter**

.
.
.

24.4 Some Quotations

Lord Rutherford, 1931:

"When we consider the life work of Faraday it is clear that his researches were guided and inspired by the strong belief that the various forces of nature were inter-related and dependent on one another. It is not too much to say that this philosophic conviction gave the impulse and driving power in most of his researches and is the key to his extraordinary success in adding to knowledge.

"The more we study the work of Faraday with the perspective of time, the more we are impressed by his unrivalled genius as an experimenter and natural philosopher. When we consider the magnitude and extent of his discoveries and their influence on the progress of science and industry, there is no honor too great to pay to the memory of Michael Faraday — one of the greatest scientific discoverers of all time."

Maxwell:

"As I proceeded with the study of Faraday, I perceived that his method of conceiving the phenomena was also a mathematical one, though not exhibited in the conventional form of mathematical symbols. I also found that these methods were capable of being expressed in the ordinary mathematical form, and thus compared with those of the professed mathematicians." — *Treatise on Electricity and Magnetism, 1873*

Faraday:

"When a mathematician engaged in investigating physical actions and results has arrived at his conclusions, may they not be expressed in common language as fully clearly and definitely as in mathematical formulae? If so, would it not be a great boon to such as I to express them so — translating them out of their hieroglyphics that we also might work upon them by experiment." — letter to James Clerk Maxwell

Maxwell:

"I was aware that there was supposed to be a difference between Faraday's way of conceiving phenomena and that of the mathematicians, so that neither he nor they were satisfied with each other's language. I had also the conviction that this discrepancy did not arise from either party being wrong. I was first convinced of this by Sir William Thomson, to whose advice and assistance, as well as to his published papers, I owe most of what I have learned on the subject." — *Treatise on Electricity and Magnetism, 1873*

Maxwell:

"...we have strong reason to conclude that light itself — including radiant heat, and other radiations if any — is an electromagnetic disturbance in the form of waves propagated through the electromagnetic field according to electromagnetic laws." — *Dynamical Theory of the Electromagnetic Field, 1864*

Einstein:

"The greatest alteration in the axiomatic basis of physics — in our conception of the structure of reality — since the foundation of theoretical physics by Newton, originated in the researches of Faraday and Maxwell on electromagnetic phenomena."

Boltzmann:

"Available energy is the main object at stake in the struggle for existence and the evolution of the world." [in D'A. W. Thompson, *On Growth and Form* (Cambridge 1917).]

"The most ordinary things are to philosophy a source of insoluble puzzles. With infinite ingenuity it constructs a concept of space or time and then finds it absolutely impossible that there be objects in this space or that processes occur during this time ... the source of this kind of logic lies in excessive confidence in the so-called laws of thought." [in B. McGuinness, *Ludwig Boltzmann, Theoretical Physics and Philosophical Problems*, (Dordrecht, 1974) 64.]

"To go straight to the deepest depth, I went for

Hegel; what unclear thoughtless flow of words I was to find there! My unlucky star led me from Hegel to Schopenhauer.... Even in Kant there were many things that I could grasp so little that given his general acuity of mind I almost suspected that he was pulling the reader's leg or was even an imposter." [in *D. Flamm. Stud. Hist. Phil. Sci.* **14** (1983) 257.]

"One should not imagine that two gases in a 0.1 liter container, initially unmixed, will mix, then again after a few days separate, then mix again, and so forth. On the contrary, one finds ... that not until a time enormously long compared to $10^{10^{10}}$ years will there be any noticeable unmixing of the gases. One may recognize that this is practically equivalent to never...."

Einstein:

"A theory is the more impressive the greater the simplicity of its premises, the more different kinds of things it relates, and the more extended its area of applicability. Therefore the deep impression that classical thermodynamics made upon me. It is the only physical theory of universal content which I am convinced will never be overthrown, within the framework of applicability of its basic concepts."

Planck:

"The general connection between energy and temperature may only be established by probability considerations. [Two systems] are in statistical equilibrium when a transfer of energy does not increase the probability."

Thomson: [toast]

"To the electron: may it never be of any use!"

Niels Bohr:

"Evidence obtained under different experimental conditions cannot be comprehended within a single picture, but must be regarded as complementary in the sense that only the totality of the phenomena exhausts the possible information about the objects."

"Notwithstanding the fundamental departure from the ideas of the classical theories of mechanics and electrodynamics involved in these postulates, it has been possible to trace a connection between the radiation emitted by the atom and the motion of the particles which exhibits a far-reaching analogy to that claimed by the classical ideas of the origin of radiation."

H.B.G. Casimir:

"Even Bohr, who concentrated more intensely and had more staying power than any of us, looked for relaxation in crossword puzzles, in sports, and in facetious discussions."

Rutherford:

"Bohr's different [from other Continental theorists] — he's a football player!"

Pauli: [writing about his days as a student at Munich]

"I was not spared the shock which every physicist accustomed to the classical way of thinking experienced when he came to know Niels Bohr's basic postulate of quantum theory for the first time."

Heisenberg:

"I learned optimism from Sommerfeld, mathematics at Göttingen, and physics from Bohr."

Louis de Broglie:

"Two seemingly incompatible conceptions can each represent an aspect of the truth.... They may serve in turn to represent the facts without ever entering into direct conflict." — *Dialectica*

"As in my conversations with my brother we always arrived at the conclusion that in the case of x-rays one had both waves and corpuscles, thus suddenly — ... it was certain in the course of summer 1923 — I got the idea that one had to extend this duality to material particles, especially to electrons. And I realised that, on the one hand, the Hamilton-Jacobi theory pointed somewhat in that direction, for it can be applied to particles and, in addition, it represents a geometrical optics; on the other hand, in quantum phenomena one obtains quantum

numbers, which are rarely found in mechanics but occur very frequently in wave phenomena and in all problems dealing with wave motion." — *from an interview in 1963*

De Broglie described himself as "...having much more the state of mind of a pure theoretician than that of an experimenter or engineer, loving especially the general and philosophical view...."

"...the statistical theories hide a completely determined and ascertainable reality behind variables which elude our experimental techniques."

Einstein: [after reading Pauli's article on relativity]

"Whoever studies this mature and grandly conceived work might not believe that its author is a twenty-one year old man."

Wolfgang Pauli:

"...a new phase of my scientific life began when I first met Niels Bohr personally for the first time. During these meetings, Bohr asked me whether I could come to Copenhagen for a year."

"The fact that the author thinks slowly is not serious, but the fact that he publishes faster than he thinks is inexcusable."

"This paper is so bad it is not even wrong." — Quoted in D. MacHale, *Comic Sections* (Dublin 1993)

"I refuse to believe that God is a weak left-hander."

Max Born:

"I am now convinced that theoretical physics is actual philosophy." — *Autobiography*

"If God has made the world a perfect mechanism, He has at least conceded so much to our imperfect intellect that in order to predict little parts of it, we need not solve innumerable differential equations, but can use dice with fair sucess." — Quoted in H.R. Pagels, *The Cosmic Code*

"The difficulty involved in the proper and adequate means of describing changes in continuous deformable bodies is the method of differential equations. ... They express mathematically the physical concept of contiguous action." — *Einstein's Theory of Relativity*

One of his research students described Born's days in Edinburgh: "When Born arrived in the morning he first used to make the round of his research students, asking them whether they had any progress to report, and giving them advice, sometimes presenting them with sheets of elaborate calculations concerning their problems which he had himself done the day before. ... The rest of the morning was spent by Born in delivering his lectures to undergraduate honours students, attending to departmental business, and doing research work of his own. Most of the latter, however, he used to carry out at home in the afternoons and evenings."

(Born develops a nonrelativistic **wave equation** for the electron in **quantum mechanics**.)

Erwin Rudolf Josef Alexander Schrödinger: [about his time at the Akademisches Gymnasium in 1898]

"I was a good student in all subjects, loved mathematics and physics, but also the strict logic of the ancient grammars, hated only memorising incidental dates and facts. Of the German poets, I loved especially the dramatists, but hated the pedantic dissection of this work."

"Especially in physics and mathematics, Schrödinger had a gift for understanding that allowed him, without any homework, immediately and directly to comprehend all the material during the class hours and to apply it. After the lecture ... it was possible for [our professor] to call Schrödinger immediately to the blackboard and to set him problems, which he solved with playful facility." — *from a student in Schrdinger's class at school*

"A few days ago I read with great interest the

ingenious thesis of Louis de Broglie, which I finally got hold of..." — *letter to Einstein, 3 November 1925*

"I have been intensely concerned these days with Louis de Broglie's ingenious theory. It is extraordinarily exciting, but still has some very grave difficulties." — *a different letter on 16 November 1925*

"To each function of the position- and momentum-coordinates in wave mechanics there may be related a matrix in such a way that these matrices, in every case satisfy the formal calculation rules of Born and Heisenberg. ... The solution of the natural boundary value problem of this differential equation in wave mechanics is completely equivalent to the solution of Heisenberg's algebraic problem." — 1926 paper

Werner Karl Heisenberg:

"To those of us who participated in the development of atomic theory, the five years following the Solvay Conference in Brussels in 1927 looked so wonderful that we often spoke of them as the golden age of atomic physics. The great obstacles that had occupied all our efforts in the preceding years had been cleared out of the way; the gate to an entirely new field, the quantum mechanics of the atomic shells stood wide open, and fresh fruits seemed ready for the picking."

"An expert is someone who knows some of the worst mistakes that can be made in his subject, and how to avoid them." — *Physics and Beyond* (New York 1971)

Paul Adrien Maurice Dirac:

"I think that there is a moral to this story, namely that it is more important to have beauty in one's equations than to have them fit experiment. If Schroedinger had been more confident of his work, he could have published it some months earlier, and he could have published a more accurate equation. It seems that if one is working from the point of view of getting beauty in one's equations, and if one has really a sound insight, one is on a sure line of progress. If there is not complete agreement between the results of one's work and experiment, one should not allow oneself to be too discouraged, because the discrepancy may well be due to minor features that are not properly taken into account and that will get cleared up with further development of the theory." — in *Scientific American*, May 1963.

"Mathematics is the tool specially suited for dealing with abstract concepts of any kind and there is no limit to its power in this field." — Quoted in P.J. Davis and R. Hersh, *The Mathematical Experience* (Boston 1981)

"In science one tries to tell people, in such a way as to be understood by everyone, something that no one ever knew before. But in poetry, it's the exact opposite." — Quoted in H. Eves, *Mathematical Circles Adieu* (Boston 1977)

"I learned to distrust all physical concepts as the basis for a theory. Instead one should put one's trust in a mathematical scheme, even if the scheme does not appear at first sight to be connected with physics. One should concentrate on getting interesting mathematics."

"Now when Heisenberg noticed that, he was really scared." — Quoted in D. MacHale, *Comic Sections* (Dublin 1993)

"I consider that I understand an equation when I can predict the properties of its solutions, without actually solving it." — Quoted in F. Wilczek & B. Devine, *Longing for the Harmonies*

"This result is too beautiful to be false." — 'The evolution of the Physicist's Picture of Nature', *Scientific American* **208**, 5 (1963)

24.5 SKIT:

"The Dreams Stuff is Made Of"

Salviati: "That's all very well and good, but there is a problem. Actually there are several problems with Maxwell's theory of electrodynamics and light. First off, the equations describe a propagating electromagnetic wave, but they don't mention what it is propagating *in*, and for that reason they don't specify what it is propagating *relative to*."

Simplicio: "What does that matter?"

Salviati: "Simplicio, you idiot, if the wave propagates past *me* at the velocity c predicted by Maxwell's electrodynamics, but *you* are moving relative to *me* at some large velocity, then the wave obviously can't be moving past *you* at that same c. But the equations say it is!"

Simplicio: "So the equations must be wrong, right?"

Salviati: "No, the equations are right. Common sense is wrong."

Sagredo: "What?!"

Salviati: "Never mind, that's Relativity. We have enough of a problem trying to make sense of the *other* unambiguous prediction of Maxwell's equations: that any time a *charge* is *accelerated*, it *radiates away energy* in the form of electromagnetic waves."

Sagredo: "So what's the problem with that? Isn't that how antennas make radio waves?"

Salviati: "Yes, but it's also why any *atom* consisting of a negatively charged electron orbiting around a heavy, positively charged nucleus will fall into it within about a billionth of a second. Too bad, you're dead."

Simplicio: "I still don't see a problem. Obviously *atoms* must not have that form."

Salviati: "Obviously. Unfortunately, they *do* have that form. Maybe we'd better visit Balmer and Rydberg, who are just starting to collect some perplexing data on the light emitted by hot atoms. . . ."

England, 1885-1890:

Balmer and Rydberg: Conducting experiments with a Bunsen burner and a nichrome wire on the other side of the stage.

[Make up some dialogue.]

Outcome: argument about the empirical formulae describing the line spectra, leading to the general idea that the light frequencies are *differences* between some maximum frequency and other frequencies that are smaller by factors of $1/n^2$ where n is an integer.

Salviati: "Anyone who ever threw bits of stuff into a campfire and noticed the different colours produced when different materials burns knows that atoms give off light when they are heated enough. Analytical chemists know that one good way to identify certain elements (especially metals) is the *flame test*, in which you vapourize bits of your mystery sample in a Bunsen burner and look at the pretty colours. The question these guys were trying to answer was, *How come those particular colours?*

"So they set up *spectrometers* to find out how much light of different colours was in the atomic spectra, and to everyone's amazement they found *line spectra!* That is, only certain very pure colours are emitted from a given element's atoms. This was completely mysterious and no one had the faintest idea what was going on, although Balmer, Ritz and Rydberg made up very successful empirical formulas that could accurately predict the wavelengths of the light emitted.

"The simplest atom, the *hydrogen atom*, was the first to be described with this empirical precision, and soon it was to be the first to be un-

derstood in terms of a radical new theory."

Sagredo: "Wait a minute. We know that if we heat up a wire it first gets red hot, then orange, then yellow and finally white-hot. How can atoms give off the same colours regardless of how hot you get them?"

Salviati: "An excellent question, Sagredo, and one which especially bothered the spectroscopists, because a few years earlier this characteristic *blackbody spectrum* had been described in some detail. But that description also eluded explanation for two decades. Let's go to Germany and visit Stefan, Boltzmann and Wien. . . ."

Simplicio: "This is pretty complicated. Are all the new ideas that made up Quantum Mechanics this . . . *obscure*?"

Salviati: "No, Simplicio, the blackbody radiation part was the worst. Now let's visit the Cavendish Laboratory at Oxford in 1897 and see what J.J. Thomson can tell us about *electrons*."

Salviati: "Thomson has just built the first television set. The first cathode ray tube, anyway; there are no TV signals to receive, so I guess it's an exaggeration to call it a TV set. But the principle is the same."

Simplicio: "You mean *cathode rays* are *electrons*?"

Salviati: "Simplicio, sometimes you surprise me! Yes, that's exactly right."

Sagredo: "So what's the big deal?"

Salviati: "Ah, well, the trouble is, Thomson's electrons turn out to have an electric charge about a thousand times bigger than they ought to, given their mass. Actually Thomson only found the *ratio* of their charge to their mass. It wasn't until 1907 that Milliken measured the electron's charge directly by putting a single extra electron on a tiny oil droplet and watching how it moved when he applied an electric field. But already in 1897 Thomson showed that electrons carried a lot of charge and very little mass. The rest of the atom has most of the mass but only the same amount of charge (only opposite) as all its electrons."

Simplicio: "The *nucleus*, right?"

Salviati: "No, Simplicio, they didn't know about the nucleus yet. People just assumed the atom must be like a *plum pudding*, with the positive charge all spread around with the mass and the electrons stuck in it like raisins."

Simplicio: "What a dumb idea! It only makes sense to think of the electrons like planets orbiting the Sun."

Sagredo: "Actually, Simplicio, it makes no sense at all."

Simplicio: "What do you mean?"

Sagredo: "Look, Simplicio, remember Electrodynamics? Any accelerated charge radiates away energy as electromagnetic waves, right? So if an electron is in an *orbit*, it is *constantly accelerated* and it *has* to radiate away all its energy and fall into the nucleus. Has to. Lasts about a billionth of a second, remember?"

Simplicio: "But atoms are stable!"

Salviati and Sagredo [in unison]: "*Duh!*"

Simplicio: "How is that possible?"

Salviati: "Now you're beginning to see the problem. Everyone knew it must have something to do with Rydberg's and Balmer's weird empirical rules about the quantization of frequencies of light emitted by atoms, but no one could make any sense of it until 1913.

"But we're getting ahead of ourselves. Before Bohr could make his outrageous (and incorrect) hypothesis about hydrogen atoms, the way had to be prepared by none other than Albert Einstein."

Salviati: "In 1902 Philipp E.A. von Lenard discovered that you can't excite electrons out of a metal using low frequency light, no matter how high you turn up the intensity. But if you use even a little bit of *higher frequency* light, out

pop lots of electrons. So it isn't the *power* of the electromagnetic radiation, it's the *frequency*."

Simplicio: "That doesn't make any sense."

Salviati: "Bingo. That's where Albert comes in. He had a knack for showing simple reasons for inexplicable results; but you had to be willing to trust logic more than common sense. In 1905, sitting at his desk in a Swiss patent office in Bern, he came up with three papers that changed the world. One was on the Special Theory of Relativity, one was on Brownian motion (that was his dissertation) and the third was an explanation of Lenard's *photoelectric effect*."

Simplicio: "I don't see how it explains atomic spectra."

Salviati: "Just remember that $E_\gamma = h\nu = \hbar\omega$ (where $\hbar \equiv h/2\pi$) and we'll skip over to Montreal where Ernest Rutherford is bouncing a beam of Marie Curie's *alpha* particles off gold atoms. . . ."

Chapter 25

The Special Theory of Relativity

Let's briefly recapitulate the situation in 1865: MAXWELL'S EQUATIONS, which correctly described all the phenomena of electromagnetism known in the mid-19$^{\text{th}}$ Century (and then some), predicted also that electromagnetic fields should satisfy the WAVE EQUATION — *i.e.*, by virtue of a changing $\vec{\boldsymbol{E}}$ creating $\vec{\boldsymbol{B}}$ and *vice versa*, the electric and magnetic fields would be able to "play off each other" and propagate through space in the form of a *wave* with all the properties of *light* (or its manifestations in shorter and longer wavelengths, which we also term "light" when discussing electromagnetic waves in general). Fine, so far.

But there are some unsettling implications of this "final" explanation of light. First of all (and the focus of this Chapter) is the omission of any reference to a *medium* that does the "wiggling" as the electromagnetic wave goes through it. Water waves propagate through water, sound waves through air, liquid or solid, plasma waves through plasmas, *etc.* This was the first time anyone had ever postulated a wave that *just propagated by itself* through *empty vacuum* (or "free space," as it is often called in this context). Moreover, the propagation velocity of light (or any electromagnetic wave) through the vacuum is given unambiguously by MAXWELL'S EQUATIONS to be $c = 2.99792458 \times 10^8$ m/s, *regardless of the motion of the observer.*

25.1 Galilean Transformations

So what? Well, this innocuous looking claim has some *very* perplexing logical consequences with regard to *relative velocities*, where we have expectations that follow, seemingly, from self-evident common sense. For instance, suppose the propagation velocity of ripples (water waves) in a calm lake is 0.5 m/s. If I am walking along a dock at 1 m/s and I toss a pebble in the lake, the guy sitting at anchor in a boat will see the ripples move by at 0.5 m/s but I will see them *dropping back* relative to me! That is, I can "outrun" the waves. In mathematical terms, if all the velocities are in the same direction (say, along x), we just *add* relative velocities: if v is the velocity of the wave relative to the water and u is my velocity relative to the water, then v', the velocity of the wave relative to *me*, is given by $v' = v - u$. This common sense equation is known as the GALILEAN VELOCITY TRANSFORMATION — a big name for a little idea, it would seem.

With a simple diagram, we can summarize the common-sense GALILEAN TRANSFORMATIONS (named after Galileo, no Biblical reference):

First of all, it is self-evident that $t' = t$, otherwise nothing would make any sense at all.[1] Nevertheless, we include this explicitly. Similarly, if the relative motion of O' with respect to O is only in the x direction, then $y' = y$

[1] By now, this phrase should alert you to the likelihood of error.

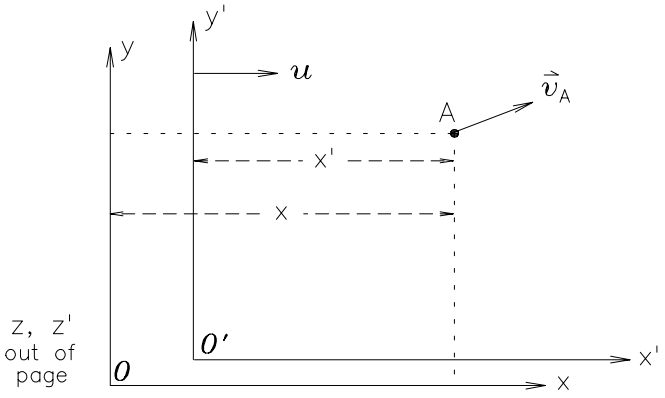


Figure 25.1 Reference frames of a "stationary" observer O and an observer O' moving in the x direction at a velocity u relative to O. The coordinates and time of an event at A measured by observer O are $\{x, y, z, t\}$ whereas the coordinates and time of the *same* event measured by O' are $\{x', y', z', t'\}$. An object at A moving at velocity $\vec{\boldsymbol{v}}_A$ relative to observer O will be moving at a different velocity $\vec{\boldsymbol{v}}'_A$ in the reference frame of O'. For convenience, we always assume that O and O' coincide initially, so that everyone agrees about the "origin:" when $t = 0$ and $t' = 0$, $x = x'$, $y = y'$ and $z = z'$.

and $z' = z$, which were true at $t = t' = 0$, must remain true at all later times. In fact, the only coordinates that *differ* between the two observers are x and x'. After a time t, the distance (x') from O' to some object A is *less* than the distance (x) from O to A by an amount ut, because that is how much *closer* O' has *moved* to A in the interim. Mathematically, $x' = x - ut$.

The *velocity* $\vec{\boldsymbol{v}}_A$ of A in the reference frame of O also looks different when viewed from O' — namely, we have to subtract the relative velocity of O' with respect to O, which we have labelled $\vec{\boldsymbol{u}}$. In this case we picked $\vec{\boldsymbol{u}}$ along $\hat{x}$, so that the vector subtraction $\vec{\boldsymbol{v}}'_A = \vec{\boldsymbol{v}}_A - \vec{\boldsymbol{u}}$ becomes just $v'_{A_x} = v_{A_x} - u$ while $v'_{A_y} = v_{A_y}$ and $v'_{A_z} = v_{A_z}$. Let's summarize all these "coordinate transformations:"

The Galilean transformations:

Coordinates:

$$x' = x - ut \quad (1)$$
$$y' = y \quad (2)$$
$$z' = z \quad (3)$$
$$t' = t \quad (4)$$

Velocities:

$$v'_{A_x} = v_{A_x} - u \quad (5)$$
$$v'_{A_y} = v_{A_y} \quad (6)$$
$$v'_{A_z} = v_{A_z} \quad (7)$$

This is all so simple and obvious that it is hard to focus one's attention on it. We take all these properties for granted — and therein lies the danger.

25.2 Lorentz Transformations

The problem is, *it doesn't work for light.* Without any *stuff* with respect to which to measure relative velocity, one person's vacuum looks exactly the same as another's, even though they may be moving past each other at enormous velocity! If so, then Maxwell's equations tell *both* observers that they should "see" the light go past them at c, even though one *observer* might be moving at $\frac{1}{2}c$ relative to the other!

The only way to make such a description *self-consistent* (not to say reasonable) is to allow *length* and *duration* to be different for observers moving relative to one another. That is, x' and t' must differ from x and t *not only* by additive constants but also by a *multiplicative factor.*

For æsthetic reasons I will reproduce here the equations that provide such coordinate transformations; the derivation will come later.

The ubiquitous factor γ is equal to 1 for

vanishingly small relative velocity u and grows without limit as $u \to c$. In fact, if u ever got as big as c then γ would "blow up" (become infinite) and then (worse yet) become *imaginary* for $u > c$.

The LORENTZ TRANSFORMATIONS:

Coordinates:

$$x' = \gamma\,(x - ut) \tag{8}$$
$$y' = y \tag{9}$$
$$z' = z \tag{10}$$
$$t' = \gamma\,(t - \frac{ux}{c^2}) \tag{11}$$

Velocities:

$$v'_{A_x} = \frac{v_{A_x} - u}{1 - uv_{A_x}/c^2} \tag{12}$$
$$v'_{A_y} = \frac{v_{A_y}}{\gamma\,(1 - uv_{A_x}/c^2)} \tag{13}$$
$$v'_{A_z} = \frac{v_{A_z}}{\gamma\,(1 - uv_{A_x}/c^2)} \tag{14}$$

where $\beta \equiv \frac{u}{c}$ (15)

and $\gamma \equiv \frac{1}{\sqrt{1-\beta^2}}$ (16)

25.3 Luminiferous Æther

This sort of nonsense convinced most people that MAXWELL'S EQUATIONS were *wrong* — or, more charitably, *incomplete*. The obvious way out of this dilemma was to assume that what we perceive (in our ignorance) as *vacuum* is actually an extremely peculiar *substance* called the "luminiferous æther" through which ordinary "solid" matter passes more or less freely but in which the "field lines" of electromagnetism are actual "ripples." (Sort of.) This recovers the rationalizing influence of a *medium* through which light propagates, at the expense of some pretty unfamiliar properties of the medium. [You can see the severity of the dilemma in the lengths to which people were willing to go to find a way out of it.] All that remained was to find a way of measuring the *observer's* velocity *relative to the æther*.

Since "solid" objects slip more or less effortlessly through the æther, this presented some problems. What was eventually settled for was to measure the *apparent speed of light* propagation in different directions; since we are moving through the æther, the light should *appear* to propagate more slowly in the direction *we* are moving, since we are then catching up with it a little.[2]

25.3.1 The Speed of Light

The speed of light is so enormous (299,792 km/s) that we scarcely notice a delay between the transmission and reception of electromagnetic waves under normal circumstances. However, the same electronic technology that raised all these issues in the first place also made it possible to perform *timing* to a precision of millionths of a second (microseconds [μs]) or even billionths of a second (nanoseconds [ns]). Today we routinely send telephone signals out to geosynchronous satellites and back (a round trip of at least 70,800 km) with the result that we often notice [and are irritated by] the delay of 0.236 seconds or more in transoceanic telephone conversations. For computer communications this delay is even more annoying, which was a strong motive for recently laying optical fiber communications cables under the Atlantic and Pacific oceans! So we are already bumping up against the limitations of the finite speed of light in our "everyday lives" (well, almost) without any involvement of the weird effects in this Chapter!

[2] Recall the image of the pebble-thrower walking along the dock and watching the ripples propagate in the pond.

25.3.2 Michelson-Morley Experiment

The famous experiment of Albert Abraham Michelson and Edward Williams Morley actually involved an *interferometer* — a device that measures how much *out of phase* two waves get when one travels a certain distance North and South while the other travels a different distance East and West. Since one of these signals may have to "swim upstream" and then downstream against the æther flowing past the Earth, it will lose a little ground overall relative to the one that just goes "across" and back, with the result that it gets out of phase by a wavelength or two. There is no need to know the *exact* phase difference, because one can simply *rotate the interferometer* and watch as one gets behind the other and then *vice versa*. When Michelson and Morley first used this ingenious device to measure the velocity of the Earth through the æther, they got an astonishing result: *the Earth was at rest!*

Did Michelson or Morley experience brief paranoid fantasies that the ergocentric doctrines of the Mediæval Church might have been right after all? Probably not, but we shall never know. Certainly they assumed they had made some mistake, since their result implied that the Earth was, at least at that moment, at rest with respect to the Universe-spanning luminiferous æther, and hence in some real sense at the centre of the Universe. However, repeating the measurement gave the same result.

Fortunately, they knew they had only to wait six months to try again, since at that time the Earth would be on the opposite side of the Sun, moving in the opposite direction relative to it (the Sun) at its orbital velocity, which should be easily detected by their apparatus. This they did, and obtained the same result. The Earth was *still* at rest relative to the æther.

Now everyone was in a bind. If they insisted in positing an æther to dispell the absurdities of propagation through a vacuum at a fixed velocity, then they had to adopt the embarrassing view that the æther actually chose the Earth, of all the heavenly bodies, to define its rest frame — *and even followed it around* in its accelerated orbital path! This was too much.

25.3.3 FitzGerald/Lorentz Æther Drag

George Francis FitzGerald and H.A. Lorentz offered a solution of sorts: in drifting through the æther, "solid" bodies were not perfectly unaffected by it but in fact suffered a common "drag" in the direction of motion that caused all the yardsticks to be "squashed" in that direction, so that the apparatus *seemed* to be unaffected only because the apparatus and the yardstick and the experimenters' eyeballs were all *contracted* by exactly the same multiplicative factor! They showed by simple arguments that said factor was in fact $\gamma = 1/\sqrt{1-\beta^2}$ where $\beta = u/c$ — *i.e.* exactly the factor defined earlier in the LORENTZ TRANSFORMATIONS, so named after one of their originators![3] Their equations were right, but their explanation (though no more outlandish than what we now believe to be correct) was wrong.

For one thing, these famous "LORENTZ CONTRACTIONS" of the lengths of meter or yardsticks were not accompanied (in their model) by any change in the relative lengths of *time* intervals — how could they be? Such an idea makes no sense! But this leads to qualitative inconsistencies in the descriptions of sequences of events as described by different observers, which also makes no sense. Physics was cornered, with no way out.

Ernst Mach, who had a notorious distaste for "fake" paradigms (he believed that Physics had no business talking about things that couldn't be experimented upon),[4] proposed that Physics had created its own dilemma by inventing a

[3] Poor FitzGerald gets less press these days, alas.

[4] Mach would have had apoplexy over today's *quarks* — but that's a story for a later Chapter!

nonexistent "æther" in the first place, and we would do well to forget it! He was right, in this case, but it took a less crusty and more optimistic genius to see how such a dismissal could be used to explain all the results at once.

25.4 Einstein's Simple Approach

At this time, Albert Einstein was working as a clerk in the patent office in Zürich, a position which afforded him lots of free time to toy with crazy ideas. Aware of this dilemma, he suggested the following approach to the problem: since we have to give up some part of our common sense, why not simply take both the experiments and MAXWELL'S EQUATIONS at face value and see what the consequences are? No matter how crazy the implications, at least we will be able to remember our starting assumptions without much effort. They are:

- The "Laws of Physics" are the same in one inertial reference frame as in another, regardless of their relative motion.[5]
- All observers will inevitably measure the same velocity of propagation for light in their own reference frame, namely c.

These two postulates are the starting points for Einstein's celebrated SPECIAL THEORY OF RELATIVITY ($\mathcal{STR}$), for which this Chapter is named.[6] The adjective "Special" is there mainly to distinguish the $\mathcal{STR}$ from the *General* Theory of Relativity, which deals with *gravity* and *accelerated reference frames*, to be covered later.

25.5 Simultaneous for Whom?

The first denizen of common sense to fall victim to the $\mathcal{STR}$ was the "obvious" notion that if two physical events occur at the same time in my reference frame, they must occur at the same time in *any* reference frame. This is not true unless they also occur at the same *place*. Let's see why.

Einstein was fond of performing imaginary experiments in his head — *Gedankenexperimenten* in German — because the resultant laboratory was larger than anything he could fit into the patent office and better equipped than even today's funding agencies could afford. Unfortunately, the laboratory of the imagination also affords the option of altering the Laws of Physics to suit one's expectations, which means that only a person with a striking penchant for honesty and introspection can work there without producing mostly fantasies. Einstein was such a person, as witnessed by the ironic fact that he used the *Gedankenexperiment* to dismantle much of our common sense and replace it with a stranger truth. Anyway, one of his devices was the laboratory aboard a fast-moving vehicle. He often spoke of *trains*, the most familiar form of transportation in Switzerland to this day; I will translate this into the *glass spaceship* moving past a "stationary" observer [someone has to be designated "at rest," although of course the choice is arbitrary].

In Fig. 25.2 both observers (O and O') must measure the same velocity (c) for the light from the flash bulb. The light propagates outward symmetrically in all directions (in particular,

[5] An *inertial reference frame* is one that is *not accelerated* — *i.e.* one that is at rest or moving at constant velocity.

[6] It is perhaps unfortunate that the theory was called "Relativity" when in fact it expresses the principle that the "Laws of Physics" are *not* relative; they are the same for *all* reference frames, moving or not! It is the *transformations* between measurements by different observers in relative motion that give weird results. When someone says, "Yeah, Einstein showed that everything is relative," every Physicist within earshot winces. On the other hand, the $\mathcal{STR}$ does explicitly rule out any *absolute* reference frame with respect to which all motion must be measured — thus elevating the negative result of the Michelson-Morley experiment to the status of a First Principle — and does imply that certain phenomena that we always thought *were* absolute, like *simultaneity*, are not! So the name "Relativity" does stimulate appropriate debate.

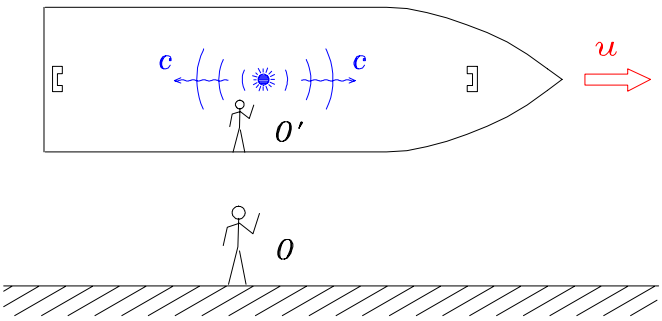


Figure 25.2 A flash bulb is set off in the centre of a glass spaceship (O') at the instant it coincides with a fixed observer O. As the spaceship moves by at velocity u relative to O, the light propagates toward the bow and stern of the ship at the same speed c in both frames.

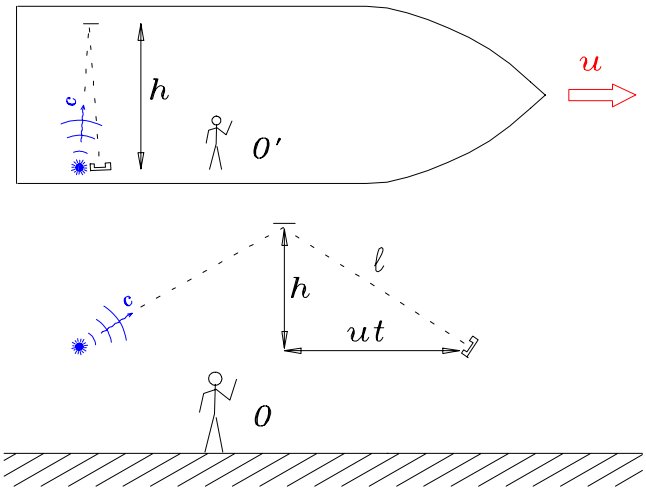


Figure 25.3 A "light clock" is constructed aboard a glass spaceship (reference frame O') as follows: the "tick" of the clock is defined by *one half* the time interval t' required for the light from a strobe light to traverse the width of the ship (a height h), bounce off a mirror and come back, a total distance of $2h$. In the reference frame of a ground-based observer O (with respect to whom the ship is travelling at a velocity u), the light is emitted a distance $2ut$ behind the place where it is detected a time $2t$ later. Since the light has further to go in the O frame (a distance $\ell = \sqrt{h^2 + u^2t^2}$), but it travels at c in both frames, t must be *longer* than t'. This effect is known as TIME DILATION.

to the right and left) from the point where the bulb went off in either frame of reference. In the O' frame, if the two detectors are equidistant from that point they will both detect the light *simultaneously*, but in the O frame the stern of the spaceship moves closer to the source of the flash while the bow moves away, so the stern detector will detect the flash *before* the bow detector!

This is not just an optical illusion or some misinterpretation of the experimental results; this is *actually what happens!* What is *simultaneous* for O' is *not* for O, and *vice versa*. Common sense notwithstanding, SIMULTANEITY is *relative*.

25.6 Time Dilation

Fig. 25.3 pictures a device used by R.P. Feynman, among others, to illustrate the phenomenon of TIME DILATION: a clock aboard a fast-moving vessel (even a normal clock) appears[7] to run *slower* when observed from the "rest frame" — the name we give to the reference frame arbitrarily chosen to be at rest. Now, if we choose to regard the ship's frame as "at rest" (as is the wont of those aboard) and the Earth as "moving," a clock on Earth will appear to be running slowly when observed from the ship! *Who is right?* The correct answer is "both," in utter disregard for common sense. This seems to create a logical paradox, which we will discuss momentarily. But first let's go beyond the qualitative statement, "The clock runs slower," and ask *how much* slower.

For this we need only a little algebra and ge-

[7] The term "appears" may suggest some sort of illusion; this is not the case. The clock aboard the spaceship *actually does* run slower in the Earth's rest frame, and *vice versa*.

ometry; nevertheless, the derivation is perilous, so watch carefully. For O', the time interval described in Fig. 25.3 is simply

$$t' = \frac{h}{c} \qquad \text{so that} \qquad h = ct'$$

whereas for O the time interval is given by

$$t = \frac{\ell}{c} \qquad \text{where} \qquad \ell^2 = h^2 + u^2t^2$$

by the Pythagorean theorem. Expanding the latter equation gives

$$t = \frac{\sqrt{h^2 + u^2t^2}}{c} \quad \text{or} \quad c^2t^2 = h^2 + u^2t^2$$

which is not a solution yet because it does not relate t to t'. We need to "plug in" $h^2 = c^2t'^2$ from earlier, to get

$$\begin{aligned} c^2t^2 &= c^2t'^2 + u^2t^2 \\ \text{or} \qquad t^2 &= t'^2 + \frac{u^2}{c^2}\,t^2 \\ \text{or} \qquad t^2\,(1-\beta^2) &= t'^2 \end{aligned}$$

where we have recalled the definition $\beta \equiv u/c$. In one last step we obtain

$$t = \frac{t'}{\sqrt{1-\beta^2}} \qquad \text{or} \qquad t = \gamma\,t'$$

where γ is defined as before: $\gamma \equiv 1/\sqrt{1-\beta^2}$. This derivation is a little crude, but it shows where γ comes from.

25.6.1 The Twin Paradox

Like most "paradoxes," this one isn't. But it sure looks like one at first glance. Suppose two identical twins part company at age twenty; the first twin hops aboard a spaceship of *very* advanced design and heads out for the distant stars, eventually travelling at velocities very close to c, while the second twin stays home at rest. They give each other going-away presents of identical watches guaranteed to keep perfect time under all conditions. At the midpoint of the voyage, while coasting (and therefore in an inertial reference frame), the first twin looks back at Earth with a *very* powerful telescope and observes the second twin's wristwatch. After correcting for some truly illusory effects, he concludes that the first twin's watch is running slower than his and that his twin on Earth must be aging more slowly as well. Meanwhile, the second twin, on Earth, is looking through *his* telescope at the *first* twin's watch (aboard the spaceship) and concludes that the *first* twin is suffering the effects of time dilation and is consequently aging more slowly than *him!* Who is right? Both, at that moment.

Aha! But now we can bring the first twin *home* after his relativistic journey and *compare ages*. Certainly they can't *both* be younger; this truly would create a logical paradox that goes beyond the mere violation of common sense!

What happens? The first twin, who went travelling, is in fact younger now than the twin who stayed home. The paradox is resolved by a meticulous use of the LORENTZ TRANSFORMATIONS, especially if we make use of the graphical gimmick of the LIGHT CONE, to be discussed later.

25.7 Einstein Contraction(?)

We can obtain the concomitant effect of LORENTZ CONTRACTION without too much trouble[8] using the following *Gedankenexperi-*

[8] I haven't shown all the false starts in which I got the wrong answer using what seemed like perfectly logical arguments.... Here's a good one:

We can obtain the concomitant effect of LORENTZ CONTRACTION in a sloppy way merely by referring back to Fig. 25.2: let x be the distance between the flash bulb and the forward detector, as measured by the observer O on the ground, and let x' be the same distance as measured by the observer O' aboard the spaceship. Assume that O stretches out a tape measure from the place where the flash bulb is set off (say, by a toggle switch on the outer hull of the spaceship which gets hit by a stick held up by O as O' flies by) to the position of the detector in the O frame at

ment, which is so simple we don't even need a Figure:

Suppose a spaceship gets a nice running start and whips by the Earth at a velocity u on the way to Planet X, a distance x away as measured in the Earth's reference frame, which we call O. [We assume that Planet X is at rest with respect to the Earth, so that there are no complications due to their relative motion.] If the spaceship just "coasts" the rest of the way at velocity u [this is what is meant by an INERTIAL REFERENCE FRAME], then by definition the time required for the voyage is $t = x/u$. But this is the time *as measured in the Earth's reference frame*, and we already know about TIME DILATION, which says that the duration t' of the trip *as measured aboard the ship* (frame O') is *shorter* than t by a factor of $1/\gamma$: $t' = t/\gamma$.

Let's look at the whole trip from the point of view of the observer O' aboard the ship: since our choice of who is at rest and who is moving is perfectly arbitrary, we can choose to consider the *ship* at rest and the Earth (and Planet X) to be hurtling past/toward the ship at velocity u. As measured in the ship's reference frame, the distance from the Earth to Planet X is x' and we must have $u = x'/t'$ by definition. But we also must have $u = x/t$ in the other frame; and by symmetry they are both talking about the same u, so

$$\frac{x'}{t'} = u = \frac{x}{t}$$

and since $t = \gamma t'$ we must also have

$$x = \gamma x'.$$

That is, the distance between fixed points, as measured by the space traveller, is *shorter* than that measured by stay-at-homes on Earth by a factor of $1/\gamma$. This is because the Earth and Planet X represent the *moving* system as measured from the ship. This effect is known as LORENTZ CONTRACTION; it has nothing whatsoever to do with "æther drag!" So one might wonder why it isn't called "Einstein contraction," since we calculated it the way Einstein would have.

Of course, the effect works both ways. The *length of the spaceship*, for instance, will be *shorter* as viewed from the Earth than it is aboard the spaceship itself, because in this case the length in question is *in* the frame that moved *with respect to* the Earth. The sense of the contraction effect can be remembered by this mnemonic:

Moving rulers are shorter. (17)

However, it is possible to conjure up situations that defy common sense and thus are often (wrongly) described as "paradoxes."

25.7.1 The Polevault Paradox

I have a favourite *Gedankenexperiment* for illustrating the peculiarities of LORENTZ CONTRACTION: picture a *polevaulter* standing beside a 10 foot long barn with a 10 foot polevault pole in her hands. Tape measures are brought out and it is confirmed to everyone's satisfaction that the pole is exactly the same length as the barn. Got the picture? Now the barn door

---

the instant of the flash. That way we don't need to worry about the *position* of the detector in the O frame when the light pulse actually arrives there some time later; we are only comparing the *length* of the spaceship in one frame with the same length in the other. [It may take a few passes of the spaceship to get this right; but hey, this is a *Gedankenexperiment*, where resources are cheap!] Then the time light takes to traverse distance x', according to O', is $t' = x'/c$, whereas the time t for the same process in the rest frame is $t = x/c$. Therefore, if (from TIME DILATION) t is *longer* than t' by a factor γ, then x must also be *longer* than x' by the same factor if both observers are using the same c.

Simple, eh? Unfortunately, I got the wrong answer! Can you figure out why?

is opened — no tricks — and our intrepid polevaulter walks back a few parsecs to begin her run up.

Suppose we permit a certain amount of fantasy in this *Gedankenexperiment* and imagine that Superwoman, a *very* adept polevaulter, can run with her pole at a velocity $u = 0.6c$. (Thus $\beta = 0.6$ and $\gamma = 1.25$ — check it yourself!) This means that as she runs past a stationary observer her 10 foot pole turns into a 8 foot pole due to LORENTZ CONTRACTION. On the other hand, in her own reference frame she is still carrying a 10 foot pole but the barn is now only 8 feet long. She runs into the barn and the attendant (Superman) slams the barn door behind her.

From Superwoman's point of view, the following sequence of events occurs: first the end of her pole smashes through the end of the barn, and then[9] (somewhat pointlessly, it seems) the barn door slams behind her. A few nanoseconds later she herself hits the end of the barn and the whole schmier explodes in a shower of elementary particles — except for Superwoman and Superman, who are (thankfully) invulnerable.

Superman sees it differently. He has no trouble shutting the barn door behind Superwoman before her polevault pole hits the other end of the barn, so he has successfully performed his assignment — to get Superwoman and her polevaulting skills hidden away inside the barn for the two nanosecond period that the scout for the Olympic Trials happens to be looking this way. What happens after that is pretty much the same as described by Superwoman.

Imagine that you have been called in to mediate the ensuing argument. Who is right? Can you counsel these two Superbeings out of a confrontation that might devastate the surrounding landscape? Or will this become the Parent of all Battles?

[9]It takes about 3.4 ns [nanoseconds, 10^{-9} s] to go 2 feet at a velocity of $0.6c$.

Well, if they want to fight they will fight, of course; but the least you can do is point out that objectively there is nothing to fight about: *they are both right!* When you think about it you will see that they have both described the same *events*; it is only the *sequence* of the events that they disagree on. And *the sequence of events is not necessarily the same* for two observers in relative motion! It all comes back to the RELATIVITY OF SIMULTANEITY and related issues. For Superwoman the pole hits the wall before the door slams, while for Superman the door slams before the pole hits the wall. Both events occur for both observers, but the sequence is different.[10]

25.8 Relativistic Travel

Numerous misconceptions have been bred by lazy science fiction ($\mathcal{SF}$) authors anxious to circumvent the limitations imposed by the $\mathcal{STR}$. Let's examine these limitations and ask whether in fact they restrict space-flight options as severely as $\mathcal{SF}$ fans have been led to believe.

The first and most familiar restriction is the familiar statement, "You can't ever go quite as fast as light." Why is this? Well, consider the behaviour of that ubiquitous scaling factor γ as $u \to c$ (*i.e.*, as $\beta \to 1$): as β gets closer and closer to unity, $(1 - \beta)$ gets closer and closer to zero, as does its square root, which means that γ "blows up" (becomes infinite) as $u \to c$. TIME DILATION causes clocks aboard fast-moving spaceships to *freeze* completely and LORENTZ CONTRACTION causes the length of the ship (in the direction of its motion) to *squash* to nothing, if $u \to c$. [As observed by Earth-bound telescopes, of course.] Worse yet, if we *could* achieve a velocity *greater*

[10]If the door were at the *far* end of the barn (where the pole hits), there could be no such disagreement, since two events at the *same place* and the same time are for all intents and purposes part of the *same event*. It is only events *separated in space* about which such differences of opinion can arise.

than c, time would *not* run backwards [or any of the other simplistic extrapolations tossed off in mediocre $\mathcal{SF}$]; rather the time-dilation / Lorentz-contraction factor γ becomes *imaginary* — in other words, *there is no such physical solution* to the LORENTZ TRANSFORMATION equations! At least not for objects with masses that are *real* in the mathematical sense. [I will deal with the hypothetical *tachyons* in a later section.] Another way of understanding why it is impossible to reach the speed of light will be evident when we begin to discuss RELATIVISTIC KINEMATICS in the next Chapter.

So there is no way to get from here to another star 10 light years distant in less than ten years — *as time is measured on Earth!* However, contrary to popular misconceptions, this does *not* eliminate the option of relativistic travel to distant stars, because the so-called "subjective time"[11] aboard the spaceship is far shorter! This is because in the traveller's reference frame the *stars* are moving and the distances between them (in the direction of motion) *shrink* due to LORENTZ CONTRACTION.

It is quite interesting to examine these effects quantitatively for the most comfortable form of relativistic travel: constant acceleration at $1g$ (9.81 m/s^2) as measured in the spaceship's rest frame, allowing shipboard life to conform to the appearance of Earth-normal gravity. I will list two versions of the "range" of such a voyage (measured in the Earth's rest frame) for different "subjective" elapsed times (measured in the ship's rest frame) — one for *arrival at rest* [the only mode of travel that could be useful for "visiting" purposes], in which one must accelerate halfway and then decelerate the rest of the way, and one for a "*flyby*," in which you don't bother to stop for a look [this could only appeal to someone interested in setting a long-distance record].

The practical limit for an *impulse* drive converting mass carried along by ship into a collimated light beam with 100% efficiency is about 10-12 years. Longer acceleration times require use of a "ram scoop" or similar device using *ambient* matter.

Now, what does this say about the real possibilities for relativistic travel? Without postulating any "unPhysical" gimmicks — *e.g.* "warp drives" or other inventions that contradict today's version of the "Laws" of Physics — we can easily compose $\mathcal{SF}$ stories in which humans (or others) can travel all through our own Galaxy without resorting to suspended animation[12] or other hypothetical future technologies.[13] There is only one catch: As Thomas Wolfe said, *You can't go home again*. Or, more precisely, you can go home but you won't recognize the old place, because all those years it took light to get to your destination and back (that you cleverly dodged by taking advantage of LORENTZ CONTRACTION) still passed normally for the folks back home, now thousands of years dead and gone.

So a wealthy misanthropic adventurer may decide to leave it all behind and go exploring, but no government will ever pay to build a reconnaissance vessel which will not return before the next election. This implies that there may well be visitors from other stars, but they would be special sorts of characters with powerful curiosities and not much interest in socializing. And we can forget about "scouts" from aggressive races bent on colonization, unless they take a *very* long view!

[11] Time measured aboard the spaceship is no more "subjective" than time on Earth, of course; this terminology suggests that the experience of the traveller is somehow bogus, which is not the case. Time *actually does* travel more slowly for the moving observer and the distance between origin and destination *actually does* get shorter.

[12] The idea of suspended animation is a good one and I find it plausible that we may one day learn to use it safely; but it does not quite fall into the category of a simple extrapolation from known technology — yet.

[13] Except for the "ramscoop" technology and the requisite shields against the thin wisp of ambient matter (protons, electrons,...) inhabiting interstellar space, which is converted into high-energy radiation by virtue of our ship's relative motion. Minor details.

Table 25.1 Distances covered (measured in Earth's rest frame) by a spaceship accelerating at a constant $1g$ (9.81 m/s^2) in its own rest frame.

Elapsed Time aboard ship (years)	**Distance Travelled** (Light Years)	
	Arriving at Rest	**"Fly-by"**
1	0.063	0.128
2	0.98	2.76
3	2.70	9.07
4	5.52	26.3
5	10.26	73.2
6	18.14	200.7
7	31.14	547.3
8	52.6	1,490
9	88	4,050
10	146	11,012
11	244	29,936
12	402	81,376
13	665	221,200
14	1,096	601,300
15	1,808	1,635,000
16	2,981	4,443,000
17	4,915	12,077,000
18	8,103	32,830,000
19	13,360	89,241,000
20	22,000	243,000,000
21	36,300	659,000,000
22	59,900	1,792,000,000
23	99,000	4,870,000,000
24	163,000	13,200,000,000
25	268,000	36,000,000,000
26	442,000	98,000,000,000
27	729,000	(present diam.
28	1,200,000	of universe
29		thought to be
30		less than about
		30,000,000,000)

25.9 Natural Units

As I mentioned in the Chapter on UNITS AND DIMENSIONS, in any context where the *speed of travel* is virtually (or, in this case, exactly) a *constant*, people automatically begin to express *distances* in *time* units. [*Q*: "How far is is from New York to Boston?" *A*: "Oh, about three hours."] This is equivalent to defining the *speed of travel* to be a dimensionless constant of magnitude 1. Relativistic Physics is no different. Anyone who has to discuss relativistic phenomena at any length will usually slip into "NATURAL UNITS" where

$$c = 1$$

and distance and time are measured in the same units. You get to pick your favourite unit — seconds, meters, light years or (as we shall see later) inverse masses! The list is endless. Then β is just "the velocity" measured in natural units and the calculations become much simpler. But you have to convert all your other units accordingly, and this can be interesting. It does take a little getting used to, but the exercise is illuminating.

25.10 A Rotational Analogy

If we compare the LORENTZ TRANSFORMATIONS with the GALILEAN TRANSFORMATIONS, several striking qualitative features are apparent: the first is the multiplicative factor γ which describes both TIME DILATION and LORENTZ CONTRACTION; the second is the fact that *time* and *space* get *mixed together* by the LORENTZ TRANSFORMATION — a blasphemy in the paradigm of classical Physics.

The latter weirdness is going to be confusing no matter what we do; is there any way to at least make it *look* familiar? What we need is an *analogy* with something that *does* "make sense" and is still intact. Fortunately there is a precedent for a transformation that *mixes coordinates*, namely the ROTATION.

25.10.1 Rotation in Two Dimensions

Suppose we have a point A in a plane with perpendicular x and y coordinate axes scribed on it, as pictured in Fig. 25.4.

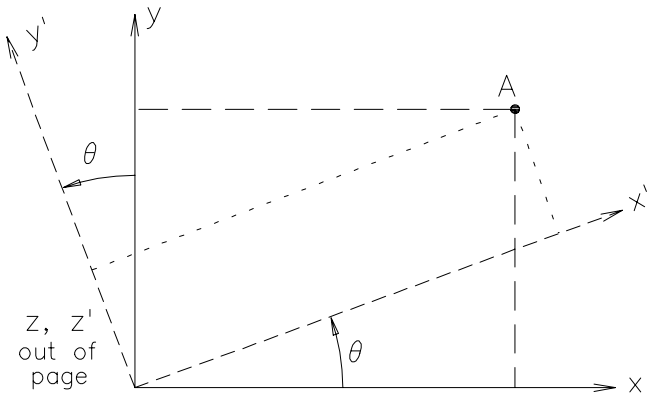


Figure 25.4 A fixed point A can be located in a plane using either of two coordinate systems O (x, y) and O' (x', y') that differ from each other by a rotation of θ about the common origin $(0, 0)$.

We can scribe a *different* pair of perpendicular coordinate axes x' and y' on the same plane surface using dashed lines by simply *rotating* the original coordinate axes by an angle θ about their common origin, the coordinates of which are $(0, 0)$ in *either* coordinate system.

Now suppose that we have the coordinates (x_A, y_A) of point A in the original coordinate system and we would like to *transform* these coordinates into the coordinates (x'_A, y'_A) of the *same point* in the new coordinate system.[14] How do we do it? By trigonometry, of course.

[14] This situation might arise if an architect suddenly discovered that his new plaza had been drawn from coordinates laid out by a surveyor who had aligned his transit to magnetic North while standing next to a large industrial electromagnet. The measurements are all OK but they have to be converted to true latitude and longitude!

You can figure this out for yourself. The transformation is

$$x' = x\cos(\theta) + y\sin(\theta) \quad (18)$$
$$y' = -x\sin(\theta) + y\cos(\theta) \quad (19)$$

25.10.2 Rotating Space into Time

If we now look at just the x and t part of the LORENTZ TRANSFORMATION [leaving out the y and z parts, which don't do much anyway], we have

$$x' = \gamma x - \gamma\beta ct \quad (20)$$
$$ct' = -\gamma\beta x + \gamma ct \quad (21)$$

— *i.e.*, the LORENTZ TRANSFORMATION "sort of" *rotates* the *space* and *time* axes in "sort of" the same way as a normal rotation of x and y. I have used ct as the time axis to keep the units explicitly the same; if we use "natural units" ($c = 1$) then we can just drop c out of the equations completely and the analogy becomes obvious. However, you should resist the temptation to think of the LORENTZ TRANSFORMATION as "just a rotation of space and time into each other." If we "boost" the O' frame by some large relative velocity in the *negative* x direction and try to plot up x' and ct' on the same graph as (x, ct) then we get a weird picture.

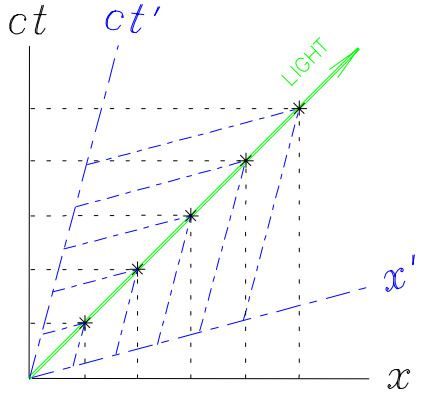


Figure 25.5 An attempt to draw (x', ct') coordinates on the same graph as the (x, ct) coordinates. The result is misleading because the spatial surface on which it is drawn obeys EUCLIDEAN geometry (the invariant length of an interval is the square root of the sum of the squares of its two perpendicular components) whereas spacetime obeys the MINKOWSKI metric: the invariant "length" of a spacetime interval (the PROPER TIME) is equal to $c^2t^2 - x^2$, not $c^2t^2 + x^2$). You may *think* of the LORENTZ TRANSFORMATION as a sort of rotation, but you can't *draw* it as a rotation, because you don't have Minkowski paper!

Proper Time and Lorentz Invariants

The most important important difference between ordinary ROTATIONS and the LORENTZ TRANSFORMATIONS is that the former preserve the RADIUS distance

$$r = \sqrt{x^2 + y^2} = \sqrt{x'^2 + y'^2} \quad (22)$$

of point A from the origin, whereas the latter preserve the PROPER TIME τ of an event:

$$c\tau = \sqrt{c^2t^2 - x^2} = \sqrt{c^2t'^2 - x'^2} \quad (23)$$

The $-$ sign in the latter is important!

In general, any quantity which we can define (like τ) that will have *the same value* in every inertial reference frame, regardless of relative motion, may be expected to become very precious to our bruised sensibilities. The *STR* has dismantled most of our common sense about which physical observables are reliable, universal constants and which depend upon the reference frame of the observer; if we can specifically identify those properties of a quantity that will guarantee its *invariance* under LORENTZ TRANSFORMATIONS, then we can at least count on such quantities to remain reliably and directly comparable for different observers. Such quantities are known as LORENTZ INVARIANTS.

The criterion for LORENTZ INVARIANCE is that the quantity in question be the *scalar product of two 4-vectors*, or any combination of such scalar products. What do we mean by *4-vectors*? {Space and time} make the classic example, but we can *define* a *4-vector* to be *any 4-component quantity that transforms like spacetime*. That is, $a_\mu = \{a_0, a_1, a_2, a_3\}$ — where a_0 is the "timelike" component (like ct) and $\{a_1, a_2, a_3\}$ are the three "spacelike" components (like x, y, z) — is a *4-vector* if a "boost" of u in the x direction gives

$$\begin{aligned} a'_0 &= \gamma(a_0 - \beta a_1) \\ a'_1 &= \gamma(a_1 - \beta a_0) \\ a'_2 &= a_2 \\ a'_3 &= a_3 \end{aligned}$$

just like for $x_\mu = \{ct, x, y, z\}$. The most important example (other than x_μ itself) is $p_\mu = \{E, p_x, p_y, p_z\}$, the ENERGY-MOMENTUM 4-vector, which we will encounter next.

25.11 Light Cones

If we picture LORENTZ TRANSFORMATIONS as ROTATIONS of time and space into each other, we can make good use of a handy and simple graphical gimmick: the "light cone."

25.12 Tachyons

Chapter 26

Relativistic Kinematics

Since MECHANICS is so intimately concerned with the relationships between *mass*, *time* and *distance*, the weird properties of the time and space revealed by the $\mathcal{STR}$ may be expected to be accompanied by some equally weird MECHANICS at relativistic velocities. This is indeed the case. On the other hand, we can rely upon Einstein's first postulate of the $\mathcal{STR}$, namely that the "Laws of Physics" are the same in one reference frame as in another. Thus most of our precious paradigms of MECHANICS (such as CONSERVATION LAWS) will still be reliable.

26.1 Momentum is Still Conserved!

For instance, MOMENTUM CONSERVATION must still hold, or else we would be able to tell one reference frame from another (in an *absolute* sense) by seeing which one got less than its share of momentum in a collision. To pursue this example, we invoke MOMENTUM CONSERVATION in a *glancing collision* between two identical billiard balls, as pictured in Fig. 26.1:

[Get ready to keep track of a lot of subscripts and primes! If you want to avoid the tedium of paying close attention to which quantity is measured in whose rest frame, skip to the formal derivation in terms of LORENTZ INVARIANTS and the 4-MOMENTUM. . . .]

Now, each of A and B is at rest in its own reference frame before the collision (A sees B approaching from the right at $-u$ whereas B sees A approaching from the left at $+u$); after the collision, each measures[1] *its own* final velocity *transverse* (perpendicular) to the initial direction of motion of the other. Out of courtesy and in the spirit of scientific cooperation, each sends a message to the other reporting this measurement. By symmetry, these messages must be identical:

$$v'_{A_\perp} \;=\; v_{B_\perp} \tag{1}$$

Using the same argument, each must report the same measurement for the transverse component of the *other's* velocity after the collision:

$$v_{A_\perp} \;=\; v'_{B_\perp} \tag{2}$$

[1] If the anthropomorphism of billiard balls bothers you, please imagine that these are *very large* "billiard balls" with cabins occupied by Physicists who make all these observations and calculations.

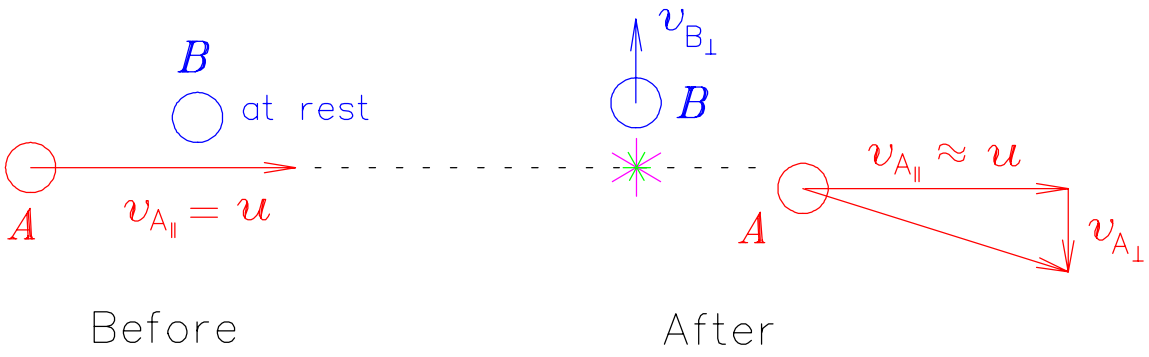


Figure 26.1 A *glancing collision* between two identical billiard balls of rest mass m, shown in the reference frame of ball B. Ball A barely touches ball B as it passes at velocity u, imparting a miniscule *transverse* velocity $v_{B_\perp}$ (perpendicular to the initial velocity of A) to ball B and picking up its own transverse velocity $v_{A_\perp}$ in the process. *Primed* quantities (like $v'_{A_\perp}$ and $v'_{B_\perp}$) are measured in A's reference frame, whereas *unprimed* quantities (like $v_{A_\parallel} = u$, $v_{A_\perp}$ and $v_{B_\perp}$) are measured in B's reference frame.

Meanwhile, MOMENTUM CONSERVATION must still hold for the transverse components in each frame:

$$\text{In } B \text{ (unprimed) frame} \qquad m\,v_{B_\perp} \;=\; m_A\,v_{A_\perp} \tag{3}$$

$$\text{and in } A \text{ (primed) frame} \qquad m'_B\,v'_{B_\perp} \;=\; m\,v'_{A_\perp}, \tag{4}$$

where the masses of the billiard balls *in their own rest frames* are written as m but I have expressly allowed for the possibility that a ball's effective mass *in the other ball's frame* may differ from its rest mass. (It helps to know the answer.) Thus m_A is the effective mass of A as seen from B's reference frame and m'_B is the effective mass of B as seen from A's reference frame.

We may now apply the LORENTZ VELOCITY TRANSFORMATION to the transverse velocity component of A:

$$v'_{A_\perp} \;=\; \frac{v_{A_\perp}}{\gamma\left(1 - uv_{A_\parallel}/c^2\right)} \;=\; \frac{\sqrt{1-\beta^2}\,v_{A_\perp}}{1-u^2/c^2} \;=\; \gamma\,v_{A_\perp} \tag{5}$$

Combining Eq. (1) with Eq. (3) gives $m\,v'_{A_\perp} \;=\; m_A\,v_{A_\perp}$ which, combined with Eq. (5), gives $m\,\gamma\,v_{A_\perp} \;=\; m_A\,v_{A_\perp}$ or $m_A \;=\; \gamma\,m$. Similarly, combining Eq. (2) with Eq. (4) gives $m'_B\,v_{A_\perp} \;=\; m\,v'_{A_\perp} \;=\; m\,\gamma\,v_{A_\perp}$ or $m'_B \;=\; \gamma\,m$.

We can express both results in a general form without any subscripts:

$$m' \;=\; \gamma\,m \tag{6}$$

> *The* EFFECTIVE MASS m' *of an object moving at a velocity* $u = \beta c$ *is* γ *times its* REST MASS m *(its mass measured in its own rest frame).*

That is, *moving masses have more inertia!*

26.1.1 Another Reason You Can't Go as Fast as Light

The preceding argument was not very rigourous, but it served to show the essential necessity for regarding the EFFECTIVE MASS of an object as a *relative* quantity. Let's see what happens as we try to accelerate a mass to the velocity of light: at first it picks up speed just as we have been trained to expect by Galileo.[2] But as β becomes appreciable, we begin to see an interesting phenomenon: *it gets harder to accelerate!* (This is, after all, what we *mean* by "effective mass.") As $\beta \to 1$, the multiplicative "mass correction factor" $\gamma \to \infty$ and eventually we can't get any more speed out of it, we just keep pumping energy into the effective mass. This immediately suggests a new way of looking at mass and energy, to be developed in the following section.

But first let's note an interesting side effect: the rate at which a constant accelerating force produces velocity changes, *as measured from a nonmoving reference frame*, slows down by a factor $1/\gamma$; but the *same* factor governs the TIME DILATION of the "speed" of the clock in the moving frame. So (as observed from a stationary frame) the change in velocity *per tick of the clock* in the moving frame is constant. This has no practical consequences that I know of, but it is sort of cute.

26.2 Mass and Energy

In the hand-waving spirit of the preceding section, let's explore the consequences of Eq. (6). The BINOMIAL EXPANSION of γ is

$$\gamma \;=\; (1-\beta^2)^{-\frac{1}{2}} \;=\; 1 \;+\; \tfrac{1}{2}\,\beta^2 \;-\; \cdots \tag{7}$$

For *small* β, we can take only the first two terms (later terms have still higher powers of $\beta \ll 1$ and can be neglected) to give the approximation

$$m' \;\approx\; (1+\tfrac{1}{2}\beta^2)\,m \qquad \text{or} \qquad m'c^2 \;\approx\; mc^2 \;+\; \tfrac{1}{2}mu^2 \tag{8}$$

The last term on the right-hand side is what we ordinarily think of as the KINETIC ENERGY T. So we can write the equation (in the limit of small velocities) as

$$T \;=\; \gamma\, mc^2 \;-\; mc^2 \tag{9}$$

It turns out that Eq. (9) is the *exact* formula for the kinetic energy at *all* velocities, despite the "handwaving" character of the derivation shown here.

We can stop right there, if we like; but the two terms on the right-hand side of Eq. (9) look so simple and similar that it is hard to resist the urge to give them names and start thinking *in terms of them*.[3] It is conventional to call $\gamma\, mc^2$ the TOTAL RELATIVISTIC ENERGY and mc^2 the REST MASS ENERGY. What do these names mean? The suggestion is that there is an irreducible energy $E_0 = mc^2$ associated with any object of mass m, even when it is sitting still! When it speeds up, its total energy changes by a multiplicative factor γ; the *difference* between the total energy $E = \gamma\, mc^2$ and E_0 is the energy due to its *motion*, namely the *kinetic energy* T.

[2] It had better! The behaviour of *slow-moving* objects did not undergo some sudden retroactive change the day Einstein wrote down these equations!

[3] This is, after all, the most ubiquitous instinct of Physicists and perhaps the main æsthetic foundation of Physics. It is certainly what I mean by "Physics as Poetry!"

26.2.1 Conversion of Mass to Energy

Einstein's association of the term mc^2 with a REST MASS ENERGY E_0 naturally led to a great deal of speculation about what might be done to *convert* mass into useable energy, since for a *little* mass you get a *lot* of energy! Let's see just how much: in *S.I.* units $1\ \text{J} \equiv 1\ \text{kg-m}^2/\text{s}^2$ so a 1 kg mass has a rest mass energy of $(1\ \text{kg}) \times (2.9979 \times 10^8\ \text{m/s})^2 = 8.9876 \times 10^{16}\ \text{J}$ — *i.e.*,

$$1\ \text{kg} \longleftrightarrow 8.9876 \times 10^{16}\ \text{J} \tag{10}$$

which is a lot of joules. To get an idea how many, remember that one WATT is a unit of *power* equal to one joule per second, so a JOULE is the same thing as a WATT-SECOND. Therefore a device converting *one millionth of a gram* (1 μg) of mass to energy *every second* would release approximately *90 megawatts* [millions of watts] of power!

Contrary to popular belief, the first conclusive demonstration of mass-energy conversion was in a controlled nuclear *reactor*. However, not long after came the more unpleasant manifestation of mass→energy conversion: the fission bomb. An unpleasant subject, but one about which it behooves us to be knowledgeable. For this, we need a new energy unit, namely the KILOTON [kt], referring to the energy released in the explosion of one thousand *tons* of TNT [*tri*nitro*to*luene], a common chemical high explosive. The basic conversion factor is

$$1\ \text{kt} \equiv \text{ a trillion CALORIES} = 4.186 \times 10^{12}\ \text{J} \tag{11}$$

which, combined with Eq. (10), gives a rest-mass equivalent of

$$1\ \text{kt} \longleftrightarrow 4.658 \times 10^{-5}\ \text{kg} \tag{12}$$

That is, one KILOTON's worth of energy is released in the conversion of 0.04658 grams [46.58 mg] of mass. Thus a MEGATON [equivalent to one million tons of TNT or 10^3 kt] is released in the conversion of 46.58 grams of mass; and the largest thermonuclear device [bomb] ever detonated, about 50 megatons' worth, converted some 2.329 kg of mass directly into raw energy.

Comment from **Daniel Rosenblatt**, 8 July 2003:

> The largest thermonuclear bomb ever detonated was the *Tsar Bomba*. While its design yield was 100 megatons, it was detonated without its ^{238}U jacket, reducing its actual yield to 50 megatons. This was because the engineers already knew how much energy the jacket would add to the explosion, and the fallout generated would have contaminated thousands of square kilometres.
>
> As a side note, the Tsar Bomba (as detonated) was the cleanest nucear weapon ever tested, deriving only about 3% of its total energy from fission.[4]

[4] As implied by Daniel Rosenblatt's comment, a large fraction of the energy released by a "fusion" bomb is actually generated by *fission* of a ^{238}U jacket around the true thermonuclear (fusion) core. While ^{238}U does not spontaneously fission from the addition of slow neutrons made in ordinary fission, and is thus immune to the chain reaction that takes place in ^{235}U or ^{239}Pu, the *fast* neutrons from the fusion reaction *will* cause ^{238}U to fission, releasing both energy and nasty radioisotopes. Why anyone would do it this way is beyond my understanding, since it generates so much more radioactive fallout that will eventually reach the country that launched the weapon in the first place. But this is not the only thing about nuclear war that defies common sense! At any rate, we can see that the term "H bomb" is a misnomer in at least two different ways.

Also, while there is no theoretical limit to the size of an H bomb, there is a practical one. A 100 megaton bomb is almost useless militarily, and is also inefficient. Bombs of this size spend most of their energy re-re-re-pulverizing the target area. They do not scale up linearly (a bomb twice the size doesn't destroy twice the area). The largest weapon in the US arsenal is the B53 at 9 megatons; the rest of the stockpile is in the 100-475 Kt range.

Nuclear Fission

Where did the energy come from? *What* mass got converted? To answer this question we must look at the processes involved on a sub-microscopic scale. First we must consider the natural tendency for oversized atomic nuclei to spontaneously *split* into smaller components.[5] This process is known as NUCLEAR FISSION and is the energy source for all presently functioning NUCLEAR REACTORS on Earth. [Also for so-called "atomic" bombs.][6]

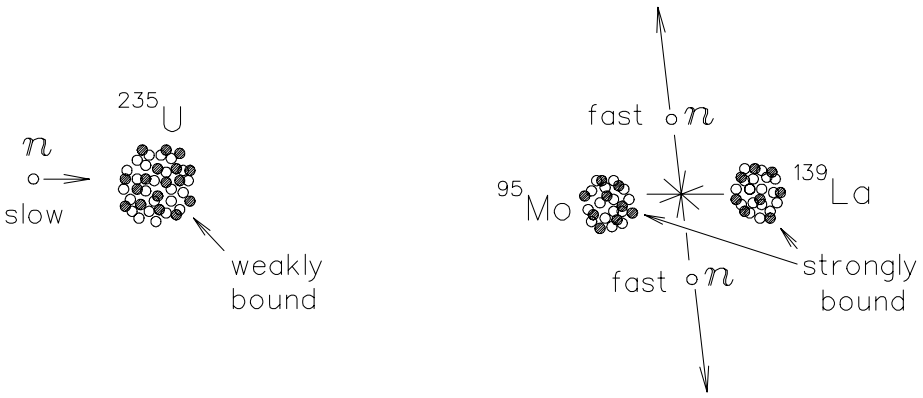


Figure 26.2 One case of the fission of ^{236}U. The net mass of the initial neutron plus the ^{235}U nucleus is 219,883 MeV/c^2. The net mass of the fission products (two neutrons, a ^{95}Mo nucleus and a ^{139}La nucleus) is 219,675 MeV/c^2 — smaller because of the stronger *binding* of the Mo and La nuclei. The "missing mass" of 208 MeV/c^2 goes into the *kinetic energy* of the *fragments* (mainly the *neutrons*), which of course adds up to 208 MeV.

The basic event in the most common variety of NUCLEAR FISSION is the spontaneous splitting of one ^{236}U nucleus into (for example) ^{95}Mo, ^{139}La and two neutrons.[7] [There are numerous other possible fission products. This is just one case.] The fraction of the total mass that gets converted

[5] I know I haven't explained what I mean by a "nucleus" yet, or even an "atom;" but here I will suspend rigourous sequence and "preview" this subject. The details are not important for this description.

[6] The name, "ATOMIC BOMB," is a frightful misnomer; the *atoms* have nothing whatsoever to do with the process involved in such horrible weapons of destruction, except insofar as their *nuclei* are the active ingredients. The correct name for the "atomic" bomb is the NUCLEAR FISSION bomb.

[7] The notation used here is ^{A}El, where the *atomic weight A* of an element is the total number of *neutrons* (uncharged nucleons) and *protons* (positively charged nucleons) in the nucleus and *El* is the *chemical symbol* for the *element* in question. "Nucleon" is just a generic name for either protons or neutrons, which have about the same mass [the neutron is slightly heavier] and the number of *protons* in a nucleus [called its *atomic number Z*] determines its net electrical charge, which in turn must be balanced by an equal number of negatively charged *electrons* in orbit about the nucleus to make up the *atom*. The *atomic number Z* therefore determines all the *chemical* properties of the atom and so defines which *element* it is. We

into kinetic energy is $208/219833 = 0.946 \times 10^{-3}$ or about a tenth of a percent. The energy liberated in the fission of one ^{236}U nucleus produced in this way is 208 MeV or 0.333×10^{-10} J. That means it takes 3×10^{10} such fissions to produce one joule of utilizable energy. Since there are 2.55×10^{21} such nuclei in one gram of pure ^{235}U metal, 3×10^{10} isn't such a large number!

What sort of *control* do we have over this process? To answer this question we must understand a bit more about the details of the CHAIN REACTION whereby an appreciable number of such fissions take place.

The ^{236}U nucleus is formed by adding one neutron to a ^{235}U nucleus, which is found in natural uranium ore on Earth at a concentration of about 0.72% [the rest is almost all ^{238}U]. Now, left to its own devices (*i.e.*, if we don't drop any slow neutrons into it) a ^{235}U nucleus will live for an average of 0.7038 billion years, eventually decaying spontaneously by α particle emission (*not* the fission reaction that produces more neutrons!) just like its brother isotope ^{238}U, whose lifetime is only about 6 times longer (4.468 billion years). If the lifetimes weren't so long, there wouldn't be any left on Earth to dig up — which might be regarded as a good thing overall, but we have to play the hand we're dealt. So an isolated ^{235}U nucleus generally sits around doing nothing and minding its own business; but when a slow *neutron* comes by (picture a ball bearing slowly rattling down through a peg board) it has a strong tendency to be *captured* by the ^{235}U nucleus to form ^{236}U, and then the action starts. This is also a little tricky, because if the ^{236}U nucleus gets a chance to settle into its *ground state* (*i.e.*, if all the jiggling and vibrating caused by absorption of a neutron has a chance to die down) then it (the ^{236}U nucleus) is also quite stable [mean lifetime = 23.42 million years] and also decays by α emission (no new neutrons). However, this is rarely the case; usually the excitations caused by absorbing that extra neutron are too much for the excited ^{236}U nucleus and it *fissions* as described earlier, releasing several not-too-fast neutrons.

What follows depends upon the neighbourbood in which the fission occurs. If the original ^{235}U nucleus is off by itself somewhere, the two neutrons just escape, rattle around until they lose enough energy to be captured by some less unstable nuclei, and the process ends. If the fission occurs right next to some *other* ^{235}U nuclei, then the outcome depends (critically!) upon the MODERATION [slowing down] of the neutrons: when they are emitted in the fission process, they are much too fast to be captured by other ^{235}U nuclei and will just escape to bury themselves eventually in some innocuous nuclei ensewhere. If, however, we run them through some MODERATOR [slower-downer] such as graphite, heavy water (deuterium oxide, D_2O) or, under extreme conditions of density and pressure, uranium metal itself, the neutrons will slow down by a sort of frictional drag until they reach the right energy to be captured efficiently by other ^{235}U nuclei. Then we get what is known as a CHAIN REACTION. One neutron is captured by a ^{235}U nucleus which splits up into fission products including fast neutrons, which are moderated until they can be captured by other ^{235}U nuclei, which then split up into fission products including fast neutrons, which are....

The moderation of the neutrons generates a lot of *heat* in the moderator (it is a sort of *friction*, after all) which can be used in turn to boil water to run steam turbines to generate electricity. [Or misused to make a large explosion.] A good fission REACTOR design (like the Canadian CANDU reactor)

could just specify Z in addition to A to know everything we need to know about the specific nucleus in question [which we call an ISOTOPE], but names are more appealing than numbers [even to Physicists!] so we use the chemical symbol [*e.g.* U = Uranium, Mo = Molybdenum, La = Lanthanum, H = Hydrogen, He = Helium and Li = Lithium] as an abbreviation for the name of the element. Sometimes you will see Z as a *subscript* on the left of the chemical symbol, as in $^{238}_{92}$U, but this is not the only convention for isotopic notation and I see no reason to confuse matters any further. There — a micro-introduction to nuclear, atomic and chemical terminology!

involves a moderator like heavy water (D_2O) which boils away when the reactor core overheats, thus stopping the moderation and automatically shutting down the reactor. A bad design (like the Soviet or American reactors) uses MODERATOR RODS that are shoved into the core mechanically and can get stuck there if the core overheats, as happened at Three Mile Island and (much worse) at Chernobyl.[8]

Potential Energy is Mass, Too!

Where did the mass "go" in the reaction we just discussed? The answer is that the BINDING ENERGY of the ^{235}U nucleus is substantially *less negative* than that of the final products.

Remember that the *gravitational potential energy* between two massive bodies is *zero* when they are infinitely far apart and becomes more and more *negative* as they get closer together? [Lower gravitational potential energy for an object at a lower height?] Well, the STRONG NUCLEAR FORCE that binds nuclei together has at least this much in common with gravity: it is *attractive* (at least at intermediate range) and therefore produces a POTENTIAL ENERGY "WELL" into which the constituents "fall" when we make up a nucleus.[9]

The other thing to realize is that *potential energy counts* in the evaluation of the total relativistic energy of an object; and if the object is at rest, then its potential energy counts in the evaluation of its REST MASS. As a result, we might expect the rest mass of a space ship to be slightly larger after it leaves the Earth than it was on Earth, simply because it has left the "gravity well" of the Earth. This is the case! However, the mass change is imperceptibly *small* in this case.

Nuclear Fusion

Actually, a large nucleus is *rarely* heavier than the sum of its constituents. If you think about it, this is the equivalent of having a ball stored at the top of a potential energy *hill*.[10] Once it moves over the edge, the process is all downhill, resulting in liberation of kinetic energy. The heaviest nuclei represent *stored-up energy* from "endothermic" (energy-absorbing) processes that

[8] There is an interesting history to the American [and presumably the Soviet] reactor design: the original version was built on a small scale to go into nuclear submarines, where it worked quite well (and was comparatively safe, considering the unlimited supply of coolant!). However, the successful submarine reactor design was simply *scaled up* to make the big land-based power reactors, a thoroughly dumb and lazy manœuver by the power industry that has led to a long series of unnecessary troubles. If the world had standardized on the CANDU design, nuclear power would have a much better reputation today, except for the irreducible (though undeserved) taint of psychological association with nuclear weapons, which has even prompted doctors to change the name of NMR (nuclear magnetic resonance) imaging machines — probably the most harmless and beneficial devices ever created by modern technology — to "MRI" (for Magnetic Resonance Imaging) just so their patients wouldn't be spooked by the boogey-word "nuclear."

[9] Note how extensively we rely on this gravitational metaphor! This is partly because we don't know any more compelling poetic technique and partly because it works so well — it is a "good" metaphor!

[10] If you think about it some more, you will realize that such a situation usually constitutes UNSTABLE EQUILIBRIUM: the tiniest push will set the ball rolling downhill, never to return of its own accord. In this case (carrying the nice metaphor a little further) there is actually a slight *depression* at the top of the hill, so that the ball can rest easy in METASTABLE EQUILIBRIUM: as long as it doesn't get to rolling around too energetically [enough to roll up over the edge of the depression], the ball will stay where it is; but if we "tickle" it enough [in this case, by dropping in a neutron] it will bounce out and from there it is all downhill again. This picture works almost perfectly in developing your intuition about metastable nuclei, except for the peculiar prediction of QUANTUM MECHANICS that the ball can get through the "barrier" without ever having enough kinetic energy to make it up over the ridge! But that's another story....

took place in SUPERNOVA explosions billions of years ago, and are in that sense correctly referred to as "supernova fossils." Anything heavier than *iron* falls into this category!

Nuclei *lighter* than iron (^{57}Fe), if they can be regarded as composed of lighter nuclei, are almost always *lighter* than the sum of their constituents, simply because their BINDING ENERGY is greater. The process of combining light nuclei to make heavier ones (up to iron) is called NUCLEAR FUSION, which also liberates kinetic energy. There are many, many varieties of nuclear fusion reactions, most of which are realized on a large scale in *stars*, whose main energy source is nuclear fusion. [A nice, romantic aspect of nuclear physics, for a change!] Our own Sun, for example, is one big fusion power plant and has *all* the pleasant and unpleasant features of the putative man-made versions, such as radiation....

Unfortunately, here on Earth we have not yet succeeded in *controlling* NUCLEAR FUSION well enough to make a reactor that will generate more energy than it takes to run, though billions of dollars have been (and will doubtless continue to be) spent in the attempt. So far all we have achieved with notable success is the *uncontrolled* thermonuclear[11] reaction [bomb] known as the "H bomb."[12] A nasty feature of thermonuclear bombs is that there doesn't seem to be an upper limit on how big one might make them. The only good thing about them (other than the questionable virtue of "deterrence") is that they are not intrinsically as "dirty" (in terms of radioactive fallout) as fission bombs, at least not "per kiloton." However, most tactical "H bombs" are actually mainly *fission* devices *triggered* by a fusion core. This makes them quite dirty. Yuk. I have said rather more than I like about this subject already.

Cold Fusion

"Wouldn't it be nice," most reasonable people would agree, "if there were a way to obtain energy from fusion of some innocuous nuclei like deuterium without the enormous temperatures of nuclear explosions or the various 'hot' controlled fusion reactors on the drawing boards." There certainly is a way to get deuterium nuclei close enough together to fuse without high temperatures — in fact I recently participated in an experiment that achieved D-D fusion at a temperature of 2.5 K: this involves forming a *molecule* of two deuterons and one negative *muon* — an unstable elementary particle which is more or less like an electron except that its mass is 207 times bigger. The heavy muon pulls the deuterons so close together that they fuse. This works. Unfortunately *it doesn't work well enough to generate more energy than it took to make the muon in the first place!* The closest anyone has come to "breakeven" using muons is more than a factor of ten too low in efficiency. Too bad. It is frustrating to come so close and then fail.

Perhaps because of this frustration, a few years ago some people deluded themselves into believing that they had coaxed deuterons into fusing by regular electrochemical means in a palladium metal matrix. Unfortunately this was bogus. Even more unfortunately, the fantasy remained so seductive

[11] We call such processes *thermo*nuclear because the positively charged nuclei don't "like" to get close enough to each other for the strong, short-range nuclear force to take over (they repell each other electrically), and to overcome this "Coulomb barrier" they are heated to such enormous temperatures that their kinetic energy is high enough to get them together and then ... *bang!* The *heating* is usually done by means of a small *fission* bomb, from what I understand.

[12] Once again, the popular terminology "H bomb" is completely misleading. The first thermonuclear bombs used a mixture of deuterium (^{2}H) and tritium (^{3}H) — two isotopes of hydrogen — as the components that fused to form heavier products, hence the name; but modern thermonuclear bombs use (I think) deuterium and lithium, which can be combined chemically into a solid form that is relatively easy to handle and not spontaneously radioactive.

that a lot of otherwise respectable scientists were willing to compromise their integrity (probably unconsciously – I hope) and generate supporting evidence from flawed experiments or muddy reasoning. Consequently, many gullible people still believe in "cold fusion." Who can blame them? If you can't trust the experts, who can you trust? Maybe the popularity of the *X-Files* and other signs of people losing their grip on reality can all be traced back to the betrayal of public trust in the "cold fusion" debacle. Oh well. I did what I could.

26.2.2 Conversion of Energy into Mass

In a NUCLEAR REACTOR, a spontaneous nuclear process results in a net *decrease* in the net mass of all the particles involved. The "missing mass" appears as the *kinetic energy* of the reaction products, which is dissipated by what amounts to friction and generates *heat* that boils water; the steam is used to spin turbines that run generators that send electrical power down the wires.

This leads to an obvious question: can we do the *opposite*? Can we take electrical power out of the wires, use it to raise the kinetic energy of some particles to enormous values, smack the particles together and *generate* some *extra* mass? Yes! This is what a PARTICLE ACCELERATOR like TRIUMF[13] does. Every such accelerator is a sort of "reactor in reverse," taking electrical power out of the grid and turning it into mass.

Such things happen *naturally*, too. Gamma rays of sufficient energy often convert into electron-positron *pairs* when they have a glancing collision with a heavy nucleus. This is pictured in Figs. 26.3 and 26.4.

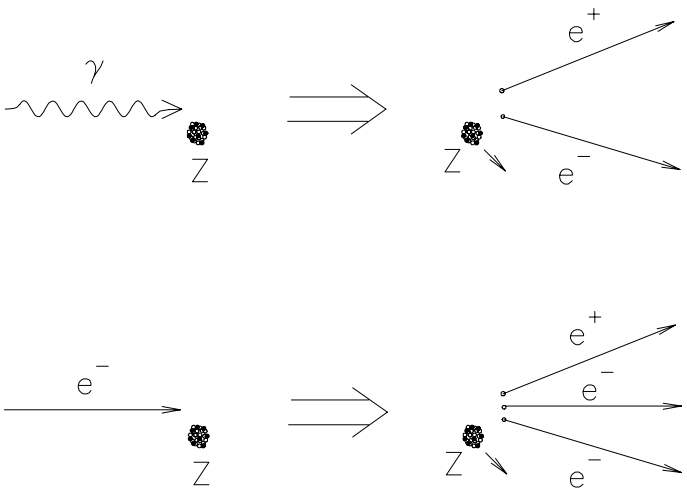


Figure 26.3 Electron-positron PAIR PRODUCTION by gamma rays (above) and by electrons (below). The positron (e^+) is the ANTIPARTICLE of the electron (e^-) [to be explained in the Chapter on Elementary Particle Physics]. The gamma ray (γ) must have an energy of at least 1.022 MeV [twice the rest mass energy of an electron] and the pair production must take place near a heavy nucleus (Z) which absorbs the momentum of the γ.

[13]The acronym TRIUMF stands for TRI-University Meson Facility, in recognition of the three B.C. Universities that originally founded to project [there are now several more, but we don't change the cute name] and the main product of the cyclotron.

There is a neat, compact way of representing such reactions by FEYNMAN DIAGRAMS[14] I will draw them "left to right" but the convention is actually to draw them "down to up." I don't know why.

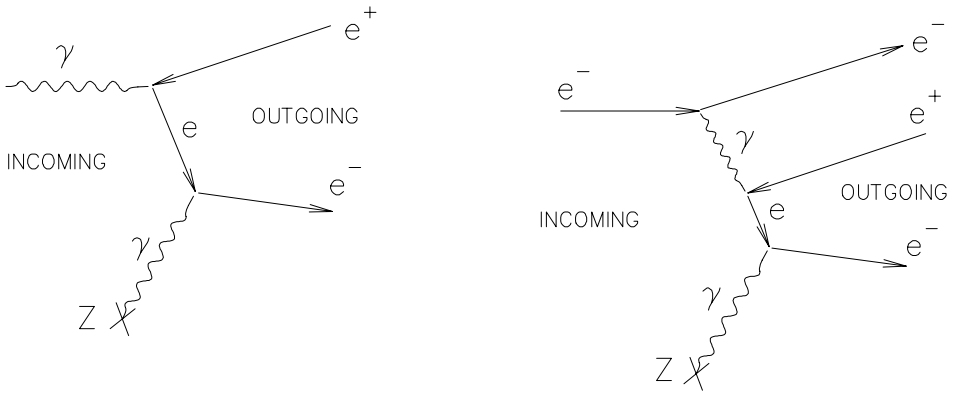


Figure 26.4 FEYNMAN DIAGRAMS for pair production by a gamma ray (left) or an electron (right). These represent the processes in the preceding sketch.

The convention in FEYNMAN DIAGRAMS is that *antiparticle* lines (e^+, for instance) are drawn in the "backward" sense as if they were propagating backward in time. This allows all "electron lines" to be *unbroken*, a graphical expression of the CONSERVATION OF ELECTRONS.[15] There are lots more elegant graphical features to FEYNMAN DIAGRAMS, but I will wait until we discuss QUANTUM FIELD THEORY in the Chapter on Elementary Particles to discuss them further.

The main point here is that the *incoming* particle(s) [γ or e^-] must have at least 1.022 MeV of kinetic energy to create a positron and an electron, both of which have rest masses of 0.511 MeV/c^2. With an *accelerator* one can give the original projectile(s) more energy [there seems to be no limit on how much, except for mundane concerns about funding resources and real estate] and thus facilitiate the creation of *heavier* particles. At TRIUMF, for instance, we accelerate protons to 520 MeV [just over half their rest mass energy of 938 MeV], which is enough to create π MESONS [mass = 139 MeV/c^2] with reasonable efficiency; the high *intensity*[16] of the TRIUMF cyclotron qualifies it for the elite club of "MESON FACTORIES," so named because they "mass produce" π mesons (or PIONS) in unprecedented numbers.

Since heavier particles can in principle decay into lighter particles like gamma rays, neutrinos, antineutrinos, electrons and positrons, almost of these "manufactured" particles are unstable. Nevertheless, they hang around long enough to be studied and sometimes their very instability is what makes them interesting, if only because it precludes finding a cache of them in a more Natural setting.

I have gotten *far* beyond the terms of reference of this Chapter here, but I wanted to "preview" some of the phenomenology of Elementary Particle Physics while focussing your attention on the

[14]This is basically what won Feynman his Nobel Prize; these simple diagrams are rigourously *equivalent* to great hairy *contour integrals* that you would not really want to see! Thus Feynman brought the Right Hemisphere to bear on elementary particle physics. Without this simple tool I wonder how far we would have come by now....

[15]Note that *gamma* particles [photons] are *not* conserved — they are always being created or destroyed!

[16]The intensity of an accelerated particle beam can be measured in particles per unit time [TRIUMF has about 10^{15} protons/sec] or, if the particles carry electric charge, in AMPERES of electrical current [TRIUMF has about 140 μA (microamperes)].

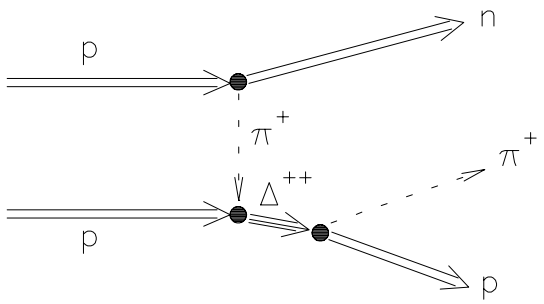


Figure 26.5 Feynman diagram for production of a π^+ meson by a collision between two protons (the most important interaction at TRIUMF).

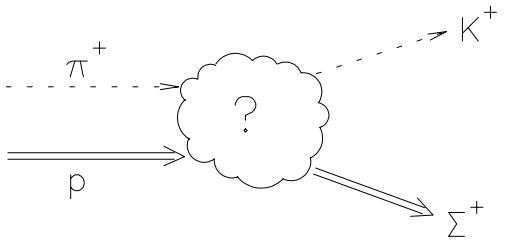


Figure 26.6 Feynman diagram for "ASSOCIATED PRODUCTION" of a K^+ meson [mass = 494 MeV/c^2 and "strangeness" $S = +1$] and a Σ^+ HYPERON [a type of BARYON with mass = 1193 MeV/c^2 and strangeness $S = -1$] in a collision between a π^+ and a proton (the pions produced at TRIUMF don't have enough energy to do this).

simple motive for building higher- and higher-energy accelerators:

> The more kinetic energy is available, the more mass can be created. The heavier the particle, the more options it is apt to have for other lighter particles to decay into, and the more unstable it can be expected to be; hence the less likely we are to observe it in Nature.[17] And the heavier the particle, the more exotic its properties might be.

So far this simple strategy has paid off in many new discoveries; of course, it may not keep working indefinitely....

26.3 Lorentz Invariants

In the previous Chapter we encountered the notion of *4-vectors*, the prototype of which is the SPACE-TIME vector, $x_\mu \equiv \{ct, \vec{\boldsymbol{x}}\} \equiv \{x_0, x_1, x_2, x_3\}$, where the "zeroth component" x_0 is *time* multiplied by the speed of light ($x_0 \equiv ct$) and the remaining three components are the three ordinary spatial

[17] The real surprises come when we find *heavy* particles that *don't* decay into lighter ones [or at least not right away]; this always means some hitherto unsuspected CONSERVED PROPERTY like "strangeness" or "charm" — but now I really *am* getting too far ahead!

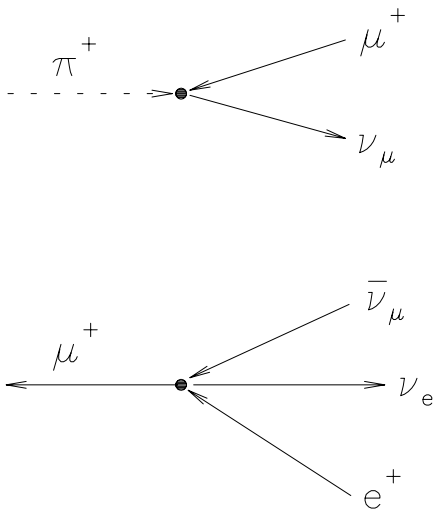


Figure 26.7 *Top:* Feynman diagram for decay of a π^+ meson [mass = 139 MeV/c^2] into a positive MUON (μ^+) [mass = 106.7 MeV/c^2] and a [massless] muon NEUTRINO (ν_μ). *Bottom:* Feynman diagram for decay of a μ^+ into a muon antineutrino ($\bar{\nu}_\mu$), a positron (e^+) and an electron neutrino (ν_e). These are the reactions I use in almost all of my research.

---

coordinates. [The notation is new but the idea is the same.] In general a vector with *Greek indices* (like x_μ) represents a *4-vector*, while a vector with *Roman indices* (like x_i) is an ordinary spatial 3-vector. We could make up any old combination of a 3-vector and an arbitrary zeroth component in the same units, but it would not be a genuine 4-vector unless it *transforms like spacetime* under LORENTZ TRANSFORMATIONS. That is, if we "boost" a 4-vector a_μ by a velocity $u = \beta c$ along the x_1 axis, we must get (just like for $x_\mu = \{ct, x, y, z\}$)

$$\begin{aligned} a'_0 &= \gamma(a_0 - \beta a_1) \\ a'_1 &= \gamma(a_1 - \beta a_0) \\ a'_2 &= a_2 \\ a'_3 &= a_3 \ . \end{aligned}$$

It can be shown[18] that the INNER or SCALAR PRODUCT of any two *4-vectors* has the agreeable property of being a LORENTZ INVARIANT — *i.e.*, it is unchanged by a LORENTZ TRANSFORMATION — *i.e.*, it has the *same value for all observers*. This comes in *very* handy in the confusing world of Relativity! We write the SCALAR PRODUCT of two *4-vectors* as follows:

$$a_\mu b^\mu \ \equiv \ \sum_{\mu=0}^{3} a_\mu b^\mu \ = \ a_0 b_0 \ - \ \vec{\boldsymbol{a}} \cdot \vec{\boldsymbol{b}} \ = \ a_0 b_0 \ - \ (a_1 b_1 + a_2 b_2 + a_3 b_3) \tag{13}$$

where the first equivalence expresses the EINSTEIN SUMMATION CONVENTION — we automatically *sum over repeated indices*. Note the $-$ sign! It is part of the definition of the "metric" of space

[18] Don't you hate that phrase? Actually this one is pretty easy to work out; why don't you do it for yourself?

and time, just like the PYTHAGOREAN THEOREM defines the "metric" of flat 3-space in Euclidean geometry.

Our first LORENTZ INVARIANT was the PROPER TIME τ of an event, which is just the *square root* of the scalar product of the *space-time 4-vector* with *itself*:

$$c\tau \;=\; \sqrt{x_\mu x^\mu} \;=\; \sqrt{c^2t^2 - \vec{\boldsymbol{x}}\cdot\vec{\boldsymbol{x}}} \tag{14}$$

We now encounter our second *4-vector*, the ENERGY-MOMENTUM *4-vector*:

$$p_\mu \;\equiv\; \{\tfrac{E}{c}, \vec{\boldsymbol{p}}\} \;\equiv\; \{\tfrac{E}{c}, p_x, p_y, p_z\} \tag{15}$$

where $cp_0 \equiv E = \gamma mc^2$ is the TOTAL RELATIVISTIC ENERGY and $\vec{\boldsymbol{p}}$ is the usual MOMENTUM 3-vector of some object in whose kinematics we are interested. [Check for yourself that all the components of this vector have the *same units*, as required.] If we take the scalar product of p_μ with itself, we get a new LORENTZ INVARIANT:

$$p^\mu p_\mu \;\equiv\; \frac{E^2}{c^2} - \vec{\boldsymbol{p}}\cdot\vec{\boldsymbol{p}} \;=\; \frac{E^2}{c^2} \;-\; p^2 \tag{16}$$

where $p^2 \equiv \vec{\boldsymbol{p}}\cdot\vec{\boldsymbol{p}}$ is the square of the magnitude of the ordinary 3-vector momentum.

It turns out[19] that the constant value of this particular LORENTZ INVARIANT is just the c^4 times the *square of the* REST MASS of the object whose momentum we are scrutinizing: $\frac{E^2}{c^2} - p^2 = m^2c^2$ or $E^2 - p^2c^2 = m^2c^4$. As a result, we can write

$$E^2 \;=\; p^2c^2 \;+\; m^2c^4 \tag{17}$$

which is a *very* useful formula relating the ENERGY E, the REST MASS m and the MOMENTUM p of a relativistic body.

Although there are lots of other LORENTZ INVARIANTS we can define by taking the scalar products of *4-vectors*, these two will suffice for my purposes; you may forget this derivation entirely if you so choose, but I will need Eq. (17) for future reference.

26.3.1 The Mass of Light

Allow me to hearken momentarily back to Newton's picture of light as *particles*.[20] Actually the following analysis pertains to *any* particles *whose rest mass is zero*. If $m = 0$ then Eq.(6) is absurd, except in the rather useless sense that we may let γ become infinite. On the other hand, Eq.(17) works fine if $m = 0$. Then we just have

$$E \;=\; pc \tag{18}$$

— that is, the ENERGY and MOMENTUM of a *massless* particle differ only by a factor of c, its speed of propagation. Although we cannot define γ because the massless particle *always* moves at c relative to *any* observer [this was, after all, one of the original postulates of the $\mathcal{STR}$], we can talk about its EFFECTIVE MASS, which is the same as its KINETIC ENERGY divided by c^2.

[19] Ouch! There's another one.

[20] This is also a preview of topics to come; as we shall see later, Newton was quite right! Light *does* come in well-defined *quanta* known as PHOTONS, particles of zero rest mass that always propagate at the speed of you-know-what!

Thus, even though light has no REST MASS (because it can never be at rest!), it *does* have an effective mass which (it turns out) has all the properties one expects from MASS — in particular, it has *weight* in a gravitational field [photons can "fall"] and exerts a gravitational attraction of its own on other masses. The classic *Gedankenexperiment* on this topic is one in which the *net mass* of a closed box with mirrored sides *increases* if it is filled with *light* bouncing back and forth off the mirrors!

Is that weird, or what?

Chapter 27

Radiation Hazards

Few issues in our uncomfortably complicated high-tech modern world are so muddled as that of *radiation hazards*. The confusion stems partly from the emotionally charged politics surrounding any subject associated with the word "nuclear" — which in turn is the result of the brutalizing terror of nuclear war that has infected the psyches of several generations of Cold War veterans — and partly from ignorance and misunderstanding of what radiation *does* and how it can be harmful — which in its own turn is the result of decades of gleeful indulgence in the thrills of grade-B sci-fi horror films. Moreover, most people seem quite content with their fantasies and "good *vs.* evil" decision-making strategies, so don't expect a deeper understanding to enhance your popularity! Nevertheless, knowledge is power and *someone* has to know what's going on, so it looks like you're it. Let me tell you what I can.[1]

[1] *Caveat!* I encourage you to distrust everything I say (and everything anyone else says) on this subject until you have seen (and believe) the data for yourself. Like most people, I am not a scholar or even an expert in the field of radiation hazards, just an amateur with strong convictions which will distort my presentation of the evidence; my only excuse for subjecting you to my opinions is that everyone else seems to be so timid about expressing any ideas on this subject that the only information you are likely to get elsewhere (without determined effort on your part) is even more politically motivated and less reliable than mine, which I acquired through informal discussions with various people who *do* have legitimate professional credentials.

27.1 What Hazards?

One thing we can all agree on is that radiation is bad for you, right? Well.... First we have to be careful to define what we mean by "radiation." Your fireplace *radiates* in the infrared (heat) and visible (light) parts of the electromagnetic (EM) spectrum; these forms of radiation are certainly beneficial as long as they don't get out of control. On the other hand, visible light in the form of a high-power laser can inflict damage, as can excessive heat or even microwave EM radiation. On the shorter-wavelength side of the EM spectrum, ultraviolet light can cause sunburn to the skin, while X-rays penetrate deeper and can do the same sort of microscopic damage as the still shorter-wavelength gamma (γ) rays emitted by ^{60}Co (cobalt) radioisotopes. Can we make general statements about all of these? Perhaps, "A little is good, but a lot is bad!" Sorry, nothing so simple. It is certainly true that we cannot maintain health without both heat and light, and a certain amount of "near ultraviolet" may be required for natural vitamin D production in the skin, but we probably have no biological need for microwave or radio frequency radiation; and all EM radiation from "far ultraviolet" upward in frequency (downward in wavelength) is exclusively and unambiguously bad for the individual.[2]

[2] Whether or not genetic mutations are beneficial for the human race as a whole is a difficult question both scientifically and ethically; I will avoid trying to answer it.

Why the big qualitative difference? What do ultraviolet, X-rays and γ-rays do that visible and infrared light don't? At last, a question to which there is a simple answer! They cause *ionization* of atoms and molecules inside cells, leaving behind a variety of free radicals — types of molecules that quickly react chemically with other nearby molecules. If the free radicals react with the DNA molecules in which are encoded all the instructions to our cells for how to act and how to reproduce, some of these instructions can get scrambled.

Surprisingly, this does not always happen. The simplest detectable damage to a DNA molecule is a "single-strand break," in which one of the strands of the double helix is broken by a chemical reaction with a radical. It is a testimony to the robustness of DNA that it is usually able to repair its own single-strand breaks in a few hours![3] If, however, the DNA molecule with a single-strand break is subjected to further damage before it has a chance to "heal itself" then it may sustain a "double-strand break" (two breaks in the same strand), which it seems to be far less able to repair. Before we go on to discuss the consequences of permanent DNA damage, it is important to note that the irreparable damage usually takes place only after a large fraction of DNA molecules have already sustained temporary damage — and that the temporary damage is mostly repaired in a fairly short time. This explains why a given "dose" of radiation is less harmful when accumulated over a long time than when delivered in the space of a few hours.[4]

What sorts of bad things are liable to happen when a DNA molecule sustains irreparable damage, scrambling some part of the instruction manual for the operation of the cell it inhabits?

- **Cell Reproductive Death** [most common] — The cell containing the defective DNA may be unable to reproduce itself, so that although it may be able to function normally for its remaining natural lifetime, when it dies a natural death it will not have a new cell to replace it. Whether this causes a problem or not depends upon whether many other nearby cells have the same malady (one by itself will never be missed!) and upon the natural lifetime of that type of cell — which ranges from a few days for hair follicles, skin and mucous membrane cells to "forever" for brain cells. Obviously, the loss of reproductive capacity is meaningless for a cell that never reproduces!

- **Genetic Mutation** [most subtle] — If the cell in question happens to be a *gamete* destined for fusion with a member of the opposite sex, the resulting individual will have some scrambled instructions in the construction manual and will probably not grow up normally. In almost every case this will be fatal to the fœtus, and in almost all the remaining cases it will be detrimental to the survival of the individual, although such mutations have presumably played a rôle in evolution to date. Note however that it is strictly impossible for any individual's genetic makeup to be *retroactively* altered by radiation (like the Hulk or Spiderman or any number of cheap sci-fi horrors), as this would require the *same* accidental scrambling to take place independently in every DNA strand in the victim's body!

  For men, there are two types of genetic damage: the sperm cells themselves have

[3] Whether this is because of multiple redundancy or context programming I do not know, but it sure is an impressive feat.

[4] I should add an extra *caveat* at this point: what I have said about single- and double-strand breaks and healing times is what I recall from sitting on the PhD committee of a student working on pion radiotherapy about ten years ago. I don't imagine it has been substantially revised since then, but I am not absolutely sure. If you want a more reliable witness I will be glad to direct you to local experts.

an active lifetime of only a few days, after which a new generation takes over; but the sperm-producing cells are never replaced and so can never repair damage to themselves. The latter applies also to women: the female gametes (eggs) are all produced early in life and, once damaged, cannot be repaired.

If the altered cell is "just any old cell" then usually the change is harmless — either the cell merely fails to do its part in the body until it dies or else the affected part of the DNA is irrelevant to the functioning of that cell in the first place — but occasionally the change is related to cell division itself, and then there can be real trouble.

- **Cancer** [most unpleasant] — Sometimes (very rarely) a damaged DNA molecule instructs a cell to mobilize all its resources and the resources of all its neighbours to reproduce as many copies of itself as possible. The offspring preserve the mandate, and a chain reaction takes place that "crashes the system." This runaway reproductive zeal of a misguided cell is what we know as CANCER, and it is the worst hazard of radiation exposure. As far as anyone knows, *any* exposure to ionizing radiation increases one's chances of developing cancer, and so we can unambiguously say that **ionizing radiation is bad for you.**

Before we go on, it is interesting to note that *all* of the most potent therapies for *treating* cancer involve either ionizing radiation or chemical reactions that cause similar DNA damage; the strategy for these "interventions" is always to cause such overwhelming DNA damage to the cancer cells that *every single cancer cell* suffers "cell reproductive death" as described above. Although there are various techniques for making the cancer cells more susceptible to the radiation or harsh chemicals than normal cells, there are inevitably many casualties among the latter. It is not unusual, for instance, to kill off (in the sense of "reproductive death") as many as 90% of the normal cells in the tissues surrounding a tumour, relying upon the fantastic healing capacity of normal tissue to bounce back from this insult. Remember, the idea is to kill 100% (!) of the cancer cells.

It provides an important perspective to realize that the radiation used to kill the cancer may deliver a "dose" to healthy tissues that is more than 10,000 times the maximum legal limit for environmental radiation exposure, and yet the increased likelihood of developing another cancer from the radiation therapy is regarded as a negligible risk relative to allowing the existing cancer to progress unchecked. Whether or not oncologists have optimized their treatment strategies is another charged issue which I will avoid, but it is clear that a large radiation dose does not necessarily "give you cancer" immediately; rather it increases your *chances* of developing cancer *in the long run*. By how much? And over how long a run? These are the quantitative statistical questions that must be answered if one is to develop a rational scheme for evaluating radiation hazards.

27.2 Why Worry, and When?

Unfortunately much of our public policy today seems to be based on the belief that if we could only eliminate the last vestiges of hazardous materials and dangerous practices from our society then none of us would ever get sick or die. This must be regarded as nonsense until medical science finds a way to halt or reverse the natural aging process — which might not be such a great idea.[5]

[5]Even if a sufficiently totalitarian regime could be instituted to forcibly prevent the population from increasing exponentially once immortality was commonplace, would such a thing be beneficial? Would life seem as precious if it were

If I were exposed to radiation that virtually guaranteed that I would develop cancer within 200 years, *but no sooner than 100 years*, would I be wise to worry? What if it raised my chances of developing cancer within 20 years by 2%? My chances of developing cancer within 20 years are roughly 20% *normally*, now that we have eliminated most other mortal dangers except for heart disease. Most people would agree that I would be foolish to allow myself to be exposed to enough radiation to increase my chances of developing cancer within 10 years by 10% (unless we mean 10% *of* 10%, in which case it is a rather small increase — one must always ask for precise explanations of statistical statements!) and yet we all routinely choose to engage in activities that are as least as hazardous, such as downhill skiing or motorcycle riding. Why do we reserve such terror for one sort of hazard when we so stoically accept others of far greater risk? Which is the healthier attitude?

27.2.1 Informed Consent *vs.* Public Policy

One answer to this question is that there are two *entirely separate* issues regarding life-threatening hazards: the first relates to *personal choice*, in which the individual has a right to decide for him/herself how much risk is justified for the sake of certain perceived benefits; the second relates to *public policy*, in which decisions may affect millions of people without their knowledge or consent. It is not unethical for me to choose to risk my life for what I conceive to be worthwhile, or even for fun (as long as I don't expect anyone else to bear the consequences); but it *is* unethical for me to subject millions of other people to the same level of risks without their consent.

not so annoyingly *short*? Again I shall bypass the thorny issues and play the hand I am dealt.

27.2.2 Cost/Benefit Analyses

Unfortunately, this does not necessarily make the issues simpler. It does not help to conclude that any global policy decision that increases the public risk *at all* is *a priori* wrong, because of unintended consequences and complex interconnections. A nuclear power plant in New York puts local residents at some risk from possible cancer due to possible radiation exposure from possible leaks due to probable bungling and/or inadequate engineering and/or substandard construction. On the other hand, a fossil fuel plant of the same size puts a different population at risk from acid rain, ozone depletion and the Greenhouse Effect. [6] And no power plant at all increases the risk of pneumonia in the area served during Winter brownouts — probably the worst hazard of the three in the short term, but one to which millennia of familiarity have hardened us!

The point is, *every* public policy decision creates risks. Even a decrease in bus fare, if it affects millions of people, will cause some people to die this year who would otherwise have lived longer. The questions must always be, "Is this likely to do any good? How much good? Is it likely to do any harm? How much harm? What are the relative probabilities of good and harm? How many people are likely to suffer from the harm? How many people are likely to benefit from the good?" And of course the two questions most popular with politicians, "Which people?" and "When?"

Time to duck the difficult issues again. I am satisfied to point out the questions; I have no more competence than the next person to offer answers. Suffice it to say that any sensible policy regarding radiation hazards, whether public or personal, must take into account that each of us is going to die, that our lifespan is frustrat-

[6] Also, surprisingly enough, from the *radioactivity* released from fossil fuels in combustion, which is far greater than that released by a nuclear power plant *in normal operation*.

ingly short no matter what we do, and that our chances of dying of cancer (radiation-induced or otherwise) are already rather high.[7] So any strategy dictated exclusively by absolute minimization of our cancer risk is somewhat silly. Still, *all other things being equal*, less (ionizing) radiation is better!

27.3 How Bad is How Much of What, and When?

Time to get quantitative. What kinds of radiation are there, how do we measure how much we get, and what effects can we expect from different exposures to different parts of our bodies over different times?

There are lots of kinds of radiation, from the EM spectrum we have already discussed to neutrons, alpha (α) particles, beta (β) "rays" (high-energy electrons) and γ-rays — all constant companions in our environment due to natural or man-made radioisotopes — to the utterly harmless neutrinos coming from our Sun, to beams of high-energy protons, electrons, positrons, pions, muons *etc.*, produced by accelerators like TRIUMF, to catastrophically destructive cosmic rays from which we are shielded by our atmosphere (except when we fly across country in an airliner) and so on *ad infinitum*. Everyone is constantly exposed to most of these types of radiation, accumulating an annual dose varying from a few hundred mR to several R. What are these units "R" and how can we gauge what they mean in practical terms? Time to get more technical.

27.3.1 Units

The basic unit of radiation dose used to be the "*rad*," defined in terms of the *energy deposited* by ionizing radiation *per unit mass* of exposed matter (*e.g.* flesh or bone):

$$1\ rad \equiv 100\ erg/g$$

(g means *gram* here.) More recently, for some reason this nice mnemonic unit has been officially supplanted by yet another "personal name *SI* unit" in honour of British physicist and radiation biologist Louis Harold Gray (1905-1965) — the "*gray*:"

$$1\ gray \equiv 100\ rad \equiv 1\ J/kg.$$

Early work on radiation hazards was based on X-ray exposure[8] and the units used were always *rœntgen* (after the scientist by that name), which are about the same as *rad* for X-rays *only*, and are virtually unused today. Later it was found that even the *rad* was too simple; different *types* of radiation (*e.g.* neutrons) were found to be more (or less) destructive than X-rays for different types of tissues, so an empirical "fudge factor" called the *R*elative *B*iological *E*ffectiveness (RBE) was invented to account for these differences (averaged over all body parts, of course, which decreased its usefulness). The RBEs of γ-rays, X-rays and β-rays (fast electrons) are all 1 by definition; thermal neutrons have an average RBE of 3; fast neutrons (on average), protons and α-rays (^{4}He nuclei) all have RBEs of 10; and fast heavy ions have an RBE of 20.[9]

A new unit was then constructed by combining the RBE with the dosage in rads, namely the *rem* (*r*œntgen *e*quivalent to *m*an), defined by

$$rem \equiv \text{ RBE} \times \ rad.$$

[7] I have been assuming 30%, but that number could be out of date; I don't think it makes much difference to my arguments.

[8] I can remember sticking my feet into the fluoroscope at the corner shoe store and looking at my foot bones inside my new shoes; it was quite popular about 40 years ago.

[9] Actually, the RBE of neutrons varies tremendously for different tissues and is a complicated function of the neutron energy because of the energy-dependence of the neutron capture cross-sections of different elements. Neutrons are very bad.

The "*R*" in the preceding paragraph stands for *rem* and the "*mR*" for *millirem* — one thousandth of a *rem*.

Today the standard international unit for measuring "effective dosage" is the *seivert*, named after Rolf Sievert (1898-1966), a pioneering Swedish radiation physicist. Converting between *rem* and *seivert* is just like converting between *rad* and *gray*:

$$1 \textit{ seivert} \equiv 100 \textit{ rem}.$$

Now that all mnemonic content has been deleted from the names of the units associated with radiation dosage, you may expect these names to stick.[10]

27.3.2 Effects

All of these units are meaningless until one has some idea of how bad one of them is for you. Here are some rules of thumb that may be off by factors of two from one case to the next:

- **Instant Death:** It takes a monumental radiation dose to kill outright, typically something like 5000 *R* (50 Grays) "whole-body" — *i.e.* half a million ergs of energy deposited in every gram of your body. This amount of energy wipes out your central nervous system (CNS) immediately when delivered all at once. Needless to say, only the military mind makes a strong distinction between this and the next level down.
- **Overnight Death:** Approximately 900 *R* (9 Grays) whole-body will accomplish the same thing as 50 Grays but it takes about a day.
- **Ugly Death:** A somewhat lower dose, around 500 *R* (5 Grays) causes severe "radiation sickness" (*i.e.* nausea, hair loss, skin lesions, *etc.*) as the body's short-lived cells fail to provide new generations to replace their normal mortality ("cell reproductive death"). It is not this trauma which usually kills, however, but the complications that arise from a lack of resistance to infection, due in turn to the lack of new generations of white blood cells. If you survive the initial radiation sickness and avoid infection, you will probably recover completely in the short term; but you are very likely to develop cancer (especially leukemia) in later years (usually some 10-20 years later!) and your offspring, if any, will have a high probability of genetic mutations.
- **Sub-Acute Exposures:** From a whole-body dose of around 100 *R* (1 Gray) delivered in less than about a week, you are unlikely to notice any immediate severe symptoms. However, you are likely to develop leukemia in 10-30 years, and there is a significant chance of genetic mutations in your offspring. A whole-body exposure of 5 *R* delivered over 1 year was believed in 1970 to represent 1.8 "doubling doses" — *i.e.* it was thought to multiply your odds of developing cancer by a factor of 2.8 if maintained year after year. At that time it was also the legal exposure limit for radiation workers in the U.S.A., set by the Atomic Energy Commission (AEC) there. Presumably quite a few people received this exposure for a few years, although it is unusual for more than a small fraction of workers to receive the maximum allowed exposure. For perspective, it is noteworthy that a series of spinal X-rays is apt to give an exposure of 1–4 *R* locally, and that an afternoon on Wreck Beach in midsummer often pro-

[10] The purpose of *SI* units is evidently to make it as difficult as possible for intelligent laypersons to understand what "experts" are talking about. I cannot imagine a more humiliating posthumous fate than to have countless generations confused by some perfectly simple unit renamed the "*brewer*" in honour of my efforts to make some field more understandable.

duces a painful sunburn that represents 10-20 R to the skin; the resultant burn is a *bona fide* radiation burn and is just as dangerous as any other kind! In fact, the overwhelming majority of all radiation-induced cancer fatalities on Earth can be attributed directly to far ultraviolet from our favourite nuclear fusion power plant in the sky: the Sun.

- **Marginal Exposures:** The average exposure from natural sources of radiation is on the order of 300 mR per year. As of 1979 this was also the Canadian legal limit for public exposure from artificial sources. Whether an extra 300 mR makes a significant difference epidemiologically in the incidence of cancer depends almost entirely on what one considers significant; however, it is a fact that the statistical difference between populations that have received such an exposure "artificially" and those who have not is smaller than the statistical differences between populations with different eating habits, who live in different regions, who have different types of jobs, *etc.* This is partly because of the wide variety in the amount and type of *natural* radiation exposure.

Before we go on to discuss *sources* of radiation, it is important to note that different organs or body parts have dramatically different resistance to radiation. The *hands*, in particular, are able to withstand radiation doses that would kill if the whole body were subjected to them! The *lens of the eye* and the *gonads* are considered to be the most vulnerable and should be protected first.

27.4 Sources of Radiation

In 1972 a detailed survey was made of average annual whole-body doses to the U.S.A. population from various sources. Occupational and miscellaneous artificial exposures averaged about 1-2 mR/y (remember, some people got enough to make up for the vast majority who got none!); global fallout from nuclear testing made up about 6 mR/y; medical exposures (X-rays, radiotherapy, *etc.*) were good for nearly 100 mR/y; and natural background (see below) averaged about 120 mR/y. The numbers have not changed much in the intervening years. One must conclude that for the average person there are only two significant sources of radiation exposure: medical and natural. Although this begs the question of "extraordinary cases" who receive larger exposures in accidents such as Chernobyl, it still helps to set perspectives for those examples.

Some medical and natural radiation sources are listed below. For medical examples I have shown the mean dose per exposure. It is important to note that these are only the *easily measured* forms of radiation — X-rays and γ-rays — that penetrate flesh (and detectors!) easily. More insidious and difficult-to-measure types will be discussed in the next Section.

- **Medical X-rays:** Chest, radiographic: 45 mR. Chest, photofluorographic: 504 mR. Spinal (per film): 1265 mR. Dental (average): 1138 mR.[11]
- **Cosmic Rays:** Sea level: 30–40 mR/y. Colorado: 120 mR/y. At 40,000 ft: 0.7 mR/h.[12]
- **Natural Terrestrial Radionuclides:** γ-radiation is fairly uniform in the U.S.A., ranging from 30 mR/y in Texas

[11] Note: medical X-rays are normally *localized* to the region being imaged; they are not "whole-body" and therefore are not as bad as they look. Still. . . .

[12] Note: that is per *hour* at a typical cruising altitude for a normal commercial jetliner; thus an average round-trip transcontinental flight yields a dose of 6-8 mR! The estimated average cosmic-ray dose for airline crew is 670 mR/y. Astronauts have it even worse.

to 115 *mR/y* in South Dakota. Guess where the uranium deposits are![13]

27.5 The Bad Stuff: Ingested Radionuclides

The information given above would seem to indicate that medical X-rays were the worst radiation hazard around, except for natural sources we can't do much about. Unfortunately this is a distortion based on the difficulty of measuring the most dangerous kind of radiation: α-emitting radionuclides (radioactive isotopes). Many heavy elements have isotopes which naturally fission into lighter elements plus a helium nucleus, with the latter being emitted with a substantial kinetic energy as an alpha "ray." The range of most α particles is only a few *cm* in air and less than a *mm* in tissue, so the damage they cause is localized. While this may be reassuring when the isotopes are at arm's length, it can be bad news if you have breathed them into your lungs or swallowed them so that they can collect in your bones, where they can do the most damage! Since there is such a wide variety of radioactive elements with assorted chemical properties, it is wise to be aware of the specific hazards associated with each. I have neither the expertise nor the space to provide a comprehensive survey here, but I can mention a few of the most common culprits.

- **Radon:** All rock contains some amount of naturally occurring *radium* which gradually decays, releasing the chemically inert noble gas *radon*. Radon in turn is a radioactive element which decays by emitting a rather low energy α particle that is quite difficult to detect since it has such a short range it can't penetrate the window of a typical Geiger counter. Thus until recently there was little known about radon in our environment, even though it is generally believed that Madame Curie died from exposure to radon emitted by the radium upon which she performed her famous experiments. It is now felt by many that radon is the most widespread and dangerous of all radiation hazards, because it accumulates in the air of any building made of rock, brick or concrete (especially those with closed circulation air conditioning!) and thence in the lungs of the people breathing that air. Lungs in fact make a superb filter for the radioactive byproducts of radon, so that one of the most effective radon detection schemes is to measure the radioactivity of the *people* who live in high-radon environments. In the lung tissue, the short-ranged α particles expend all their energy where it does the most harm, raising the incidence of lung disease and cancer. Rocks from different regions have a tremendous range of radium content, so that a stone house may be perfectly safe in one city and hazardous in another.[14]

- **Potassium and Carbon:** Radioisotopes of potassium and carbon are continually created in the atmosphere by cosmic ray bombardment; these isotopes build up to a constant level in all living tissues, only to decay away in a few thousand years after death. This means that the most radioactive component in your household is probably you![15] It also provides a handy method of estimating the time since formerly living matter was alive (^{14}C and potassium-argon dating).

[13] I don't have the numbers for the Okanagen, but I believe they are even higher than for South Dakota.

[14] I think Vancouver is just slightly on the hazardous side; but in the Okanagen, where there are concentrated uranium ore deposits, I might choose to live in a wooden house. However, you should check out the latest data before you jump to any conclusions.

[15] Married folks who sleep together pick up a few extra *mR/y* from their spouses!

- **Man-made Radionuclides:** There are too many of these to make a comprehensive list here.[16] The most famous is plutonium, ^{239}Pu, the stuff of which fission bombs are made. Plutonium is both a deadly chemical poison and a nasty radioisotope. If a miniscule grain is caught in your lungs or other tissues, it may not do much damage to your body as a whole, but it exposes the tissue immediately around it to a huge dose of radiation, drastically increasing the likelihood of cancer in that tissue. Cancer is just as deadly no matter where it begins, which makes the ingestion of radionuclides the worst possible sort of radiation hazard.

It is important to note that the food chain may serve to concentrate "harmless" levels of radionuclides in (*e.g.*) sea water to a level which is worthy of our concern. Were it not for this effect, and the fact that the waste products of nuclear fission include a large variety of radionuclides with various chemical properties that naturally occurring isotopes do not exhibit, it would be a sensible strategy to dispose of radioactive waste by diluting it and spreading it far and wide in the oceans — since the net radioactivity of reactor fuel actually *decreases* in the process of digging up the uranium, burning it in a reactor and storing the spent fuel rods for 10 years until the short-lived isotopes decay away. Because of the biological concentration effect, however, it is wiser to seek safe long-term containments for radioactive waste.

27.6 Protection

By far the best shielding against radioactivity is Gauss' Law: the intensity of a point source falls off as the square of its distance from the observer. All localized sources are labelled with their activity *at a given distance*, for instance "10 *mr/h* at 1 *m*." If one keeps at least 10 *m* away from such a source, one will receive less than 0.1 *mR* per hour, which is not worrisome.[17] Other safety measures include lead aprons, which are effective *only* for X-rays and γ-rays, and thick concrete shielding for neutrons and high-energy charged particles (these are much in evidence at TRIUMF).

27.7 Conclusions

Draw your own. Please.

Just try to keep in mind that neither extreme attitude ("There's nothing to worry about." *vs* "The only acceptable risk is no risk at all.") represents much of a commitment to the public good. Radiation hazards are subtle and complex, but the benefits of major sources of environmental radiation (*e.g.* medical X-rays) are important. They often save lives by endangering them; the deciding factor must involve relative probabilities and cost/benefit analyses, which may seem cold-blooded but are essential if you really want to do as little harm and as much good as you can.

Remember, if you let someone else decide for you, then you forfeit your right to righteous indignation if you later disapprove of their decision.

[16] One may feel that there are simply too many, period!

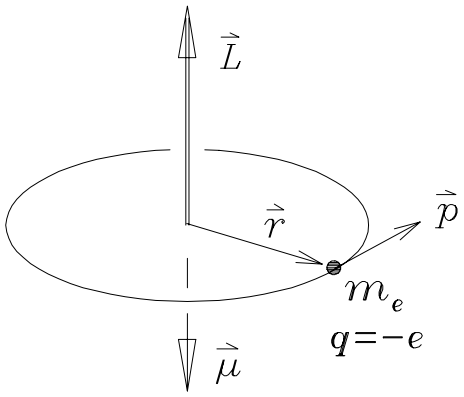
[17] Needless to say, one should never *touch* a radioactive source, because $1/r^2$ can be very large as $r \to 0$.

Chapter 28

Spin

One of the most mysterious yet simple aspects of quantum mechanics ($\mathcal{QM}$) is the behaviour of ANGULAR MOMENTUM, both of the familiar "orbital" variety and of the "intrinsic" variety called SPIN. The latter type has a copyright on the term "spin" in the language of Physics; even though in more colloquial usage "spin" is just another name for angular momentum, in Physics it has a special meaning. But before we can make that distinction, we must first remind ourselves of the definition and behaviour of angular momentum and see how it is tied to the MAGNETIC MOMENT of a charged particle.

28.1 Orbital Angular Momentum



A particle of mass m in a circular orbit of radius r has an angular momentum $\vec{\boldsymbol{L}} = \vec{\boldsymbol{r}} \times \vec{\boldsymbol{p}}$, where $\vec{\boldsymbol{p}} = m\vec{\boldsymbol{v}}$ (in the nonrelativistic limit) is the particle's momentum. Although $\vec{\boldsymbol{r}}$ and $\vec{\boldsymbol{p}}$ are constantly changing direction, $\vec{\boldsymbol{L}}$ is a constant in the absence of any *torques* on the system. If the particle happens to carry an *electric charge* as well as a mass (the case shown being an *electron* with mass m_e and charge $-e$) then the circulation of that charge constitutes a *current loop* which in turn generates a MAGNETIC MOMENT $\vec{\boldsymbol{\mu}}$ which is inextricably "locked" to the angular momentum: $\vec{\boldsymbol{\mu}} = -\mu_B \vec{\boldsymbol{L}}/\hbar$, where $\mu_B \equiv e\hbar/2m_e = 9.2741 \times 10^{-24}\ J/T$ is the BOHR MAGNETON for the case of the electron orbit. Because the potential energy of a magnetic dipole moment $\vec{\boldsymbol{\mu}}$ in a uniform magnetic field $\vec{\boldsymbol{B}}$ is given by $V_B = -\vec{\boldsymbol{\mu}} \cdot \vec{\boldsymbol{B}}$, the orientation of the orbit in a magnetic field determines the contribution of its magnetic interaction to the total energy of the state: $E_B = (\mu_B/\hbar)\vec{\boldsymbol{L}} \cdot \vec{\boldsymbol{B}}$. This contribution is much smaller than the difference between "shells" with different principle quantum numbers n, and so it is called "FINE STRUCTURE" in atomic spectroscopy.

28.1.1 Back to Bohr

One of Niels Bohr's main contributions to Physics was his assertion (backed up by experiment) that *angular momentum is quantized* — it can only occur in integer multiples of $\hbar$. Erwin Schrödinger showed why this was true for the wave functions of the hydrogen atom, but by that time Bohr's principle had been elevated to an empirical "law" of Physics that went well

beyond the realm of atoms. Schrödinger also showed the peculiar nature of the quantization of $\vec{\boldsymbol{L}}$: first, its *magnitude* obeys $|\vec{\boldsymbol{L}}| = \hbar\sqrt{\ell(\ell+1)}$ where ℓ can only have integer values from zero to $(n-1)$, n being the PRINCIPLE QUANTUM NUMBER for which $E_n = -E_\circ/n^2$ in the case of hydrogen; second, its *projection* onto the z axis obeys $L_z = m_\ell\hbar$ where m_ℓ can take on only integer values from $-\ell$ to $+\ell$. Note that Bohr's original prescription for angular momentum quantization (integer multiples of $\hbar$) is actually applicable to the z *component* of $\vec{\boldsymbol{L}}$ — its projection onto the z *quantization axis*, which is chosen arbitrarily unless there is a magnetic field applied, in which case $\hat{z}$ is always chosen *along the field*, $\vec{\boldsymbol{B}} = B\hat{z}$.

28.1.2 Magnetic Interactions

The reason m_ℓ is called the MAGNETIC QUANTUM NUMBER (and the reason m is used for it, rather than some other letter) is that when one imposes a *magnetic field* $\vec{\boldsymbol{B}}$ on an atom, the energy levels E_n (determined only by n in the absence of $\vec{\boldsymbol{B}}$) are "*split*" by the ZEEMAN ENERGY $E_B = \mu_B m_\ell B$ due to the interaction potential of the magnetic moment with the field (see above).

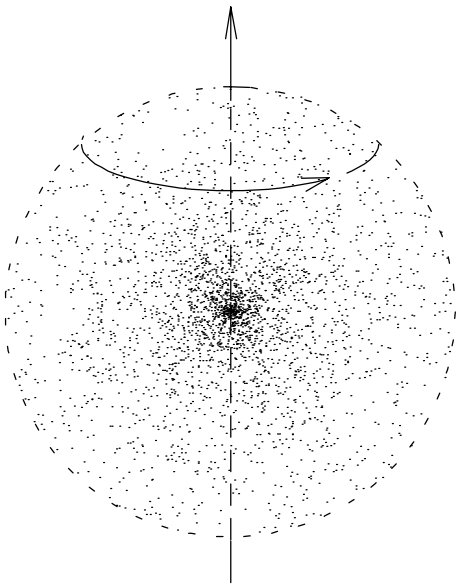
In 1925 Goudsmit and Uhlenbeck reported that, in addition to the "splittings" predicted by the quantization of the orbital angular momentum eigenstates of the electrons in an applied magnetic field, there were *additional* splittings of roughly the same magnitude that could only be explained in terms of some "extra" angular momentum associated with *the electrons themselves*. This was relevant to a previous result that had mystified the community:

In 1922, Stern and Gerlach had done an experiment on various neutral atoms passing through a region of large magnetic field *gradient*, the effect of which is to exert on the passing atoms a *net force* that is proportional to the component of their angular momentum along the axis of the gradient. This allowed Stern and Gerlach to experimentally verify that "spin 1" atoms (with $\ell = 1$) did indeed have three and only three possible values of $L_z = m_\ell\hbar$: $m_\ell = +1, 0$ or -1; and similarly for other integer ℓ. However, their experiments on neutral silver atoms revealed *two* possible projections of the angular momentum along the z axis, a range of options incompatible with the rules $\ell = 0, 1, \cdots (n-1)$ and $m_\ell = -\ell, \cdots 0 \cdots \ell$. The discoveries of Goudsmit and Uhlenbeck suggested that the electron itself might have an *intrinsic* angular momentum that was (somehow) *half* as large as the smallest allowable nonzero *orbital* angular momentum — what we now call "spin $\frac{1}{2}$."

28.2 Intrinsic Spin

The following description is bogus. That it, this is not "really" what intrinsic angular momentum is all about; but it is possible to understand it in "common sense" terms, so we can use it as a mnemonic technique. Many $\mathcal{QM}$ concepts are introduced *via* this sort of "cheating" until students get comfortable enough with them to define them rigourously. (The truth about SPIN, like much of $\mathcal{QM}$, can never be made to seem sensible; it can only be gotten used to!) Imagine a big fuzzy ball of mass spinning about an axis. While you're at it, imagine some electric charge sprinkled in, a certain amount of charge for every little bit of mass. (If you like, you can think of a cloud of particles, each of which has the same charge-to-mass ratio, all orbiting about a common axis.) Each little mass element contributes a bit of angular momentum and a proportional bit of magnetic moment, so that $\vec{\boldsymbol{L}} = \sum \vec{\boldsymbol{r}} \times \vec{\boldsymbol{p}}$ (summed over all the mass elements) and, as for a single particle, $\vec{\boldsymbol{\mu}} = (\text{constant}) \times \vec{\boldsymbol{L}}$. If the charge-to-mass ratio happens to be the same as for an *electron*, then (constant) $= \mu_B$, the Bohr magneton.

Now imagine that, like a figure skater pulling



in her/his arms to spin faster, the little bits of charge and mass collapse together, making r smaller everywhere. To conserve angular momentum (which is *always* conserved!) the momentum p has to get bigger — the bits must spin faster. The relationship between L and μ is such that μ also remains constant as this happens.

Eventually the constituents can shrink down to a *point* spinning infinitely fast. Obviously we get into a bit of trouble here with both relativity and quantum mechanics; nevertheless, this is (sort of) how we think (privately) of an *electron*: although we have never been able to find any evidence for "bits" within an electron, we are able to rationalize its possession of an *irreducible, intrinsic angular momentum* (or "SPIN") in this way.

Such *intrinsic* angular momentum is *a property of the particle itself* as well as a dynamical variable that behaves just like orbital angular momentum. It is given a special label ($\vec{\boldsymbol{S}}$ instead of $\vec{\boldsymbol{L}}$) just to emphasize its difference. Like $\vec{\boldsymbol{L}}$, it is *quantized* — *i.e.* it only comes in integer multiples of a fundamental quantum of intrinsic angular momentum — but (here comes the weird part!) that quantum can be either $\hbar$, as for $\vec{\boldsymbol{L}}$, or $\frac{1}{2}\hbar$!

In the following, s is the "spin quantum number" analogous to the "orbital quantum number" ℓ such that the spin angular momentum $\vec{\boldsymbol{S}}$ has a magnitude $|\vec{\boldsymbol{S}}| = \hbar\sqrt{s(s+1)}$ and a z component $S_z = m_s\hbar$ where $\hat{z}$ is the chosen spin quantization axis. The magnetic quantum number for spin $\frac{1}{2}$ has only two possible values, spin "up" ($m_s = +\frac{1}{2}$) and spin "down" ($m_s = -\frac{1}{2}$). This is the explanation of the Stern-Gerlach result for silver atoms: with no orbital angular momentum at all, the Ag atoms have a single "extra" electron whose spin determines their overall angular momentum and magnetic moment.

28.3 Identical Particles:

The Platonic Ideal lives!

Plato taught something along these lines: every "real" chair is merely an "imprint" of (or an imperfect approximation to) a single "ideal" chair. Similarly for tables, glasses and certainly for the wine in the glass!

As things turned out, when it came to elementary particles Plato didn't go quite far enough. When physicists talk of "*The* electron..." or "*The* neutron..." they are not referring to one particular electron or neutron; they are expressing their (experimentally verified) belief that *every electron is exactly identical* to every other electron! Not just very similar, but *indistinguishable even in principle*. That is, every electron is a *perfect* "imprint" of the "ideal" electron — it *is*, in fact, the ideal electron! This is true of *all* elementary particles, and in QUANTUM FIELD THEORY reaches its ultimate expression: the *number* of (*e.g.*) electrons in existence is just another "quantum number" of the unique and solitary "electron *field*." All evidence suggests that this description is "true" — *i.e.* it predicts what is observed, and its negation would predict things that are not ob-

served.

28.3.1 Spin and Statistics

What are some of these predictions? Probably the most unambiguous example is the dramatic effect of the indistinguishable nature of the particles on *scattering probabilities*. However, another example was historically more important and is more obviously essential to the qualitative properties of our universe. This has to do with ATOMIC PHYSICS and the PAULI EXCLUSION PRINCIPLE for electrons. Said principle was actually surmised from atomic physics data in the first place, and was later proved mathematically. We will simply state it.

BOSONS

All elementary particles with *integer spin* ($s = 0, 1, 2, 3, \cdots$) are called BOSONS because they obey BOSE-EINSTEIN STATISTICS — namely, *an unlimited number of identical bosons can be in exactly the same state*. (Here "state" means that *all* properties are *fully specified*.) Examples of bosons are photons, ^{4}He atoms, pions, ...

FERMIONS

All elementary particles with *half-integer spin* ($s = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \cdots$) are called FERMIONS because they obey FERMI-DIRAC STATISTICS — namely, *no two identical fermions can be in exactly the same quantum state*. (This rule includes the PAULI EXCLUSION PRINCIPLE, but it carries over into many other branches of physics.) Examples of fermions are electrons, protons, neutrons, muons, quarks, ...

The name FERMION honours Enrico Fermi, who (with Paul A.M. Dirac) described the special properties of this type of particles.

28.4 Chemistry and The Periodic Table

Let us consider the consequences of the PAULI EXCLUSION PRINCIPLE for the chemical properties of atoms: for each allowed wave funtion (or "orbital") of an electron in the Coulomb potential, fully specified by the quantum numbers (n, ℓ, m_ℓ), there are *two* (and *only* two) possible *spin* states for the electron: spin "up" ($m_s = +\frac{1}{2}$) and spin "down" ($m_s = -\frac{1}{2}$). Since the Pauli principle only excludes occupation of *exactly* the same state by two electrons, this means that *each orbital* may be occupied by at most *two* electrons; and if both electrons are present in that orbital, their spins are necessarily "paired" — equal and opposite in direction, thus perfectly cancelling each other and contributing nothing to the net angular momentum or magnetic moment.

Determining the chemical properties of the elements thus becomes a simple matter of *counting states*: for each "shell" (n) we see how many different orbitals there are with different values of ℓ and m_ℓ and then multiply by two for the two m_s possibilities to get the total number of electrons that will "fit" into that shell. Of course, this simple picture contains some crude approximations that we must expect to break down before things get very complicated, and they do. We assume, for example, that for a positive charge of $+Ze$ on the nucleus, each time we "add another electron" it goes into the lowest available (unfilled) energy state and effectively reduces the effective charge of the nucleus by one. That is, the first electron "sees" a positive charge of $+Ze$ but the second electron "sees" a positive charge of $+(Z-1)e$, and so on. This is not true, of course, for electrons in the same n shell; they all pretty much see the same apparent charge on the "core" of the atom (and see each others' charges as well). One important effect of such "screening" is that states of higher ℓ (whose wavefunctions do not pene-

trate as deeply into the core) "see" on average a lower effective charge on the core and therefore are slightly less bound (higher energy) than their low-ℓ neighbours. This provides the rule that (up to a point) lower ℓ states are filled first. A collection of states with the same n and ℓ is called a "subshell."

If we want to do this *right* for an arbitrary many-electron atom, we have several very difficult problems to solve: first, for large Z the innermost electrons have kinetic energies comparable to their rest mass energy; thus our nonrelativistic Schrödinger equation is inadequate and one must resort to more advanced descriptions. Second, the approximation that one shell simply "screens" part of the nucleus' charge from the next shell is not that great, especially for wavefunctions that have a significant probability density near the nucleus; moreover, electrons in the *same* shell definitely "see" each other and are affected by that interaction. One way of approaching this problem is to start with an *initial guess* based upon this approximation and then recalculate each electron wavefunction including the effects of the other electrons in their wavefunctions; then go back and correct again with the new guess as a starting point. Eventually this "Hartree-Fock" method should converge on the *actual* wavefunctions for all the electrons.... Finally, the coupling of orbital and spin angular momentum contributions can follow several possible scenarios, depending upon which couplings are the strongest; the final result must always be a single overall net angular momentum (and magnetic moment) for the atom as a whole; but for large atoms it is very difficult to predict even its magnitude; the only thing we can be sure of is that if there are an even number of electrons, protons and neutrons making up the atom, it will be a BOSON.

Having issued these *caveats*, we can go back to our "state counting":

- For each "shell" characterized by an energy E_n (starting at $n = 1$, where n is the PRINCIPLE QUANTUM NUMBER) we have n possible values of ℓ, ranging from $\ell = 0$ to $\ell = (n - 1)$.
- For each ORBITAL QUANTUM NUMBER ℓ, we have $2\ell+1$ possible values of the MAGNETIC QUANTUM NUMBER m_ℓ, ranging from $m_\ell = -\ell$ to $m_\ell = +\ell$.
- For each orbital specified by (n, ℓ, m_ℓ), there are two possible values of m_s: spin up ($m_s = +\frac{1}{2}$) and spin down ($m_s = -\frac{1}{2}$).

Thus there are a total of $N_n = \sum_{\ell=0}^{n-1} 2(2\ell + 1)$ possible fully-specified states for electrons in the n^{th} energy shell. For the first shell, $N_1 = 2$; for the second shell, $N_2 = 8$.

Since (at least for small n) it is a long way (in energy) between shells, a "closed shell" is especially stable — it "likes" neither to give up an electron nor to acquire an extra one from another atom; it has "zero valence." This means an atom with $N_1 = 2$ electrons [helium] is very unreactive chemically, and so is one with $N_1 + N_2 = 10$ electrons [neon]. This goes on until the breakdown of our assumption that all the electrons for one shell are filled in before any electrons from the *next* shell. This breakdown comes at the $n = 3, \ell = 2$ subshell. The $n = 4, \ell = 0$ subshell is actually *lower* in energy than the $n = 3, \ell = 2$ subshell and therefore fills first, despite having a higher n. At this point our simple qualitative rules fail and we must rely on empirical information to further understand the chemical properties of the elements.

28.4.1 Chemical Reactions

Two atoms can "react" chemically in either of two ways. In the IONIC reaction, one can give up an electron (becoming a positive ion or *cation* [so called because it is attracted to an electric *cat*hode]) to the other, which becomes a

negative ion or *anion* [because it is attracted to an electric *an*ode]; the two oppositely charged ions are then attracted to each other by the Coulomb interaction but do not have to stay together — sort of like today's high school dances. When the two atoms can't decide which would rather be the donor or the receptor of electrons, they can form the ambiguous or COVALENT bond, in which they *share* one or more electrons; in this case everyone stays electrically neutral but no drifting away from one's partner is allowed.

Chapter 29

Small Stuff

From Democritos through the Alchemists of the Middle Ages and Mendeleyev's Periodic Table of the Elements all the way to modern Elementary Particle Physics, one of the first duties of "Natural Philosophers" has been to make up *lists* of all possible *constituents of matter* — preferably (for the sake of simplicity) including only the *irreducible* components.

This notion may well be obsolete in the literal physical sense, but the concept lives on; and it is tempting (if misleading) to describe Elementary Particle Physics as the art of inventing the simplest possible *classification scheme* for the "zoo" of known "elementary" particles.

Objects or entities can only be *classified* in terms of their *properties*. Thus the first task is to *define* all the (known) intrinsic properties of matter as concisely as possible, invent ways of *measuring* how much of each property a given particle has, and do the experiments. Of course, this is a highly *iterative* process — after each round of experiments the theorists have to go back to their drawing boards and revise the Ultimate Classification Scheme — but the idea is still the same. My task is now to summarize in one Chapter over half a century of progress along these lines. Naturally I will omit as many of the false starts and backtracks as possible, to make it look as if the present scheme[1] is correct and was obvious from the outset.

29.1 High Energy Physics

Before we begin to construct a classification scheme for the "elementary" particles, we need to have some feeling for the phenomenology involved — and maybe even a bit of historical perspective.

In some sense High Energy Physics (the experimental discipline) began when the first cyclotron capable of producing *pions* "artificially" was built by Ernest Orlando Lawrence at Berkeley in the early 1940's.[2] However, *high energy physics* (the behaviour of Nature) began in the instant of creation of the Universe — and it will be a long time before we are able to study the interactions of

[1] Actually, to be honest, this is not the present scheme. It is the one I learned 30 years ago, beefed up with the tidbits I have absorbed since then. Nowadays people talk about the "*Standard Model*," a more elegant presentation of the dog's breakfast you will get from this Chapter — but not, I think, really a different story. Some of the lower limits on the masses of as yet undiscovered particles will have doubled or tripled recently, so don't take the *numbers* in the tables too seriously.

[2] Lawrence's 184 inch Cyclotron, the biggest *solid pole-tip magnet* synchrocyclotron ever built, was originally conceived as a giant *mass spectrometer* for separating the isotopes of uranium for the first fission bomb; however, a far more efficient method was invented soon after it was built, and "the 184" went into service as a pion and muon producer. Many Ph.D. theses (including my own in 1972) were written on experiments performed at the 184 until it was dismantled in the 1980's to

matter at the energies and densities of those first few femtoseconds.[3] I will compromise by dating HIGH ENERGY PHYSICS (the modern human endeavour) from the hypothesis of Hideki Yukawa in 1935 that the STRONG nuclear force must be mediated by the exchange of particles of intermediate [between electrons and protons] mass, which he therefore named "MESONS" [as in *meso*zoic or *Meso*potamia(?)]. Where did he ever get such an idea?

29.1.1 $\mathcal{QED}$

It began with the FEYNMAN DIAGRAM first shown in the Chapter on RELATIVISTIC KINEMATICS. In Fig. 29.1 I show the Feynman diagrams for single and double photon exchange in QUANTUM ELECTRODYNAMICS or $\mathcal{QED}$, for which Richard P. Feynman shared a Nobel Prize. As before, I will draw Feynman diagrams "left to right" instead of the conventional "down to up." The idea of $\mathcal{QED}$ was (and is) that *all* electromagnetic interactions between charged particles can be described in terms of the *exchange of photons* created by one particle and destroyed by another. The simplest case is the "first-order" diagram in Fig. 29.1, where two electrons exchange a *single* photon. The next (second-order) process is a factor of α^2 less important, where $\alpha \approx \frac{1}{137}$ is the FINE STRUCTURE CONSTANT (not a very mnemonic name any more), which is (sort of) the *strength* of the $\mathcal{QED}$ "vertex" (the point where the photon begins or ends). Because each successive diagram (single photon exchange, double photon exchange, triple photon exchange, *etc.*) is a factor of about 19,000 less important than the one before, $\mathcal{QED}$ is a PERTURBATION THEORY that *converges very rapidly*. That is, you can get a pretty accurate result with very few diagrams.

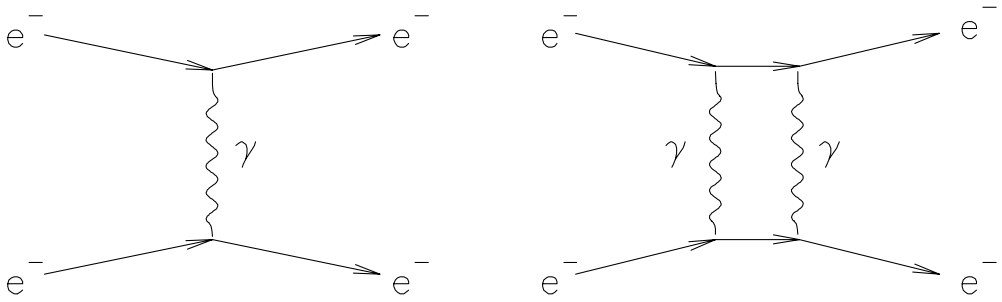


Figure 29.1 FEYNMAN DIAGRAMS for electromagnetic electron-electron scattering in first order (left) and second order (right).

Each diagram, you see, is *rigourously equivalent* to a big messy integral which is definitely less appealing to the Right Hemisphere; but the big integral can be evaluated to give the correct formula for the interaction of the two electrons *to that order in* $\mathcal{QED}$, properly taking into account all the ramifications of QUANTUM FIELD THEORY. Which is...? Let's take another step back for better

make room for what was then the world's most intense Synchrotron Light Source on the same site at what has been called the Lawrence Berkeley Laboratory (LBL) since the end of the 1960's. [Before that it was called the Lawrence Radiation Laboratory (LRL); the name was changed to avoid association with the other LRL branch in Livermore (now known as LLL, the Lawrence Livermore Laboratory) where weapons research is conducted, and to expunge that fearsome word "Radiation." Spineless politicians!]

[3] I refer, of course, to the "BIG BANG" scenario, which is almost universally regarded as the best model of cosmogony [a fancy word for Creation]. Perhaps I will get to say a few words about the Big Bang in a Chapter on GENERAL RELATIVITY.

perspective.

29.1.2 Plato's Particles

When Quantum Mechanics was first developed, it was formulated in a *nonrelativistic limit* — *i.e.*, the particles involved were presumed not to have enough kinetic energy to create *other* particles. Because, if they did, then not only the quantum states of each particle, but *the number of particles present*, would have to be described by the theory. You can see that the combination of Quantum Mechanics with Relativity makes RELATIVISTIC QUANTUM MECHANICS a rather more complicated sort of problem.

Quantum mechanical equations were found for bosons (the KLEIN-GORDON EQUATION) and for fermions (the DIRAC EQUATION) which obeyed the correct relativistic transformations, but now the WAVE FUNCTIONS [ϕ for bosons, ψ for fermions] could not be interpreted as simply as before — in terms of the probability amplitude for a *single particle*. Now they had to be interpreted as the probability amplitude of the FIELD of the corresponding particle, for which the *number of such particles* was merely a quantum number of the field.[4]

As a result, when a Particle Physicist speaks of "THE ELECTRON," (s)he is referring to the electron FIELD, an absolutely literal example of the Platonic Ideal, in which the disposition (and even the *number*) of actual individual electrons is merely a *state* of THE ELECTRON [field]. An actual *single* individual particle in the laboratory is rarely the source of much information about the complete set of all its identical siblings.

A given Feynman diagram therefore represents *one possible case* of the numbers and types of particles present in an interaction with a specified initial and final state. It is one possible *manifestation* of the FIELDS.

29.1.3 The Go-Betweens

A common feature of all such Feynman diagrams is the VIRTUAL PARTICLE(S) being exchanged [created on one side and annihilated on the other] between the interacting particles. They are called "virtual" because they never manifest themselves directly outside the scattering region; of course, in most cases the same sorts of particles *can* be "knocked clear" of the collision by appropriate combinations of momenta, but then the diagram has a different topology. For instance, in Fig. 29.2 the right-hand diagram involves a simple rotation of the left-hand diagram by 90° and so it describes in some sense "the same physics" — but the process depicted, in which a positron and an electron "temporarily annihilate" into a photon and then that photon immediately converts into a new e^+e^- pair, is nominally quite different from the electron-electron scattering in the left diagram. Any $\mathcal{QED}$ adept would automatically think of both as being more or less the same thing.

How is it possible to create a particle "out of nothing" as pictured in these diagrams? Only by virtue of the time-energy version of HEISENBERG'S UNCERTAINTY PRINCIPLE, which says that you

[4] Just to give a hint of how this works, ψ is now composed of some complex exponential wave functions multiplied by *creation* and *annihilation operators* that respectively increase and decrease the number of particles of that species by one. The creation and annihilation operators obey an algebra that corresponds to the statistical properties of the particle — *e.g.*, for fermions no two can be in the same state, *etc.* I will resist the temptation to show any of the equations, which are actually very compact but (as one might expect) have an extremely high "interpretation density."

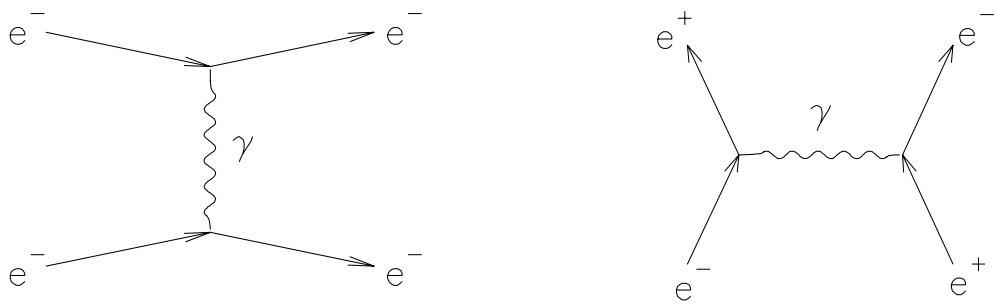


Figure 29.2 Left: Feynman diagram for electron-electron scattering by single photon exchange. Right: "CROSSING SYMMETRY" diagram for electron-positron scattering in the "*s*-channel" by virtual photon annihilation and pair production.

---

can "cheat" energy conservation by an uncertainty ΔE, but only for a short time Δt such that

$$\Delta t \, \Delta E \geq \frac{\hbar}{2} \tag{1}$$

The bigger the "cheat," the shorter the time.

For photons, with no rest mass, a minimum of energy has to be "embezzled" from the "energy bank" to create a virtual photon; as a result it can travel as far as it needs to find another charged particle to absorb [annihilate] it. A heavier particle, on the other hand, cannot live for long without either being reabsorbed by the emitting particle or finding a receiver to annihilate it; otherwise the UNCERTAINTY PRINCIPLE is violated. This brings us back to Yukawa.

Around Yukawa's time every physicist knew that atomic nuclei were composed of NUCLEONS (protons and neutrons) confined to an extremely small volume. The problem with this picture is that the protons are all positively charged and the neutrons are (as the name suggests) neutral, so that such a nucleus entails keeping positive charges *very* close to each other — something that COULOMB REPULSION would rather they didn't do! Therefore (reasoned Yukawa) there must be a "STRONG" *attractive* force between NUCLEONS that was able to overpower the electrostatic repulsion.

But if the STRONG force were *long-range* like the ELECTROMAGNETIC force, then *all* nucleons *everywhere* would "reach out to someone" and fall together into one gigantic nucleus! This appears not to be the case, luckily for us. Therefore (reasoned Yukawa) the STRONG force must be *short-range*.

Now, we have just finished describing what would make a force have a short range — namely, the EXCLUSION PRINCIPLE: if the VIRTUAL QUANTA (particles) mediating the force are moderately *massive* [*i.e.*, "mesons"] then they require a big "cheat" of energy conservation to be created in the first place, and must be annihilated again very soon to have existed at all. Yukawa compared the known size of nuclei (about 10^{-15} m) with the UNCERTAINTY PRINCIPLE, assuming propagation at roughly the speed of light, and deduced that the MESONS mediating the STRONG force must have a mass of about 130 MeV/c^2.

A few years later, MUONS were discovered in high-energy COSMIC RAYS,[5] and the Physics world was quick to acclaim them as Yukawa's MESONS. Unfortunately, they were wrong; the muon is a

[5]Muons are the main component of cosmic rays that make it to the Earth's surface — all the more strongly interacting

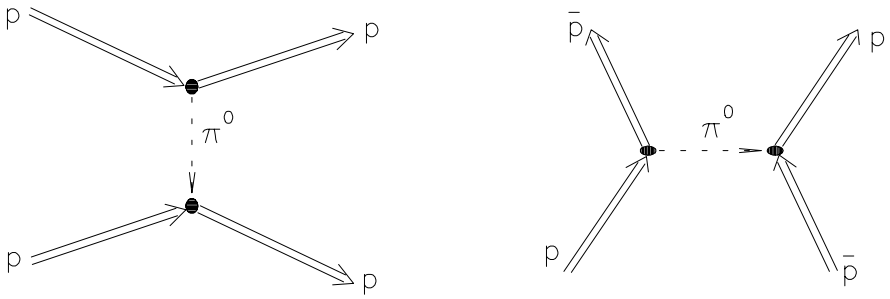


Figure 29.3 Left: Feynman diagram for proton-proton scattering by single pion exchange. Right: "CROSSING SYMMETRY" diagram for proton-antiproton scattering in the "s-channel" by virtual annihilation into a π^0 followed by proton pair production. Note the similarity with the Feynman diagrams for $\mathcal{QED}$, where the pion's rôle is played by a photon.

LEPTON, like the electron or the neutral NEUTRINOS, which accounts for its penetration through the atmosphere (leptons do not interact strongly).[6] This quickly became clear, and shortly thereafter the true "nuclear glue" meson, the PION, was discovered in *very* high-altitude cosmic ray experiments and at the 184 inch Cyclotron in Berkeley. Then HIGH ENERGY PHYSICS began in earnest.

The Perturbation Paradigm Stumbles

It didn't take long for the theory of strong interactions to run into problems. The essence of the difficulty lies in the very word "strong." The strength of an interaction can be calibrated by the magnitude of the dimensionless *coupling constant* applied at each *vertex* [wherever a virtual particle is created or annihilated] in a Feynman diagram such as Fig. 29.1. As explained earlier, each such vertex in $\mathcal{QED}$ has a strength of $\alpha \approx \frac{1}{137}$, which makes "higher order diagrams" rapidly insignificant — great for calculating with a PERTURBATION THEORY!

Unfortunately, the "strength" of a vertex in STRONG INTERACTIONS is on the order of 1. This means that the single pion exchange diagram shown on the left in Fig. 29.3 or Fig. 29.4 is in principle *no more likely* than the incomprehensible mess on the right in Fig. 29.4, involving manifold exchanges of pions and other mesons, as well as creation and annihilation of baryon-antibaryon pairs.[7] Worse yet, *this is only one example* of the seemingly endless variety of possible diagrams one must in principle consider in order to make an accurate calculation of "simple" nucleon-nucleon scattering!

Of course, it wasn't *quite* that bad. Handy "sum rules" were discovered that explained why single pion exchange usually got you pretty close to the right answer, but *in principle* one had to make an almost infinitely difficult calculation in order to get the sort of *precise* predictions that Perturbation Theorists had come to expect from their experiences with $\mathcal{QED}$. Moreover, there were conceptual

---

particles are absorbed or re-scattered in the atmosphere, which makes a pretty good shield. In fact, if you take a transcontinental trip at 30,000 feet altitude, you pick up about 50 mR of ionizing radiation from cosmic rays that are *not* absorbed because you are above most of the shield!

[6] In case you wondered, I am skipping over a lot of agonizing reevaluation and painstaking experiments that led to the discoveries that justify using the "modern" names for all these particles; the muon was called a "mesotron" for years and is still sometimes referred to as a "mu meson" in the USSR. But why sacrifice simplicity for mere historical accuracy?

[7] I haven't bothered to label all the particles; see if you can find any violations of local conservation laws.

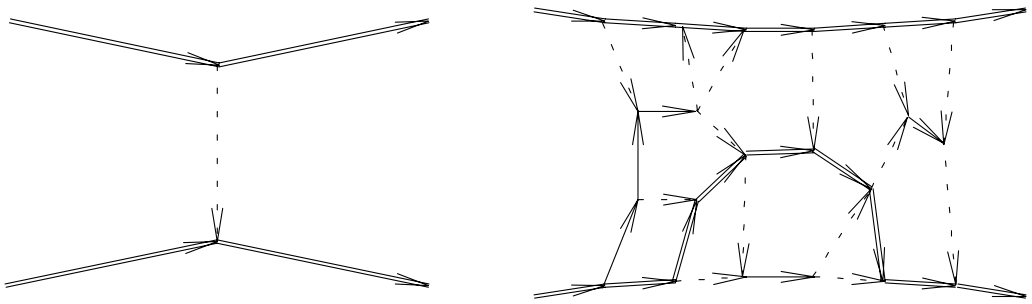


Figure 29.4 Left: Feynman diagram for single pion exchange. Right: A far more complicated Feynman diagram that is in principle no less important!

---

nightmares to sweat out — if you look closely at Fig. 29.4, for instance, you will notice that a proton can emit a pion [OK, there are pions inside protons] which can turn into a proton-antiproton pair [OK, there are protons inside pions... Wait a minute!] and so on. Like many nightmares, this revealed an unexplored avenue of understanding: in the 1960's and 70's, Geoffrey Chew and his Theory group at Berkeley developed a non-perturbative theory of strong interactions that contained the "BOOTSTRAP PRINCIPLE:" every hadron is made up of combinations of all the other hadrons (and itself). Although I never could understand Chew's models, they represented a genuinely new paradigm that gained a good deal of purchase on the problem when suddenly the attention of the Particle Physics community was diverted by a revival of Perturbation Theory in the form of a QUARK model, about which I will say more later; since then Chew's approach has been sadly neglected, which I suspect is a great loss to Physics. Still, if we can get answers more easily by "recycling an old paradigm," the outcome is inevitable.

Weak Interactions

Skipping ahead to the 1980's, the virtual quanta mediating the WEAK INTERACTION (the force next weakest to the gravitational force) have only recently been identified directly in immense experiments at the biggest accelerators. These are the $W^{\pm}$ and Z^0 "INTERMEDIATE VECTOR BOSONS" whose masses are shown in Table 29.1.3. What can you conclude about the *range* of the WEAK INTERACTION?

In a unification of the WEAK and ELECTROMAGNETIC interactions that won acclaim for numerous theorists in the past two decades or so, the γ and the W and Z bosons have been shown to be merely different aspects of the same "ELECTROWEAK" force, despite their gross dissimilarities in mass and lifetime.[8]

[8]This theory now forms the core of what is known as "THE STANDARD MODEL" of elementary particles — a name which reveals a certain disaffection, since no one is particularly excited at the prospect of serving the Establishment prejudices connoted by a "standard model." Particle Physicists, like most free thinkers, prefer the self-image of a romantic revolutionary challenging established conventions and "standard models" everywhere. Not surprisingly, a great deal of experimental effort goes into "tests of the Standard Model" which the experimenters openly hope will throw a monkey wrench into the works.

Table 29.1 The INTERMEDIARY particles that convey various forces between other elementary particles.

Particle		Mass (GeV/c^2)	Interaction mediated	Lifetime (s)
graviton	(?)	0	*gravity*	stable
photon	γ	0	*electromagnetism*	stable
vector boson	$W^{\pm}$	80.6	*weak*	2.93×10^{-25}
vector boson	Z^0	91.2	〃	2.60×10^{-25}
pion (mainly)	π	0.139	*strong*	$\pi^{\pm}: \ 2.6 \times 10^{-8}$ $\pi^0: \ 8.3 \times 10^{-17}$
gluon	g	0?	*superstrong*	?
Higgs boson	H^0	> 24	*ultrastrong*	?
Higgs boson	$H^{\pm}$	> 35	〃	?

29.1.4 The Zero-Body Problem

Before I depart from QUANTUM FIELD THEORY, let me point out a rather amusing consequence of being able to create almost anything you like out of nothing, provided you only do it for a very short time: As you may have heard, no one has ever found a completely satisfactory *general* solution for the THREE-BODY PROBLEM in Classical Mechanics — *i.e.*, the detailed behaviour of 3 particles all mutually interacting; however, the TWO-BODY PROBLEM (2 particles orbiting or scattering off one another) was "solved." RELATIVISTIC QUANTUM FIELD THEORY makes the 2-body problem into a *many*-body problem by virtue of all those virtual quanta being exchanged. Worse yet, the ONE-BODY PROBLEM (a single particle hanging around lonely in empty space) is similarly complicated by its tendency to emit and reabsorb a "cloud" of virtual quanta — not a trivial matter, since most "bare" particles are thought to acquire many of their "dressed" properties (such as mass) by virtue of such "renormalization."

Worst of all, the ZERO-BODY PROBLEM (*i.e.*, the vacuum) is now poorly understood, since there is truly *no such thing* as "empty space" — it is constantly filled with virtual electron-positron pairs (for example) popping into and out of existence, and these short-lived virtual quanta have the capacity to interact with each other and external particles! For example, there is a measurable effect on the H atom energy levels due to "VACUUM POLARIZATION," in which the virtual e^+e^- pairs actually notice the presence of passing "real" electrons and interact with them before disappearing again.[9]

[9]There is an even more dramatic consequence in the neighbourhood of a very small BLACK HOLE whose *tidal forces* (the *gradient* of the gravitational field between one place and another) is so intense that one of the virtual particles of a pair can

Simple, eh?

29.1.5 The Seven(?) Forces

Although I have not yet defined what I mean by half the terms in Table 29.1.5, this is a convenient place to summarize the known and hypothetical interactions of matter. It is conventional to group "superweak"[10] together with the ELECTROWEAK interaction (which "unifies" the WEAK and ELECTROMAGNETIC forces) and to put "superstrong" and "ultrastrong" in with the STRONG interaction so that you should not be surprised to hear that there are only *three* "official" forces — gravity, electroweak and strong. However, there is a certain amount of freedom in semantics here....

Table 29.2 Interactions of the elementary particles. A "yes" means that the types of particle indicated at the left are directly coupled to the force above; "no" means the opposite; three asterisks ($***$) means that the particle in question is the *intermediary* for that force.

PARTICLE(s)		Gravity	Super-weak	Weak	Electro-magnetic	Strong	Super-strong	Ultra-strong
gravitons		$***$						
photons	γ	yes	?	no	$***$	no	no	no
neutrinos	ν_e, ν_μ, ν_τ	yes	?	yes	no	no	no	no
leptons	e, μ, τ	yes	?	yes	yes	no	no	no
mesons	$\pi, K, \ldots$	yes	?	yes	yes	yes	no	no
baryons	$p, n, \Lambda, \ldots$	yes	?	yes	yes	yes	no	no
neutral kaons	$K^0, \bar{K}^0$	yes	yes	yes	yes	yes	no	no
vector bosons	W, Z	yes	?	$***$	yes	no	no	no
quarks	u, d, s, c, b, t	yes	?	yes	yes	no	yes	no
gluons	g	yes					$***$	
(hypothetical)	T, V	yes						yes
Higgs bosons	H	yes	?					$***$
Relative strength		10^{-40}	?	10^{-4}	$\frac{1}{137}$	1	10-100	$> 10^{10}$?

fall into the black hole while the other is ejected and becomes a "real" particle — leading to intense radiation that can be described as the *explosive annihilation* of the miniature black hole. This explains why there are no *small* black holes around any more, only *big* ones whose gravitational gradient is very gentle at the Schwartzschild radius. I will define these terms in the Chapter on GENERAL RELATIVITY.

[10]The "superweak" force is a name coined to describe a *really* esoteric interaction which appears to affect *only* the decays of strange neutral mesons (if it exists at all).

29.1.6 Particle Detectors

Turning back to the hardware of High Energy Physics (HEP), I should point out that it is not enough to build accelerators capable of delivering enough energy to a collision to create more massive and more exotic particles — one must also have some way to "see" those particles once they are created. This is in principle rather challenging, since they are all apt to be moving at near light speed and are certainly too small to detect with visible light; moreover, usually they don't last very long — the heavier the particle, the larger the variety of lighter particles into which it might decay! This rule-of-thumb works quite well in general, so that exceptions (long-lived heavy particles) stand out rather dramatically; more on this later.

In practice it is surprisingly easy to "see" elementary particles, once you get used to a new way of "seeing." The basis of all particle detectors is that *charged* particles cause *ionization* where they pass through matter.[11] The ions they leave behind form a "track" that can be detected in several ways.

Scintillating!

The "workhorse" of experimental HEP is the *scintillation counter*. This simple device works as follows: the ionization of certain types of molecules causes photochemical reactions that liberate visible light called "scintillation" light.[12] This light is conveyed through a clear liquid, plastic or crystalline matrix, bouncing off polished exterior surfaces *via* total internal reflection until it reaches the *photocathode* of a vacuum tube where the photons liberate electrons *via* the PHOTOELECTRIC EFFECT. These electrons are then accelerated by high voltages in the tube until they strike a "first dynode" where each electron knocks loose about ten additional electrons which are accelerated in turn to the "second dynode" where they in turn each knock loose another ten electrons each, and so on down a cascade of up to 18 dynodes. As a result, that one electron originally liberated by the incoming photon can produce a pulse of 10^{18} electrons at the "anode" or the tube, which is (mnemonically, for once) called a PHOTOMULTIPLIER TUBE. These amazing devices have been refined over a period of nearly half a century until some have "quantum efficiencies" approaching 100% (they can fairly reliably detect *single photons*) and (most importantly) generate electrical pulses a few ns (nanoseconds, billionths of a second) wide whose arrival at a bank of fast electronics is correlated with the time the original ionizing particle hit the detector within a fraction of a ns. This means High Energy Physicists can routinely do *timing* with a resolution comparable to the length of time it takes light to go 10 cm! Without this impressive *timing* capability it would be very difficult to do *any* modern HEP experiments. Interestingly enough, this part of the technology has not improved significantly in several decades.

[11] Neutral particles either convert into charged particles (which do ionize the medium) or else are conspicuous in their invisibility!

[12] One example is old-fashioned "mothballs" — if you take a handful of mothballs into a very dark closet (you must get rid of *all* ambient light!) and wait for your eyes to adjust, you should be able to see tiny flashes of light every few seconds as cosmic ray muons zap the mothballs. There are many apocryphal stories about graduate students in closets with mothballs and manual counters in the early days of nuclear physics. . . .

Clouds, Bubbles and Wires

Although one can build arrays of scintillation counters that act like "pixels" in computer graphics and can tell where particles go within an uncertainty of the size of the individual counters, this is very expensive and not usually very precise. Moreover, it was not how the business of "tracking" elementary particles got started.

The earliest "position-sensitive detectors" took advantage of the tendency of liquid droplets to form (or "nucleate") on *ions* when a gas (like air) is "supersaturated" with a vapour (like water or alcohol) that would like to precipitate but can't quite make up its mind where to start. The result, once the process is finished, is a *cloud* of liquid droplets, hence the name "CLOUD CHAMBER." But this final state is not very useful. It is the situation *just after a fast ionizing particle passes through* the saturated gas that is interesting — the left-behind *ions* nucleate a trail of liquid droplets like a string of beads, and one can see (and/or take a picture of) that trail at that moment, to "see the track" of the particle. If it is passing through a magnetic field, the *curvature* of the track reveals its *momentum* and the *density* of the track reveals its *charge* and its *speed*, from which one learns its mass and just about everything about it that can be measured directly. This device was used for many of the early cosmic ray experiments.

The trouble with CLOUD CHAMBERS is that they don't have very fine resolution and the droplets start *falling* as soon as they form. Moreover, even a saturated gas has a rather low density, so if one is looking for interactions of a beam particle with other nuclei the events are spread out over too large a volume to photograph efficiently. Another method still used today is to place a stack of PHOTOGRAPHIC EMULSIONS in the path of the beam and to examine the resulting tracks of silver particles created by the ionizing particle. The problem with this technique is that the emulsion is not reusable — one "takes an exposure" and then the emulsions must be dissected and painstakingly examined with a microscope. Too much work. What was really needed was a sort of "high density cloud chamber" that "healed" soon after each track had been photographed.

The apocryphal story is that a HEP experimenter sat staring glumly into his beer glass one night after wishing for such a device, and noticed that the bubbles always seemed to form in the same places. He sprinkled in a few grains of salt and, sure enough, the bubbles formed on the salt grains. "Eureka!" he cried, leaping up, "the bubbles form on *ions*!" And off he went to build the first bubble chamber.[13]

The idea of the BUBBLE CHAMBER is that a liquid (usually liquid hydrogen) can be abruptly *decompressed*, causing it to "want" to boil, but (like the supersaturated vapour) it can't make up its mind where to start first.[14] If the decompression is done just as ionizing particles pass through the liquid, the ions in their tracks will nucleate the first bubbles of vapour and a clear, sharp track can be seen and photographed; then the liquid is quickly recompressed, the bubbles go away, and the chamber is ready for another "event."

Such liquid hydrogen BUBBLE CHAMBERS are still in use today, but they had their heyday back in the 1950's and 1960's when higher energy accelerators introduced Particle Physicists to the "Hadron

[13] Probably this was a bar frequented by many HEP types, so such behaviour went unremarked.

[14] If you have access to a microwave oven, you can observe this effect for yourself: take a cup of cold water and slowly increase the cooking time (replacing it with new cold water each time) until it is just starting to boil as the timer runs out. Then do one more with a slightly decreased cooking time, take out the cup and drop in a few grains of sugar or salt — the dissolved gases will abruptly come out of solution around these "nucleation centres" to make a stream of bubbles for a short time.

Zoo" of strongly-interacting particles. The most gratifying aspect of a bubble chamber picture is that you can make a big copy of it and put it on your wall, where anyone can point to the different tracks and say, "There goes a pion," or, "This short gap here is a Lambda." The picture appeals to the all-important Visual Cortex, leading to such familiar phrases as, "Seeing is believing," and, "A picture is worth a thousand words." [I won't attack these comforting myths this time; I like bubble chamber pictures too!]

The trouble came when experimenters set out to *measure* the curvatures and densities of millions of tracks in bubble chamber pictures. This involves more than just patience; in the 1960's an army of "scanners" was hired by the big HEP labs to filter hundreds of thousands of bubble chamber pictures looking for certain topological configurations of tracks that were of interest to the experimenter; a lexicon of "vees" and "three-prongs" was built up and eventually these people could recognize events containing different types of elementary particles more efficiently than any Physicist — for, almost without exception, the scanners were nonscientists selected for their rare talents of patience and pattern recognition. It was a fascinating sociological phenomenon, but it cost enormous sums for the salaries of these people and Physicists would always rather buy fancy equipment than create mere jobs. So, as electronics and computers grew in power and shrank in price, it was inevitable that the experiments pressing the limits of HEP technology would seek an "electronic bubble chamber" that could be read out, analyzed and tabulated all by computers.

The result was the WIRE CHAMBER, which again uses the ionization caused by charged particles but this time detects the ion's charges directly with sensitive electronics. There are many versions of this technology, but almost all involve thousands of tiny *wires* strung through a target volume at extremely precise positions and maintained at high voltage so that any ions formed will drift toward one or more of the wires and form a pulse that can be read out at the ends of the wire and interpreted. Such devices can "track" particles through huge volumes to a fraction of a mm and can analyze hundreds or even thousands of events per second, with one "event" containing dozens or even hundreds of particle tracks.

Today's large HEP experiments all involve scintillation counters, wire chamber arrays and other components, each especially sensitive to one or another type of particles, and require on-line computers that must be built specially to handle the enormous flow of information;[15] an ubiquitous feature of *really* high energy particle physics is that there are enormous numbers of particles in the "final state" after two extremely high energy projectiles collide head-on. It is easy to see why this is: the more energy you have, the more mass you can create. It also follows that the heavier the particle, the more ways it has to decay, so the heaviest particles should have the shortest lifetimes. When this rule is *not* obeyed, we have cause to get suspicious.

29.2 Why Do They Live So Long?

If a heavy particle is free to decay into lighter particles, then why isn't the universe filled with *only* the lightest particles? Why, for instance, doesn't an electron (mass 0.511 MeV/c^2) decay into photons (zero mass), with the excess mass appearing as kinetic energy? Well, to begin with, the

[15] For decades, HEP has "driven" the leading edge of supercomputer hardware and software development. Today's computing environment is rapidly becoming more driven by the personal workstation, which is probably a more healthy arrangement, but it is certainly true that we would not have the computer technology we do without the demand created by HEP from about 1950 to about 1980.

electron has "spin $\frac{1}{2}$" (*i.e.*, an intrinsic angular momentum of $\frac{1}{2}\hbar$), while a photon has "spin 1" (*i.e.*, 1 $\hbar$). There is no way to combine several spin 1 objects to make a spin $\frac{1}{2}$ object, so ANGULAR MOMENTUM CONSERVATION forbids an electron to decay into photons. What else? Well, the electron is *charged*, and the photons aren't! So what? Well, electric charge $\mathcal{Q}$ is a CONSERVED QUANTITY — not only is the total amount of charge in the universe constant, but the net charge *in any reaction* must also remain unchanged *at every step*.

OK, the electron is stable. But why can't the *proton* decay into a *positron* (the antiparticle of an electron), which has the same charge and the same spin as the proton? It could also give off two photons with opposite spins, satisfying all the criteria mentioned so far. Well, protons must have some special property that we will call BARYON NUMBER because only *heavy* particles like the proton have it. So far as we know, baryon number manifests itself *only* as a CONSERVED QUANTITY in the interactions of elementary particles. We define the baryon number of a proton to be 1 and that of electrons and photons to be zero. Baryon number is conserved just like electric charge, and this accounts for the stability of protons: the proton is the lightest baryon, so there is nothing for it to decay into!

The next lightest baryon is the *neutron*, and it does indeed decay (slowly) into a proton, an electron (to compensate for the charge of the proton) and an electron antineutrino to compensate for the electron number.[16] Huh? What's "ELECTRON NUMBER?" It's yet another CONSERVED QUANTITY that the *weak interaction* governing neutron decay has to keep account of. We know it exists only because neutrons *don't* decay into just a proton and an electron. The electron NEUTRINO is a sort of chargeless, *massless* version of an electron that has almost no interaction with matter at all — a typical neutrino can pass through the Earth (and a lot more planets besides!) without much chance of touching anything!

How about MUONS? Everyone says these are "sort of like heavy electrons," so why can't a muon decay into an electron and a photon?[17] The muon *does* decay into an electron plus an electron antineutrino and a muon neutrino, but *not* into an electron and a photon. This is because the muon has another different CONSERVED QUANTITY called — you guessed it – MUON NUMBER which is a different FLAVOUR[18] from ELECTRON NUMBER. Naturally, the muon NEUTRINO has muon number too, and is therefore unmistakable for an *electron* neutrino. But only because it never appears where an electron neutrino might.

Is all this perfectly clear? No? I don't blame you. Just remember, whenever a particle refuses to decay into lighter particles for no apparent reason, it is presumed to be because of some new CONSERVED QUANTITY that one has and the others don't. The assignment of names to these ephemeral quantities which Nature seems to hold in such reverence is pretty much arbitrary, so their "discoverers" get to think up names they think are mnemonic, allusive or just cute. There are some examples that are a little embarrassing.

For instance, while discovering hordes of new short-lived heavy particles in the 1950's, people ran across a heavy, spinless, uncharged particle called the neutral KAON which decayed (as expected)

[16] It just barely makes it, mass-wise, which partly accounts for the slowness of the decay.

[17] As a matter of fact, this is still an open question — experiments have recently pushed the upper limit on the "branching ratio" for $\mu \to e\gamma$ (*i.e.*, the fraction of the time muons decay into electrons and photons) to less than one part in 10^{11} and more experiments are underway, because several theories demand that such "flavour-violating" decays must exist at some level.

[18] No, I'm not kidding, the official name for the difference between muons and electrons (and, later on, TAU leptons) is "lepton FLAVOUR."

into lighter PIONS but *very slowly*, suggesting that kaons must have some new property which the STRONG interaction (that should make kaons decay *very rapidly* into pions) could not "violate" but the WEAK interaction could. This new quantity, conserved in strong interactions but not necessarily in weak interactions, was called "STRANGENESS" for reasons that were obvious but hopelessly parochial. I hate this one, because it takes over a perfectly good English word that one might want to use in the same sentence.[19]

It gets worse. But I have introduced far too many new particles and mentioned far too many jargoney names without explaining what they are supposed to mean; I will come back to the literary tastes of Particle Physicists after I have outlined some of the currently used classification schemes.

29.3 Particle Taxonomy

The most efficient classification scheme is a succession of *orthogonal binary dichotomies* in which (if possible) roughly half the items to be classified go on each side of every successive distinction. These may be drawn as "Venn diagrams" in which a circle (representing everything) has a line drawn through the middle.

The first distinction does not even come close to splitting up all the "elementary" particles into two equal groups, but at least it is *unequivocal*. This is the question of whether the particle is STRONGLY INTERACTING or not. If it is affected by the strong interaction, it is called a HADRON. If *not*, it is called a LEPTON. [Both of these have Greek roots. Look them up if you're curious.]

29.3.1 Leptons

The LEPTONS make a short list and are easy to classify by the three known "FLAVOURS" — e, μ and τ. Each type experiences GRAVITY, the ELECTROWEAK interaction and apparently nothing else.

29.3.2 Hadrons

The remaining strongly-interacting HADRONS make a huge "zoo" of mostly short-lived particles of almost every shape and size. However, these too can be separated cleanly into two dichotomous categories: the half-integer spin BARYONS (so named because they tend to be more *heavy*), which are all FERMIONS — *i.e.*, each type obeys its own version of the PAULI EXCLUSION PRINCIPLE — and the zero or integer spin MESONS (so named because they tend to me *medium* heavy), which are all BOSONS — *i.e.*, you can put as many as you like in the same state at the same time. We now know lots of interesting things about the BARYONS and MESONS, but the modern *definitions* of these classes of hadrons are in terms of their *spins*.

Integer spin hadrons are bosons and are all called *mesons*; Half-integer spin hadrons are fermions; those which are not *quarks* are called *baryons*. All baryons have a "baryon number" $\mathcal{B} = 1$; mesons have none. The "hypercharge" $\mathcal{Y}$ of a particle is the sum of its baryon number and its strangeness:

[19] Kirk: "Boy, this particle sure looks *strange*." Spock: "Not at all, Captain. If you look more closely, I believe you'll find it's *charmed*."

Table 29.3 The LEPTONS (particles with only weak and sometimes electromagnetic interactions). All leptons have spin $\mathcal{J} = \frac{1}{2}\hbar$ and are therefore *fermions*. Each "generation" of lepton has its own distinctive "flavour" (electron, muon, tau) and is governed by its own conserved "lepton number." For each particle there corresponds an *antiparticle* of the same mass and spin but with opposite values of electric charge and lepton number of the corresponding flavour.

PARTICLE(s)		**Mass** (MeV/c^2)	**Charge** $\mathcal{Q}/e$	**Lifetime** (s)	**Principle Decay Modes**
electron	e	0.511	-1	$> 6 \times 10^{29}$	none
e neutrino	ν_e	$< 1.7 \times 10^{-5}$	0	∞	none
muon	μ	105.66	-1	2.2×10^{-6}	$\mu^- \to e^- + \bar{\nu}_e + \nu_\mu$ $\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$
μ neutrino	ν_μ	< 0.27	0	∞	none
tau	τ	1784	-1	3.03×10^{-13}	$\tau^- \to (\mu, e)^- + \bar{\nu}_{(\mu,e)} + \nu_\tau$ $\tau^- \to (\text{hadron})^- + (\text{neutrals}) + \nu_\tau$
τ neutrino	ν_τ	< 35	0	∞	none

$\mathcal{Y} = \mathcal{B} + \mathcal{S}$. Quarks all have $\mathcal{B} = \frac{1}{3}$ as well as fractional electric charge because it takes 3 to make one baryon; otherwise they follow the same rules. For each particle (including quarks) there corresponds an *antiparticle* of the same mass, spin, parity and isospin, but with opposite values of electric charge, strangeness, baryon number and hypercharge.

Generally speaking, all the heavy hadrons are *very short-lived* because the interaction governing their decay into lighter hadrons is, after all, *strong*. I have already mentioned a notable exception to this rule, namely the *strange mesons*, which take far longer than they should to decay into pions. In the 1950's this led to the coining of the term STRANGENESS to describe that strange (grrr...) property of K mesons (for instance) that could not be "swept under the rug" by the strong interaction. By checking to see what other particles *could* decay into kaons, and in the company of what else, a STRANGENESS was assigned to each of the hadrons. Then a strange [Oops! Can't use that!] — an odd [Ouch! That implies a PARITY quantum number] — a peculiar [Whew!] pattern began to manifest itself when the particles were grouped together according to the known *quantifiable properties* of SPIN, CHARGE, STRANGENESS and MASS.

The various hadrons are first separated into collections that all have the same *spin*, such as the SCALAR [zero spin] MESONS or the VECTOR [spin 1] MESONS or the SPIN-$\frac{1}{2}$ BARYONS or the SPIN-$\frac{3}{2}$ BARYONS. It is immediately evident that the *masses* of all the particles in any one of these groups are roughly similar, whereas two different groups tend to have significantly different masses. This arouses some suspicion. Then we notice that, within these groups, the particles with the most *strangeness* tend to be the *heaviest*.

Next we notice that if we *plot* the particles in a group on a graph of the two other quantifiable properties — charge $\mathcal{Q}$ and strangeness $\mathcal{S}$ — they form arrangements that are remarkably *similar* in shape![20] The hexagonal arrangement with two particles at the centre appears in each of the

[20]The shapes are a little crooked in this representation. The HYPERCHARGE $\mathcal{Y}$ and ISOSPIN $\mathcal{I}$ (whose "projection" $\mathcal{I}_3$ along

Table 29.4 Some of the HADRONS (strongly interacting particles).

Name		Mass ($\mathrm{MeV/c^2}$)	Lifetime (s)	Spin $\mathcal{J}^{\mathcal{P}}$ [$\hbar$]	Charge $\mathcal{Q}/e$	Isospin $\mathcal{I}$	Strangeness $\mathcal{S}$
QUARKS:							
"up"	u	411?	∞?	$\frac{1}{2}$	$+\frac{2}{3}$	$\frac{1}{2}$	0
"down"	d	411?	∞?	$\frac{1}{2}$	$-\frac{1}{3}$	$\frac{1}{2}$	0
"strange"	s	558?	∞?	$\frac{1}{2}$	$-\frac{1}{3}$	0	-1
MESONS:							
pion	π	139	$\pi^{\pm}:\ 2.6\times10^{-8}$ $\pi^0:\ 8.3\times10^{-17}$	0^-	$-1, 0, +1$	1	0
kaon	K	495	$K^{\pm}:\ 1.2\times10^{-8}$ $K^0:$ ambiguous	0^-	$-1, 0, +1$	$\frac{1}{2}$	$K^0, K^+:\ +1$ $\bar{K}^0, K^-:\ -1$
eta	η	549	8.9×10^{-15}	0^-	0	0	0
rho	ρ	770	4.3×10^{-24}	1^-	$-1, 0, +1$	1	0
omega	ω	783	6.58×10^{-23}	1^-	0	0	0
phi	ϕ	1020	1.6×10^{-22}	0^-	0	0	0
	K^*	892	1.33×10^{-23}	1^-	$-1, 0, +1$	$\frac{1}{2}$	$K^{*0}, K^{*+}:\ +1$ $\bar{K}^{*0}, K^{*-}:\ -1$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
BARYONS:							
nucleon	N	938	proton $(p):\infty$ neutron $(n):920$	$\frac{1}{2}^+$	$0, +1$	$\frac{1}{2}$	0
lambda	Λ	1116	2.6×10^{-10}	$\frac{1}{2}^+$	0	0	-1
sigma	Σ	1190	$\Sigma^{\pm}:\approx10^{-10}$ $\Sigma^0:<10^{-14}$	$\frac{1}{2}^+$	$-1, 0, +1$	1	-1
cascade	Ξ	1320	$\approx 2\times10^{-10}$	$\frac{1}{2}^+$	$-1, 0$	$\frac{1}{2}$	-2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
delta	Δ	1232	5×10^{-24}	$\frac{3}{2}^+$	$-1, 0, +1, +2$	$\frac{3}{2}$	0
	Σ^*	1383	1.6×10^{-23}	$\frac{3}{2}^+$	$-1, 0, +1$	1	-1
	Ξ^*	1530	6.6×10^{-23}	$\frac{3}{2}^+$	$-1, 0$	$\frac{1}{2}$	-2
Omega	Ω	1672	1.3×10^{-10}	$\frac{3}{2}^+$	-1	0	-3
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

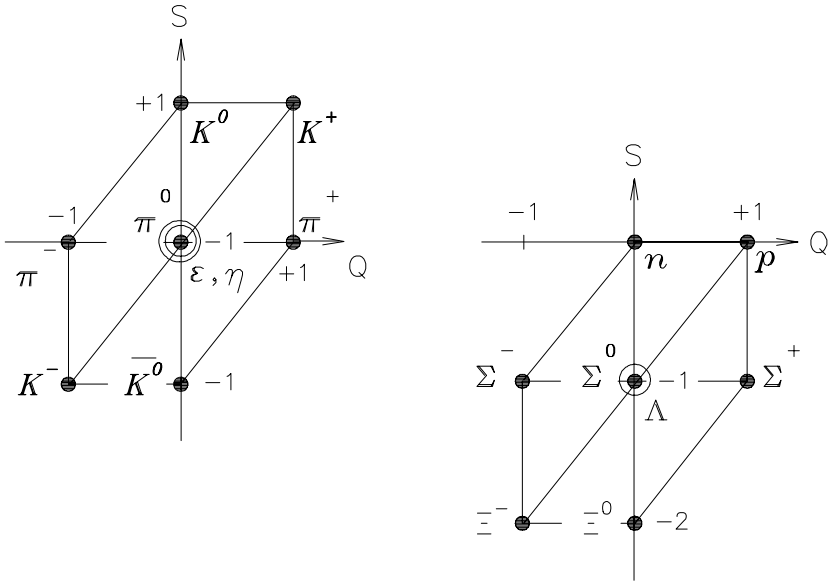


Figure 29.5 Murray Gell-Mann's "Eightfold Way." Left: the SCALAR MESONS. Right: the SPIN-$\frac{1}{2}$ BARYONS. Note the striking similarity of the grouping when strangeness $\mathcal{S}$ is plotted against charge $\mathcal{Q}$. The VECTOR (spin 1) MESONS form a group exactly like the SCALAR MESONS on the left, further reinforcing the pattern.

---

first three groupings listed above; Murray Gell-Mann decided that THIS MUST MEAN SOMETHING about the *constituents* of these particles, just as the regular groupings of elements in the PERIODIC TABLE meant something about the constituents of *atoms*. Because of the number of particles in the pattern, because of his eclectic intellect and because he wanted to make up a catchy name for his theory that people would want to talk about just to sound savvy, Murray named this pattern the EIGHTFOLD WAY after the spiritual/behavioural prescription in Buddhism. More cuteness.

29.3.3 Quarks

Fair enough, obviously these *symmetries* were trying to tell us something about the composition of hadrons. What? Well, needless to say, Gell-Mann *et al.* did not immediately come up with a simple nuts-and-bolts assembly manual; instead, they developed an abstract mathematical description called $SU(3)$ analogous to the description of *spin* for electrons, $SU(2)$. [If you're interested, the acronym stands for *S*imple *U*nitary group of order 2 or 3.] I won't attempt to elaborate, but you can see why something like this was needed — as for the $\hat{z}$ component of spin, the projections of the three $SU(3)$ operators along God-only-knows what axes in God-only-knows what dimensions[21] cannot have a continuum of possible values but only a fixed number of discrete or *quantized* values.

God-only-knows what axis is the same as its charge $\mathcal{Q}$, within a constant) were invented partly to make the diagrams of $\mathcal{Y}$ *vs.* $\mathcal{I}_3$ nicely symmetric with the origin at the centre of each arrangement. I haven't bothered.

[21] Honest, we don't have the faintest idea whether there is actually some *space* in which ISOSPIN actually refers to *rotations* about some *axis*, we only know that isospin *transforms that way*. If there is such a space, none of its dimensions are our familiar x, y or z directions. Very weird.

What is *actually refers to* is totally unknown. Or, more properly, it refers to just what it says; if that means nothing to us, well, that's just because our empirical personal experience of the space of $SU(3)$ is so limited that we don't relate to it very well. What do "normal" space and time *actually refer to*?

Anyway, someone inevitably formulated a simpler instruction manual for assembling hadrons. This was to give the requisite properties to three (there are more now, but hold off on that) *really* fundamental *component* particles called "QUARKS."[22] All MESONS are composed of a *quark-antiquark pair* whereas BARYONS are composed of *three quarks* held together by a "SUPERSTRONG" force mediated by a new type of intermediary called "GLUONS" (g) [more cuteness, but who can argue...].

Table 29.5 The known (or suspected) "generations" of QUARKS All quarks have a "baryon number" $\mathcal{B} = \frac{1}{3}$ as well as fractional electric charge because it takes 3 to make one baryon. The "hypercharge" $\mathcal{Y}$ of any particle is the sum of its baryon number and its strangeness: $\mathcal{Y} = \mathcal{B} + \mathcal{S}$. For each quark there corresponds an *antiquark* of the same mass, spin, parity and isospin, but with opposite values of electric charge, strangeness, baryon number and hypercharge.

Name		**Mass** (MeV/c^2)	**Lifetime** (s)	**Spin** $\mathcal{J}^{\mathcal{P}}$ [$\hbar$]	**Charge** $\mathcal{Q}/e$	**Isospin** $\mathcal{I}$	**Strangeness** $\mathcal{S}$
"up"	u	411?	∞?	$\frac{1}{2}$	$+\frac{2}{3}$	$\frac{1}{2}$	0
"down"	d	411?	∞?	$\frac{1}{2}$	$-\frac{1}{3}$	$\frac{1}{2}$	0
"strange"	s	558?	∞?	$\frac{1}{2}$	$-\frac{1}{3}$	0	-1
"charm"	c	≥ 1500?	∞?	$\frac{1}{2}$	$+\frac{2}{3}$	0	0
"bottom"	b	?	∞?	$\frac{1}{2}$	$-\frac{1}{3}$	0	0
"top"	t	?	∞?	$\frac{1}{2}$	$+\frac{2}{3}$	0	0
$c\bar{c}$	J/ψ	3100	0.97×10^{-20}	1^-	0	0	0
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Colour

The three original quarks are "up" (u), "down" (d) and of course "strange" (s). Each is a spin-$\frac{1}{2}$ FERMION but it took some time to understand how three similar quarks could coexist in the same state within a baryon. (The extension of the PAULI EXCLUSION PRINCIPLE forbids this.) The resolution of this dilemma was to propose (and later believe) that *each* quark comes in three different complementary "COLOURS" (call them red, green and blue) that have to be combined to make the composite particle (meson or baryon) *colourless* (white) just the way the three colours on a TV monitor must all be lit up at once to produce a white "pixel." Of course, we have no

[22] See James Joyce's *Finnegan's Wake* for the origin of the term "quark" — it was originally a nonsense syllable, which makes it a pretty good choice for its present application. At least the commandeering of the word "quark" by Particle Physics did not inconvenience any users of the English language.

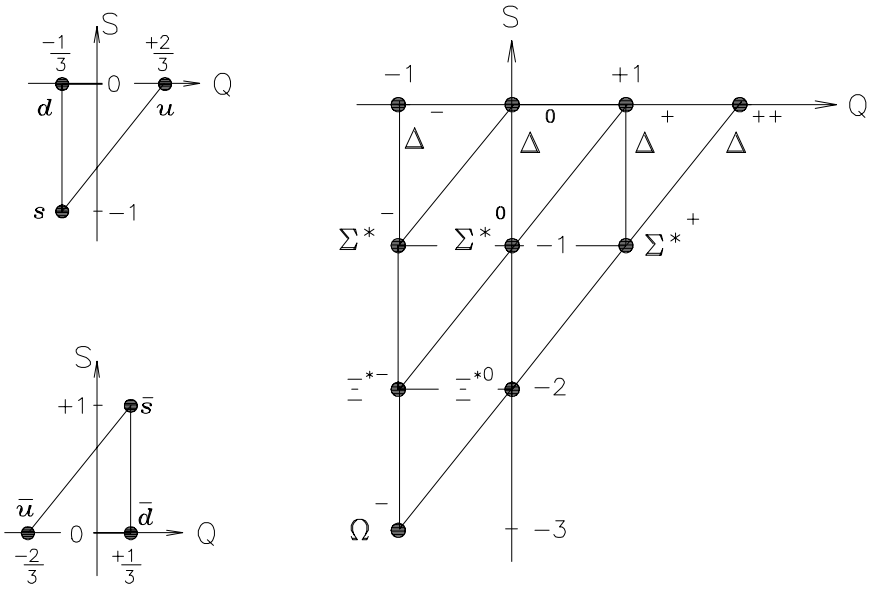


Figure 29.6 Upper left: the three lowest-mass QUARKS. Lower left: the corresponding ANTIQUARKS. Right: the SPIN-$\frac{3}{2}$ BARYONS. The Ω^- (strangeness -3) was predicted by a "quark content" analysis and later found experimentally, convincing everyone that the $SU(3)$ model was correct.

---

idea what COLOUR *is* — it certainly has nothing whatsoever to do with the wavelengths of visible light! — but by now you should be comfortably disconnected from the world of empirical personal experience, so the fact that the metaphor of colour gives us a handy way of getting right answers should suffice.

Using this *quark model* with *gluon exchange* [gluons are *colour changers*, they convert a quark from one colour to another when emitted or absorbed] in a fashion exactly analogous to $\mathcal{QED}$, theorists are now able to accurately describe much of the structure of hadrons, thereby rescuing PERTURBATION THEORY from the ashes of strong interactions, where it failed miserably.[23] The new theory inevitably became known as Quantum ChromoDynamics (or $\mathcal{QCD}$) by analogy with $\mathcal{QED}$ except with colour (Greek *chromos*) in place of electric charge.

Why Quarks are Hidden

If quarks are "real" particles and not just a cute mnemonic metaphor for some esoteric mathematics,[24] we ought to be able to "see" one in a bubble chamber or other device "watching" a high energy scattering event. Unfortunately, this can never be. The reason is intriguing.

The "SUPERSTRONG" force between quarks is transmitted by the exchange of GLUONS [a nice

[23] Unfortunately, the genuinely new paradigms that were springing up to deal with this crisis (*e.g.* Geoffrey Chew's BOOTSTRAP THEORY, in which each hadron is composed of small amounts of all the others [think about it!]) have been neglected since the development of $\mathcal{QCD}$.

[24] Of course, ENERGY is "just a cute mnemonic metaphor for some esoteric mathematics," if we think back to Classical Mechanics; but we have gotten so used to ENERGY that we don't think of it that way any more, whereas QUARKS are still... well, *weird*.

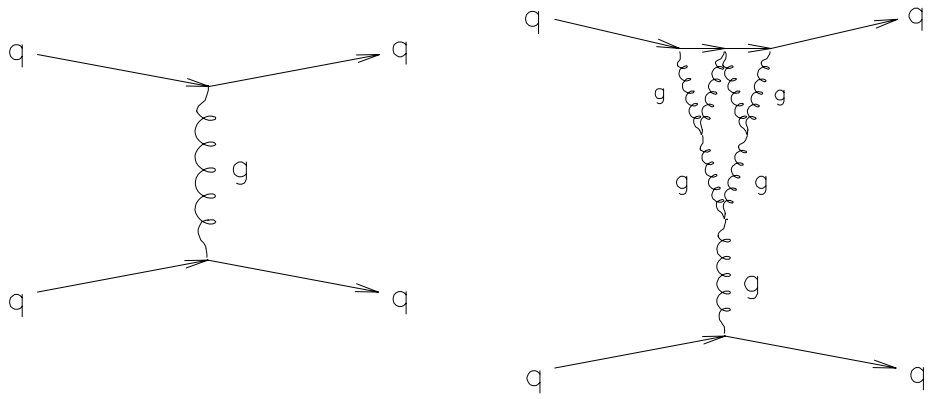


Figure 29.7 Left: $\mathcal{QCD}$ in first order — two quarks exchange a single GLUON at close range. Right: if the two quarks get too far apart, the original gluon gets an chance to *branch* into several gluons, strengthening the attractive force.

---

descriptive name, for once!] which are massless, like photons, but have one trick up their sleeves that photons don't: they can "branch" (one gluon coupling to two gluons, and so on) if given enough room. Thus, while the electromagnetic force drops off as $1/r^2$, the SUPERSTRONG or $\mathcal{QCD}$ force actually *increases* with increasing distance between the two quarks! Once the distance gets big enough — as in a high-energy collision — the energy stored in the gluon field is so intense that quark-antiquark *pairs* are created out of the vacuum between the quarks and the original quark grabs the new antiquark to become a MESON, while the new quark takes the place of the old one in the hadron that has collided.

Thus, try as we might, we can never create a free quark. We can never "see" these ubiquitous particles that make up everything around us except leptons. This is very frustrating and for years led many Particle Physicists to insist that quarks were just figments of theorists' imaginations. But of course the paradigm works too well to be abandoned and the skeptics have by now pretty much given up.

29.4 More Quarks

Elementary Particle Physics seemed to be "converging" at last on a simple description in terms of a manageable number of *really elementary* constituents until around 1964, when some rogues suggested that if there were 6 leptons (counting the neutrinos) then there ought to be 6 quarks too, Nature being endowed with frugality and æsthetics just like Mathematicians. Actually the argument may have been more convincing than that, but I didn't understand it. This might not have raised many eyebrows except that in 1974 two huge groups of Particle Physicists led by Burton Richter and Samuel Ting *simultaneously* (or so close that no one could claim the other had stolen the idea) discovered a new meson that was both very heavy (3100 MeV/c^2) and extremely stable (0.97×10^{-20} s). [Well, for a particle that *heavy*, 10^{-20} s is a *long* time!] This particle, which has the unique disadvantage of *two names* — J and ψ — because of the unusual circumstances of its discovery and the enormous egos required for undertaking and directing such huge experiments, was

immediately recognized to be the manifestation of a new kind of quark, the c quark, which had yet another weird property conserved by strong interactions. In an unsuccessful attempt to compensate for the callousness with which useful words had been ripped off from the English language in the past, the new property was named (groan) "CHARM."

Now there is a whole new *menagerie* of CHARMED particles to complicate matters; and (skipping ahead to today) another[25] of the predicted 6 quarks has been found as well. It is the b quark, and what the "b" stands for makes an interesting story.

The final[26] two quarks were originally posited to manifest two additional conserved properties called TRUTH (t) and BEAUTY (b). This, however, was too much even for the Particle Physics community. Whether we were finally exercising some restraint or had merely become embarrassed by newspaper headlines reading, "CERN Physicists hunt for Naked Bottom," or "Still no Truth in Quark Hunts," shall never be known. It was, however, decided to retroactively change the names of the new quarks (and their corresponding properties) to "TOP" and "BOTTOM" — which, you will note, have the same first letter as the old names, so that the old publications written by Particle Physicists who forbear to use the full names were still valid.

Now, personally, I think this was a mistake. No one is fooled by this attempt to pretend Particle Physicists are not crazy megalomaniacs, and now we have to try to remember the difference between UP-DOWN and TOP-BOTTOM. Perhaps newly discovered particles should be submitted to a panel of English scholars for naming, but this would take some of the fun out of Particle Physics, and if it isn't fun then what is there to keep it going? Hmmm....

29.5 Where Will It End?

Many people have been quick to point out that things don't ever seem to get any better. First we had the elements to explain, then nuclei; there was a pleasant time when the world consisted only of photons, electrons, neutrinos, protons, neutrons and pions — but we had to spoil it by looking more closely and making higher energy accelerators. Then the "hadron zoo" collapsed to three quarks and the gluon, and things were looking up again; but now there are *six* quarks (one of which, the t, still hasn't been observed) and as many leptons, and at least 4 different intermediaries.

Is this just another round of simplification followed by more complexity at a deeper level? Possibly. It has been proposed that quarks and leptons may themselves be composite particles, and further that every particle must have a "SUPERSYMMETRIC" (or "SUSY") partner with the opposite sort of *statistics* — for each fermion there must be a supersymmetric boson, and *vice versa*.[27] There is no shortage of new theories, nor is arrogance in short supply — one model called "SUPERSTRINGS" has been touted as a TOE (Theory Of Everything) by the New York Times (which loves to get into

[25] A Fermilab consortium has also announced a "body of evidence" for the sixth and heaviest quark, the t quark. Most Physicists now are of the opinion that they are probably right, but the CERN LHC is still being built largely to make lots of t quarks to confirm its mass and other properties. Darn, I am getting ahead of myself again. Must be those pesky tachyons.

[26] There is now actual experimental evidence that there are *only six* quarks — or at least that any further quarks "generations" are so massive as to have no observable consequences in any experiments we might perform on Earth. If you want to know more about this story, ask a real Particle Physicist!

[27] The SUSY partner of the *photon* is the *photino*, the SUSY partner of the *graviton* is the *gravitino*, the SUSY partner of the $W^{\pm}$ *boson* is (I am not making this up!) the *wino*, and so on. This is not a joke, but no one knows if it is "real" either. That is, we do not yet know if Nature contains phenomena for which there is no other known explanation.

these debates).[28] There is, however, a small *practical* problem.

All the Grand Unification Theories (or GUTs) predict wonderful simplifications at enormously high energies on the scale of the first moments of the Big Bang — Cosmologists work closely with Particle Physicists these days — but such energies cannot be achieved on Earth. Gigantic accelerators, like the LHC at CERN (in Switzerland and France[29] or the ill-fated SSC (Superconducting Super-Collider) in the USA, cost billions of dollars and take up thousands of square kilometers of space. Particle Physicists hope they will find the next "round" of new structure at these energies, but there are plausible theories that predict the next "interesting" break will come at stupendous energies far beyond those feasible on Earth.[30] If this is true, experimental Particle Physics may not end forever [we may one day build a synchrotron in orbit about the Sun] but the present socioeconomic structures will not be able to support further pushes toward higher energy. Particle Physics will then be forced to go back and take longer, harder looks at the particles already observed, and the "*Excelsior!*" school of Particle Physics will be at an end.

Still, it's been a great ride!

[28] My personal opinion is that such extravagant claims miss the point of Physics almost entirely. We know, for example, that the ordinary properties of solids are governed completely by $\mathcal{QED}$, the most perfectly understood physical theory in the history of Humanity, but we are still discovering unexpected qualitative behaviour of solids as we explore the seemingly endless variety of ways that large numbers of simple units (like electrons) can interact collectively with other simple units (like phonons or positive ions). To understand the components out of which things are built is *not* the same as understanding the things! So-called "naïve Reductionism" is alive and well in certain overly arrogant elementary particle Physicists....

[29] The LHC is a *big* accelerator!

[30] Let me tell you about my design for an accelerator in geosynchronous orbit....

Chapter 30

General Relativity & Cosmology

As Elementary Particle Physicists direct their attention "down" toward the indescribably tiny, so Cosmologists turn their gaze "upward" toward the unfathomably huge. Of course, these days both are increasingly likely to be incarnate in the same individual — I'll get to that later. As one who has never looked through a telescope larger than I could carry, I am certain to give short shrift to the magnificent observational science of ASTRONOMY, which provides COSMOLOGY (a theoretical discipline) with all its data. But a summary of the former without good colour plates of star fields and nebulae would be a terrible waste anyway, so I hope I have motivated the curious to go out and read a good Astronomy book on their own. Moreover, I am so ignorant of General Relativity and most of the fine points of Cosmology that I really have no business writing about either. Therefore I must content myself with a justification in terms of my "unique point of view," whereby I excuse the following distortions.

30.1 Astronomy

Having just declared my intention *not* to cover ASTRONOMY, here I start right in with it! Well, I want to make a few abstract generalizations about the subject. The first is a commentary on the idea of an *observational* science in a Quantum Mechanical *millieu*. Until recently, all astronomical observations were made by detecting *light* emitted by distant objects a long time ago. Nowadays Astronomers detect the full range of the electromagnetic spectrum, from long-wavelength radio waves to gamma rays, as well as the odd *neutrino*,[1] but the qualitative picture hasn't changed: a virtual quantum is emitted at a distant source and absorbed here on Earth; by measuring the relative intensity of such quanta arriving from different directions, we get a picture (literally) of the Universe around us. On the one hand, we cannot detect the photons without annihilating them; in this sense the act of MEASUREMENT interferes with the system being measured, as Quantum Mechanics has taught us to expect. On the other hand, it is reasonable to expect that our interference is only with the photons themselves, not with their distant emitters; and in this sense the Astronomer is an awfully good approximation to the classical OBSERVER.

The next philosophical point is that the photons we detect on Earth may have been "in transit" for

[1]The Sudbury Neutrino Observatory (SNO), now under construction in a Canadian mine shaft, will revolutionize this technology; nevertheless, the best one can hope for is some rough estimate of the direction of the source of individual neutrinos. The pesky critters just don't interact much! (Which is why they get here at all!)

millions or even billions of years, depending upon how far away their source was when they were emitted. Thus as we look *outward* to the distant galaxies we are also looking *backward* in time. Sort of. So if we see the same sort of SPECTRUM (including, for instance, the ubiquitous hydrogen atom emission lines) from a star in another galaxy as we do from Sol, it means that the "Laws of Physics" are pretty much the same here and now as they were there and then. This gives a comforting sense of stability and permanence, even if our individual destinies are short and unknown.

In recent decades humans have developed the technical ability to *go and have a closer look* at other bodies in our own Solar System; this is absolutely delightful and has rekindled interest in Astronomy among the people who end up paying for it, better yet! However, it probably will come to be known by a different name (*e.g.* PLANETOLOGY) simply because of the increased scope of the Experimenter's capacity to interfere with the Observed. Ultimately, humans will again set foot on other worlds [as we did back in 1969 and the early 1970's — doesn't anyone remember?!] and carry the Laboratory to the stars where whole new categories of information can be gathered. However, the sheer distance of other stars makes patience a virtue in such plans....

30.1.1 Tricks of the Trade

Since Astronomers began to chart the heavens (probably before recorded history as we know it), they have been learning *tricks* for finding out more about the stars than would seem possible, given their limited experimental tools. I don't know many of these, but I can point out a few of the important ones.

Parallax

When you watch a distant object out of the corner of your eye, you can keep it in view without turning your head even though you walk some distance at the same time, as long as you walk in a straight line. However, if the object is about the same distance away as the length of your walk, you will end up looking over your shoulder if you insist on keeping an eye on it. This is the essence of PARALLAX, the shift of the apparent direction of a source as the observer changes position — which might not seem to be much help to Astronomers, until you realize that *the Earth moves* quite some distance every year in its path about Sol. By carefully measuring the angular shift in a star's position throughout a year, Astronomers can gauge its distance from the Earth out to an impressive range.

Spectroscopy

Meanwhile, looking at the SPECTRUM of light from a star can tell us (a) how *hot* it is [recall the BLACKBODY spectrum] and (b) what sort of *atoms* are in its "chromosphere" [the hot surface that we see]. Finally, the sheer *brightness* of the star (combined with a knowledge of its *distance* and *temperature*) tells us how *big* it is.

30.1.2 Astrophysics

Putting together lots of such information has allowed a large number of stars to be *catalogued*, with the result that certain combinations of brightness and spectral "signatures" can be generally assigned to stars of a given age, size and character even before their distance is known empirically by PARALLAX measurements. In this way a great deal has been learned about STELLAR EVOLUTION and (by inference) about the nuclear reactions in the cores of stars. This is the science of ASTROPHYSICS, which differs from ASTRONOMY in that the latter seeks mainly to *observe* while the former seeks to *explain* the stars.

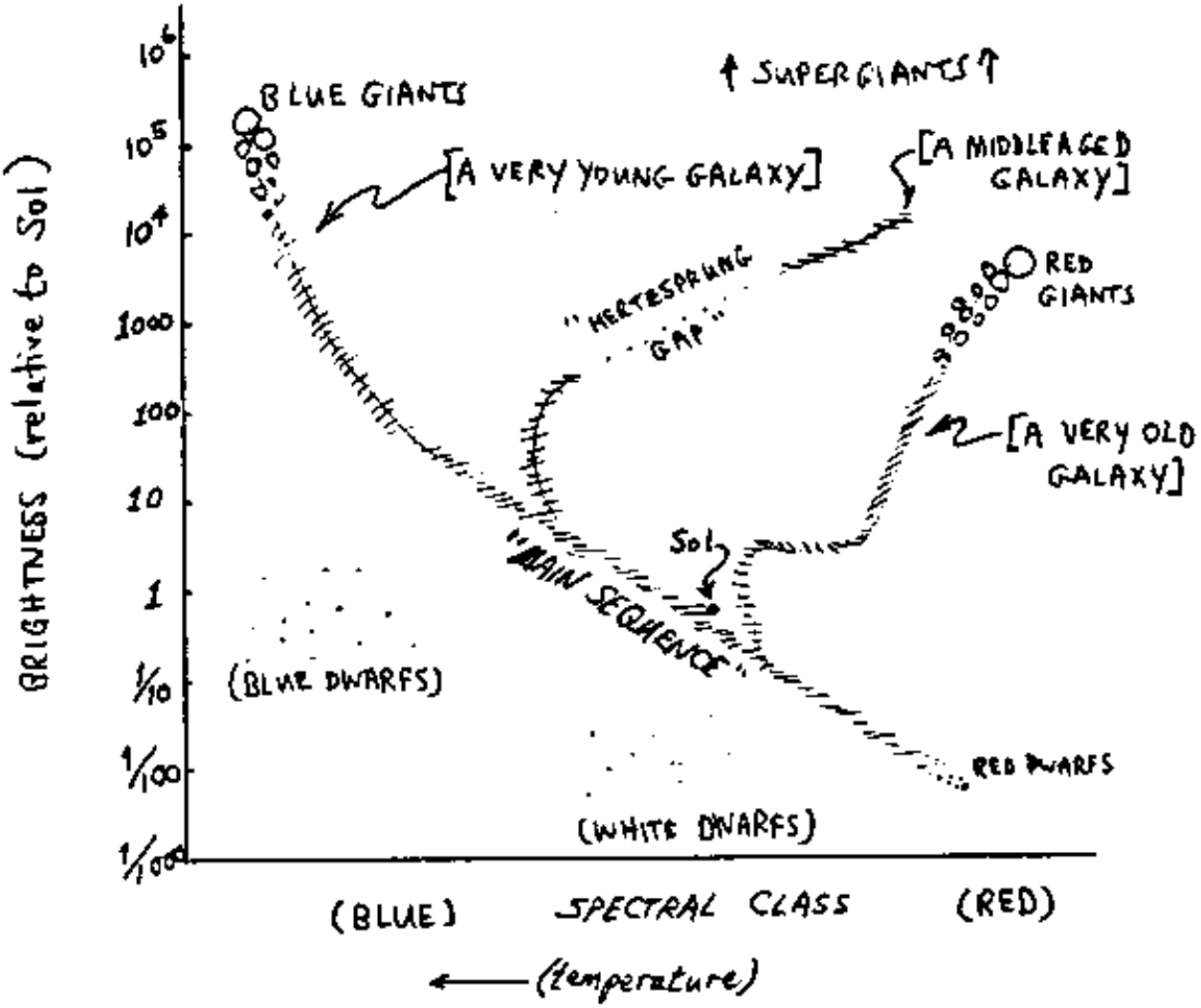


Figure 30.1 A cartoon version of the HERTZSPRUNG-RUSSELL DIAGRAM showing the common categories of stars arranged by their SPECTRAL CLASS (colour) and BRIGHTNESS. Suns are plotted as points or circles. For a given galaxy or star cluster, the distribution of suns on this diagram characterizes the age and evolution of the cluster.

Astrophysical theories of stellar evolution are wondrous detailed, which suggests that I omit further

attempts to describe them here. It is important to note, however, that much of the edifice of COSMOLOGY rests upon the internal consistency and predictive power of these theories.

30.2 Bang!

As spectroscopists began to study more and more distant stars, they noticed a peculiar effect: the Doppler effect for light from distant stars [apparent in the H atom line spectrum, for instance] was not randomly scattered between red and blue shifts, as might be expected for a Universe full of stars "milling about." Instead, Hubble discovered that *the more distant the star, the bigger the* RED SHIFT. That is, all the other stars are, on average, moving *away from us*; and *the more distant the star, the faster it is receding*.

It was a relatively easy matter to estimate from HUBBLE'S CONSTANT how far away a star would have to be in order to be receding from us at the speed of light; the answer was in the neighbourhood of 10-20 billion light years. Since none can be moving any *faster* than the speed of light, this sets a crude limit on the *size of the Universe*.

Moreover, if this has been going on for 10-20 billion years, then all those stars and galaxies are *shrapnel* from an *explosion* 10-20 billion years ago that sent us all flying apart at velocities up to the speed of light. This scenario is known as the BIG BANG model of the origin (and subsequent evolution) of the Universe.

What a picture! In the moment of Creation, all the matter in the Universe was at a single point, after which [to use the refined understatement of Cosmologists] "it began to expand." Initially the energy density was rather high, obviating all our notions about ELEMENTARY PARTICLES, the heaviest of which looks like empty space by comparison. Only after the Universe had expanded and cooled by many, many orders of magnitude was it possible for the particles we know to "freeze out" and begin to go their separate ways.

Modern Cosmologists spend a great deal of their time worrying about the details of the "Early Universe," meaning the period from "$t = 0$" of the BIG BANG until today's elementary particles condensed from the primal fireball. This explains (as promised) why there is often not much separation between COSMOLOGY and ELEMENTARY PARTICLE PHYSICS — basically, the *big* was once *small*.

30.2.1 Crunch?

This raises the question: Will it be small again someday? Is the present trajectory of matter in the Universe an "escape trajectory" so that the Universe will keep on expanding indefinitely, or is there enough mass present to *bind* the Universe — slowing down the "shrapnel" by gravitational attraction until it stops and begins to fall inexorably inwards. . . ?

To the best of my knowledge (which isn't all that impressive), opinion is divided. No one has been able to account for enough mass to keep the Universe "closed" (bound), so that it looks like a "BIG CRUNCH" is not in store for us. On the other hand, a careful analysis of the present distribution of matter in the Universe suggests (or so I am told) that a "wide open" Universe (forever expanding) is not compatible with its present homogeneity. In fact, the theorists would be happiest with a

perfect balance so that the Universe can't quite make up its mind whether it is bound or not! [This would appeal to anyone, but I think they actually have arguments why it must be so.] If this is the case, we must be missing two things: (1) a lot of mass that doesn't interact much and hence is known as DARK MATTER composed of Weakly Interacting Massive Particles or WIMPs; (2) any idea of the mechanism that ensures such incredibly "fine tuning" of the so-called COSMOLOGICAL CONSTANT — if it were infinitesimally *larger*, the Universe would have collapsed back upon itself in a matter of seconds, while a slightly *smaller* value would have us lost in empty space by now.

Am I out of my depth here, or what?

30.3 Cosmology and Special Relativity

So far I have been sweeping the worst confusion under the rug.

First off, when we talk about "the Universe today," we mean "what we see today." This isn't quite fair, since the light we detect from distant objects was emitted a *long* time ago, maybe almost at the beginning of time! We have no way of knowing, even in principle, what those objects have been up to since then. Maybe they are all gone by now.

This creates a problem with *energy conservation*: since every star is in a different inertial reference frame from every other, what is *simultaneous* for one is not for another; in that case, how does one talk about *energy conservation* on a Cosmic scale? *When* do the books get balanced, according to *whose* perspective? I don't know of any resolution for this confusion. Perhaps energy conservation is an obsolete concept on the *large scale*.

30.3.1 I am the Centre of the Universe!

On the other hand, the BIG BANG picture does make it possible to resolve an old conflict between Ergocentric and Heliocentric Cosmologies. All the "bits of shrapnel" were once in the same place and have been flying apart ever since; in the crude approximation that their trajectories are non-interacting (*i.e.* disregarding the little deflections caused by gravitational attractions between neighbours), each one is perfectly justified in regarding itself as *at rest* while the others are all in motion. If I am at rest now, then (in this approximation) I have been at rest all along, and am still at the centre of the Universe where the BIG BANG took place, whereas all you other bits are flying off to infinity.

Even if you insist upon a *geometrical* definition of the "centre of the Universe," I am still at its centre, for what can we possibly mean by the geometrical centre but the point equidistant from all the most rapidly receding bits — namely, photons and other massless particles moving at the speed of light. Since these were all emitted initially from the same point where I was then, and all are moving away in every direction at the same speed (guaranteed by the $\mathcal{STR}$), this is still the centre.

Of course, every other fragment is equally entitled to the same point of view — we are *all* at the centre of the Universe, as viewed in our own reference frame![2]

[2] Once again Physics comes around to the same conclusion that has been reached by Psychology.

30.4 Gravity

Cosmology is intimately involved with gravity, about which we may have a lot of instincts but not much accurate knowledge. Here's where we finally get down to the hard part. The first trick is to understand the only interaction that really matters in today's Universe: gravity. To do it right, of course, we must formulate a *relativistic* theory, since all those distant stars are moving away from us at velocities approaching the speed of light. Enter Albert Einstein, again.

30.4.1 Einstein Again

Encouraged by his successes with Special Relativity and Quantum Mechanics, Albert tackled the thorny problem of General Relativity (the behaviour of Physics in *accelerated reference frames*) with his characteristic *élan*. The first difficulty was in distinguishing between *truly* accelerated frames (like a compartment in a rocket) and frames that only *seem* to be accelerated (like where you are sitting). Consider: if you can't look out the window, how do you *know* you are being pressed into the seat of your chair by the Earth's *gravity*, as opposed to being in a rocket somewhere in interstellar space accelerating "up" at 9.81 m/s^2? Well, yes, you walked into the room from outside and sat down just a short while ago; but suppose you had lost your memory? How can you *tell* (by experiment) which is the case?

The Correspondence Principle

Einstein, following his usual æsthetics of simplicity, assumed the "dilemma" was its own solution — namely, *you can't* tell an accelerated reference frame from a reference frame in a gravitational field. This is known as the correspondence principle:

> No experiment performed in a closed system can tell whether it is in an *accelerated* reference frame or a reference frame in a *gravitational field*.

If you wake up in a closed box and you experience "weight" (as one normally does on Earth), *there is no way* to be sure you are actually being attracted by gravity, as opposed to being in a spaceship (far from any stars or planets) which is accelerating at one "gee." What's more, if Einstein is right, *no matter how clever you are* you will not be able to measure *any* phenomenon from which you can tell the difference. The two cases are *perfectly equivalent*, hence the name of the Principle.[3]

So far this Principle agrees with experiments, which has led people to look for ways to make the statement, "A gravitational field is *the same thing* as an accelerated reference frame," sound reasonable. To make any progress along these lines we have to turn to an analysis of our notion of "acceleration" — *i.e.* of the nature of space and time, and therefore of *geometry*.

30.4.2 What is Straight?

If we want to do geometry, the first thing we need is a *straightedge*. Any straight line will do. What shall we use? Well, modern surveyors are mighty fond of *lasers* for the simple reason that

[3] You could open the door and look out, of course, but that would be cheating; besides, how do you know the view is not just an excellent illusion?

light travels in a straight line. (If light doesn't, what does?!) At least in empty space this must be true. So if we like we can *define* a "straight line" in 3-space (x, y, z) to be the path of a ray of light. We call this path a GEODESIC of space for an important reason that is best explained by analogy [like most topics in Relativity].

Consider *air travel* on Earth. Most intercontinental flights take routes called "great circles" which may go over the North Pole *etc.* This is because these are *the shortest paths between two points* on the Earth, subject to the *constraint* that one must travel essentially in *two dimensions* along the *surface* of the Earth. Such lines, the shortest distances between points subject to the constraint that you must travel along a certain surface, are *in general* called GEODESICS, and now we begin to see the connection.

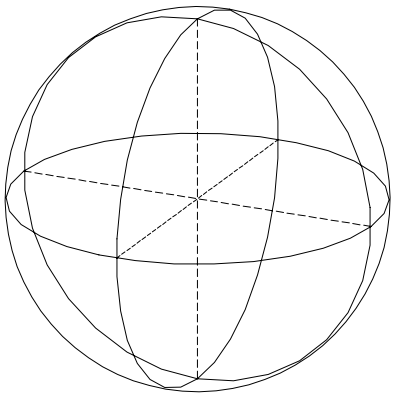


Figure 30.2 "Great circle" routes on the Earth are GEODESICS of the Earth's surface (a 2-D HYPERSURFACE *embedded* in 3-D SPACE); geometrical figures drawn on this HYPERSURFACE do not obey Euclidian geometry!

When we wander around the Earth's surface like "bugs on a balloon," we imagine that [neglecting the odd bump here and there] we live in a 2-D space (North-South and East-West). In fact, we are simply restricted by practical considerations to a 2-D *surface* [within a few miles of altitude] "*embedded*" in a 3-D *space*. The analogous situation can arise for a 3-D HYPERSURFACE embedded in a 4-D *space-time* continuum. Such a hypersurface contains the GEODESICS along which light travels.

30.4.3 Warp Factors

Before we go much further with the hard stuff, let's see if there is any way to *know* whether we are constrained to such a *curved* or "warped" [hyper]surface.

For the "bug on the balloon" there certainly is: simply check whether Euclidean geometry (trigonometry, *etc.*) works properly on figures in the "plane" of the Earth's surface. As an extreme example, note that two "straight lines" which cross at one point on the Earth *will cross again* on the other side! Also note that one can make a "triangle" out of *great circles* in which *all three angles* are 90°! [Just make the length of each side equal to $\frac{1}{4}$ the circumference of the Earth.] And so on.

π as a Parameter

If we like, we can ever be quantitative about the degree of *curvature* of our embedded hypersurface. Picture the following construction: attach a string of length r to a fixed centre and tie a pencil to the other end. Keeping the string tight, draw a circle around the centre with radius r. Now take out a measuring device and run it around the perimeter to measure the circumference of the circle, ℓ. The ratio $\frac{\ell}{2r}$ can be defined to be Π. If the hypersurface to which we are confined is "flat," then Π will be equal to the value we know, $\pi = 3.14159\cdots$; but if we are on a *curved* (or "warped") hypersurface then we will get a "wrong" answer, $\Pi < \pi$.

Minkowski Space and Metrics

Star Trek notwithstanding, this is what is meant by "warped space." Our apparently "flat" (*i.e.* Euclidean) 3-D (x, y, z) universe is *embedded* in a 4-D (t, x, y, z) space called "MINKOWSKI SPACE." *Light* always follows a *geodesic* — the "shortest" distance between two points *constrained* to a given 3-D *hypersurface* — and *we can tell* if this hypersurface is *curved* in a 4-D analogy of the curvature of the Earth's 2-D surface in 3-D, because if it is, *Euclidean geometry will fail.*

⟵ H. Minkowski

This occurs (it turns out) *in any gravitational field.* Hence the terminology that has been popularized by various *SF* authors: "Gravity warps space."

Another way of putting this is to say that the METRIC of Minkowski space changes in a gravitational field. A detailed mathematics of *tensor calculus* has been worked out to describe this effect quantitatively; I don't understand a bit of it, so you will be spared.

30.4.4 Supernovae and Neutron Stars

Despite my ignorance, I can't resist trying to explain what happens in the presence of *really strong* gravitational fields. A typical scenario has a large sun (at least 10 times as big as ours, usually; relax!) cooling off until the gravitational attraction is strong enough to supply the *energy of confinement* necessary to overcome the UNCERTAINTY PRINCIPLE that normally prevents electrons from being confined inside protons. Then the reaction $e^- p \to \nu_e n$ (a sort of inverse neutron beta-decay) begins to convert hydrogen atoms to neutrons, emitting neutrinos as they go. The neutrons further enhance the gravitational energy density until there is a sudden chain reaction producing a SUPERNOVA (the most violent explosion known) that blows off the exterior of the star (which is now rich in heavy elements because of all the neutrons being generated)[4] and leaves behind a NEUTRON STAR — basically a giant atomic nucleus that doesn't fission because *gravity* holds it together.

Neutron stars are generally spinning very rapidly and have enormous magnetic fields "locked in" to their spin, so that the fields sweep up nearby charged particles and turn them into a beacon

[4] If it weren't for supernovae, there wouldn't be any heavy elements floating around the Galaxy to make planets out of and none of us would be here! Think of yourself as a sort of "supernova fossil."

emitting electromagnetic radiation synchronized with the spinning star. Such beacons are "seen" on Earth as regularly pulsing radio sources or "PULSARS," many of which are now known. Most nebulae (the remnants of supernovae) contain neutron stars at their cores.

The phenomenology of neutron stars is itself a huge and fascinating subject about which I know too little. Let's both go look them up and read more about them!

30.4.5 Black Holes

If the neutron star is massive enough, then the gravitational force can grow strong enough even to overcome the hard-core repulsion between quarks and compress the neutrons themselves, making the gravitational force even stronger until *no force can resist* the GRAVITATIONAL COLLAPSE, at which point the entire mass of the star compresses (theoretically) to a single point called the SINGULARITY. We can't tell anything about the singularity for a simple reason: nothing that gets close to it can ever get away again.

The easy, handwaving way to see why is as follows: at any distance from a massive object, any other object will be *in orbit* about it providing it executes circular motion at just the right speed. As you get closer, the *orbital velocity* gets higher. Now, for a sufficiently heavy object, there is some radius at which the nominal orbital velocity is *the speed of light*. From inside that radius, called the SCHWARZSCHILD RADIUS (r_S), *not even light can escape* but is inexorably drawn "down" into the singularity. Thus all light (or anything else!) falling on such an object's Schwarzschild radius will be perfectly absorbed, which accounts for the name, "BLACK HOLE."

We can easily estimate r_S using a crude classical approximation: for a masss m in a circular orbit about a mass M, $F = ma$ gives $GMm/r^2 = mv^2/r$ which reduces to $GM/r = v^2$ or $r = GM/v^2$. If $v \to c$ this becomes $r_S = GM/c^2$. This result is actually off by a factor of 2: the actual Schwarzschild radius is twice as large as predicted by this dumb derivation:

$$\textsf{True} \quad r_S = 2\frac{GM}{c^2}.$$

I have tried to find a simple explanation for this extra factor of two, but failed. Simply using the "effective mass" γm in place of m makes no difference, for instance, because it appears on both sides of the equation the same way. However, I don't feel too bad, because apparently it took Einstein about seven years to get it right. [The time it took him to develop his General Theory of Relativity, which explains that extra factor properly.]

A more rigourous description is beyond me, but I can repeat what I've heard and list some of the phenomenology attributed to BLACK HOLES, of which there are two types: the SCHWATZSCHILD (non-rotating) black hole and the KERR black hole, which *spins*. Presumably all real black holes are of the latter category, since virtually every star has some angular momentum, but there is probably a criterion for how *fast* it must spin to qualify as a Kerr black hole.

Schwarzschild Black Holes

⟵ K. Schwarzschild

One of the interesting features of GENERAL RELATIVITY is that *time slows down* as you approach the Schwarzschild radius of a black hole. Not to *you*, of course; your subjective experience of time is unaffected, but an outside observer would see your clock moving slower and slower (and turning redder and redder) as you fell into the black hole, until (paradoxically) you stopped completely (and were red-shifted out of sight) at r_S. Your own experience would depend upon the *mass* of the black hole. If it were big enough, the trip in free fall through r_S would be rather uneventful — you wouldn't notice much of anything unusual, unless of course you tried to get out again.

If, on the other hand, you approached a *small* black hole, the *tidal forces* [the gravitational *gradient*] would tear you apart before you even reached r_S. This has some interesting consequences which I will discuss later.

The transformation between "outside" and "inside" coordinates has an interesting feature: while it is strictly impossible for anything *inside* r_S to come *out*, one can imagine extending the mathematics of the relativistic transformation from outside to inside, at least formally. The result would be that "inside time" is in the *opposite direction* from "outside time." This would mean that what we see as matter falling inexorably *into* a black hole must "look" to the interior inhabitants (if any) like an *expansion* of matter *away* from the singularity — a sort of BIG BANG. Which raises an interesting question about *our* BIG BANG: are we inside a BLACK HOLE in someone else's Universe? Hmm.... And are the BLACK HOLES in *our* Universe time-reversed BIG BANGs for the inhabitants (if any) of their interiors? Hmmmmmm.... Unfortunately, this sophistry is probably all wrong. If you want a proper, correct and comprehensible description of phenomena at the Schwarzschild radius, go talk to Bill Unruh!

Kerr Black Holes

Well, moving right along, I should repeat what I've heard about KERR (*spinning*) BLACK HOLES. The problem with SCHWARZSCHILD BLACK HOLES is, of course, that exploring them is strictly a one-way trip; once you pass through their Schwarzschild radius, you are doomed to fall right on in to the singularity.

Not so, apparently, with a KERR BLACK HOLE if it is spinning fast enough. Then the singularity is in a *ring* (sort of) and you can in principle plot a trajectory through the *middle* of the ring (or something like that) and *come out the other side*. Except that "the other side" may not have any resemblance to where-when you went in on this side! This has already been used as a great gimmick for *SF* stories involving time travel and other apparent logical paradoxes. I don't understand it at all, and I doubt very much that anyone else does, but one can always postulate that someone will, someday, and use it for practical(?) purposes. After all, as Arthur C. Clarke says, "Any sufficiently advanced technology is indistinguishable from magic."

Wormholes?

⟵ John Archibald Wheeler

Another favourite gimmick of "hard *SF*" authors [those who try to make their stories consistent with the known "Laws of Physics"] is the WORMHOLE, a sort of "space warp" analogous to the BLACK HOLE but topologically more interesting. One can distort [*e.g.* fold] a 2-D surface (like a sheet of paper) embedded in a 3-D space until two apparently distant points are "actually" quite close together in the higher-dimensional continuum. Then a simple *puncture* across both sides will provide a "shrt cut" and drastically change the CONNECTEDNESS [a formal term in the mathematics of TOPOLOGY, believe it or not] of space. In a similar (?) fashion, one can imagine (?) a gravitational anomaly creating a "WORMHOLE" making a "short cut" connection between two nominally distant regions of 3-D space. Great potential for space travel, right?

Sorry. John Archibald Wheeler, who has played a major rôle in the development of all this weird Gravitation stuff, proved a long time ago that *wormholes always pinch off* spontaneously before anything (even a signal propagating at the speed of light!) can get through them. Of course, this fact doesn't stop *Star Trek Deep Space 9* from having a lot of fun with the idea anyway.

Exploding Holes!

Another feature of *small* BLACK HOLES is that they are *unstable*. This was explained in some detail by Bill Unruh in the UBC Physics Department. The basic idea is that for a *small* black hole the TIDAL FORCES at the Schwarzschild radius are so enormous that they can *tear apart the vacuum* — that is, pull one of the partners in a "virtual pair" or "bubble" down into the black hole while the other escapes as radiation. The resultant energy loss is deducted from the *mass* of the black hole, making it still *smaller*. This is a runaway process that ends in a rather impressive explosion. Not to worry, all the small "primordial" black holes (made in the BIG BANG) have by now decayed. On the other hand, a *marginally larger* primordial black hole might have taken until now to get down to a size where the radiation really starts taking off. . . .

Mutability

What CONSERVATION LAWS do BLACK HOLES respect? Not many. Mass-energy, angular momentum and electric charge are the only properties of what falls in that remain properties of the black hole itself. That means that all other "conserved" properties of matter, like BARYON NUMBER, are "MUTABLE" in the final analysis.

One consequence is that *protons* might experience GRAVITATIONAL DECAY in which they collapse

into a *very* tiny black hole, only to immediately explode into (probably) a positron and some gamma rays. The estimated lifetime of protons against such a fate is $\sim 10^{45}$ years, which is not too worrisome.

Other consequences are more interesting, but only philosophically: the *interior* of a black hole [with which we can never communicate] may have entirely different properties — or even different "Laws of Physics" — than what we drop into it. Wheeler has taken this idea much further than I can follow, but it does make for interesting thinking. Good luck.

30.4.6 Gravitational Redshifts and Twisted Time

In addition to the "ordinary" redshifts of distant stars caused by the relativistic Doppler shift due to the fact that they are actually receding from the observer on Earth, there is a GRAVIATIONAL REDSHIFT of the light from *near* a large mass M when observed from a position *far* from the source, even if the source and observer are at rest relative to one another. This is not too surprising if we recall that a gravitational field has to be indistinguishable from an accelerated reference frame, and an accelerated object cannot be at rest for long! But an easier way to see the result is to remember that a massless particle like a PHOTON still has an *effective mass* $m' = E/c^2$ where (if I may borrow a hitherto undemonstrated result from QUANTUM MECHANICS) $E = h\nu$ for a photon. Here ν is the *frequency* of the light and $h = 6.626 \times 10^{-34}$ J-s is PLANCK'S CONSTANT. Anyway, if the energy of a photon far from M is $E_\infty = h\nu_\infty$ (at $r \to \infty$) then its effective mass there is $m'_\infty = h\nu_\infty/c^2$ and as the photon "falls" toward M it should pick up kinetic energy until at a finite distance r its energy is $E = E_\infty + GMm'/r$ where the new effective mass is $m' = E/c^2$. Thus $E = E_\infty + (GM/c^2)E/r$ and if we collect the terms proportional to E we get $E_\infty = E(1 - r_\circ/r)$ where $r_\circ \equiv GM/c^2$. Dividing through by h/c^c gives the formula for the GRAVITATIONAL REDSHIFT,

$$\frac{\nu_\infty}{\nu} = 1 - \frac{r_S}{r} \qquad \text{where} \qquad r_S = 2\frac{GM}{c^2}.$$

(I have fudged in that extra factor of 2 that turns $r_\circ$ into the correct SCHWARZSCHILD RADIUS r_S). This derivation is completely bogus, of course, but it does indicate why there is a gravitational redshift.

Given that any mechanism for generating electromagnetic waves constitutes a "clock" of sorts, the waves emitted by such a device constitute a signal from it telling distant observers about the passage of time at the origin. (Think of each wave crest as a "tick" of the clock.) The very existence of a GRAVITATIONAL REDSHIFT therefore implies that time passes slower for the clock that is closer to the mass — a result that was referred to earlier without proof.