科目：物性機能科学ＩＩ

１～７目

プリント　参考するもの

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

R.ミケレット

1

Quantum Behavior

Note: This chapter is almost exactly the same as Chapter 37 of Volume I.

1–1 Atomic mechanics

"Quantum mechanics" is the description of the behavior of matter and light in all its details and, in particular, of the happenings on an atomic scale. Things on a very small scale behave like nothing that you have any direct experience about. They do not behave like waves, they do not behave like particles, they do not behave like clouds, or billiard balls, or weights on springs, or like anything that you have ever seen.

Newton thought that light was made up of particles, but then it was discovered that it behaves like a wave. Later, however (in the beginning of the twentieth century), it was found that light did indeed sometimes behave like a particle. Historically, the electron, for example, was thought to behave like a particle, and then it was found that in many respects it behaved like a wave. So it really behaves like neither. Now we have given up. We say: "It is like *neither*."

There is one lucky break, however—electrons behave just like light. The quantum behavior of atomic objects (electrons, protons, neutrons, photons, and so on) is the same for all, they are all "particle waves," or whatever you want to call them. So what we learn about the properties of electrons (which we shall use for our examples) will apply also to all "particles," including photons of light.

The gradual accumulation of information about atomic and small-scale behavior during the first quarter of this century, which gave some indications about how small things do behave, produced an increasing confusion which was finally resolved in 1926 and 1927 by Schrödinger, Heisenberg, and Born. They finally obtained a consistent description of the behavior of matter on a small scale. We take up the main features of that description in this chapter.

Because atomic behavior is so unlike ordinary experience, it is very difficult to get used to, and it appears peculiar and mysterious to everyone—both to the novice and to the experienced physicist. Even the experts do not understand it the way they would like to, and it is perfectly reasonable that they should not, because all of direct, human experience and of human intuition applies to large objects. We know how large objects will act, but things on a small scale just do not act that way. So we have to learn about them in a sort of abstract or imaginative fashion and not by connection with our direct experience.

In this chapter we shall tackle immediately the basic element of the mysterious behavior in its most strange form. We choose to examine a phenomenon which is impossible, *absolutely* impossible, to explain in any classical way, and which has in it the heart of quantum mechanics. In reality, it contains the *only* mystery. We cannot make the mystery go away by "explaining" how it works. We will just *tell* you how it works. In telling you how it works we will have told you about the basic peculiarities of all quantum mechanics.

1–2 An experiment with bullets

To try to understand the quantum behavior of electrons, we shall compare and contrast their behavior, in a particular experimental setup, with the more familiar behavior of particles like bullets, and with the behavior of waves like water waves. We consider first the behavior of bullets in the experimental setup shown diagrammatically in Fig. 1–1. We have a machine gun that shoots a stream of bullets. It is not a very good gun, in that it sprays the bullets (randomly) over a fairly large angular spread, as indicated in the figure. In front of the gun we have

a wall (made of armor plate) that has in it two holes just about big enough to let a bullet through. Beyond the wall is a backstop (say a thick wall of wood) which will "absorb" the bullets when they hit it. In front of the wall we have an object which we shall call a "detector" of bullets. It might be a box containing sand. Any bullet that enters the detector will be stopped and accumulated. When we wish, we can empty the box and count the number of bullets that have been caught. The detector can be moved back and forth (in what we will call the x-direction). With this apparatus, we can find out experimentally the answer to the question: "What is the probability that a bullet which passes through the holes in the wall will arrive at the backstop at the distance x from the center?" First, you should realize that we should talk about probability, because we cannot say definitely where any particular bullet will go. A bullet which happens to hit one of the holes may bounce off the edges of the hole, and may end up anywhere at all. By "probability" we mean the chance that the bullet will arrive at the detector, which we can measure by counting the number which arrive at the detector in a certain time and then taking the ratio of this number to the *total* number that hit the backstop during that time. Or, if we assume that the gun always shoots at the same rate during the measurements, the probability we want is just proportional to the number that reach the detector in some standard time interval.

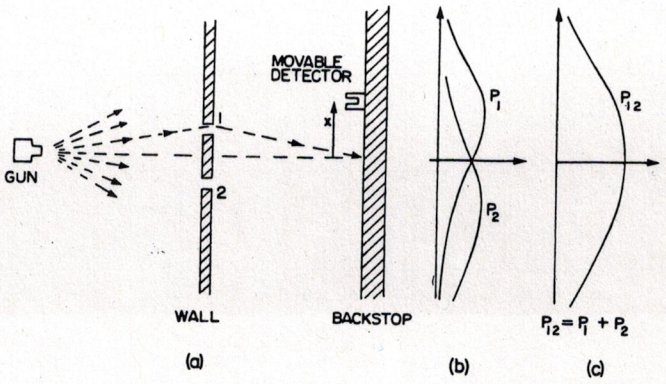

Fig. 1–1. Interference experiment with bullets.

For our present purposes we would like to imagine a somewhat idealized experiment in which the bullets are not real bullets, but are *indestructible* bullets—they cannot break in half. In our experiment we find that bullets always arrive in lumps, and when we find something in the detector, it is always one whole bullet. If the rate at which the machine gun fires is made very low, we find that at any given moment either nothing arrives, or one and only one—exactly one—bullet arrives at the backstop. Also, the size of the lump certainly does not depend on the rate of firing of the gun. We shall say: "Bullets *always* arrive in identical lumps." What we measure with our detector is the probability of arrival of a lump. And we measure the probability as a function of x. The result of such measurements with this apparatus (we have not yet done the experiment, so we are really imagining the result) are plotted in the graph drawn in part (c) of Fig. 1–1. In the graph we plot the probability to the right and x vertically, so that the x-scale fits the diagram of the apparatus. We call the probability P_{12} because the bullets may have come either through hole 1 or through hole 2. You will not be surprised that P_{12} is large near the middle of the graph but gets small if x is very large. You may wonder, however, why P_{12} has its maximum value at $x = 0$. We can understand this fact if we do our experiment again after covering up hole 2, and once more while covering up hole 1. When hole 2 is covered, bullets can pass only through hole 1, and we get the curve marked P_1 in part (b) of the figure. As you would expect, the maximum of P_1 occurs at the value of x which is on a straight line with the gun and hole 1. When hole 1 is closed, we get the symmetric curve P_2 drawn in the figure. P_2 is the probability distribution for bullets that pass through hole 2. Comparing parts (b) and (c) of Fig. 1–1, we find the important result that

$$P_{12} = P_1 + P_2. \tag{1.1}$$

The probabilities just add together. The effect with
the effects with each hole open alone. We shall ca
"*no interference*," for a reason that you will see lat
come in lumps, and their probability of arrival shows no interference.

1–3 An experiment with waves

Now we wish to consider an experiment with water waves. The apparatus is shown diagrammatically in Fig. 1–2. We have a shallow trough of water. A small object labeled the "wave source" is jiggled up and down by a motor and makes circular waves. To the right of the source we have again a wall with two holes, and beyond that is a second wall, which, to keep things simple, is an "absorber," so that there is no reflection of the waves that arrive there. This can be done by building a gradual sand "beach." In front of the beach we place a detector which can be moved back and forth in the x-direction, as before. The detector is now a device which measures the "intensity" of the wave motion. You can imagine a gadget which measures the height of the wave motion, but whose scale is calibrated in proportion to the *square* of the actual height, so that the reading is proportional to the intensity of the wave. Our detector reads, then, in proportion to the *energy* being carried by the wave—or rather, the rate at which energy is carried to the detector.

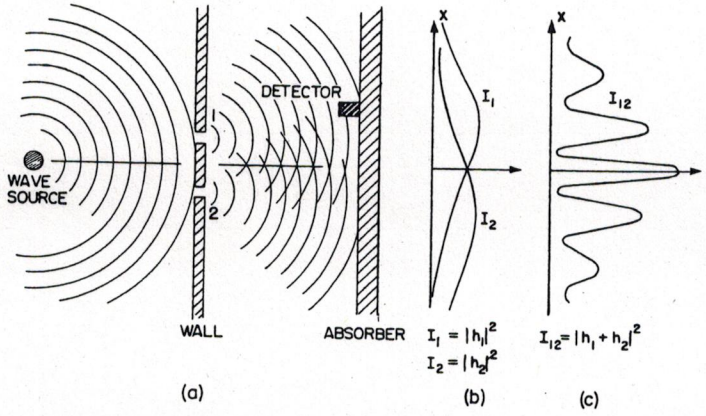

Fig. 1–2. Interference experiment with water waves.

With our wave apparatus, the first thing to notice is that the intensity can have *any* size. If the source just moves a very small amount, then there is just a little bit of wave motion at the detector. When there is more motion at the source, there is more intensity at the detector. The intensity of the wave can have any value at all. We would *not* say that there was any "lumpiness" in the wave intensity.

Now let us measure the wave intensity for various values of x (keeping the wave source operating always in the same way). We get the interesting-looking curve marked I_{12} in part (c) of the figure.

We have already worked out how such patterns can come about when we studied the interference of electric waves in Volume I. In this case we would observe that the original wave is diffracted at the holes, and new circular waves spread out from each hole. If we cover one hole at a time and measure the intensity distribution at the absorber we find the rather simple intensity curves shown in part (b) of the figure. I_1 is the intensity of the wave from hole 1 (which we find by measuring when hole 2 is blocked off) and I_2 is the intensity of the wave from hole 2 (seen when hole 1 is blocked).

The intensity I_{12} observed when both holes are open is certainly *not* the sum of I_1 and I_2. We say that there is "interference" of the two waves. At some places (where the curve I_{12} has its maxima) the waves are "in phase" and the wave peaks add together to give a large amplitude and, therefore, a large intensity. We say that the two waves are "interfering constructively" at such places. There will be such constructive interference wherever the distance from the detector to one hole is a whole number of wavelengths larger (or shorter) than the distance from the detector to the other hole.

At those places where the two waves arrive at the detector with a phase difference of π (where they are "out of phase") the resulting wave motion at the detector will be the difference of the two amplitudes. The waves "interfere destructively," and we get a low value for the wave intensity. We expect such low values wherever the distance between hole 1 and the detector is different from the distance between hole 2 and the detector by an odd number of half-wavelengths. The low values of I_{12} in Fig. 1–2 correspond to the places where the two waves interfere destructively.

You will remember that the quantitative relationship between I_1, I_2, and I_{12} can be expressed in the following way: The instantaneous height of the water wave at the detector for the wave from hole 1 can be written as (the real part of) $h_1 e^{i\omega t}$, where the "amplitude" h_1 is, in general, a complex number. The intensity is proportional to the mean squared height or, when we use the complex numbers, to the absolute value squared $|h_1|^2$. Similarly, for hole 2 the height is $h_2 e^{i\omega t}$ and the intensity is proportional to $|h_2|^2$. When both holes are open, the wave heights add to give the height$(h_1 + h_2)e^{i\omega t}$ and the intensity $|h_1 + h_2|^2$. Omitting the constant of proportionality for our present purposes, the proper relations for *interfering waves* are

$$I_1 = |h_1|^2, \qquad I_2 = |h_2|^2, \qquad I_{12} = |h_1 + h_2|^2. \tag{1.2}$$

You will notice that the result is quite different from that obtained with bullets (Eq. 1–1). If we expand $|h_1 + h_2|^2$ we see that

$$|h_1 + h_2|^2 = |h_1|^2 + |h_2|^2 + 2|h_1||h_2| \cos \delta, \tag{1.3}$$

where δ is the phase difference between h_1 and h_2. In terms of the intensities, we could write

$$I_{12} = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta. \tag{1.4}$$

The last term in (1.4) is the "interference term." So much for water waves. The intensity can have any value, and it shows interference.

1–4 An experiment with electrons

Now we imagine a similar experiment with electrons. It is shown diagrammatically in Fig. 1–3. We make an electron gun which consists of a tungsten wire heated by an electric current and surrounded by a metal box with a hole in it. If the wire is at a negative voltage with respect to the box, electrons emitted by the wire will be accelerated toward the walls and some will pass through the hole. All the electrons which come out of the gun will have (nearly) the same energy. In front of the gun is again a wall (just a thin metal plate) with two holes in it. Beyond the wall is another plate which will serve as a "backstop." In front of the backstop we place a movable detector. The detector might be a geiger counter or, perhaps better, an electron multiplier, which is connected to a loudspeaker.

We should say right away that you should not try to set up this experiment (as you could have done with the two we have already described). This experiment

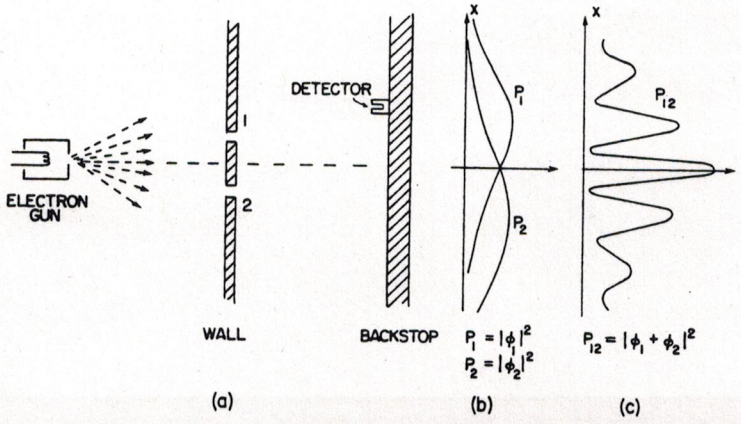

Fig. 1–3. Interference experiment with electrons.

has never been done in just this way. The trouble is that the apparatus would have to be made on an impossibly small scale to show the effects we are interested in. We are doing a "thought experiment," which we have chosen because it is easy to think about. We know the results that *would* be obtained because there *are* many experiments that have been done, in which the scale and the proportions have been chosen to show the effects we shall describe.

The first thing we notice with our electron experiment is that we hear sharp "clicks" from the detector (that is, from the loudspeaker). And all "clicks" are the same. There are *no* "half-clicks."

We would also notice that the "clicks" come very erratically. Something like: click . . . . . click-click . . . click . . . . . click . . . . click-click . . . . . click . . . , etc., just as you have, no doubt, heard a geiger counter operating. If we count the clicks which arrive in a sufficiently long time—say for many minutes—and then count again for another equal period, we find that the two numbers are very nearly the same. So we can speak of the *average rate* at which the clicks are heard (so-and-so-many clicks per minute on the average).

As we move the detector around, the *rate* at which the clicks appear is faster or slower, but the size (loudness) of each click is always the same. If we lower the temperature of the wire in the gun, the rate of clicking slows down, but still each click sounds the same. We would notice also that if we put two separate detectors at the backstop, one *or* the other would click, but never both at once. (Except that once in a while, if there were two clicks very close together in time, our ear might not sense the separation.) We conclude, therefore, that whatever arrives at the backstop arrives in "lumps." All the "lumps" are the same size: only whole "lumps" arrive, and they arrive one at a time at the backstop. We shall say: "Electrons always arrive in identical lumps."

Just as for our experiment with bullets, we can now proceed to find experimentally the answer to the question: "What is the relative probability that an electron 'lump' will arrive at the backstop at various distances x from the center?" As before, we obtain the relative probability by observing the rate of clicks, holding the operation of the gun constant. The probability that lumps will arrive at a particular x is proportional to the average rate of clicks at that x.

The result of our experiment is the interesting curve marked P_{12} in part (c) of Fig. 1–3. Yes! That is the way electrons go.

1–5 The interference of electron waves

Now let us try to analyze the curve of Fig. 1–3 to see whether we can understand the behavior of the electrons. The first thing we would say is that since they come in lumps, each lump, which we may as well call an electron, has come either through hole 1 or through hole 2. Let us write this in the form of a "Proposition":

Proposition A: Each electron *either* goes through hole 1 *or* it goes through hole 2.

Assuming Propositon A, all electrons that arrive at the backstop can be divided into two classes: (1) those that come through hole 1, and (2) those that come through hole 2. So our observed curve must be the sum of the effects of the electrons which come through hole 1 and the electrons which come through hole 2. Let us check this idea by experiment. First, we will make a measurement for those electrons that come through hole 1. We block off hole 2 and make our counts of the clicks from the detector. From the clicking rate, we get P_1. The result of the measurement is shown by the curve marked P_1 in part (b) of Fig. 1–3. The result seems quite reasonable. In a similar way, we measure P_2, the probability distribution for the electrons that come through hole 2. The result of this measurement is also drawn in the figure.

The result P_{12} obtained with *both* holes open is clearly not the sum of P_1 and P_2, the probabilities for each hole alone. In analogy with our water-wave experi-

ment, we say: "There is interference."

$$\textit{For electrons:} \qquad P_{12} \neq P_1 + P_2. \tag{1.5}$$

How can such an interference come about? Perhaps we should say: "Well, that means, presumably, that it is *not true* that the lumps go either through hole 1 or hole 2, because if they did, the probabilities should add. Perhaps they go in a more complicated way. They split in half and . . ." But no! They cannot, they always arrive in lumps . . . "Well, perhaps some of them go through 1, and then they go around through 2, and then around a few more times, or by some other complicated path . . . then by closing hole 2, we changed the chance that an electron that *started out* through hole 1 would finally get to the backstop . . ." But notice! There are some points at which very few electrons arrive when *both* holes are open, but which receive many electrons if we close one hole, so *closing* one hole *increased* the number from the other. Notice, however, that at the center of the pattern, P_{12} is more than twice as large as $P_1 + P_2$. It is as though closing one hole *decreased* the number of electrons which come through the other hole. It seems hard to explain *both* effects by proposing that the electrons travel in complicated paths.

It is all quite mysterious. And the more you look at it the more mysterious it seems. Many ideas have been concocted to try to explain the curve for P_{12} in terms of individual electrons going around in complicated ways through the holes. None of them has succeeded. None of them can get the right curve for P_{12} in terms of P_1 and P_2.

Yet, surprisingly enough, the *mathematics* for relating P_1 and P_2 to P_{12} is extremely simple. For P_{12} is just like the curve I_{12} of Fig. 1–2, and *that* was simple. What is going on at the backstop can be described by two complex numbers that we can call ϕ_1 and ϕ_2 (they are functions of x, of course). The absolute square of ϕ_1 gives the effect with only hole 1 open. That is, $P_1 = |\phi_1|^2$. The effect with only hole 2 open is given by ϕ_2 in the same way. That is, $P_2 = |\phi_2|^2$. And the combined effect of the two holes is just $P_{12} = |\phi_1 + \phi_2|^2$. The *mathematics* is the same as that we had for the water waves! (It is hard to see how one could get such a simple result from a complicated game of electrons going back and forth through the plate on some strange trajectory.)

We conclude the following: The electrons arrive in lumps, like particles, and the probability of arrival of these lumps is distributed like the distribution of intensity of a wave. It is in this sense that an electron behaves "sometimes like a particle and sometimes like a wave."

Incidentally, when we were dealing with classical waves we defined the intensity as the mean over time of the square of the wave amplitude, and we used complex numbers as a mathematical trick to simplify the analysis. But in quantum mechanics it turns out that the amplitudes *must* be represented by complex numbers. The real parts alone will not do. That is a technical point, for the moment, because the formulas look just the same.

Since the probability of arrival through both holes is given so simply, although it is not equal to $(P_1 + P_2)$, that is really all there is to say. But there are a large number of subtleties involved in the fact that nature does work this way. We would like to illustrate some of these subtleties for you now. First, since the number that arrives at a particular point is *not* equal to the number that arrives through 1 plus the number that arrives through 2, as we would have concluded from Proposition A, undoubtedly we should conclude that *Proposition A is false*. It is *not* true that the electrons go *either* through hole 1 or hole 2. But that conclusion can be tested by another experiment.

1–6 Watching the electrons

We shall now try the following experiment. To our electron apparatus we add a very strong light source, placed behind the wall and between the two holes, as shown in Fig. 1–4. We know that electric charges scatter light. So when an

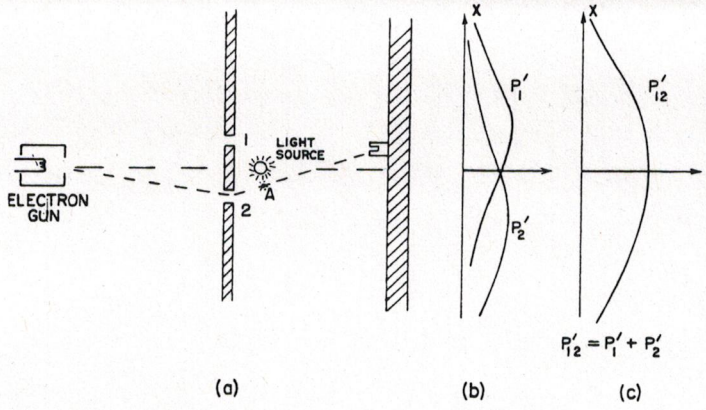

Fig. 1–4. A different electron experiment.

electron passes, however it does pass, on its way to the detector, it will scatter some light to our eye, and we can *see* where the electron goes. If, for instance, an electron were to take the path via hole 2 that is sketched in Fig. 1–4, we should see a flash of light coming from the vicinity of the place marked A in the figure. If an electron passes through hole 1, we would expect to see a flash from the vicinity of the upper hole. If it should happen that we get light from both places at the same time, because the electron divides in half . . . Let us just do the experiment!

Here is what we see: *every* time that we hear a "click" from our electron detector (at the backstop), we *also see* a flash of light *either* near hole 1 *or* near hole 2, but *never* both at once! And we observe the same result no matter where we put the detector. From this observation we conclude that when we look at the electrons we find that the electrons go either through one hole or the other. Experimentally, Proposition A is necessarily true.

What, then, is wrong with our argument *against* Proposition A? Why *isn't* P_{12} just equal to $P_1 + P_2$? Back to experiment! Let us keep track of the electrons and find out what they are doing. For each position (x-location) of the detector we will count the electrons that arrive and *also* keep track of which hole they went through, by watching for the flashes. We can keep track of things this way: whenever we hear a "click" we will put a count in Column 1 if we see the flash near hole 1, and if we see the flash near hole 2, we will record a count in Column 2. Every electron which arrives is recorded in one of two classes: those which come through 1 and those which come through 2. From the number recorded in Column 1 we get the probability P_1' that an electron will arrive at the detector via hole 1; and from the number recorded in Column 2 we get P_2', the probability that an electron will arrive at the detector via hole 2. If we now repeat such a measurement for many values of x, we get the curves for P_1' and P_2' shown in part (b) of Fig. 1–4.

Well, that is not too surprising! We get for P_1' something quite similar to what we got before for P_1 by blocking off hole 2; and P_2' is similar to what we got by blocking hole 1. So there is *not* any complicated business like going through both holes. When we watch them, the electrons come through just as we would expect them to come through. Whether the holes are closed or open, those which we see come through hole 1 are distributed in the same way whether hole 2 is open or closed.

But wait! What do we have *now* for the *total* probability, the probability that an electron will arrive at the detector by any route? We already have that information. We just pretend that we never looked at the light flashes, and we lump together the detector clicks which we have separated into the two columns. We *must* just *add* the numbers. For the probability that an electron will arrive at the backstop by passing through *either* hole, we do find $P_{12}' = P_1 + P_2$. That is, although we succeeded in watching which hole our electrons come through, we no longer get the old interference curve P_{12}, but a new one, P_{12}', showing no interference! If we turn out the light P_{12} is restored.

We must conclude that *when we look at the electrons* the distribution of them on the screen is different than when we do not look. Perhaps it is turning on our light source that disturbs things? It must be that the electrons are very delicate, and the light, when it scatters off the electrons, gives them a jolt that changes their

motion. We know that the electric field of the light acting on a charge will exert a force on it. So perhaps we *should* expect the motion to be changed. Anyway, the light exerts a big influence on the electrons. By trying to "watch" the electrons we have changed their motions. That is, the jolt given to the electron when the photon is scattered by it is such as to change the electron's motion enough so that if it *might* have gone to where P_{12} was at a maximum it will instead land where P_{12} was a minimum; that is why we no longer see the wavy interference effects.

You may be thinking: "Don't use such a bright source! Turn the brightness down! The light waves will then be weaker and will not disturb the electrons so much. Surely, by making the light dimmer and dimmer, eventually the wave will be weak enough that it will have a negligible effect." O.K. Let's try it. The first thing we observe is that the flashes of light scattered from the electrons as they pass by does *not* get weaker. *It is always the same-sized flash.* The only thing that happens as the light is made dimmer is that sometimes we hear a "click" from the detector but see *no flash at all.* The electron has gone by without being "seen." What we are observing is that light *also* acts like electrons, we *knew* that it was "wavy," but now we find that it is also "lumpy." It always arrives—or is scattered—in lumps that we call "photons." As we turn down the *intensity* of the light source we do not change the *size* of the photons, only the *rate* at which they are emitted. *That* explains why, when our source is dim, some electrons get by without being seen. There did not happen to be a photon around at the time the electron went through.

This is all a little discouraging. If it is true that whenever we "see" the electron we see the same-sized flash, then those electrons we see are *always* the disturbed ones. Let us try the experiment with a dim light anyway. Now whenever we hear a click in the detector we will keep a count in three columns: in Column (1) those electrons seen by hole 1, in Column (2) those electrons seen by hole 2, and in Column (3) those electrons not seen at all. When we work up our data (computing the probabilities) we find these results: Those "seen by hole 1" have a distribution like P_1'; those "seen by hole 2" have a distribution like P_2' (so that those "seen by either hole 1 or 2" have a distribution like P_{12}'); and those "not seen at all" have a "wavy" distribution just like P_{12} of Fig. 1–3! *If the electrons are not seen, we have interference!*

That is understandable. When we do not see the electron, no photon disturbs it, and when we do see it, a photon has disturbed it. There is always the same amount of disturbance because the light photons all produce the same-sized effects and the effect of the photons being scattered is enough to smear out any interference effect.

Is there not *some* way we can see the electrons without disturbing them? We learned in an earlier chapter that the momentum carried by a "photon" is inversely proportional to its wavelength ($p = h/\lambda$). Certainly the jolt given to the electron when the photon is scattered toward our eye depends on the momentum that photon carries. Aha! If we want to disturb the electrons only slightly we should not have lowered the *intensity* of the light, we should have lowered its *frequency* (the same as increasing its wavelength). Let us use light of a redder color. We could even use infrared light, or radiowaves (like radar), and "see" where the electron went with the help of some equipment that can "see" light of these longer wavelengths. If we use "gentler" light perhaps we can avoid disturbing the electrons so much.

Let us try the experiment with longer waves. We shall keep repeating our experiment, each time with light of a longer wavelength. At first, nothing seems to change. The results are the same. Then a terrible thing happens. You remember that when we discussed the microscope we pointed out that, due to the *wave nature* of the light, there is a limitation on how close two spots can be and still be seen as two separate spots. This distance is of the order of the wavelength of light. So now, when we make the wavelength longer than the distance between our holes, we see a *big* fuzzy flash when the light is scattered by the electrons. We can no longer tell which hole the electron went through! We just know it went somewhere! And it is just with light of this color that we find that the jolts given to the electron

are small enough so that P'_{12} begins to look like P_{12}—that we begin to get some interference effect. And it is only for wavelengths much longer than the separation of the two holes (when we have no chance at all of telling where the electron went) that the disturbance due to the light gets sufficiently small that we again get the curve P_{12} shown in Fig. 1–3.

In our experiment we find that it is impossible to arrange the light in such a way that one can tell which hole the electron went through, and at the same time not disturb the pattern. It was suggested by Heisenberg that the then new laws of nature could only be consistent if there were some basic limitation on our experimental capabilities not previously recognized. He proposed, as a general principle, his *uncertainty principle*, which we can state in terms of our experiment as follows: "It is impossible to design an apparatus to determine which hole the electron passes through, that will not at the same time disturb the electrons enough to destroy the interference pattern." If an apparatus is capable of determining which hole the electron goes through, it *cannot* be so delicate that it does not disturb the pattern in an essential way. No one has ever found (or even thought of) a way around the uncertainty principle. So we must assume that it describes a basic characteristic of nature.

The complete theory of quantum mechanics which we now use to describe atoms and, in fact, all matter, depends on the correctness of the uncertainty principle. Since quantum mechanics is such a successful theory, our belief in the uncertainty principle is reinforced. But if a way to "beat" the uncertainty principle were ever discovered, quantum mechanics would give inconsistent results and would have to be discarded as a valid theory of nature.

"Well," you say, "what about Proposition A? Is it true, or is it *not* true, that the electron either goes through hole 1 or it goes through hole 2?" The only answer that can be given is that we have found from experiment that there is a certain special way that we have to think in order that we do not get into inconsistencies. What we must say (to avoid making wrong predictions) is the following. If one looks at the holes or, more accurately, if one has a piece of apparatus which is capable of determining whether the electrons go through hole 1 or hole 2, then one *can* say that it goes either through hole 1 or hole 2. *But*, when one does *not* try to tell which way the electron goes, when there is nothing in the experiment to disturb the electrons, then one may *not* say that an electron goes either through hole 1 or hole 2. If one does say that, and starts to make any deductions from the statement, he will make errors in the analysis. This is the logical tightrope on which we must walk if we wish to describe nature successfully.

If the motion of all matter—as well as electrons—must be described in terms of waves, what about the bullets in our first experiment? Why didn't we see an interference pattern there? It turns out that for the bullets the wavelengths were so tiny that the interference patterns became very fine. So fine, in fact, that with any detector of finite size one could not distinguish the separate maxima and minima. What we saw was only a kind of average, which is the classical curve. In Fig. 1–5 we have tried to indicate schematically what happens with large-scale objects. Part (a) of the figure shows the probability distribution one might predict for bullets, using quantum mechanics. The rapid wiggles are supposed to represent the interference pattern one gets for waves of very short wavelength. Any physical detector, however, straddles several wiggles of the probability curve, so that the measurements show the smooth curve drawn in part (b) of the figure.

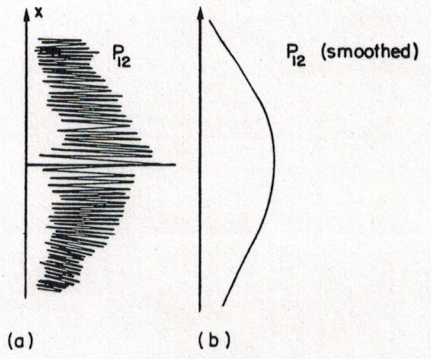

Fig. 1–5. Interference pattern with bullets: (a) actual (schematic), (b) observed.

1–7 First principles of quantum mechanics

We will now write a summary of the main conclusions of our experiments. We will, however, put the results in a form which makes them true for a general class of such experiments. We can write our summary more simply if we first define an "ideal experiment" as one in which there are no uncertain external influences, i.e., no jiggling or other things going on that we cannot take into ac-

count. We would be quite precise if we said: "An ideal experiment is one in which all of the initial and final conditions of the experiment are completely specified." What we will call "an event" is, in general, just a specific set of initial and final conditions. (For example: "an electron leaves the gun, arrives at the detector, and nothing else happens.") Now for our summary.

SUMMARY

(1) The probability of an event in an ideal experiment is given by the square of the absolute value of a complex number ϕ which is called the probability amplitude:

$$\begin{aligned} P &= \text{probability}, \\ \phi &= \text{probability amplitude}, \\ P &= |\phi|^2. \end{aligned} \tag{1.6}$$

(2) When an event can occur in several alternative ways, the probability amplitude for the event is the sum of the probability amplitudes for each way considered separately. There is interference:

$$\begin{aligned} \phi &= \phi_1 + \phi_2, \\ P &= |\phi_1 + \phi_2|^2. \end{aligned} \tag{1.7}$$

(3) If an experiment is performed which is capable of determining whether one or another alternative is actually taken, the probability of the event is the sum of the probabilities for each alternative. The interference is lost:

$$P = P_1 + P_2. \tag{1.8}$$

One might still like to ask: "How does it work? What is the machinery behind the law?" No one has found any machinery behind the law. No one can "explain" any more than we have just "explained." No one will give you any deeper representation of the situation. We have no ideas about a more basic mechanism from which these results can be deduced.

We would like to emphasize a very important difference between classical and quantum mechanics. We have been talking about the probability that an electron will arrive in a given circumstance. We have implied that in our experimental arrangement (or even in the best possible one) it would be impossible to predict exactly what would happen. We can only predict the odds! This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will happen in a definite circumstance. Yes! physics *has* given up. *We do not know how to predict what would happen in a given circumstance*, and we believe now that it is impossible—that the only thing that can be predicted is the probability of different events. It must be recognized that this is a retrenchment in our earlier ideal of understanding nature. It may be a backward step, but no one has seen a way to avoid it.

We make now a few remarks on a suggestion that has sometimes been made to try to avoid the description we have given: "Perhaps the electron has some kind of internal works—some inner variables—that we do not yet know about. Perhaps that is why we cannot predict what will happen. If we could look more closely at the electron, we could be able to tell where it would end up." So far as we know, that is impossible. We would still be in difficulty. Suppose we were to assume that inside the electron there is some kind of machinery that determines where it is going to end up. That machine must *also* determine which hole it is going to go through on its way. But we must not forget that what is inside the electron should not be dependent on what *we* do, and in particular upon whether we open or close one of the holes. So if an electron, before it starts, has already made up its mind (a) which hole it is going to use, and (b) where it is going to land, we should find P_1 for those electrons that have chosen hole 1, P_2 for those that have chosen hole 2, *and necessarily* the sum $P_1 + P_2$ for those that arrive through the two holes. There seems to be no way around this. But we have verified experimentally that that is not the case. And no one has figured a way out of this puzzle. So at the

present time we must limit ourselves to computing probabilities. We say "at the present time," but we suspect very strongly that it is something that will be with us forever—that it is impossible to beat that puzzle—that this is the way nature really *is*.

1–8 The uncertainty principle

This is the way Heisenberg stated the uncertainty principle originally: If you make the measurement on any object, and you can determine the x-component of its momentum with an uncertainty Δp, you cannot, at the same time, know its x-position more accurately than $\Delta x = h/\Delta p$, where h is a definite fixed number given by nature. It is called "Planck's constant," and is approximately 6.63×10^{-34} joule-seconds. The uncertainties in the position and momentum of a particle at any instant must have their product greater than Planck's constant. This is a special case of the uncertainty principle that was stated above more generally. The more general statement was that one cannot design equipment in any way to determine which of two alternatives is taken, without, at the same time, destroying the pattern of interference.

Let us show for one particular case that the kind of relation given by Heisenberg must be true in order to keep from getting into trouble. We imagine a modification of the experiment of Fig. 1–3, in which the wall with the holes consists of a plate mounted on rollers so that it can move freely up and down (in the x-direction), as shown in Fig. 1–6. By watching the motion of the plate carefully we can try to tell which hole an electron goes through. Imagine what happens when the detector is placed at $x = 0$. We would expect that an electron which passes through hole 1 must be deflected downward by the plate to reach the detector. Since the vertical component of the electron momentum is changed, the plate must recoil with an equal momentum in the opposite direction. The plate will get an upward kick. If the electron goes through the lower hole, the plate should feel a downward kick. It is clear that for every position of the detector, the momentum received by the plate will have a different value for a traversal via hole 1 than for a traversal via hole 2. So! Without disturbing the electrons *at all*, but just by watching the *plate*, we can tell which path the electron used.

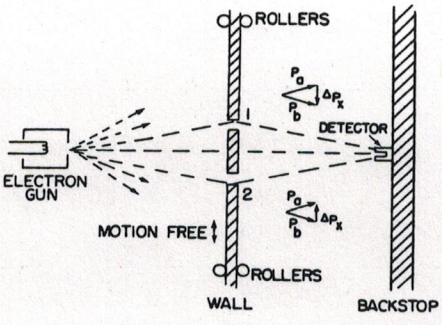

Fig. 1–6. An experiment in which the recoil of the wall is measured.

Now in order to do this it is necessary to know what the momentum of the screen is, before the electron goes through. So when we measure the momentum after the electron goes by, we can figure out how much the plate's momentum has changed. But remember, according to the uncertainty principle we cannot at the same time know the position of the plate with an arbitrary accuracy. But if we do not know exactly *where* the plate is, we cannot say precisely where the two holes are. They will be in a different place for every electron that goes through. This means that the center of our interference pattern will have a different location for each electron. The wiggles of the interference pattern will be smeared out. We shall show quantitatively in the next chapter that if we determine the momentum of the plate sufficiently accurately to determine from the recoil measurement which hole was used, then the uncertainty in the x-position of the plate will, according to the uncertainty principle, be enough to shift the pattern observed at the detector up and down in the x-direction about the distance from a maximum to its nearest minimum. Such a random shift is just enough to smear out the pattern so that no interference is observed.

The uncertainty principle "protects" quantum mechanics. Heisenberg recognized that if it were possible to measure the momentum and the position simultaneously with a greater accuracy, the quantum mechanics would collapse. So he proposed that it must be impossible. Then people sat down and tried to figure out ways of doing it, and nobody could figure out a way to measure the position and the momentum of anything—a screen, an electron, a billiard ball, anything—with any greater accuracy. Quantum mechanics maintains its perilous but still correct existence.

7

The Dependence of Amplitudes on Time

Review: Chapter 17, Vol. I, *Space-Time*
Chapter 48, Vol. I, *Beats*

7–1 Atoms at rest; stationary states

We want now to talk a little bit about the behavior of probability amplitudes in time. We say a "little bit," because the actual behavior in time necessarily involves the behavior in space as well. Thus, we get immediately into the most complicated possible situation if we are to do it correctly and in detail. We are always in the difficulty that we can either treat something in a logically rigorous but quite abstract way, or we can do something which is not at all rigorous but which gives us some idea of a real situation—postponing until later a more careful treatment. With regard to energy dependence, we are going to take the second course. We will make a number of statements. We will not try to be rigorous—but will just be telling you things that have been found out, to give you some feeling for the behavior of amplitudes as a function of time. As we go along, the precision of the description will increase, so don't get nervous that we seem to be picking things out of the air. It is, of course, all out of the air—the air of experiment and of the imagination of people. But it would take us too long to go over the historical development, so we have to plunge in somewhere. We could plunge into the abstract and deduce everything—which you would not understand—or we could go through a large number of experiments to justify each statement. We choose to do something in between.

An electron alone in empty space can, under certain circumstances, have a certain definite energy. For example, if it is standing still (so it has no translational motion, no momentum, or kinetic energy), it has its rest energy. A more complicated object like an atom can also have a definite energy when standing still, but it could also be internally excited to another energy level. (We will describe later the machinery of this.) We can often think of an atom in an excited state as having a definite energy, but this is really only approximately true. An atom doesn't stay excited forever because it manages to discharge its energy by its interaction with the electromagnetic field. So there is some amplitude that a new state is generated—with the atom in a lower state, and the electromagnetic field in a higher state, of excitation. The total energy of the system is the same before and after, but the energy of the *atom* is reduced. So it is not precise to say an excited atom has a *definite* energy; but it will often be convenient and not too wrong to say that it does.

[Incidentally, why does it go one way instead of the other way? Why does an atom radiate light? The answer has to do with entropy. When the energy is in the electromagnetic field, there are so many different ways it can be—so many different places where it can wander—that if we look for the equilibrium condition, we find that in the most probable situation the field is excited with a photon, and the atom is de-excited. It takes a very long time for the photon to come back and find that it can knock the atom back up again. It's quite analogous to the classical problem: Why does an accelerating charge radiate? It isn't that it "wants" to lose energy, because, in fact, when it radiates, the energy of the world is the same as it was before. Radiation or absorption goes in the direction of increasing *entropy*.]

Nuclei can also exist in different energy levels, and in an approximation which disregards the electromagnetic effects, we can say that a nucleus in an excited state stays there. Although we know that it doesn't stay there forever, it is often useful to start out with an approximation which is somewhat idealized and easier to think about. Also it is often a legitimate approximation under certain circumstances. (When we first introduced the classical laws of a falling body, we did not include friction, but there is almost never a case in which there isn't *some* friction.)

Then there are the subnuclear "strange particles," which have various masses. But the heavier ones disintegrate into other light particles, so again it is not correct to say that they have a precisely definite energy. That would be true only if they lasted forever. So when we make the approximation that they have a definite energy, we are forgetting the fact that they must blow up. For the moment, then, we will intentionally forget about such processes and learn later how to take them into account.

Suppose we have an atom—or an electron, or any particle—which at rest would have a definite energy E_0. By the energy E_0 we mean the mass of the whole thing times c^2. This mass includes any internal energy; so an excited atom has a mass which is different from the mass of the same atom in the ground state. (The *ground* state means the state of lowest energy.) We will call E_0 the "energy at rest."

For an atom *at rest*, the quantum mechanical *amplitude* to find an atom at a place is the *same everywhere*; it does *not* depend on position. This means, of course, that the *probability* of *finding* the atom anywhere is the same. But it means even more. The *probability* could be independent of position, and still the *phase* of the *amplitude* could vary from point to point. But for a particle at rest, the complete amplitude is identical everywhere. It does, however, depend on the *time*. For a particle in a state of definite energy E_0, the amplitude to find the particle at (x, y, z) at the time t is

$$ae^{-i(E_0/\hbar)t}, \tag{7.1}$$

where a is some constant. The amplitude to be at any point in space is the same for all points, but depends on time according to (7.1). We shall simply assume this rule to be true.

Of course, we could also write (7.1) as

$$ae^{-i\omega t}, \tag{7.2}$$

with

$$\hbar\omega = E_0 = Mc^2,$$

where M is the rest mass of the atomic state, or particle. There are three different ways of specifying the energy: by the frequency of an amplitude, by the energy in the classical sense, or by the inertia. They are all equivalent; they are just different ways of saying the same thing.

You may be thinking that it is strange to think of a "particle" which has equal amplitudes to be found throughout all space. After all, we usually imagine a "particle" as a small object located "somewhere." But don't forget the uncertainty principle. If a particle has a definite energy, it has also a definite momentum. If the uncertainty in momentum is zero, the uncertainty relation, $\Delta p\,\Delta x = \hbar$, tells us that the uncertainty in the position must be infinite, and that is just what we are saying when we say that there is the same amplitude to find the particle at all points in space.

If the internal parts of an atom are in a different state with a different total energy, then the variation of the amplitude with time is different. If you don't know in which state it is, there will be a certain amplitude to be in one state and a certain amplitude to be in another—and each of these amplitudes will have a different frequency. There will be an interference between these different components —like a beat-note—which can show up as a varying probability. Something will be "going on" inside of the atom—even though it is "at rest" in the sense that its center of mass is not drifting. However, if the atom has one definite energy, the amplitude is given by (7.1), and the absolute square of this amplitude does not depend on time. You see, then, that if a thing has a definite energy and if you ask any *probability* question about it, the answer is independent of time. Although the *amplitudes* vary with time, if the energy is *definite* they vary as an imaginary exponential, and the absolute value doesn't change.

That's why we often say that an atom in a definite energy level is in a *stationary state*. If you make any measurements of the things inside, you'll find that nothing (in probability) will change in time. In order to have the probabilities change in

time, we have to have the interference of two amplitudes at two different frequencies, and that means that we cannot know what the energy is. The object will have one amplitude to be in a state of one energy and another amplitude to be in a state of another energy. That's the quantum mechanical description of something when its *behavior* depends on time.

If we have a "condition" which is a mixture of two different states with different energies, then the amplitude for each of the two states varies with time according to Eq. (7.2), for instance, as

$$e^{-i(E_1/\hbar)t} \quad \text{and} \quad e^{-i(E_2/\hbar)t}. \tag{7.3}$$

And if we have some combination of the two, we will have an interference. But notice that if we added a constant to both energies, it wouldn't make any difference. If somebody else were to use a different scale of energy in which all the energies were increased (or decreased) by a constant amount—say, by the amount A—then the amplitudes in the two states would, from his point of view, be

$$e^{-i(E_1+A)t/\hbar} \quad \text{and} \quad e^{-i(E_2+A)t/\hbar}. \tag{7.4}$$

All of his amplitudes would be multiplied by the same factor $e^{-i(A/\hbar)t}$, and all linear combinations, or interferences, would have the same factor. When we take the absolute squares to find the probabilities, all the answers would be the same. The choice of an origin for our energy scale makes no difference; we can measure energy from any zero we want. For relativistic purposes it is nice to measure the energy so that the rest mass is included, but for many purposes that aren't relativistic it is often nice to subtract some standard amount from all energies that appear. For instance, in the case of an atom, it is usually convenient to subtract the energy $M_s c^2$, where M_s is the mass of all the *separate* pieces—the nucleus and the electrons—which is, of course, different from the mass of the atom. For other problems it may be useful to subtract from all energies the amount $M_g c^2$, where M_g is the mass of the whole atom *in the ground* state; then the energy that appears is just the excitation energy of the atom. So, sometimes we may shift our zero of energy by some very large constant, but it doesn't make any difference, provided we shift all the energies in a particular calculation by the same constant. So much for a particle standing still.

7–2 Uniform motion

If we suppose that the relativity theory is right, a particle at rest in one inertial system can be in uniform motion in another inertial system. In the rest frame of the particle, the probability amplitude is the same for all x, y, and z but varies with t. The *magnitude* of the amplitude is the same for all t, but the *phase* depends on t. We can get a kind of a picture of the behavior of the amplitude if we plot lines of equal phase—say, lines of zero phase—as a function of x and t. For a particle at rest, these equal-phase lines are parallel to the x-axis and are equally spaced in the t-coordinate, as shown by the dashed lines in Fig. 7–1.

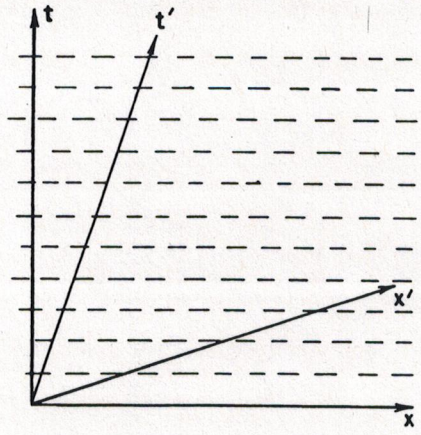

Fig. 7–1. Relativistic transformation of the amplitude of a particle at rest in the x-t systems.

In a different frame—x', y', z', t'—that is moving with respect to the particle in, say, the x-direction, the x' and t' coordinates of any particular point in space are related to x and t by the Lorentz transformation. This transformation can be represented graphically by drawing x' and t' axes, as is done in Fig. 7–1. (See Chapter 17, Vol. I, Fig. 17–2.) You can see that in the x'-t' system, points of equal phase† have a different spacing along the t'-axis, so the frequency of the time variation is different. Also there is a variation of the phase with x', so the probability amplitude must be a function of x'.

† We are assuming that the phase should have the same value at corresponding points in the two systems. This is a subtle point, however, since the phase of a quantum mechanical amplitude is, to a large extent, arbitrary. A complete justification of this assumption requires a more detailed discussion involving interferences of two or more amplitudes.

Under a Lorentz transformation for the velocity v, say along the negative x-direction, the time t is related to the time t' by

$$t = \frac{t' - x'v/c^2}{\sqrt{1 - v^2/c^2}},$$

so our amplitude now varies as

$$e^{-(i/\hbar)E_0 t} = e^{-(i/\hbar)(E_0 t'/\sqrt{1-v^2/c^2} - E_0 vx'/c^2\sqrt{1-v^2/c^2})}.$$

In the prime system it varies in space as well as in time. If we write the amplitude as

$$e^{-(i/\hbar)(E'_p t' - p'x')},$$

we see that $E'_p = E_0/\sqrt{1 - v^2/c^2}$ is the energy computed classically for a particle of rest energy E_0 travelling at the velocity v, and $p' = E'_p v/c^2$ is the corresponding particle momentum.

You know that $x_\mu = (t, x, y, z)$ and $p_\mu = (E, p_x, p_y, p_z)$ are four-vectors, and that $p_\mu x_\mu = Et - \boldsymbol{p} \cdot \boldsymbol{x}$ is a scalar invariant. In the rest frame of the particle, $p_\mu x_\mu$ is just Et; so if we transform to another frame, Et will be replaced by

$$E't' - \boldsymbol{p}' \cdot \boldsymbol{x}'.$$

Thus, the probability amplitude of a particle which has the momentum $\boldsymbol{p}$ will be proportional to

$$e^{-(i/\hbar)(E_p t - \boldsymbol{p}\cdot\boldsymbol{x})}, \tag{7.5}$$

where E_p is the energy of the particle whose momentum is p, that is,

$$E_p = \sqrt{(pc)^2 + E_0^2}, \tag{7.6}$$

where E_0 is, as before, the rest energy. For nonrelativistic problems, we can write

$$E_p = M_s c^2 + W_p, \tag{7.7}$$

where W_p is the energy over and above the rest energy $M_s c^2$ of the parts of the atom. In general, W_p would include both the kinetic energy of the atom as well as its binding or excitation energy, which we can call the "internal" energy. We would write

$$W_p = W_{\text{int}} + \frac{p^2}{2M}, \tag{7.8}$$

and the amplitudes would be

$$e^{-(i/\hbar)(W_p t - \boldsymbol{p}\cdot\boldsymbol{x})}. \tag{7.9}$$

Because we will generally be doing nonrelativistic calculations, we will use this form for the probability amplitudes.

Note that our relativistic transformation has given us the variation of the amplitude of an atom which moves in space without any additional assumptions. The wave number of the space variations is, from (7.9),

$$k = \frac{p}{\hbar}; \tag{7.10}$$

so the wavelength is

$$\lambda = \frac{2\pi}{k} = \frac{h}{p}. \tag{7.11}$$

This is the same wavelength we have used before for particles with the momentum p. This formula was first arrived at by de Broglie in just this way. For a moving particle, the *frequency* of the amplitude variations is still given by

$$\hbar\omega = W_p. \tag{7.12}$$

The absolute square of (7.9) is just 1, so for a particle in motion with a *definite energy*, the probability of finding it is the same everywhere and does not change with time. (It is important to notice that the amplitude is a *complex* wave. If we used a real sine wave, the square would vary from point to point, which would not be right.)

We know, of course, that there are situations in which particles move from place to place so that the probability depends on position and changes with time. How do we describe such situations? We can do that by considering amplitudes which are a superposition of two or more amplitudes for states of definite energy. We have already discussed this situation in Chapter 48 of Vol. I—even for probability amplitudes! We found that the sum of two amplitudes with different wave numbers k (that is, momenta) and frequencies ω (that is, energies) gives interference humps, or beats, so that the square of the amplitude varies with space and time. We also found that these beats move with the so-called "group velocity" given by

$$v_g = \frac{\Delta\omega}{\Delta k},$$

where Δk and $\Delta\omega$ are the differences between the wave numbers and frequencies for the two waves. For more complicated waves—made up of the sum of many amplitudes all near the same frequency—the group velocity is

$$v_g = \frac{d\omega}{dk}. \tag{7.13}$$

Taking $\omega = E_p/\hbar$ and $k = p/\hbar$, we see that

$$v_g = \frac{dE_p}{dp}. \tag{7.14}$$

Using Eq. (7.6), we have

$$\frac{dE_p}{dp} = c^2\,\frac{p}{E_p}. \tag{7.15}$$

But $E_p = Mc^2$, so

$$\frac{dE_p}{dp} = \frac{p}{M}, \tag{7.16}$$

which is just the classical velocity of the particle. Alternatively, if we use the non-relativistic expressions, we have

$$\omega = \frac{W_p}{\hbar} \quad \text{and} \quad k = \frac{p}{\hbar},$$

and

$$\frac{d\omega}{dk} = \frac{dW}{dp} = \frac{d}{dp}\left(\frac{p^2}{2M}\right) = \frac{p}{M}, \tag{7.17}$$

which is again the classical velocity.

Our result, then, is that if we have several amplitudes for pure energy states of nearly the same energy, their interference gives "lumps" in the probability that move through space with a velocity equal to the velocity of a classical particle of that energy. We should remark, however, that when we say we can add two amplitudes of different wave number together to get a beat-note that will correspond to a moving particle, we have introduced something new—something that we cannot deduce from the theory of relativity. We said what the amplitude did for a particle standing still and then deduced what it would do if the particle were moving. But we *cannot* deduce from these arguments what would happen when there are *two* waves moving with different speeds. If we stop one, we cannot stop the other. So we have added tacitly the *extra* hypothesis that not only is (7.9) a *possible* solution, but that there can also be solutions with all kinds of p's for the same system, and that the different terms will interfere.

3

Probability Amplitudes

3–1 The laws for combining amplitudes

When Schrödinger first discovered the correct laws of quantum mechanics, he wrote an equation which described the amplitude to find a particle in various places. This equation was very similar to the equations that were already known to classical physicists—equations that they had used in describing the motion of air in a sound wave, the transmission of light, and so on. So most of the time at the beginning of quantum mechanics was spent in solving this equation. But at the same time an understanding was being developed, particularly by Born and Dirac, of the basically new physical ideas behind quantum mechanics. As quantum mechanics developed further, it turned out that there were a large number of things which were not directly encompassed in the Schrödinger equation—such as the spin of the electron, and various relativistic phenomena. Traditionally, all courses in quantum mechanics have begun in the same way, retracing the path followed in the historical development of the subject. One first learns a great deal about classical mechanics so that he will be able to understand how to solve the Schrödinger equation. Then he spends a long time working out various solutions. Only after a detailed study of this equation does he get to the "advanced" subject of the electron's spin.

We had also originally considered that the right way to conclude these lectures on physics was to show how to solve the equations of classical physics in complicated situations—such as the description of sound waves in enclosed regions, modes of electromagnetic radiation in cylindrical cavities, and so on. That was the original plan for this course. However, we have decided to abandon that plan and to give instead an introduction to the quantum mechanics. We have come to the conclusion that what are usually called the advanced parts of quantum mechanics are, in fact, quite simple. The mathematics that is involved is particularly simple, involving simple algebraic operations and no differential equations or at most only very simple ones. The only problem is that we must jump the gap of no longer being able to describe the behavior *in detail* of particles in space. So this is what we are going to try to do: to tell you about what conventionally would be called the "advanced" parts of quantum mechanics. But they are, we assure you, by all odds the simplest parts—in a deep sense of the word—as well as the most basic parts. This is frankly a pedagogical experiment; it has never been done before, as far as we know.

In this subject we have, of course, the difficulty that the quantum mechanical behavior of things is quite strange. Nobody has an everyday experience to lean on to get a rough, intuitive idea of what will happen. So there are two ways of presenting the subject: We could either describe what can happen in a rather rough physical way, telling you more or less what happens without giving the precise laws of everything; or we could, on the other hand, give the precise laws in their abstract form. But, then because of the abstractions, you wouldn't know what they were all about, physically. The latter method is unsatisfactory because it is completely abstract, and the first way leaves an uncomfortable feeling because one doesn't know exactly what is true and what is false. We are not sure how to overcome this difficulty. You will notice, in fact, that Chapters 1 and 2 showed this problem. The first chapter was relatively precise; but the second chapter was a rough description of the characteristics of different phenomena. Here, we will try to find a happy medium between the two extremes.

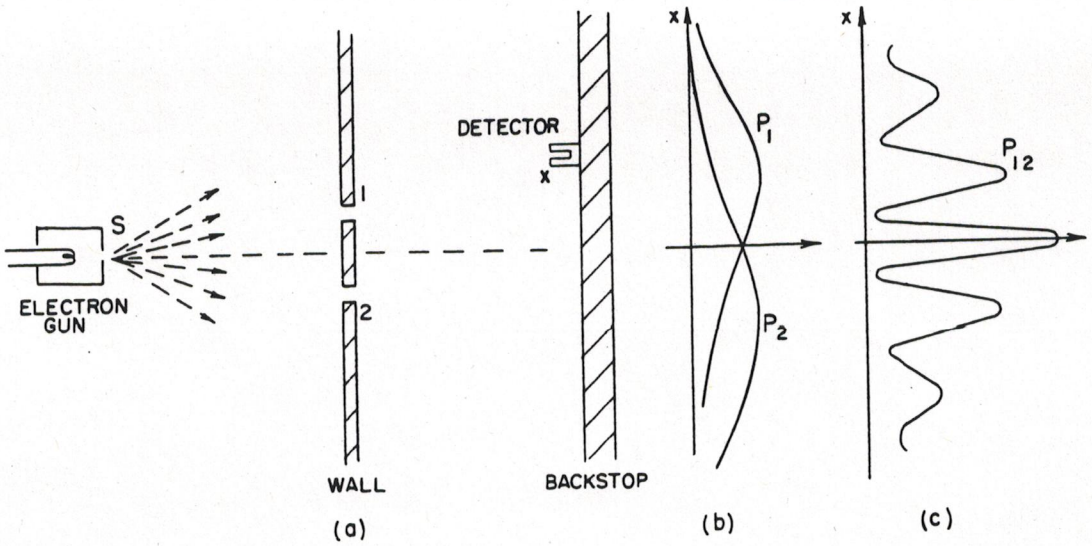

Fig 3–1. Interference experiment with electrons.

We will begin in this chapter by dealing with some general quantum mechanical ideas. Some of the statements will be quite precise, others only partially precise. It will be hard to tell you as we go along which is which, but by the time you have finished the rest of the book, you will understand in looking back which parts hold up and which parts were only explained roughly. The chapters which follow this one will not be so imprecise. In fact, one of the reasons we have tried carefully to be precise in the succeeding chapters is so that we can show you one of the most beautiful things about quantum mechanics—how much can be deduced from so little.

We begin by discussing again the superposition of *probability amplitudes*. As an example we will refer to the experiment described in Chapter 1, and shown again here in Fig. 3–1. There is a source s of particles, say electrons; then there is a wall with two slits in it; after the wall, there is a detector located at some position x. We ask for the probability that a particle will be found at x. Our *first general principle* in quantum mechanics is that the *probability* that a particle will arrive at x, when let out at the source s, can be represented quantitatively by the absolute square of a complex number called a *probability amplitude*—in this case, the "amplitude that a particle from s will arrive at x." We will use such amplitudes so frequently that we will use a shorthand notation—invented by Dirac and generally used in quantum mechanics—to represent this idea. We write the probability amplitude this way:

$$\langle \text{Particle arrives at } x \mid \text{particle leaves } s \rangle. \tag{3.1}$$

In other words, the two brackets $\langle\ \rangle$ are a sign equivalent to "the amplitude that"; the expression at the *right* of the vertical line always gives the *starting* condition, and the one at the left, the *final* condition. Sometimes it will also be convenient to abbreviate still more and describe the initial and final conditions by single letters. For example, we may on occasion write the amplitude (3.1) as

$$\langle x \mid s \rangle. \tag{3.2}$$

We want to emphasize that such an amplitude is, of course, just a single number—a *complex* number.

We have already seen in the discussion of Chapter 1 that when there are two ways for the particle to reach the detector, the resulting probability is not the sum of the two probabilities, but must be written as the absolute square of the sum of two amplitudes. We had that the probability that an electron arrives at the detector when both paths are open is

$$P_{12} = |\phi_1 + \phi_2|^2. \tag{3.3}$$

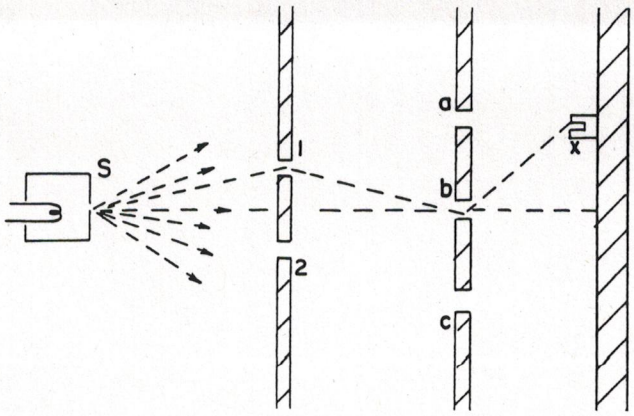

Fig. 3–2. A more complicated interference experiment.

We wish now to put this result in terms of our new notation. First, however, we want to state our *second general principle* of quantum mechanics: When a particle can reach a given state by two possible routes, the total amplitude for the process is the *sum of the amplitudes* for the two routes considered separately. In our new notation we write that

$$\langle x \mid s\rangle_{\text{both holes open}} = \langle x \mid s\rangle_{\text{through 1}} + \langle x \mid s\rangle_{\text{through 2}}. \tag{3.4}$$

Incidentally, we are going to suppose that the holes 1 and 2 are small enough that when we say an electron goes through the hole, we don't have to discuss which part of the hole. We could, of course, split each hole into pieces with a certain amplitude that the electron goes to the top of the hole and the bottom of the hole and so on. We will suppose that the hole is small enough so that we don't have to worry about this detail. That is part of the roughness involved; the matter can be made more precise, but we don't want to do so at this stage.

Now we want to write out in more detail what we can say about the amplitude for the process in which the electron reaches the detector at x by way of hole 1. We can do that by using our *third general principle*: When a particle goes by some particular route the amplitude for that route can be written as the *product* of the *amplitude* to go part way with the *amplitude* to go the rest of the way. For the setup of Fig. 3–1 the amplitude to go from s to x by way of hole 1 is equal to the amplitude to go from s to 1, multiplied by the amplitude to go from 1 to x.

$$\langle x \mid s\rangle_{\text{via 1}} = \langle x \mid 1\rangle\langle 1 \mid s\rangle. \tag{3.5}$$

Again this result is not completely precise. We should also include a factor for the amplitude that the electron will get through the hole at 1; but in the present case it is a simple hole, and we will take this factor to be unity.

You will note that Eq. (3.5) appears to be written in reverse order. It is to be read from right to left: The electron goes from s to 1 and then from 1 to x. In summary, if events occur in succession—that is, if you can analyze one of the routes of the particle by saying it does this, then it does this, then it does that—the resultant amplitude for that route is calculated by multiplying in succession the amplitude for each of the successive events. Using this law we can rewrite Eq. (3.4) as

$$\langle x \mid s\rangle_{\text{both}} = \langle x \mid 1\rangle\langle 1 \mid s\rangle + \langle x \mid 2\rangle\langle 2 \mid s\rangle.$$

Now we wish to show that just using these principles we can calculate a much more complicated problem like the one shown in Fig. 3–2. Here we have two walls, one with two holes, 1 and 2, and another which has three holes, a, b, and c. Behind the second wall there is a detector at x, and we want to know the amplitude for a particle to arrive there. Well, one way you can find this is by calculating the superposition, or interference, of the waves that go through; but you can also do it by saying that there are six possible routes and superposing an amplitude for each. The electron can go through hole 1, then through hole a, and then to x; or it could go through hole 1, then through hole b, and then to x; and so on. According to our second principle, the amplitudes for alternative routes add, so we should

be able to write the amplitude from s to x as a sum of six separate amplitudes. On the other hand, using the third principle, each of these separate amplitudes can be written as a product of three amplitudes. For example, one of them is the amplitude for s to 1, times the amplitude for 1 to a, times the amplitude for a to x. Using our shorthand notation, we can write the complete amplitude to go from s to x as

$$\langle x \mid s\rangle = \langle x \mid a\rangle\langle a \mid 1\rangle\langle 1 \mid s\rangle + \langle x \mid b\rangle\langle b \mid 1\rangle\langle 1 \mid s\rangle + \cdots + \langle x \mid c\rangle\langle c \mid 2\rangle\langle 2 \mid s\rangle.$$

We can save writing by using the summation notation

$$\langle x \mid s\rangle = \sum_{\substack{i=1,2\\ \alpha=a,b,c}} \langle x \mid \alpha\rangle\langle \alpha \mid i\rangle\langle i \mid s\rangle. \tag{3.6}$$

In order to make any calculations using these methods, it is, naturally, necessary to know the amplitude to get from one place to another. We will give a rough idea of a typical amplitude. It leaves out certain things like the polarization of light or the spin of the electron, but aside from such features it is quite accurate. We give it so that you can solve problems involving various combinations of slits. Suppose a particle with a definite energy is going in empty space from a location $\boldsymbol{r}_1$ to a location $\boldsymbol{r}_2$. In other words, it is a free particle with no forces on it. Except for a numerical factor in front, the amplitude to go from $\boldsymbol{r}_1$ to $\boldsymbol{r}_2$ is

$$\langle \boldsymbol{r}_2 \mid \boldsymbol{r}_1\rangle = \frac{e^{i\boldsymbol{p}\cdot\boldsymbol{r}_{12}/\hbar}}{r_{12}}, \tag{3.7}$$

where $\boldsymbol{r}_{12} = \boldsymbol{r}_2 - \boldsymbol{r}_1$, and $\boldsymbol{p}$ is the momentum which is related to the energy E by the relativistic equation

$$p^2c^2 = E^2 - (m_0c^2)^2,$$

or the nonrelativistic equation

$$\frac{p^2}{2m} = \text{Kinetic energy}.$$

Equation (3.7) says in effect that the particle has wavelike properties, the amplitude propagating as a wave with a wave number equal to the momentum divided by $\hbar$.

In the most general case, the amplitude and the corresponding probability will also involve the time. For most of these initial discussions we will suppose that the source always emits the particles with a given energy so we will not need to worry about the time. But we could, in the general case, be interested in some other questions. Suppose that a particle is liberated at a certain place P at a certain time, and you would like to know the amplitude for it to arrive at some location, say $\boldsymbol{r}$, at some later time. This could be represented symbolically as the amplitude $\langle \boldsymbol{r}, t = t_1 \mid P, t = 0\rangle$. Clearly, this will depend upon both $\boldsymbol{r}$ and t. You will get different results if you put the detector in different places and measure at different times. This function of $\boldsymbol{r}$ and t, in general, satisfies a differential equation which is a wave equation. For example, in a nonrelativistic case it is the Schrödinger equation. One has then a wave equation analogous to the equation for electromagnetic waves or waves of sound in a gas. However, it must be emphasized that the wave function that satisfies the equation is not like a real wave in space; one cannot picture any kind of reality to this wave as one does for a sound wave.

Although one may be tempted to think in terms of "particle waves" when dealing with one particle, it is not a good idea, for if there are, say, two particles, the amplitude to find one at $\boldsymbol{r}_1$ and the other at $\boldsymbol{r}_2$ is not a simple wave in three-dimensional space, but depends on the *six* space variables $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. If we are, for example, dealing with two (or more) particles, we will need the following additional principle: Provided that the two particles do not interact, the amplitude that one particle will do one thing *and* the other one something else is the product of the two amplitudes that the two particles would do the two things separately. For example, if $\langle a \mid s_1\rangle$ is the amplitude for particle 1 to go from s_1 to a, and $\langle b \mid s_2\rangle$

is the amplitude for particle 2 to go from s_2 to b, the amplitude that *both* things will happen together is

$$\langle a \mid s_1 \rangle \langle b \mid s_2 \rangle.$$

There is one more point to emphasize. Suppose that we didn't know where the particles in Fig. 3–2 come from before arriving at holes 1 and 2 of the first wall. We can still make a prediction of what will happen beyond the wall (for example, the amplitude to arrive at x) provided that we are given two numbers: the amplitude to have arrived at 1 and the amplitude to have arrived at 2. In other words, because of the fact that the amplitude for successive events multiplies, as shown in Eq. (3.6), all you need to know to continue the analysis is two numbers—in this particular case $\langle 1 \mid s \rangle$ and $\langle 2 \mid s \rangle$. These two complex numbers are enough to predict all the future. That is what really makes quantum mechanics easy. It turns out that in later chapters we are going to do just such a thing when we specify a starting condition in terms of two (or a few) numbers. Of course, these numbers depend upon where the source is located and possibly other details about the apparatus, but given the two numbers, we do not need to know any more about such details.

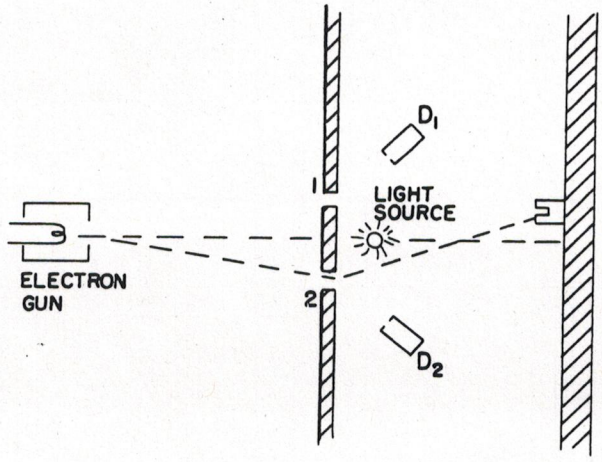

Fig. 3–3. An experiment to determine which hole the electron goes through.

3–2 The two-slit interference pattern

Now we would like to consider a matter which was discussed in some detail in Chapter 1. This time we will do it with the full glory of the amplitude idea to show you how it works out. We take the same experiment shown in Fig. 3–1, but now with the addition of a light source behind the two holes, as shown in Fig. 3–3. In Chapter 1, we discovered the following interesting result. If we looked behind slit 1 and saw a photon scattered from there, then the distribution obtained for the electrons at x in coincidence with these photons was the same as though slit 2 were closed. The total distribution for electrons that had been "seen" at either slit 1 or slit 2 was the sum of the separate distributions and was completely different from the distribution with the light turned off. This was true at least if we used light of short enough wavelength. If the wavelength was made longer so we could not be sure at which hole the scattering had occurred, the distribution became more like the one with the light turned off.

Let's examine what is happening by using our new notation and the principles of combining amplitudes. To simplify the writing, we can again let ϕ_1 stand for the amplitude that the electron will arrive at x by way of hole 1, that is,

$$\phi_1 = \langle x \mid 1 \rangle \langle 1 \mid s \rangle.$$

Similarly, we'll let ϕ_2 stand for the amplitude that the electron gets to the detector by way of hole 2:

$$\phi_2 = \langle x \mid 2 \rangle \langle 2 \mid s \rangle.$$

These are the amplitudes to go through the two holes and arrive at x if there is no light. Now if there is light, we ask ourselves the question: What is the amplitude for the process in which the electron starts at s and a photon is liberated by the

light source L, ending with the electron at x and a photon seen behind slit 1? Suppose that we observe the photon behind slit 1 by means of a detector D_1, as shown in Fig. 3–3, and use a similar detector D_2 to count photons scattered behind hole 2. There will be an amplitude for a photon to arrive at D_1 and an electron at x, and also an amplitude for a photon to arrive at D_2 and an electron at x. Let's try to calculate them.

Although we don't have the correct mathematical formula for all the factors that go into this calculation, you will see the spirit of it in the following discussion. First, there is the amplitude $\langle 1 \mid s\rangle$ that an electron goes from the source to hole 1. Then we can suppose that there is a certain amplitude that while the electron is at hole 1 it scatters a photon into the detector D_1. Let us represent this amplitude by a. Then there is the amplitude $\langle x \mid 1\rangle$ that the electron goes from slit 1 to the electron detector at x. The amplitude that the electron goes from s to x via slit 1 *and* scatters a photon into D_1 is then

$$\langle x \mid 1\rangle\, a\, \langle 1 \mid s\rangle.$$

Or, in our previous notation, it is just $a\phi_1$.

There is also some amplitude that an electron going through slit 2 will scatter a photon into counter D_1. You say, "That's impossible; how can it scatter into counter D_1 if it is only looking at hole 1?" If the wavelength is long enough, there are diffraction effects, and it is certainly possible. If the apparatus is built well and if we use photons of short wavelength, then the amplitude that a photon will be scattered into detector 1, from an electron at 2 is very small. But to keep the discussion general we want to take into account that there is always some such amplitude, which we will call b. Then the amplitude that an electron goes via slit 2 *and* scatters a photon into D_1 is

$$\langle x \mid 2\rangle\, b\, \langle 2 \mid s\rangle = b\phi_2.$$

The amplitude to find the electron at x and the photon in D_1 is the sum of two terms, one for each possible path for the electron. Each term is in turn made up of two factors: first, that the electron went through a hole, and second, that the photon is scattered by such an electron into detector 1; we have

$$\left\langle \begin{matrix}\text{electron at } x \\ \text{photon at } D_1\end{matrix} \,\middle|\, \begin{matrix}\text{electron from } s \\ \text{photon from } L\end{matrix} \right\rangle = a\phi_1 + b\phi_2. \tag{3.8}$$

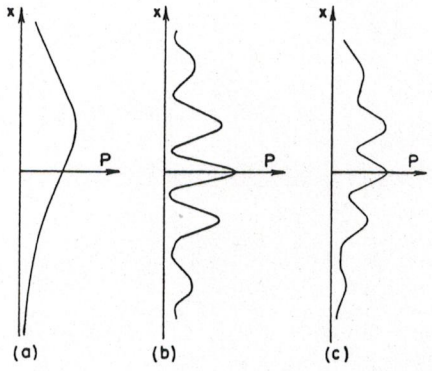

Fig. 3–4. The probability of counting an electron at x in coincidence with a photon at D in the experiment of Fig. 3 3: (a) for $b = 0$; (b) for $b = a$; (c) for $0 < b < a$.

We can get a similar expression when the photon is found in the other detector D_2. If we assume for simplicity that the system is symmetrical, then a is also the amplitude for a photon in D_2 when an electron passes through hole 2, and b is the amplitude for a photon in D_2 when the electron passes through hole 1. The corresponding total amplitude for a photon at D_2 and an electron at x is

$$\left\langle \begin{matrix}\text{electron at } x \\ \text{photon at } D_2\end{matrix} \,\middle|\, \begin{matrix}\text{electron from } s \\ \text{photon from } L\end{matrix} \right\rangle = a\phi_2 + b\phi_1. \tag{3.9}$$

Now we are finished. We can easily calculate the probability for various situations. Suppose that we want to know with what probability we get a count in D_1 and an electron at x. That will be the absolute square of the amplitude given in Eq. (3.8), namely, just $|a\phi_1 + b\phi_2|^2$. Let's look more carefully at this expression. First of all, if b is zero—which is the way we would like to design the apparatus—then the answer is simply $|\phi_1|^2$ diminished in total amplitude by the factor $|a|^2$. This is the probability distribution that you would get if there were only one hole—as shown in the graph of Fig. 3–4(a). On the other hand, if the wavelength is very long, the scattering behind hole 2 into D_1 may be just about the same as for hole 1. Although there may be some phases involved in a and b, we can ask about a simple case in which the two phases are equal. If a is practically equal to b, then the total probability becomes $|\phi_1 + \phi_2|^2$ multiplied by $|a|^2$, since the common factor a can be taken out. This, however, is just the probability

distribution we would have gotten without the photons at all. Therefore, in the case that the wavelength is very long—and the photon detection ineffective—you return to the original distribution curve which shows interference effects, as shown in Fig. 3–4(b). In the case that the detection is partially effective, there is an interference between a lot of ϕ_1 and a little of ϕ_2, and you will get an intermediate distribution such as is sketched in Fig. 3–4(c). Needless to say, if we look for coincidence counts of photons at D_2 and electrons at x, we will get the same kinds of results. If you remember the discussion in Chapter 1, you will see that these results give a quantitative description of what was described there.

Now we would like to emphasize an important point so that you will avoid a common error. Suppose that you only want the amplitude that the electron arrives at x, *regardless* of whether the photon was counted at D_1 or D_2. Should you add the amplitudes given in Eqs. (3.8) and (3.9)? No! You must *never add amplitudes for different and distinct final states.* Once the photon is accepted by one of the photon counters, we can always determine which alternative occurred if we want, without any further disturbance to the system. Each alternative has a probability completely independent of the other. To repeat, do not add amplitudes for different *final* conditions, where by "final" we mean at that moment the *probability* is desired—that is, when the experiment is "finished." You do add the amplitudes for the different *indistinguishable* alternatives inside the experiment, before the complete process is finished. At the end of the process you may say that you "don't want to look at the photon." That's your business, but you still do not add the amplitudes. Nature does not know what you are looking at, and she behaves the way she is going to behave whether you bother to take down the data or not. So here we must not add the amplitudes. We first square the amplitudes for all possible different final events and then sum. The correct result for an electron at x and a photon at either D_1 or D_2 is

$$\left|\left\langle \begin{matrix}\text{e at } x \\ \text{ph at } D_1\end{matrix} \middle| \begin{matrix}\text{e from } s \\ \text{ph from } L\end{matrix} \right\rangle\right|^2 + \left|\left\langle \begin{matrix}\text{e at } x \\ \text{ph at } D_2\end{matrix} \middle| \begin{matrix}\text{e from } s \\ \text{ph from } L\end{matrix} \right\rangle\right|^2$$
$$= |a\phi_1 + b\phi_2|^2 + |a\phi_2 + b\phi_1|^2. \qquad (3.10)$$

3–3 Scattering from a crystal

Our next example is a phenomenon in which we have to analyze the interference of probability amplitudes somewhat carefully. We look at the process of the scattering of neutrons from a crystal. Suppose we have a crystal which has a lot of atoms with nuclei at their centers, arranged in a periodic array, and a neutron beam that comes from far away. We can label the various nuclei in the crystal by an index i, where i runs over the integers 1, 2, 3, . . . N, with N equal to the total number of atoms. The problem is to calculate the probability of getting a neutron into a counter with the arrangement shown in Fig. 3–5. For any particular atom i, the amplitude that the neutron arrives at the counter C is the amplitude that the neutron gets from the source S to nucleus i, multiplied by the amplitude a that it gets scattered there, multiplied by the amplitude that it gets from i to the counter C. Let's write that down:

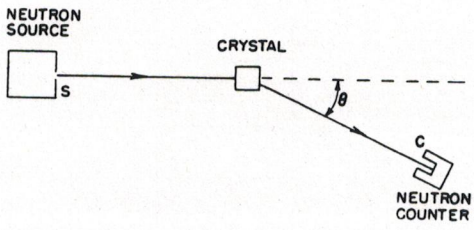

Fig. 3–5. Measuring the scattering of neutrons by a crystal.

$$\langle \text{neutron at } C \mid \text{neutron from } S\rangle_{\text{via } i} = \langle C \mid i\rangle\, a\, \langle i \mid S\rangle. \qquad (3.11)$$

In writing this equation we have assumed that the scattering amplitude a is the same for all atoms. We have here a large number of apparently indistinguishable routes. They are indistinguishable because a low-energy neutron is scattered from a nucleus without knocking the atom out of its place in the crystal—no "record" is left of the scattering. According to the earlier discussion, the total amplitude for a neutron at C involves a sum of Eq. (3.11) over all the atoms:

$$\langle \text{neutron at } C \mid \text{neutron from } S\rangle = \sum_{i=1}^{N} \langle C \mid i\rangle\, a\, \langle i \mid S\rangle. \qquad (3.12)$$

8

The Hamiltonian Matrix

Review: Chapter 49, Vol. I, *Modes*

8–1 Amplitudes and vectors

Before we begin the main topic of this chapter, we would like to describe a number of mathematical ideas that are used a lot in the literature of quantum mechanics. Knowing them will make it easier for you to read other books or papers on the subject. The first idea is the close mathematical resemblance between the equations of quantum mechanics and those of the scalar product of two vectors. You remember that if χ and ϕ are two states, the amplitude to start in ϕ and end up in χ can be written as a sum over a complete set of base states of the amplitude to go from ϕ into one of the base states and then from that base state out again into χ:

$$\langle\chi \mid \phi\rangle = \sum_{\text{all } i} \langle\chi \mid i\rangle\langle i \mid \phi\rangle. \tag{8.1}$$

We explained this in terms of a Stern-Gerlach apparatus, but we remind you that there is no need to have the apparatus. Equation (8.1) is a mathematical law that is just as true whether we put the filtering equipment in or not—it is not always necessary to imagine that the apparatus is there. We can think of it simply as a formula for the amplitude $\langle\chi \mid \phi\rangle$.

We would like to compare Eq. (8.1) to the formula for the dot product of two vectors $\boldsymbol{B}$ and $\boldsymbol{A}$. If $\boldsymbol{B}$ and $\boldsymbol{A}$ are ordinary vectors in three dimensions, we can write the dot product this way:

$$\sum_{\text{all } i} (\boldsymbol{B} \cdot \boldsymbol{e}_i)(\boldsymbol{e}_i \cdot \boldsymbol{A}), \tag{8.2}$$

with the understanding that the symbol $\boldsymbol{e}_i$ stands for the three unit vectors in the x, y, and z-directions. Then $\boldsymbol{B} \cdot \boldsymbol{e}_1$ is what we ordinarily call B_x; $\boldsymbol{B} \cdot \boldsymbol{e}_2$ is what we ordinarily call B_y; and so on. So Eq. (8.2) is equivalent to

$$B_xA_x + B_yA_y + B_zA_z,$$

which is the dot product $\boldsymbol{B} \cdot \boldsymbol{A}$.

Comparing Eqs. (8.1) and (8.2), we can see the following analogy: The states χ and ϕ correspond to the two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$. The base states i correspond to the special vectors $\boldsymbol{e}_i$ to which we refer all other vectors. Any vector can be represented as a linear combination of the three "base vectors" $\boldsymbol{e}_i$. Furthermore, if you know the coefficients of each "base vector" in this combination—that is, its three components—you know everything about a vector. In a similar way, any quantum mechanical state can be described completely by the amplitude $\langle i \mid \phi\rangle$ to go into the base states; and if you know these coefficients, you know everything there is to know about the state. Because of this close analogy, what we have called a "state" is often also called a "state vector."

Since the base vectors $\boldsymbol{e}_i$ are all at right angles, we have the relation

$$\boldsymbol{e}_i \cdot \boldsymbol{e}_j = \delta_{ij}. \tag{8.3}$$

This corresponds to the relations (5.25) among the base states i,

$$\langle i \mid j\rangle = \delta_{ij}. \tag{8.4}$$

You see now why one says that the base states i are all "orthogonal."

There is one minor difference between Eq. (8.1) and the dot product. We have that

$$\langle \phi \,|\, \chi \rangle = \langle \chi \,|\, \phi \rangle^*. \tag{8.5}$$

But in vector algebra,

$$\boldsymbol{A} \cdot \boldsymbol{B} = \boldsymbol{B} \cdot \boldsymbol{A}.$$

With the complex numbers of quantum mechanics we have to keep straight the order of the terms, whereas in the dot product, the order doesn't matter.

Now consider the following vector equation:

$$\boldsymbol{A} = \sum_i \boldsymbol{e}_i (\boldsymbol{e}_i \cdot \boldsymbol{A}). \tag{8.6}$$

It's a little unusual, but correct. It means the same thing as

$$\boldsymbol{A} = \sum_i A_i \boldsymbol{e}_i = A_x \boldsymbol{e}_x + A_y \boldsymbol{e}_y + A_z \boldsymbol{e}_z. \tag{8.7}$$

Notice, though, that Eq. (8.6) involves a quantity which is *different* from a dot product. A dot product is just a *number*, whereas Eq. (8.6) is a *vector* equation. One of the great tricks of vector analysis was to abstract away from the equations the idea of a *vector* itself. One might be similarly inclined to abstract a thing that is the analog of a "vector" from the quantum mechanical formula Eq. (8.1)—and one can indeed. We remove the $\langle \chi \,|$ from both sides Eq. (8.1) and write the following equation (don't get frightened—it's just a notation and in a few minutes you will find out what the symbols mean):

$$|\, \phi \rangle = \sum_i |\, i \rangle \langle i \,|\, \phi \rangle. \tag{8.8}$$

One thinks of the bracket $\langle \chi \,|\, \phi \rangle$ as being divided into two pieces. The second piece $|\, \phi \rangle$ is often called a *ket*, and the first piece $\langle \chi \,|$ is called a *bra* (put together, they make a "bra-ket"—a notation proposed by Dirac); the half-symbols $\langle \chi \,|$ and $|\, \phi \rangle$ are also called *state vectors*. In any case, they are *not* numbers, and, in general, we want the results of our calculations to come out as numbers; so such "unfinished" quantities are only part-way steps in our calculations.

It happens that until now we have written all our results in terms of numbers. How have we managed to avoid vectors? It is amusing to note that even in ordinary vector algebra we *could* make all equations involve only numbers. For instance, instead of a vector equation like

$$\boldsymbol{F} = m\boldsymbol{a},$$

we could always have written

$$\boldsymbol{C} \cdot \boldsymbol{F} = \boldsymbol{C} \cdot (m\boldsymbol{a}).$$

We have then an equation between dot products that is true for *any* vector $\boldsymbol{C}$. But if it is true for any $\boldsymbol{C}$, it hardly makes sense at all to keep writing the $\boldsymbol{C}$!

Now look at Eq. (8.1). It is an equation that is true for *any* χ. So to save writing, we should just leave *out* the χ and write Eq. (8.8) instead. It has the same information *provided* we understand that it should always be "finished" by "multiplying on the left by"—which simply means reinserting—some $\langle \chi \,|$ on both sides. So Eq. (8.8) means exactly the same thing as Eq. (8.1)—no more, no less. When you want numbers, you put in the $\langle \chi \,|$ you want.

Maybe you have already wondered about the ϕ in Eq. (8.8). Since the equation is true for *any* ϕ, why do we keep *it*? Indeed, Dirac suggests that the ϕ also can just as well be abstracted away, so that we have only

$$| = \sum_i |\, i \rangle \langle i \,|. \tag{8.9}$$

And this is the great law of quantum mechanics! (There is no analog in vector analysis.) It says that if you put *in* any two states χ and ϕ on the left and right of both sides, you *get back* Eq. (8.1). It is not really very useful, but it's a nice reminder that the equation is true for any two states.

8–2 Resolving state vectors

Let's look at Eq. (8.8) again; we can think of it in the following way. Any state vector $|\phi\rangle$ can be represented as a linear combination with suitable coefficients of a set of base "vectors"—or, if you prefer, as a superposition of "unit vectors" in suitable proportions. To emphasize that the coefficients $\langle i|\phi\rangle$ are just ordinary (complex) numbers, suppose we write

$$\langle i|\phi\rangle = C_i.$$

Then Eq. (8.8) is the same as

$$|\phi\rangle = \sum_i |i\rangle C_i. \tag{8.10}$$

We can write a similar equation for any other state vector, say $|\chi\rangle$, with, of course, different coefficients—say D_i. Then we have

$$|\chi\rangle = \sum_i |i\rangle D_i. \tag{8.11}$$

The D_i are just the amplitudes $\langle i|\chi\rangle$.

Suppose we had started by abstracting the ϕ from Eq. (8.1). We would have had

$$\langle\chi| = \sum_i \langle\chi|i\rangle\langle i|. \tag{8.12}$$

Remembering that $\langle\chi|i\rangle = \langle i|\chi\rangle^*$, we can write this as

$$\langle\chi| = \sum_i D_i^* \langle i|. \tag{8.13}$$

Now the interesting thing is that we can just *multiply* Eq. (8.13) and Eq. (8.10) to get back $\langle\chi|\phi\rangle$. When we do that, we have to be careful of the summation indices, because they are quite distinct in the two equations. Let's first rewrite Eq. (8.13) as

$$\langle\chi| = \sum_j D_j^* \langle j|,$$

which changes nothing. Then putting it together with Eq. (8.10), we have

$$\langle\chi|\phi\rangle = \sum_{ij} D_j^* \langle j|i\rangle C_i. \tag{8.14}$$

Remember, though, that $\langle j|i\rangle = \delta_{ij}$, so that in the sum we have left only the terms with $j = i$. We get

$$\langle\chi|\phi\rangle = \sum_i D_i^* C_i, \tag{8.15}$$

where, of course, $D_i^* = \langle i|\chi\rangle^* = \langle\chi|i\rangle$, and $C_i = \langle i|\phi\rangle$. Again we see the close analogy with the dot product

$$\boldsymbol{A}\cdot\boldsymbol{B} = \sum_i A_i B_i.$$

The only difference is the complex conjugate on D_i. So Eq. (8.15) says that if the state vectors $\langle\chi|$ and $|\phi\rangle$ are expanded in terms of the base vectors $\langle i|$ or $|i\rangle$, the amplitude to go from ϕ to χ is given by the kind of dot product in Eq. (8.15). This equation is, of course, just Eq. (8.1) written with different symbols. So we have just gone in a circle to get used to the new symbols.

We should perhaps emphasize again that while space vectors in three dimensions are described in terms of *three* orthogonal unit vectors, the base vectors $|i\rangle$ of the quantum mechanical states must range over the complete set applicable to any particular problem. Depending on the situation, two, or three, or five, or an infinite number of base states may be involved.

We have also talked about what happens when particles go through an apparatus. If we start the particles out in a certain state ϕ, then send them through

an apparatus, and afterward make a measurement to see if they are in state χ, the result is described by the amplitude

$$\langle \chi \mid A \mid \phi \rangle. \tag{8.16}$$

Such a symbol doesn't have a close analog in vector algebra. (It is closer to tensor algebra, but the analogy is not particularly useful.) We saw in Chapter 5, Eq. (5.32), that we could write (8.16) as

$$\langle \chi \mid A \mid \phi \rangle = \sum_{ij} \langle \chi \mid i \rangle \langle i \mid A \mid j \rangle \langle j \mid \phi \rangle. \tag{8.17}$$

This is just an example of the fundamental rule Eq. (8.9), used twice.

We also found that if another apparatus B was added in series with A, then we could write

$$\langle \chi \mid BA \mid \phi \rangle = \sum_{ijk} \langle \chi \mid i \rangle \langle i \mid B \mid j \rangle \langle j \mid A \mid k \rangle \langle k \mid \phi \rangle. \tag{8.18}$$

Again, this comes directly from Dirac's method of writing Eq. (8.9)—remember that we can always place a bar (|), which is just like the factor 1, between B and A.

Incidentally, we can think of Eq. (8.17) in another way. Suppose we think of the particle entering apparatus A in the state ϕ and coming out of A in the state ψ ("psi"). In other words, we could ask ourselves this question: Can we find a ψ such that the amplitude to get from ψ to χ is always identically and everywhere the same as the amplitude $\langle \chi \mid A \mid \phi \rangle$? The answer is yes. We want Eq. (8.17) to be replaced by

$$\langle \chi \mid \psi \rangle = \sum_{i} \langle \chi \mid i \rangle \langle i \mid \psi \rangle. \tag{8.19}$$

We can clearly do this if

$$\langle i \mid \psi \rangle = \sum_{j} \langle i \mid A \mid j \rangle \langle j \mid \phi \rangle = \langle i \mid A \mid \phi \rangle, \tag{8.20}$$

which determines ψ. "But it doesn't determine ψ," you say; "it only determines $\langle i \mid \psi \rangle$." However, $\langle i \mid \psi \rangle$ *does* determine ψ, because if you have all the coefficients that relate ψ to the base states i, then ψ is uniquely defined. In fact, we can play with our notation and write the last term of Eq. (8.20) as

$$\langle i \mid \psi \rangle = \sum_{j} \langle i \mid j \rangle \langle j \mid A \mid \phi \rangle. \tag{8.21}$$

Then, since this equation is true for all i, we can write simply

$$\mid \psi \rangle = \sum_{j} \mid j \rangle \langle j \mid A \mid \phi \rangle. \tag{8.22}$$

Then we can say: "The state ψ is what we get if we start with ϕ and go through the apparatus A."

One final example of the tricks of the trade. We start again with Eq. (8.17). Since it is true for any χ and ϕ, we can drop them both! We then get†

$$A = \sum_{ij} \mid i \rangle \langle i \mid A \mid j \rangle \langle j \mid. \tag{8.23}$$

What does it mean? It means no more, no less, than what you get if you put back the ϕ and χ. As it stands, it is an "open" equation and incomplete. If we multiply it "on the left" by $\mid \phi \rangle$, it becomes

$$A \mid \phi \rangle = \sum_{ij} \mid i \rangle \langle i \mid A \mid j \rangle \langle j \mid \phi \rangle, \tag{8.24}$$

---

† You might think we should write $|A|$ instead of just A. But then it would look like the symbol for "absolute value of A," so the bars are usually dropped. In general, the bar (|) behaves much like the factor one.

which is just Eq. (8.22) all over again. In fact, we could have just dropped the j's from that equation and written

$$|\psi\rangle = A\,|\phi\rangle. \tag{8.25}$$

The symbol A is neither an amplitude, nor a vector; it is a new kind of thing called an *operator*. It is something which "operates on" a state to produce a new state—Eq. (8.25) says that $|\psi\rangle$ is what results if A operates on $|\phi\rangle$. Again, it is still an open equation until it is completed with some bra like $\langle\chi|$ to give

$$\langle\chi|\psi\rangle = \langle\chi|A|\phi\rangle. \tag{8.26}$$

The operator A is, of course, described completely if we give the matrix of amplitudes $\langle i|A|j\rangle$—also written A_{ij}—in terms of any set of base vectors.

We have really added nothing new with all of this new mathematical notation. One reason for bringing it all up was to show you the way of writing pieces of equations, because in many books you will find the equations written in the incomplete forms, and there's no reason for you to be paralyzed when you come across them. If you prefer, you can always add the missing pieces to make an equation between numbers that will look like something more familiar.

Also, as you will see, the "bra" and "ket" notation is a very convenient one. For one thing, we can from now on identify a state by giving its state vector. When we want to refer to a state of definite momentum p we can say: "the state $|p\rangle$". Or we may speak of some arbitrary state $|\psi\rangle$. For consistency we will always use the ket, writing $|\psi\rangle$, to identify a state. (It is, of course an arbitrary choice; we could equally well have chosen to use the bra, $\langle\psi|$.)

8–3 What are the base states of the world?

We have discovered that any state in the world can be represented as a superposition—a linear combination with suitable coefficients—of base states. You may ask, first of all, *what* base states? Well, there are many different possibilities. You can, for instance, project a spin in the z-direction or in some other direction. There are many, many different *representations*, which are the analogs of the different *coordinate systems* one can use to represent ordinary vectors. Next, *what* coefficients? Well, that depends on the physical circumstances. Different sets of coefficients correspond to different physical conditions. The important thing to know about is the "space" in which you are working—in other words, what the base states mean physically. So the first thing you have to know about, in general, is what the base states are like. Then you can understand how to describe a situation in terms of these base states.

We would like to look ahead a little and speak a bit about what the general quantum mechanical description of nature is going to be—in terms of the now current ideas of physics, anyway. First, one decides on a particular representation for the base states—different representations are always possible. For example, for a spin one-half particle we can use the plus and minus states with respect to the z-axis. But there's nothing special about the z-axis—you can take any other axis you like. For consistency we'll always pick the z-axis, however. Suppose we begin with a situation with one electron. In addition to the two possibilities for the spin ("up" and "down" along the z-direction), there is also the momentum of the electron. We pick a set of base states, each corresponding to one value of the momentum. What if the electron doesn't have a definite momentum? That's all right; we're just saying what the *base* states are. If the electron hasn't got a definite momentum, it has some amplitude to have one momentum and another amplitude to have another momentum, and so on. And if it is not necessarily spinning up, it has some amplitude to be spinning up going at this momentum, and some amplitude to be spinning down going at that momentum, and so on. The complete description of an electron, *so far as we know*, requires only that the base states be described by the *momentum* and the *spin*. So one acceptable set of base states $|i\rangle$ for a single electron refer to different values of the momentum and

whether the spin is up or down. Different mixtures of amplitudes—that is, different combinations of the C's describe different circumstances. What any particular electron is doing is described by telling with what amplitude it has an up-spin or a down-spin and one momentum or another—for all possible momenta. So you can see what is involved in a complete quantum mechanical description of a single electron.

What about systems with more than one electron? Then the base states get more complicated. Let's suppose that we have two electrons. We have, first of all, four possible states with respect to spin: both electrons spinning up, the first one down and the second one up, the first one up and the second one down, or both down. Also we have to specify that the first electron has the momentum p_1, and the second electron, the momentum p_2. The base states for two electrons require the specification of two momenta and two spin characters. With seven electrons, we have to specify seven of each.

If we have a proton and an electron, we have to specify the spin direction of the proton and its momentum, and the spin direction of the electron and its momentum. At least that's approximately true. *We do not really know* what the correct representation is for the world. It is all very well to start out by supposing that if you specify the spin in the electron and its momentum, and likewise for a proton, you will have the base states; but what about the "guts" of the proton? Let's look at it this way. In a hydrogen atom which has one proton and one electron, we have many different base states to describe—up and down spins of the proton and electron and the various possible momenta of the proton and electron. Then there are different combinations of amplitudes C_i which together describe the character of the hydrogen atom in different states. But suppose we look at the whole hydrogen atom as a "particle." If we didn't know that the hydrogen atom was made out of a proton and an electron, we might have started out and said: "Oh, I know what the base states are—they correspond to a particular momentum of the hydrogen atom." No, because the hydrogen atom has internal parts. It may, therefore, have various states of different internal energy, and describing the real nature requires more detail.

The question is: Does a proton have internal parts? Do we have to describe a proton by giving all possible states of protons, and mesons, and strange particles? We don't know. And even though we suppose that the electron is simple, so that all we have to tell about it is its momentum and its spin, maybe tomorrow we will discover that the electron also has inner gears and wheels. It would mean that our representation is incomplete, or wrong, or approximate—in the same way that a representation of the hydrogen atom which describes only its momentum would be incomplete, because it disregarded the fact that the hydrogen atom could have become excited inside. If an electron could become excited inside and turn into something else like, for instance, a muon, then it would be described not just by giving the states of the new particle, but presumably in terms of some more complicated internal wheels. The *main problem in the study of the fundamental particles today* is to discover what are the correct representations for the description of nature. At the present time, we *guess* that for the electron it is enough to specify its momentum and spin. We also guess that there is an idealized proton which has its π-mesons, and k-mesons, and so on, that all have to be specified. Several dozen particles—that's crazy! The question of what *is* a fundamental particle and what *is not* a fundamental particle—a subject you hear so much about these days—is the question of what is the final *representation* going to look like in the ultimate quantum mechanical description of the world. Will the electron's momentum still be the right thing with which to describe nature? Or even, should the whole question be put this way at all! This question must always come up in any scientific investigation. At any rate, we see a problem—how to find a representation. We don't know the answer. We don't even know whether we have the "right" problem, but if we do, we must first attempt to find out whether any particular particle is "fundamental" or not.

In the nonrelativistic quantum mechanics—if the energies are not too high, so that you don't disturb the inner workings of the strange particles and so forth—

you can do a pretty good job without worrying about these details. You can just decide to specify the momenta and spins of the electrons and of the nuclei; then everything will be all right. In most chemical reactions and other low-energy happenings, nothing goes on in the nuclei; they don't get excited. Furthermore, if a hydrogen atom is moving slowly and bumping quietly against other hydrogen atoms—never getting excited inside, or radiating, or anything complicated like that, but staying always in the ground state of energy for internal motion—you can use an approximation in which you talk about the hydrogen atom as one object, or particle, and not worry about the fact that it *can* do something inside. This will be a good approximation as long as the kinetic energy in any collision is well below 10 electron volts—the energy required to excite the hydrogen atom to a different internal state. We will often be making an approximation in which we do not include the possibility of inner motion, thereby decreasing the number of details that we have to put into our base states. Of course, we then omit some phenomena which would appear (usually) at some higher energy, but by making such approximations we can simplify very much the analysis of physical problems. For example, we can discuss the collision of two hydrogen atoms at low energy—or any chemical process—without worrying about the fact that the atomic nuclei could be excited. To summarize, then, when we can neglect the effects of any internal excited states of a particle we can choose a base set which are the states of definite momentum and z-component of angular momentum.

One problem then in describing nature is to find a suitable representation for the base states. But that's only the beginning. We still want to be able to say what "happens." If we know the "condition" of the world at one moment, we would like to know the condition at a later moment. So we also have to find the laws that determine how things change with time. We now address ourselves to this second part of the framework of quantum mechanics—how states change with time.

8–4 How states change with time

We have already talked about how we can represent a situation in which we put something through an apparatus. Now one convenient, delightful "apparatus" to consider is merely a wait of a few minutes; that is, you prepare a state ϕ, and then before you analyze it, you just let it sit. Perhaps you let it sit in some particular electric or magnetic field—it depends on the physical circumstances in the world. At any rate, whatever the conditions are, you let the object sit from time t_1 to time t_2. Suppose that it is let out of your first apparatus in the condition ϕ at t_1. And then it goes through an "apparatus," but the "apparatus" consists of just delay until t_2. During the delay, various things could be going on—external forces applied or other shenanigans—so that something is happening. At the end of the delay, the amplitude to find the thing in some state χ is no longer exactly the same as it would have been without the delay. Since "waiting" is just a special case of an "apparatus," we can describe what happens by giving an amplitude with the same form as Eq. (8.17). Because the operation of "waiting" is especially important, we'll call it U instead of A, and to specify the starting and finishing times t_1 and t_2, we'll write $U(t_2, t_1)$. The amplitude we want is

$$\langle \chi \mid U(t_2, t_1) \mid \phi \rangle. \tag{8.27}$$

Like any other such amplitude, it can be represented in some base system or other by writing it

$$\sum_{ij} \langle \chi \mid i \rangle \langle i \mid U(t_2, t_1) \mid j \rangle \langle j \mid \phi \rangle. \tag{8.28}$$

Then U is completely described by giving the whole set of amplitudes—the matrix

$$\langle i \mid U(t_2, t_1) \mid j \rangle. \tag{8.29}$$

We can point out, incidentally, that the matrix $\langle i \mid U(t_2, t_1) \mid j \rangle$ gives much more detail than may be needed. The high-class theoretical physicist working in

high-energy physics considers problems of the following general nature (because it's the way experiments are usually done). He starts with a couple of particles, like a proton and a proton, coming together from infinity. (In the lab, usually one particle is standing still, and the other comes from an accelerator that is practically at infinity on atomic level.) The things go crash and out come, say, two k-mesons, six π-mesons, and two neutrons in certain directions with certain momenta. What's the amplitude for this to happen? The mathematics looks like this: The ϕ-state specifies the spins and momenta of the incoming particles. The χ would be the question about what comes out. For instance, with what amplitude do you get the six mesons going in such-and-such directions, and the two neutrons going off in these directions, with their spins so-and-so. In other words, χ would be specified by giving all the momenta, and spins, and so on of the final products. Then the job of the theorist is to calculate the amplitude (8.27). However, he is really only interested in the special case that t_1 is $-\infty$ and t_2 is $+\infty$. (There is no experimental evidence on the details of the process, only on what comes in and what goes out.) The limiting case of $U(t_2, t_1)$ as $t_1 \to -\infty$ and $t_2 \to +\infty$ is called S, and what he wants is

$$\langle \chi \,|\, S \,|\, \phi \rangle.$$

Or, using the form (8.28), he would calculate the matrix

$$\langle i \,|\, S \,|\, j \rangle,$$

which is called the *S-matrix*. So if you see a theoretical physicist pacing the floor and saying, "All I have to do is calculate the S-matrix," you will know what he is worried about.

How to analyze—how to specify the laws for—the S-matrix is an interesting question. In relativistic quantum mechanics for high energies, it is done one way, but in nonrelativistic quantum mechanics it can be done another way, which is very convenient. (This other way can also be done in the relativistic case, but then it is not so convenient.) It is to work out the U-matrix for a small interval of time—in other words for t_2 and t_1 close together. If we can find a sequence of such U's for successive intervals of time we can watch how things go as a function of time. You can appreciate immediately that this way is not so good for relativity, because you don't want to have to specify how everything looks "simultaneously" everywhere. But we won't worry about that—we're just going to worry about nonrelativistic mechanics.

Suppose we think of the matrix U for a delay from t_1 until t_3 which is greater than t_2. In other words, let's take three successive times: t_1 less than t_2 less than t_3. Then we claim that the matrix that goes between t_1 and t_3 is the *product* in succession of what happens when you delay from t_1 until t_2 and then from t_2 until t_3. It's just like the situation when we had two apparatuses B and A in series. We can then write, following the notation of Section 5–6,

$$U(t_3, t_1) = U(t_3, t_2) \cdot U(t_2, t_1). \tag{8.30}$$

In other words, we can analyze any time interval if we can analyze a sequence of short time intervals in between. We just multiply together all the pieces; that's the way that quantum mechanics is analyzed nonrelativistically.

Our problem, then, is to understand the matrix $U(t_2, t_1)$ for an infinitesimal time interval—for $t_2 = t_1 + \Delta t$. We ask ourselves this: If we have a state ϕ now, what does the state look like an infinitesimal time Δt later? Let's see how we write that out. Call the state at the time t, $|\, \psi(t) \rangle$ (we show the time dependence of ψ to be perfectly clear that we mean the condition at the time t). Now we ask the question: What is the condition after the small interval of time Δt later? The answer is

$$|\, \psi(t + \Delta t) \rangle = U(t + \Delta t, t) \,|\, \psi(t) \rangle. \tag{8.31}$$

This means the same as we meant by (8.25), namely, that the amplitude to

find χ at the time $t + \Delta t$, is

$$\langle \chi \,|\, \psi(t + \Delta t)\rangle = \langle \chi \,|\, U(t + \Delta t, t) \,|\, \psi(t)\rangle. \tag{8.32}$$

Since we're not yet too good at these abstract things, let's project our amplitudes into a definite representation. If we multiply both sides of Eq. (8.31) by $\langle i \,|$, we get

$$\langle i \,|\, \psi(t + \Delta t)\rangle = \langle i \,|\, U(t + \Delta t, t) \,|\, \psi(t)\rangle. \tag{8.33}$$

We can also resolve the $|\, \psi(t)\rangle$ into base states and write

$$\langle i \,|\, \psi(t + \Delta t)\rangle = \sum_j \langle i \,|\, U(t + \Delta t, t) \,|\, j\rangle\langle j \,|\, \psi(t)\rangle. \tag{8.34}$$

We can understand Eq. (8.34) in the following way. If we let $C_i(t) = \langle i \,|\, \psi(t)\rangle$ stand for the amplitude to be in the base state i at the time t, then we can think of this amplitude (just a *number*, remember!) varying with time. Each C_i becomes a function of t. And we also have some information on *how* the amplitudes C_i vary with time. Each amplitude at $(t + \Delta t)$ is proportional to *all of the other* amplitudes at t multiplied by a set of coefficients. Let's call the U-matrix U_{ij}, by which we mean

$$U_{ij} = \langle i \,|\, U \,|\, j\rangle.$$

Then we can write Eq. (8.34) as

$$C_i(t + \Delta t) = \sum_j U_{ij}(t + \Delta t, t) C_j(t). \tag{8.35}$$

This, then, is how the dynamics of quantum mechanics is going to look.

We don't know much about the U_{ij} yet, except for one thing. We know that if Δt goes to zero, nothing can happen—we should get just the original state. So, $U_{ii} \to 1$ and $U_{ij} \to 0$, if $i \neq j$. In other words, $U_{ij} \to \delta_{ij}$ for $\Delta t \to 0$. Also, we can suppose that for small Δt, each of the coefficients U_{ij} should differ from δ_{ij} by amounts proportional to Δt; so we can write

$$U_{ij} = \delta_{ij} + K_{ij}\,\Delta t. \tag{8.36}$$

However, it is usual to take the factor $(-i/\hbar)$† out of the coefficients K_{ij}, for historical and other reasons; we prefer to write

$$U_{ij}(t + \Delta t, t) = \delta_{ij} - \frac{i}{\hbar} H_{ij}(t)\,\Delta t. \tag{8.37}$$

It is, of course, the same as Eq. (8.36) and, if you wish, just defines the coefficients $H_{ij}(t)$. The terms H_{ij} are just the derivatives with respect to t_2 of the coefficients $U_{ij}(t_2, t_1)$, evaluated at $t_2 = t_1 = t$.

Using this form for U in Eq. (8.35), we have

$$C_i(t + \Delta t) = \sum_j \left[\delta_{ij} - \frac{i}{\hbar} H_{ij}(t)\,\Delta t\right] C_j(t). \tag{8.38}$$

Taking the sum over the δ_{ij} term, we get just $C_i(t)$, which we can put on the other side of the equation. Then dividing by Δt, we have what we recognize as a derivative

$$\frac{C_i(t + \Delta t) - C_i(t)}{\Delta t} = -\frac{i}{\hbar} \sum_j H_{ij}(t) C_j(t)$$

or

$$i\hbar \frac{dC_i(t)}{dt} = \sum_j H_{ij}(t) C_j(t). \tag{8.39}$$

---

† We are in a bit of trouble here with notation. In the factor $(-i/\hbar)$, the i means the imaginary unit $\sqrt{-1}$, and *not* the *index* i that refers to the ith base state! We hope that you won't find it too confusing.

You remember that $C_i(t)$ is the amplitude $\langle i \mid \psi \rangle$ to find the state ψ in one of the base states i (at the time t). So Eq. (8.39) tells us how each of the coefficients $\langle i \mid \psi \rangle$ varies with time. But that is the same as saying that Eq. (8.39) tells us how the state ψ varies with time, since we are describing ψ in terms of the amplitudes $\langle i \mid \psi \rangle$. The variation of ψ in time is described in terms of the matrix H_{ij}, which has to include, of course, the things we are doing to the system to cause it to change. If we know the H_{ij}—which contains the physics of the situation and can, in general, depend on the time—we have a complete description of the behavior in time of the system. Equation (8.39) is then the quantum mechanical law for the dynamics of the world.

(We should say that we will always take a set of base states which are fixed and do not vary with time. There are people who use base states that also vary. However, that's like using a rotating coordinate system in mechanics, and we don't want to get involved in such complications.)

8–5 The Hamiltonian matrix

The idea, then, is that to describe the quantum mechanical world we need to pick a set of base states i and to write the physical laws by giving the matrix of coefficients H_{ij}. Then we have everything—we can answer any question about what will happen. So we have to learn what the rules are for finding the H's to go with any physical situation—what corresponds to a magnetic field, or an electric field, and so on. And that's the hardest part. For instance, for the new strange particles, we have no idea what H_{ij}'s to use. In other words, no one knows the *complete* H_{ij} for the whole world. (Part of the difficulty is that one can hardly hope to discover the H_{ij} when no one even knows what the base states are!) We do have excellent approximations for nonrelativistic phenomena and for some other special cases. In particular, we have the forms that are needed for the motions of electrons in atoms—to describe chemistry. But we don't know the full true H for the whole universe.

The coefficients H_{ij} are called *the Hamiltonian matrix* or, for short, just *the Hamiltonian.* (How Hamilton, who worked in the 1830's, got his name on a quantum mechanical matrix is a tale of history.) It would be much better called the *energy matrix*, for reasons that will become apparent as we work with it. So *the* problem is: Know your Hamiltonian!

The Hamiltonian has one property that can be deduced right away, namely, that

$$H_{ij}^* = H_{ji}. \tag{8.40}$$

This follows from the condition that the total probability that the system is in *some* state does not change. If you start with a particle—an object or the world—then you've still got it as time goes on. The total probability of finding it somewhere is

$$\sum_i |C_i(t)|^2,$$

which must not vary with time. If this is to be true for any starting condition ϕ, then Eq. (8.40) must also be true.

As our first example, we take a situation in which the physical circumstances are not changing with time; we mean the *external* physical conditions, so that H is independent of time. Nobody is turning magnets on and off. We also pick a system for which only one base state is required for the description; it is an approximation we could make for a hydrogen atom at rest, or something similar. Equation (8.39) then says

$$i\hbar \frac{dC_1}{dt} = H_{11}C_1. \tag{8.41}$$

Only one equation—that's all! And if H_{11} is constant, this differential equation is easily solved to give

$$C_1 = (\text{const})e^{-(i/\hbar)H_{11}t}. \tag{8.42}$$

This is the time dependence of a state with a definite energy $E = H_{11}$. You see why H_{ij} ought to be called the energy matrix. It is the generalization of the energy for more complex situations.

Next, to understand a little more about what the equations mean, we look at a system which has two base states. Then Eq. (8.39) reads

$$i\hbar \frac{dC_1}{dt} = H_{11}C_1 + H_{12}C_2,$$
$$i\hbar \frac{dC_2}{dt} = H_{21}C_1 + H_{22}C_2. \tag{8.43}$$

If the H's are again independent of time, you can easily solve these equations. We leave you to try for fun, and we'll come back and do them later. Yes, you can solve the quantum mechanics without knowing the H's, so long as they are independent of time.

8–6 The ammonia molecule

We want now to show you how the dynamical equation of quantum mechanics can be used to describe a particular physical circumstance. We have picked an interesting but simple example in which, by making some reasonable guesses about the Hamiltonian, we can work out some important—and even practical—results. We are going to take a situation describable by two states: the ammonia molecule.

The ammonia molecule has one nitrogen atom and three hydrogen atoms located in a plane below the nitrogen so that the molecule has the form of a pyramid, as drawn in Fig. 8–1(a). Now this molecule, like any other, has an infinite number of states. It can spin around any possible axis; it can be moving in any direction; it can be vibrating inside, and so on, and so on. It is, therefore, not a two-state system at all. But we want to make an approximation that all other states remain fixed, because they don't enter into what we are concerned with at the moment. We will consider only that the molecule is spinning around its axis of symmetry (as shown in the figure), that it has zero translational momentum, and that it is vibrating as little as possible. That specifies all conditions except one: *there are still the two possible positions for the nitrogen atom*—the nitrogen may be on one side of the plane of hydrogen atoms or on the other, as shown in Fig. 8–1(a) and (b). So we will discuss the molecule as though it were a two-state system. We mean that there are only two states we are going to really worry about, all other things being assumed to stay put. You see, even if we know that it is spinning with a certain angular momentum around the axis and that it is moving with a certain momentum and vibrating in a definite way, there are still two possible states. We will say that the molecule is in the state $|\,1\rangle$ when the nitrogen is "up," as in Fig. 8–1(a), and is in the state $|\,2\rangle$ when the nitrogen is "down," as in (b). The states $|\,1\rangle$ and $|\,2\rangle$ will be taken as the set of base states for our analysis of the behavior of the ammonia molecule. At any moment, the actual state $|\,\psi\rangle$ of the molecule can be represented by giving $C_1 = \langle 1\,|\,\psi\rangle$, the amplitude to be in state $|\,1\rangle$, and $C_2 = \langle 2\,|\,\psi\rangle$, the amplitude to be in state $|\,2\rangle$. Then, using Eq. (8.8) we can write the state vector $|\,\psi\rangle$ as

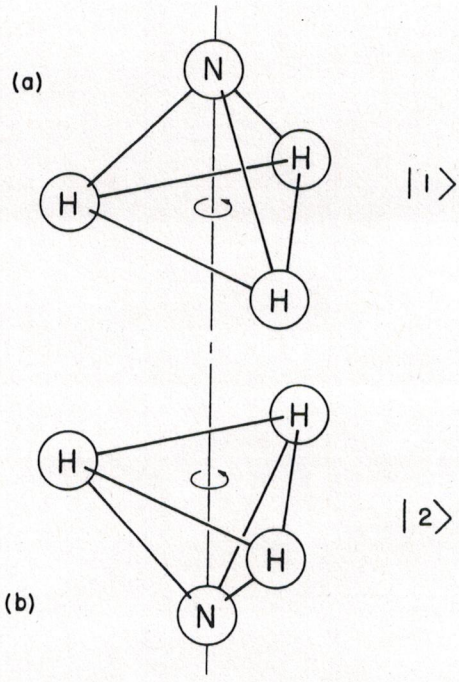

Fig. 8–1. Two equivalent geometric arrangements of the ammonia molecule.

$$|\,\psi\rangle = |\,1\rangle\langle 1\,|\,\psi\rangle + |\,2\rangle\langle 2\,|\,\psi\rangle$$

or

$$|\,\psi\rangle = |\,1\rangle C_1 + |\,2\rangle C_2. \tag{8.44}$$

Now the interesting thing is that if the molecule is known to be in some state at some instant, it will *not* be in the same state a little while later. The two C-coefficients will be changing with time according to the equations (8.43)—which hold for any two-state system. Suppose, for example, that you had made some observation—or had made some selection of the molecules—so that you *know* that the molecule is *initially* in the state $|\,1\rangle$. At some later time, there is some chance that it will be found in state $|\,2\rangle$. To find out what this chance is, we have to solve the differential equation which tells us how the amplitudes change with time.

This is the time dependence of a state with a definite energy $E = H_{11}$. You see why H_{ij} ought to be called the energy matrix. It is the generalization of the energy for more complex situations.

Next, to understand a little more about what the equations mean, we look at a system which has two base states. Then Eq. (8.39) reads

$$i\hbar \frac{dC_1}{dt} = H_{11}C_1 + H_{12}C_2,$$
$$i\hbar \frac{dC_2}{dt} = H_{21}C_1 + H_{22}C_2. \tag{8.43}$$

If the H's are again independent of time, you can easily solve these equations. We leave you to try for fun, and we'll come back and do them later. Yes, you can solve the quantum mechanics without knowing the H's, so long as they are independent of time.

8–6 The ammonia molecule

We want now to show you how the dynamical equation of quantum mechanics can be used to describe a particular physical circumstance. We have picked an interesting but simple example in which, by making some reasonable guesses about the Hamiltonian, we can work out some important—and even practical—results. We are going to take a situation describable by two states: the ammonia molecule.

The ammonia molecule has one nitrogen atom and three hydrogen atoms located in a plane below the nitrogen so that the molecule has the form of a pyramid, as drawn in Fig. 8–1(a). Now this molecule, like any other, has an infinite number of states. It can spin around any possible axis; it can be moving in any direction; it can be vibrating inside, and so on, and so on. It is, therefore, not a two-state system at all. But we want to make an approximation that all other states remain fixed, because they don't enter into what we are concerned with at the moment. We will consider only that the molecule is spinning around its axis of symmetry (as shown in the figure), that it has zero translational momentum, and that it is vibrating as little as possible. That specifies all conditions except one: *there are still the two possible positions for the nitrogen atom*—the nitrogen may be on one side of the plane of hydrogen atoms or on the other, as shown in Fig. 8–1(a) and (b). So we will discuss the molecule as though it were a two-state system. We mean that there are only two states we are going to really worry about, all other things being assumed to stay put. You see, even if we know that it is spinning with a certain angular momentum around the axis and that it is moving with a certain momentum and vibrating in a definite way, there are still two possible states. We will say that the molecule is in the state $|\,1\rangle$ when the nitrogen is "up," as in Fig. 8–1(a), and is in the state $|\,2\rangle$ when the nitrogen is "down," as in (b). The states $|\,1\rangle$ and $|\,2\rangle$ will be taken as the set of base states for our analysis of the behavior of the ammonia molecule. At any moment, the actual state $|\,\psi\rangle$ of the molecule can be represented by giving $C_1 = \langle 1\,|\,\psi\rangle$, the amplitude to be in state $|\,1\rangle$, and $C_2 = \langle 2\,|\,\psi\rangle$, the amplitude to be in state $|\,2\rangle$. Then, using Eq. (8.8) we can write the state vector $|\,\psi\rangle$ as

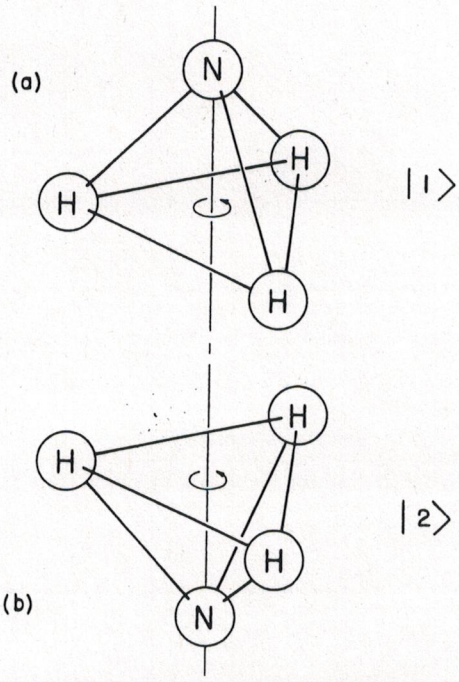

Fig. 8–1. Two equivalent geometric arrangements of the ammonia molecule.

$$|\,\psi\rangle = |\,1\rangle\langle 1\,|\,\psi\rangle + |\,2\rangle\langle 2\,|\,\psi\rangle$$

or

$$|\,\psi\rangle = |\,1\rangle C_1 + |\,2\rangle C_2. \tag{8.44}$$

Now the interesting thing is that if the molecule is known to be in some state at some instant, it will *not* be in the same state a little while later. The two C-coefficients will be changing with time according to the equations (8.43)—which hold for any two-state system. Suppose, for example, that you had made some observation—or had made some selection of the molecules—so that you *know* that the molecule is *initially* in the state $|\,1\rangle$. At some later time, there is some chance that it will be found in state $|\,2\rangle$. To find out what this chance is, we have to solve the differential equation which tells us how the amplitudes change with time.

The only trouble is that we don't know what to use for the coefficients H_{ij} in Eq. (8.43). There are some things we *can* say, however. Suppose that once the molecule was in the state $|1\rangle$ there was no chance that it could ever get into $|2\rangle$, and vice versa. Then H_{12} and H_{21} would both be zero, and Eq. (8.43) would read

$$i\hbar \frac{dC_1}{dt} = H_{11}C_1, \qquad i\hbar \frac{dC_2}{dt} = H_{22}C_2.$$

We can easily solve these two equations; we get

$$C_1 = (\text{const})e^{-(i/\hbar)H_{11}t}, \qquad C_2 = (\text{const})e^{-(i/\hbar)H_{22}t}. \tag{8.45}$$

These are just the amplitudes for *stationary* states with the energies $E_1 = H_{11}$ and $E_2 = H_{22}$. We note, however, that for the ammonia molecule the two states $|1\rangle$ and $|2\rangle$ have a definite symmetry. If nature is at all reasonable, the matrix elements H_{11} and H_{22} must be equal. We'll call them both E_0, because they correspond to the energy the states would have if H_{12} and H_{21} were zero. But Eqs. (8.45) do not tell us what ammonia really does. It turns out that it is possible for the nitrogen to push its way through the three hydrogens and flip to the other side. It is quite difficult; to get half-way through requires a lot of energy. How can it get through if it hasn't got enough energy? There is *some* amplitude that it *will* penetrate the energy barrier. It is possible in quantum mechanics to sneak quickly across a region which is illegal energetically. There is, therefore, some small amplitude that a molecule which starts in $|1\rangle$ will get to the state $|2\rangle$. The coefficients H_{12} and H_{21} are not really zero. Again, by symmetry, they should both be the same—at least in magnitude. In fact, we already know that, in general, H_{ij} must be equal to the complex conjugate of H_{ji}, so they can differ only by a phase. It turns out, as you will see, that there is no loss of generality if we take them equal to each other. For later convenience we set them equal to a negative number; we take $H_{12} = H_{21} = -A$. We then have the following pair of equations:

$$i\hbar \frac{dC_1}{dt} = E_0C_1 - AC_2, \tag{8.46}$$

$$i\hbar \frac{dC_2}{dt} = E_0C_2 - AC_1. \tag{8.47}$$

These equations are simple enough and can be solved in any number of ways. One convenient way is the following. Taking the sum of the two, we get

$$i\hbar \frac{d}{dt}(C_1 + C_2) = (E_0 - A)(C_1 + C_2),$$

whose solution is

$$C_1 + C_2 = ae^{-(i/\hbar)(E_0 - A)t}. \tag{8.48}$$

Then, taking the difference of (8.46) and (8.47), we find that

$$i\hbar \frac{d}{dt}(C_1 - C_2) = (E_0 + A)(C_1 - C_2),$$

which gives

$$C_1 - C_2 = be^{-(i/\hbar)(E_0 + A)t}. \tag{8.49}$$

We have called the two integration constants a and b; they are, of course, to be chosen to give the appropriate starting condition for any particular physical problem. Now, by adding and subtracting (8.48) and (8.49), we get C_1 and C_2:

$$C_1(t) = \frac{a}{2}e^{-(i/\hbar)(E_0 - A)t} + \frac{b}{2}e^{-(i/\hbar)(E_0 + A)t}, \tag{8.50}$$

$$C_2(t) = \frac{a}{2}e^{-(i/\hbar)(E_0 - A)t} - \frac{b}{2}e^{-(i/\hbar)(E_0 + A)t}. \tag{8.51}$$

They are the same except for the sign of the second term.

We have the solutions; now what do they mean? (The trouble with quantum mechanics is not only in solving the equations but in understanding what the solutions mean!) First, notice that if $b = 0$, both terms have the same frequency $\omega = (E_0 - A)/\hbar$. If everything changes at one frequency, it means that the system is in a state of definite energy—here, the energy $(E_0 - A)$. So there is a stationary state of this energy in which the two amplitudes C_1 and C_2 are equal. We get the result that *the ammonia molecule has a definite energy* $(E_0 - A)$ if there are equal amplitudes for the nitrogen atom to be "up" and to be "down."

There is another stationary state possible if $a = 0$; both amplitudes then have the frequency $(E_0 + A)/\hbar$. So there is another state with the definite energy $(E_0 + A)$ if the two amplitudes are equal but with the opposite sign; $C_2 = -C_1$. These are the only two states of definite energy. We will discuss the states of the ammonia molecule in more detail in the next chapter; we will mention here only a couple of things.

We conclude that *because* there is some chance that the nitrogen atom can flip from one position to the other, the energy of the molecule is not just E_0, as we would have expected, but that there are *two* energy levels $(E_0 + A)$ and $(E_0 - A)$. Every one of the possible states of the molecule, whatever energy it has, is "split" into two levels. We say *every* one of the states because, you remember, we picked out one particular state of rotation, and internal energy, and so on. For each possible condition of that kind there is a doublet of energy levels because of the flip-flop of the molecule.

Let's now ask the following question about an ammonia molecule. Suppose that at $t = 0$, we *know* that a molecule is in the state $|\,1\rangle$ or, in other words, that $C_1(0) = 1$ and $C_2(0) = 0$. What is the probability that the molecule will be found in the state $|\,2\rangle$ at the time t, or will still be found in state $|\,1\rangle$ at the time t? Our starting condition tells us what a and b are in Eqs. (8.50) and (8.51). Letting $t = 0$, we have that

$$C_1(0) = \frac{a+b}{2} = 1, \qquad C_2(0) = \frac{a-b}{2} = 0.$$

Clearly, $a = b = 1$. Putting these values into the formulas for $C_1(t)$ and $C_2(t)$ and rearranging some terms, we have

$$C_1(t) = e^{-(i/\hbar)E_0 t}\left(\frac{e^{(i/\hbar)At} + e^{-(i/\hbar)At}}{2}\right),$$

$$C_2(t) = e^{-(i/\hbar)E_0 t}\left(\frac{e^{(i/\hbar)At} - e^{-(i/\hbar)At}}{2}\right).$$

We can rewrite these as

$$C_1(t) = e^{-(i/\hbar)E_0 t}\cos\frac{At}{\hbar}, \tag{8.52}$$

$$C_2(t) = ie^{-(i/\hbar)E_0 t}\sin\frac{At}{\hbar}. \tag{8.53}$$

The two amplitudes have a magnitude that varies harmonically with time.

The probability that the molecule is found in state $|\,2\rangle$ at the time t is the absolute square of $C_2(t)$:

$$|C_2(t)|^2 = \sin^2\frac{At}{\hbar}. \tag{8.54}$$

The probability starts at zero (as it should), rises to one, and then oscillates back and forth between zero and one, as shown in the curve marked P_2 of Fig. 8–2. The probability of being in the $|\,1\rangle$ state does not, of course, stay at one. It "dumps" into the second state until the probability of finding the molecule in the first state is zero, as shown by the curve P_1 of Fig. 8–2. The probability sloshes back and forth between the two.

A long time ago we saw what happens when we have two equal pendulums with a slight coupling. (See Chapter 49, Vol. I.) When we lift one back and let go,

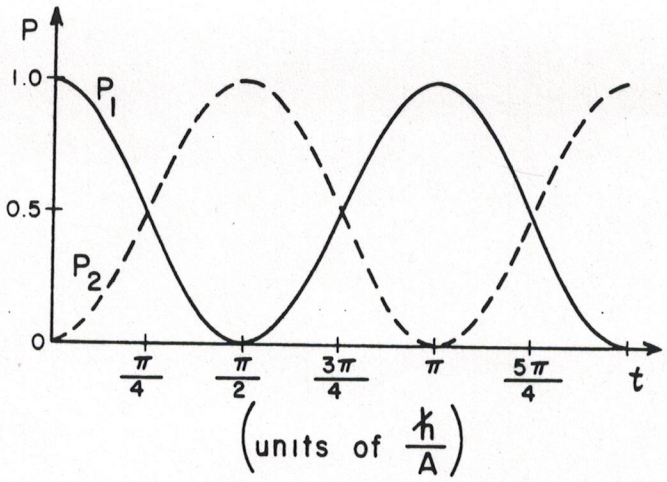

Fig. 8–2. The probability P_1 that an ammonia molecule in state $|1\rangle$ at $t = 0$ will be found in state $|1\rangle$ at t. The probability P_2 that it will be found in state $|2\rangle$.

it swings, but then gradually the other one starts to swing. Pretty soon the second pendulum has picked up all the energy. Then, the process reverses, and pendulum number one picks up the energy. It is exactly the same kind of a thing. The speed at which the energy is swapped back and forth depends on the coupling between the two pendulums—the rate at which the "oscillation" is able to leak across. Also, you remember, with the two pendulums there are two special motions—each with a definite frequency—which we call the fundamental modes. If we pull both pendulums out together, they swing together at one frequency. On the other hand, if we pull one out one way and the other out the other way, there is another stationary mode also at a definite frequency.

Well, here we have a similar situation—the ammonia molecule is mathematically like the pair of pendulums. These are the two frequencies—$(E_0 + A)/\hbar$ and $(E_0 - A)/\hbar$—for when they are oscillating together, or oscillating opposite.

The pendulum analogy is not much deeper than the principle that the same equations have the same solutions. The linear equations for the amplitudes (8.39) are very much like the linear equations of harmonic oscillators. (In fact, this is the reason behind the success of our classical theory of the index of refraction, in which we replaced the quantum mechanical atom by a harmonic oscillator, even though, classically, this is not a reasonable view of electrons circulating about a nucleus.) If you pull the nitrogen to one side, then you get a *superposition* of these two frequencies, and you get a kind of *beat* note, because the system is *not* in one or the other states of definite frequency. The splitting of the energy levels of the ammonia molecule is, however, strictly a quantum mechanical effect.

The splitting of the energy levels of the ammonia molecule has important practical applications which we will describe in the next chapter. At long last we have an example of a practical physical problem that you can understand with the quantum mechanics!

科目：物性機能科学 II

8～14 目

プリント　参考するもの

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

R. ミケレット

13

Propagation in a Crystal Lattice

13–1 States for an electron in a one-dimensional lattice

You would, at first sight, think that a low-energy electron would have great difficulty passing through a solid crystal. The atoms are packed together with their centers only a few angstroms apart, and the effective diameter of the atom for electron scattering is roughly an angstrom or so. That is, the atoms are large, relative to their spacing, so that you would expect the mean free path between collisions to be of the order of a few angstroms—which is practically nothing. You would expect the electron to bump into one atom or another almost immediately. Nevertheless, it is a ubiquitous phenomenon of nature that if the lattice is perfect, the electrons are able to travel through the crystal smoothly and easily—almost as if they were in a vacuum. This strange fact is what lets metals conduct electricity so easily; it has also permitted the development of many practical devices. It is, for instance, what makes it possible for a transistor to imitate the radio tube. In a radio tube electrons move freely through a vacuum, while in the transistor they move freely through a crystal lattice. The machinery behind the behavior of a transistor will be described in this chapter; the next one will describe the application of these principles in various practical devices.

The conduction of electrons in a crystal is one example of a very common phenomenon. Not only can electrons travel through crystals, but other "things" like atomic excitations can also travel in a similar manner. So the phenomenon which we want to discuss appears in many ways in the study of the physics of the solid state.

You will remember that we have discussed many examples of two-state systems. Let's now think of an electron which can be in either one of two positions, in each of which it is in the same kind of environment. Let's also suppose that there is a certain amplitude to go from one position to the other, and, of course, the same amplitude to go back, just as we have discussed for the hydrogen molecular ion in Section 10–1. The laws of quantum mechanics then give the following results. There are two possible states of definite energy for the electron. Each state can be described by the amplitude for the electron to be in each of the two basic positions. In either of the definite-energy states, the magnitudes of these two amplitudes are constant in time, and the phases vary in time with the same frequency. On the other hand, if we start the electron in one position, it will later have moved to the other, and still later will swing back again to the first position. The amplitude is analogous to the motions of two coupled pendulums.

Now consider a perfect crystal lattice in which we imagine that an electron can be situated in a kind of "pit" at one particular atom and with some particular energy. Suppose also that the electron has some amplitude to move into a different pit at one of the nearby atoms. It is something like the two-state system—but with an additional complication. When the electron arrives at the neighboring atom, it can afterward move on to still another position as well as return to its starting point. Now we have a situation analogous not to *two* coupled pendulums, but to an *infinite number* of pendulums all coupled together. It is something like what you see in one of those machines—made with a long row of bars mounted on a torsion wire—that is used in first-year physics to demonstrate wave propagation.

If you have a harmonic oscillator which is coupled to another harmonic oscillator, and that one to another, and so on . . . , and if you start an irregularity in one place, the irregularity will propagate as a wave along the line. The same situation exists if you place an electron at one atom of a long chain of atoms.

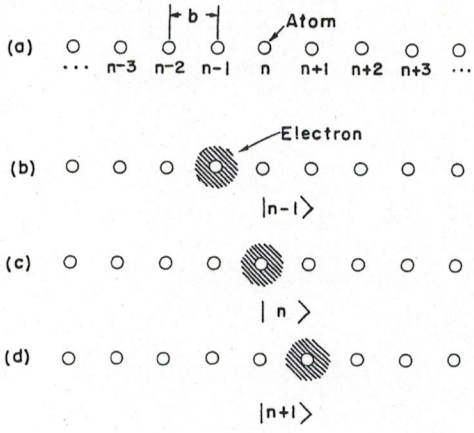

Fig. 13–1. The base states of an electron in a one-dimensional crystal.

Usually, the simplest way of analyzing the mechanical problem is not to think in terms of what happens if a pulse is started at a definite place, but rather in terms of steady-wave solutions. There exist certain patterns of displacement which propagate through the crystal as a wave of a single, fixed frequency. Now the same thing happens with the electron—and for the same reason, because it's described in quantum mechanics by similar equations.

You must appreciate one thing, however; the amplitude for the electron to be at a place is an *amplitude*, not a probability. If the electron were simply leaking from one place to another, like water going through a hole, the behavior would be completely different. For example, if we had two tanks of water connected by a tube to permit some leakage from one to the other, then the levels would approach each other exponentially. But for the electron, what happens is amplitude leakage and not just a plain probability leakage. And it's a characteristic of the imaginary term—the i in the differential equations of quantum mechanics—which changes the exponential solution to an oscillatory solution. What happens there is quite different from the leakage between interconnected tanks.

We want now to analyze quantitatively the quantum mechanical situation. Imagine a one-dimensional system made of a long line of atoms as shown in Fig. 13–1(a). (A crystal is, of course, three-dimensional but the physics is very much the same; once you understand the one-dimensional case you will be able to understand what happens in three dimensions.) Next, we want to see what happens if we put a single electron on this line of atoms. Of course, in a real crystal there are already millions of electrons. But most of them (nearly all for an insulating crystal) take up positions in some pattern of motion each around its own atom—and everything is quite stationary. However, we now want to think about what happens if we put an *extra* electron in. We will not consider what the other ones are doing because we suppose that to change their motion involves a lot of excitation energy. We are going to add an electron as if to produce one slightly bound negative ion. In watching what the *one* extra electron does we are making an approximation which disregards the mechanics of the inside workings of the atoms.

Of course the electron could then move to another atom, transferring the negative ion to another place. We will suppose that just as in the case of an electron jumping between two protons, the electron can jump from one atom to the neighbor on either side with a certain amplitude.

Now how do we describe such a system? What will be reasonable base states? If you remember what we did when we had only two possible positions, you can guess how it will go. Suppose that in our line of atoms the spacings are all equal and that we number the atoms in sequence, as shown in Fig. 13–1(a). One of the base states is that the electron is at atom number 6, another base state is that the electron is at atom number 7, or at atom number 8, and so on. We can describe the nth base state by saying that the electron is at atom number n. Let's say that this is the base state $|\,n\rangle$. Figure 13–1 shows what we mean by the three base states

$$|\,n-1\rangle, \quad |\,n\rangle, \quad \text{and} \quad |\,n+1\rangle.$$

Using these base states, any state $|\,\phi\rangle$ of our one-dimensional crystal can be described by giving all the amplitudes $\langle n\,|\,\phi\rangle$ that the state $|\,\phi\rangle$ is in one of the base states—which means the amplitude that it is located at one particular atom. Then we can write the state $|\,\phi\rangle$ as a superposition of the base states

$$|\,\phi\rangle = \sum_n |\,n\rangle\langle n\,|\,\phi\rangle. \tag{13.1}$$

Next, we are going to suppose that when the electron is at one atom, there is a certain amplitude that it will leak to the atom on either side. And we'll take the simplest case for which it can only leak to the nearest neighbors—to get to the next-nearest neighbor, it has to go in two steps. We'll take that the amplitudes for the electron jump from one atom to the next is $iA/\hbar$ (per unit time).

For the moment we would like to write the amplitude $\langle n \mid \phi \rangle$ to be on the nth atom as C_n. Then Eq. (13.1) will be written

$$|\phi\rangle = \sum_n |n\rangle C_n. \tag{13.2}$$

If we knew each of the amplitudes C_n at a given moment, we could take their absolute squares and get the probability that you would find the electron if you looked at atom n at that time.

What will the situation be at some later time? By analogy with the two-state systems we have studied, we would propose that the Hamiltonian equations for this system should be made up of equations like this:

$$i\hbar \frac{dC_n(t)}{dt} = E_0 C_n(t) - AC_{n+1}(t) - AC_{n-1}(t). \tag{13.3}$$

The first coefficient on the right, E_0, is, physically, the energy the electron would have if it couldn't leak away from one of the atoms. (It doesn't matter what we call E_0; as we have seen many times, it represents really nothing but our choice of the zero of energy.) The next term represents the amplitude per unit time that the electron is leaking into the nth pit from the $(n + 1)$st pit; and the last term is the amplitude for leakage from the $(n - 1)$st pit. As usual, we'll assume that A is a constant (independent of t).

For a full description of the behavior of any state $|\phi\rangle$, we would have one equation like (13.3) for every one of the amplitudes C_n. Since we want to consider a crystal with a very large number of atoms, we'll assume that there are an indefinitely large number of states—that the atoms go on forever in both directions. (To do the finite case, we will have to pay special attention to what happens at the ends.) If the number N of our base states is indefinitely large, then also our full Hamiltonian equations are infinite in number! We'll write down just a sample:

$$\begin{aligned}
&\vdots \qquad\qquad\qquad\qquad \vdots \\
&i\hbar \frac{dC_{n-1}}{dt} = E_0 C_{n-1} - AC_{n-2} - AC_n, \\
&i\hbar \frac{dC_n}{dt} = E_0 C_n - AC_{n-1} - AC_{n+1}, \\
&i\hbar \frac{dC_{n+1}}{dt} = E_0 C_{n+1} - AC_n - AC_{n+2}, \\
&\vdots \qquad\qquad\qquad\qquad \vdots
\end{aligned} \tag{13.4}$$

13–2 States of definite energy

We could study many things about an electron in a lattice, but first let's try to find the states of definite energy. As we have seen in earlier chapters this means that we have to find a situation in which the amplitudes all change at the same frequency if they change with time at all. We look for solutions of the form

$$C_n = a_n e^{-iEt/\hbar}. \tag{13.5}$$

The complex number a_n tell us about the non-time-varying part of the amplitude to find the electron at the nth atom. If we put this trial solution into the equations of (13.4) to test them out, we get the result

$$Ea_n = E_0 a_n - Aa_{n+1} - Aa_{n-1}. \tag{13.6}$$

We have an infinite number of such equations for the infinite number of unknowns a_n—which is rather petrifying.

All we have to do is take the determinant . . . but wait! Determinants are fine when there are 2, 3, or 4 equations. But if there are a large number—or an infinite number—of equations, the determinants are not very convenient. We'd better just try to solve the equations directly. First, let's label the atoms by their

positions; we'll say that the atom n is at x_n and the atom $(n + 1)$ is at x_{n+1}. If the atomic spacing is b—as in Fig. 13–1—we will have that $x_{n+1} = x_n + b$. By choosing our origin at atom zero, we can even have it that $x_n = nb$. We can rewrite Eq. (13.5) as

$$C_n = a(x_n)e^{-iEt/\hbar}, \tag{13.7}$$

and Eq. (13.6) would become

$$Ea(x_n) = E_0a(x_n) - Aa(x_{n+1}) - Aa(x_{n-1}). \tag{13.8}$$

Or, using the fact that $x_{n+1} = x_n + b$, we could also write

$$Ea(x_n) = E_0a(x_n) - Aa(x_n + b) - Aa(x_n - b). \tag{13.9}$$

This equation is somewhat similar to a differential equation. It tells us that a quantity, $a(x)$, at one point, (x_n), is related to the same physical quantity at some neighboring points, $(x_n \pm b)$. (A differential equation relates the value of a function at a point to the values at infinitesimally nearby points.) Perhaps the methods we usually use for solving differential equations will also work here; let's try.

Linear differential equations with constant coefficients can always be solved in terms of exponential functions. We can try the same thing here; let's take as a trial solution

$$a(x_n) = e^{ikx_n}. \tag{13.10}$$

Then Eq. (13.9) becomes

$$Ee^{ikx_n} = E_0e^{ikx_n} - Ae^{ik(x_n+b)} - Ae^{ik(x_n-b)}. \tag{13.11}$$

We can now divide out the common factor e^{ikx_n}; we get

$$E = E_0 - Ae^{ikb} - Ae^{-ikb}. \tag{13.12}$$

The last two terms are just equal to $(2A \cos kb)$, so

$$E = E_0 - 2A \cos kb. \tag{13.13}$$

We have found that for *any* choice at all for the constant k there is a solution whose energy is given by this equation. There are various possible energies depending on k, and each k corresponds to a different solution. There are an infinite number of solutions—which is not surprising, since we started out with an infinite number of base states.

Let's see what these solutions mean. For each k, the a's are given by Eq. (13.10). The amplitudes C_n are then given by

$$C_n = e^{ikx_n}e^{-(i/\hbar)Et}, \tag{13.14}$$

where you should remember that the energy E also depends on k as given in Eq. (13.13). The *space dependence* of the amplitudes is e^{ikx_n}. The amplitudes oscillate as we go along from one atom to the next.

We mean that, in space, the amplitude goes as a *complex* oscillation—the *magnitude* is the same at every atom, but the phase at a given time advances by the amount (ikb) from one atom to the next. We can visualize what is going on by plotting a vertical line to show just the real part at each atom as we have done in Fig. 13–2. The envelope of these vertical lines (as shown by the broken-line curve)

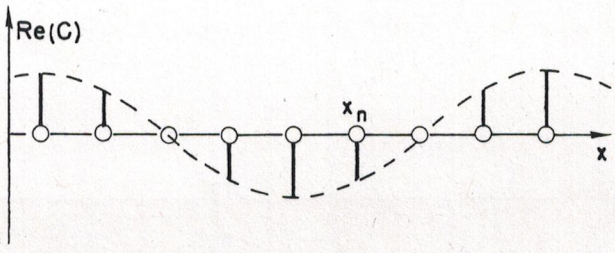

Fig. 13–2. Variation of the real part of C_n with x_n.

is, of course, a cosine curve. The imaginary part of C_n is also an oscillating function, but is shifted 90° in phase so that the absolute square (which is the sum of the squares of the real and imaginary parts) is the same for all the C's.

Thus if we pick a k, we get a stationary state of a particular energy E. And for any such state, the electron is equally likely to be found at every atom—there is no preference for one atom or the other. Only the phase is different for different atoms. Also, as time goes on the phases vary. From Eq. (13.14) the real and imaginary parts propagate along the crystal as waves—namely as the real or imaginary parts of

$$e^{i[kx_n - (E/\hbar)t]}. \tag{13.15}$$

The wave can travel toward positive or negative x depending on the sign we have picked for k.

Notice that we have been assuming that the number k that we put in our trial solution, Eq. (13.10), was a real number. We can see now why that must be so if we have an infinite line of atoms. Suppose that k were an imaginary number, say ik'. Then the amplitudes a_n would go as $e^{k'x_n}$, which means that the amplitude would get larger and larger as we go toward large x's—or toward large negative x's if k' is a negative number. This kind of solution would be O.K. if we were dealing with line of atoms that ended, but cannot be a physical solution for an infinite chain of atoms. It would give infinite amplitudes—and, therefore, infinite probabilities—which can't represent a real situation. Later on we will see an example in which an imaginary k does make sense.

The relation between the energy E and the wave number k as given in Eq. (13.13) is plotted in Fig. 13–3. As you can see from the figure, the energy can go from $(E_0 - 2A)$ at $k = 0$ to $(E_0 + 2A)$ at $k = \pm\pi/b$. The graph is plotted for positive A; if A were negative, the curve would simply be inverted, but the range would be the same. The significant result is that any energy is possible within a certain range or "band" of energies, but no others. According to our assumptions, if an electron in a crystal is in a stationary state, it can have no energy other than values in this band.

According to Eq. (13.13), the smallest k's correspond to low-energy states—$E \approx (E_0 - 2A)$. As k increases in magnitude (toward either positive or negative values) the energy at first increases, but then reaches a maximum at $k = \pm\pi/b$, as shown in Fig. 13–3. For k's larger than π/b, the energy would start to decrease again. But we do not really need to consider such values of k, because they do not give new states—they just repeat states we already have for smaller k. We can see that in the following way. Consider the lowest energy state for which $k = 0$. The coefficient $a(x_n)$ is the same for all x_n. Now we would get the same energy for $k = 2\pi/b$. But then, using Eq. (13.10), we have that

$$a(x_n) = e^{i(2\pi/b)x_n}.$$

However, taking x_0 to be at the origin, we can set $x_n = nb$; then $a(x_n)$ becomes

$$a(x_n) = e^{i2\pi n} = 1.$$

The state described by these $a(x_n)$ is physically the same state we got for $k = 0$. It does not represent a different solution.

As another example, suppose that k were $\pi/4b$. The real part of $a(x_n)$ would vary as shown by curve 1 in Fig. 13–4. If k were seven times larger ($k = 7\pi/4$), the real part of $a(x_n)$ would vary as shown by curve 2 in the figure. (The complete

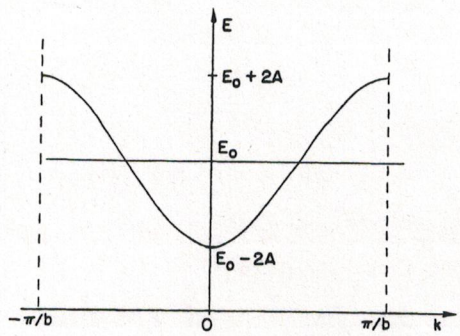

Fig. 13–3. The energy of the stationary states as a function of the parameter k.

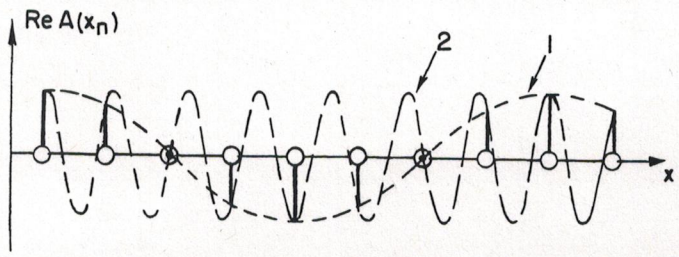

Fig. 13–4. Two values of k which represent the same physical situation; curve 1 is for $k = \pi/4$, curve 2 is for $k = 7\pi/4$.

cosine curves don't mean anything, of course; all that matters is their values *at the points* x_n. The curves are just to help you see how things are going.) You see that both values of k give the same amplitudes at all of the x_n's.

The upshot is that we have all the possible solutions of our problem if we take only k's in a certain limited range. We'll pick the range between $-\pi/b$ and $+\pi/b$—the one shown in Fig. 13–3. In this range, the energy of the stationary states increases uniformly with an increase in the magnitude of k.

One side remark about something you can play with. Suppose that the electron cannot only jump to the nearest neighbor with amplitude $iA/\hbar$, but also has the possibility to jump in one direct leap to the *next nearest* neighbor with some other amplitude $iB/\hbar$. You will find that the solution can again be written in the form $a_n = e^{ikx_n}$—this type of solution is universal. You will also find that the stationary states with wave number k have an energy equal to $(E_0 - 2A\cos kb - 2B\cos 2kb)$. This shows that the shape of the curve of E against k is not universal, but depends upon the particular assumptions of the problem. It is not always a cosine wave—it's not even necessarily symmetrical about some horizontal line. It is true, however, that the curve always repeats itself outside of the interval from $-\pi/b$ to π/b, so you never need to worry about other values of k.

Let's look a little more closely at what happens for small k—that is, when the variations of the amplitudes from one x_n to the next are quite slow. Suppose we choose our zero of energy by defining $E_0 = 2A$; then the minimum of the curve in Fig. 13–3 is at the zero of energy. For small enough k, we can write that

$$\cos kb \approx 1 - k^2b^2/2,$$

and the energy of Eq. (13.13) becomes

$$E = Ak^2b^2. \tag{13.16}$$

We have that the energy of the state is proportional to the square of the wave number which describes the spatial variations of the amplitudes C_n.

13–3 Time-dependent states

In this section we would like to discuss the behavior of states in the one-dimensional lattice in more detail. If the amplitude for an electron to be at x_n is C_n, the probability of finding it there is $|C_n|^2$. For the *stationary* states described by Eq. (13.12), this probability is the same for all x_n and does not change with time. How can we represent a situation which we would describe roughly by saying an electron of a certain energy is localized in a certain region—so that it is more likely to be found at one place than at some other place? We can do that by making a superposition of several solutions like Eq. (13.12) with slightly different values of k—and, therefore, slightly different energies. Then at $t = 0$, at least, the amplitude C_n will vary with position because of the interference between the various terms, just as one gets beats when there is a mixture of waves of different wavelengths (as we discussed in Chapter 48, Vol. I). So we can make up a "wave packet" with a predominant wave number k_0, but with various other wave numbers near k_0.†

In our superposition of stationary states, the amplitudes with different k's will represent states of slightly different energies, and, therefore, of slightly different frequencies; the interference pattern of the total C_n will, therefore, also vary with time—there will be a pattern of "beats." As we have seen in Chapter 48 of Volume I, the peaks of the beats [the place where $|C(x_n)|^2$ is large] will move along in x as time goes on; they move with the speed we have called the "group velocity." We found that this group velocity was related to the variation of k with frequency by

$$v_{\text{group}} = \frac{d\omega}{dk}; \tag{13.17}$$

† Provided we do not try to make the packet too narrow.

the same derivation would apply equally well here. An electron state which is a "clump"—namely one for which the C_n vary in space like the wave packet of Fig. 13–5—will move along our one-dimensional "crystal" with the speed v equal to $d\omega/dk$, where $\omega = E/\hbar$. Using (13.16) for E, we get that

$$v = \frac{2Ab^2}{\hbar} k. \tag{13.18}$$

In other words, the electrons move along with a speed proportional to the typical k. Equation (13.16) then says that the energy of such an electron is proportional to the square of its velocity—*it acts like a classical particle.* So long as we look at things on a scale gross enough that we don't see the fine structure, our quantum mechanical picture begins to give results like classical physics. In fact, if we solve Eq. (13.18) for k and substitute into (13.16), we can write

$$E = \tfrac{1}{2} m_{\text{eff}}\, v^2, \tag{13.19}$$

where m_{eff} is a constant. The extra "energy of motion" of the electron in a packet depends on the velocity just as for a classical particle. The constant m_{eff}—called the "effective mass"—is given by

$$m_{\text{eff}} = \frac{\hbar^2}{2Ab^2}. \tag{13.20}$$

Also notice that we can write

$$m_{\text{eff}}\, v = \hbar k. \tag{13.21}$$

If we choose to call $m_{\text{eff}}\, v$ the "momentum," it is related to the wave number k in the way we have described earlier for a free particle.

Don't forget that m_{eff} has nothing to do with the real mass of an electron. It may be quite different—although in real crystals it often happens to turn out to be the same general order of magnitude, about 2 to 20 times the free-space mass of the electron.

We have now explained a remarkable mystery—how an electron in a crystal (like an extra electron put into germanium) can ride right through the crystal and flow perfectly freely even though it has to hit all the atoms. It does so by having its amplitudes going pip-pip-pip from one atom to the next, working its way through the crystal. That is how a solid can conduct electricity.

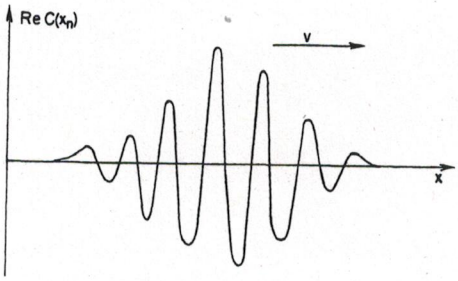

Fig. 13–5. The real part of $C(x_n)$ as a function of x for a superposition of several states of similar energy. (The spacing b is very small on the scale of x shown.)

13–4 An electron in a three-dimensional lattice

Let's look for a moment at how we could apply the same ideas to see what happens to an electron in three dimensions. The results turn out to be very similar. Suppose we have a rectangular lattice of atoms with lattice spacings of a, b, c in the three directions. (If you want a cubic lattice, take the three spacings all equal.) Also suppose that the amplitude to leap in the x-direction to a neighbor is $(iA_x/\hbar)$, to leap in the y-direction is $(iA_y/\hbar)$, and to leap in the z-direction is $(iA_z/\hbar)$. Now how should we describe the base states? As in the one-dimensional case, one base state is that the electron is at the atom whose locations are x, y, z, where (x, y, z) is one of the lattice points. Choosing our origin at one atom, these points are all at

$$x = n_x a, \qquad y = n_y b, \qquad \text{and} \qquad z = n_z c,$$

where n_x, n_y, n_z are any three integers. Instead of using subscripts to indicate such points, we will now just use x, y, and z, understanding that they take on only their values at the lattice points. Thus the base state is represented by the symbol $|\,\text{electron at } x, y, z\rangle$, and the amplitude for an electron in some state $|\,\psi\rangle$ to be in this base state is $C(x, y, z) = \langle\text{electron at } x, y, z\,|\,\psi\rangle$.

As before, the amplitudes $C(x, y, z)$ may vary with time. With our assumptions, the Hamiltonian equations should be like this:

$$i\hbar \frac{dC(x,y,z)}{dt} = E_0 C(x,y,z) - A_x C(x+a,y,z) - A_x C(x-a,y,z) - A_y C(x,y+b,z) - A_y C(x,y-b,z) - A_z C(x,y,z+c) - A_z C(x,y,z-c). \quad (13.22)$$

It looks rather long, but you can see where each term comes from.

Again we can try to find a stationary state in which all the C's vary with time in the same way. Again the solution is an exponential:

$$C(x,y,z) = e^{-iEt/\hbar} e^{i(k_x x + k_y y + k_z z)}. \quad (13.23)$$

If you substitute this into (13.22) you see that it works, provided that the energy E is related to k_x, k_y, and k_z in the following way:

$$E = E_0 - 2A_x \cos k_x a - 2A_y \cos k_y b - 2A_z \cos k_z c. \quad (13.24)$$

The energy now depends on the *three* wave numbers k_x, k_y, k_z, which, incidentally, are the components of a three-dimensional vector $\boldsymbol{k}$. In fact, we can write Eq. (13.23) in vector notation as

$$C(x,y,z) = e^{-iEt/\hbar} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}} \quad (13.25)$$

The amplitude varies as a complex *plane wave* in three dimensions, moving in the direction of $\boldsymbol{k}$, and with the wave number $k = (k_x^2 + k_y^2 + k_z^2)^{1/2}$.

The energy associated with these stationary states depends on the three components of $\boldsymbol{k}$ in the complicated way given in Eq. (13.24). The nature, of the variation of E with $\boldsymbol{k}$ depends on relative signs and magnitudes of A_x, A_y, and A_z. If these three numbers are all positive, and if we are interested in small values of k, the dependence is relatively simple.

Expanding the cosines as we did before to get Eq. (13.16), we can now get that

$$E = E_{\min} + A_x a^2 k_x^2 + A_y b^2 k_y^2 + A_z c^2 k_z^2. \quad (13.26)$$

For a simple cubic lattice with lattice spacing a we expect that A_x and A_y and A_z would be equal—say all are just A—and we would have just

$$E = E_{\min} + Aa^2(k_x^2 + k_y^2 + k_z^2),$$

or

$$E = E_{\min} + Aa^2 k^2. \quad (13.27)$$

This is just like Eq. (13.16). Following the arguments used there, we would conclude that an electron packet in *three* dimensions (made up by superposing many states with nearly equal energies) also moves like a classical particle with some effective mass.

In a crystal with a lower symmetry than cubic (or even in a cubic crystal in which the state of the electron at each atom is not symmetrical) the three coefficients A_x, A_y, and A_z are different. Then the "effective mass" of an electron localized in a small region *depends on its direction of motion.* It could, for instance, have a different inertia for motion in the x-direction than for motion in the y-direction. (The details of such a situation are sometimes described in terms of an "effective mass tensor.")

13–5 Other states in a lattice

According to Eq. (13.24) the electron states we have been talking about can have energies only in a certain "band" of energies which covers the energy range from the minimum energy

$$E_0 - 2(A_x + A_y + A_z)$$

to the maximum energy

$$E_0 + 2(A_x + A_y + A_z).$$

Other energies are possible, but they belong to a different class of electron states. For the states we have described, we imagined base states in which an electron is placed on an atom of the crystal in some particular state, say the lowest energy state.

If you have an atom in empty space, and add an electron to make an ion, the ion can be formed in many ways. The electron can go on in such a way as to make the state of lowest energy, or it can go on to make one or another of many possible "excited states" of the ion each with a definite energy above the lowest energy. The same thing can happen in a crystal. Let's suppose that the energy E_0 we picked above corresponds to base states which are ions of the lowest possible energy. We could also imagine a new set of base states in which the electron sits near the nth atom in a different way—in one of the excited states of the ion—so that the energy E_0 is now quite a bit higher. As before there is some amplitude A (different from before) that the electron will jump from its excited state at one atom to the same excited state at a neighboring atom. The whole analysis goes as before; we find a band of possible energies centered at a higher energy. There can, in general, be many such bands each corresponding to a different level of excitation.

There are also other possibilities. There may be some amplitude that the electron jumps from an excited condition at one atom to an unexcited condition at the next atom. (This is called an interaction between bands.) The mathematical theory gets more and more complicated as you take into account more and more bands and add more and more coefficients for leakage between the possible states. No new ideas are involved, however; the equations are set up much as we have done in our simple example.

We should remark also that there is not much more to be said about the various coefficients, such as the amplitude A, which appear in the theory. Generally they are very hard to calculate, so in practical cases very little is known theoretically about these parameters and for any particular real situation we can only take values determined experimentally.

There are other situations where the physics and mathematics are almost exactly like what we have found for an electron moving in a crystal, but in which the "object" that moves is quite different. For instance, suppose that our original crystal—or rather linear lattice—was a line of neutral atoms, each with a loosely bound outer electron. Then imagine that we were to remove one electron. Which atom has lost its electron? Let C_n now represent the amplitude that the electron *is missing* from the atom at x_n. There will, in general, be some amplitude $iA/\hbar$ that the electron at a neighboring atom—say the $(n-1)$st atom—will jump to the nth leaving the $(n-1)$st atom without its electron. This is the same as saying that there is an amplitude A for the "missing electron" to jump from the nth atom to the $(n-1)$st atom. You can see that the equations will be exactly the same—of course, the value of A need not be the same as we had before. Again we will get the same formulas for the energy levels, for the "waves" of probability which move through the crystal with the group velocity of Eq. (13.18), for the effective mass, and so on. Only now the waves describe the behavior of the *missing electron*—or "hole" as it is called. So a "hole" acts just like a particle with a certain mass m_{eff}. You can see that this particle will appear to have a positive charge. We'll have some more to say about such holes in the next chapter.

As another example, we can think of a line of identical *neutral* atoms one of which has been put into an excited state—that is, with more than its normal ground state energy. Let C_n be the amplitude that the nth atom has the excitation. It can interact with a neighboring atom by handing over to it the extra energy and returning to the ground state. Call the amplitude for this process $iA/\hbar$. You can see that it's the same mathematics all over again. Now the object which moves is called an *exciton*. It behaves like a neutral "particle" moving through the crystal, carrying the excitation energy. Such motion may be involved in certain biological

processes such as vision, or photosynthesis. It has been guessed that the absorption of light in the retina produces an "exciton" which moves through some periodic structure (such as the layers in the rods we described in Chapter 36, Vol. 1; see Fig. 36–5) to be accumulated at some special station where the energy is used to induce a chemical reaction.

13–6 Scattering from imperfections in the lattice

We want now to consider the case of a single electron in a crystal which is not perfect. Our earlier analysis says that perfect crystals have perfect conductivity —that electrons can go slipping through the crystal, as in a vacuum, without friction. One of the most important things that can stop an electron from going on forever is an imperfection or irregularity in the crystal. As an example, suppose that somewhere in the crystal there is a missing atom; or suppose that someone put one wrong atom at one of the atomic sites so that things there are different than at the other atomic sites. Say the energy, E_0 or the amplitude A could be different. How would we describe what happens then?

To be specific, we will return to the one-dimensional case and we will assume that atom number "zero" is an "impurity" atom and has a different value of E_0 than any of the other atoms. Let's call this energy $(E_0 + F)$. What happens? When an electron arrives at atom "zero" there is some probability that the electron is scattered backwards. If a wave packet is moving along and it reaches a place where things are a little bit different, some of it will continue onward and some of it will bounce back. It's quite difficult to analyze such a situation using a wave packet, because everything varies in time. It is much easier to work with steady-state solutions. So we will work with stationary states, which we will find can be made up of continuous waves which have transmitted and reflected parts. In three dimensions we would call the reflected part the scattered wave, since it would spread out in various directions.

We start out with a set of equations which are just like the ones in Eq. (13.6) except that the equation for $n = 0$ is different from all the rest. The five equations for $n = -2, -1, 0, +1$, and $+2$ look like this:

$$\begin{aligned}
&\vdots \qquad\qquad\qquad \vdots \\
Ea_{-2} &= E_0a_{-2} - Aa_{-1} - Aa_{-3}, \\
Ea_{-1} &= E_0a_{-1} - Aa_0 - Aa_{-2}, \\
Ea_0 &= (E_0 + F)a_0 - Aa_1 - Aa_{-1}, \\
Ea_1 &= E_0a_1 - Aa_2 - Aa_0, \\
Ea_2 &= E_0a_2 - Aa_3 - Aa_1, \\
&\vdots \qquad\qquad\qquad \vdots
\end{aligned} \tag{13.28}$$

There are, of course, all the other equations for $|n|$ is greater than 2. They will look just like Eq. (13.6).

For the general case, we really ought to use a different A for the amplitude that the electron jumps to or from atom "zero," but the main features of what goes on will come out of a simplified example in which all the A's are equal.

Equation (13.10) would still work as a solution for all of the equations except the one for atom "zero"—it isn't right for that one equation. We need a different solution which we can cook up in the following way. Equation (13.10) represents a wave going in the positive x-direction. A wave going in the negative x-direction would have been an equally good solution. It would be written

$$a(x_n) = e^{-ikx_n}.$$

The most general solution we could have taken for Eq. (13.6) would be a com-

bination of a forward and a backward wave, namely

$$a_n = \alpha e^{ikx_n} + \beta e^{-ikx_n}. \tag{13.29}$$

This solution represents a complex wave of amplitude α moving in the $+x$-direction and a wave of amplitude β moving in the $-x$-direction.

Now take a look at the set of equations for our new problem—the ones in (13.28) together with those for all the other atoms. The equations involving a_n's with $n \leq -1$ are all satisfied by Eq. (13.29), with the condition that k is related to E and the lattice spacing b by

$$E = E_0 - 2A \cos kb. \tag{13.30}$$

The physical meaning is an "incident" wave of amplitude α approaching atom "zero" (the "scatterer") from the left, and a "scattered" or "reflected" wave of amplitude β going back toward the left. We do not loose any generality if we set the amplitude α of the incident wave equal to 1. Then the amplitude β is, in general, a complex number.

We can say all the same things about the solutions of a_n for $n \geq 1$. The coefficients could be different, so we would have for them

$$a_n = \gamma e^{ikx_n} + \delta e^{-ikx_n}, \qquad \text{for} \qquad n \geq 1. \tag{13.31}$$

Here, γ is the amplitude of a wave going to the right and δ a wave coming from the right. We want to consider the *physical* situation in which a wave is originally started only from the left, and there is only a "transmitted" wave that comes out beyond the scatterer—or impurity atom. We will try for a solution in which $\delta = 0$. We can, certainly, satisfy all of the equations for the a_n except for the middle three in Eq. (13.28) by the following trial solutions.

$$\begin{aligned} a_n \text{ (for } n < 0) &= e^{ikx_n} + \beta e^{-ikx_n}, \\ a_n \text{ (for } n > 0) &= \gamma e^{ikx_n}. \end{aligned} \tag{13.32}$$

The situation we are talking about is illustrated in Fig. 13–6.

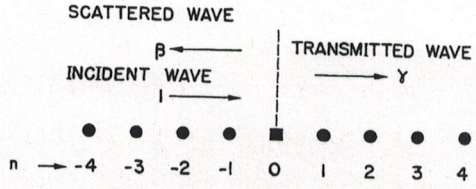

Fig. 13–6. Waves in a one-dimensional lattice with one "impurity" atom at $n = 0$.

By using the formulas in Eq. (13.32) for a_{-1} and a_{+1}, the three middle equations of Eq. (13.28) will allow us to solve for a_0 and also for the two coefficients β and γ. So we have found a complete solution. Setting $x_n = nb$, we have to solve the three equations

$$\begin{aligned} (E - E_0)\{e^{ik(-b)} + \beta e^{-ik(-b)}\} &= -A\{a_0 + e^{ik(-2b)} + \beta e^{-ik(-2b)}\}, \\ (E - E_0 - F)a_0 &= -A\{\gamma e^{ikb} + e^{ik(-b)} + \beta e^{-ik(-b)}\}, \\ (E - E_0)\gamma e^{ikb} &= -A\{\gamma e^{ik(2b)} + a_0\}. \end{aligned} \tag{13.33}$$

Remember that E is given in terms of k by Eq. (13.30). If you substitute this value for E into the equations, and remember that $\cos x = \frac{1}{2}(e^{ix} + e^{-ix})$, you get from the first equation that

$$a_0 = 1 + \beta, \tag{13.34}$$

and from the third equation that

$$a_0 = \gamma. \tag{13.35}$$

These are consistent only if

$$\gamma = 1 + \beta. \tag{13.36}$$

This equation says that the transmitted wave (γ) is just the original incident wave (1) with an added wave (β) equal to the reflected wave. This is not always true, but happens to be so for a scattering at one atom only. If there were a clump of impurity atoms, the amount added to the forward wave would not necessarily be the same as the reflected wave.

We can get the amplitude β of the reflected wave from the middle equation of Eq. (13.33); we find that

$$\beta = \frac{-F}{F - 2iA \sin kb}. \tag{13.37}$$

We have the complete solution for the lattice with one unusual atom.

You may be wondering how the transmitted wave can be "more" than the incident wave as it appears in Eq. (13.34). Remember, though, that β and γ are complex numbers and that the number of particles (or rather, the probability of finding a particle) in a wave is proportional to the absolute square of the amplitude. In fact, there will be "conservation of electrons" only if

$$|\beta|^2 + |\gamma|^2 = 1. \tag{13.38}$$

You can show that this is true for our solution.

13–7 Trapping by a lattice imperfection

There is another interesting situation that can arise if F is a negative number. If the energy of the electron is lower at the impurity atom (at $n = 0$) than it is anywhere else, then the electron can get caught on this atom. That is, if $(E_0 + F)$ is below the bottom of the band at $(E_0 - 2A)$, then the electron can get "trapped" in a state with $E < E_0 - 2A$. Such a solution cannot come out of what we have done so far. We can get this solution, however, if we permit the trial solution we took in Eq. (13.10) to have an imaginary number for k. Let's set $k = \pm i\kappa$. Again, we can have different solutions for $n < 0$ and for $n > 0$. A possible solution for $n < 0$ might be

$$a_n \text{ (for } n < 0) = ce^{+\kappa x_n}. \tag{13.39}$$

We have to take a plus sign in the exponent; otherwise the amplitude would get indefinitely large for large negative values of n. Similarly, a possible solution for $n > 0$ would be

$$a_n \text{ (for } n > 0) = c'e^{-\kappa x_n}. \tag{13.40}$$

If we put these trial solutions into Eq. (13.28) all but the middle three are satisfied provided that

$$E = E_0 - A(e^{\kappa b} + e^{-\kappa b}). \tag{13.41}$$

Since the sum of the two exponential terms is always greater than 2, this energy is below the regular band, and is what we are looking for. The remaining three equations in Eq. (13.28) are satisfied if $a_0 = c = c'$ and if κ is chosen so that

$$A(e^{\kappa b} - e^{-\kappa b}) = -F. \tag{13.42}$$

Combining this equation with Eq. (13.41) we can find the energy of the trapped electron; we get

$$E = E_0 - \sqrt{4A^2 + F^2}. \tag{13.43}$$

The trapped electron has a unique energy—located somewhat below the conduction band.

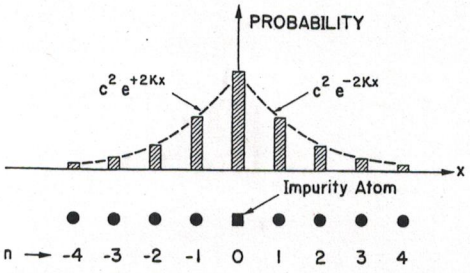

Fig. 13–7. The relative probabilities of finding a trapped electron at atomic sites near the trapping impurity atom.

Notice that the amplitudes we have in Eq. (13.39) and (13.40) do *not* say that the trapped electron sits right on the impurity atom. The probability of finding the electron at nearby atoms is given by the square of these amplitudes. For one particular choice of the parameters it might vary as shown in the bar graph of Fig. 13–7. The probability is greatest for finding the electron on the impurity atom. For nearby atoms the probability drops off exponentially with the distance from the impurity atom. This is another example of "barrier penetration." From the point-of-view of classical physics the electron doesn't have enough energy to get away from the energy "hole" at the trapping center. But quantum mechanically it can leak out a little way.

14

Semiconductors

14–1 Electrons and holes in semiconductors

One of the remarkable and dramatic developments in recent years has been the application of solid state science to technical developments in electrical devices such as transistors. The study of semiconductors led to the discovery of their useful properties and to a large number of practical applications. The field is changing so rapidly that what we tell you today may be incorrect next year. It will certainly be incomplete. And it is perfectly clear that with the continuing study of these materials many new and more wonderful things will be possible as time goes on. You will not need to understand this chapter for what comes later in this volume, but you may find it interesting to see that at least something of what you are learning has some relation to the practical world.

There are large numbers of semiconductors known, but we'll concentrate on those which now have the greatest technical application. They are also the ones that are best understood, and in understanding them we will obtain a degree of understanding of many of the others. The semiconductor substances in most common use today are silicon and germanium. These elements crystallize in the diamond lattice, a kind of cubic structure in which the atoms have tetrahedral bonding with their four nearest neighbors. They are insulators at very low temperatures—near absolute zero—although they do conduct electricity somewhat at room temperature. They are not metals; they are called *semiconductors.*

If we somehow put an extra electron into a crystal of silicon or germanium which is at a low temperature, we will have just the situation we described in the last chapter. The electron will be able to wander around in the crystal jumping from one atomic site to the next. Actually, we have looked only at the behavior of electrons in a rectangular lattice, and the equations would be somewhat different for the real lattice of silicon or germanium. All of the essential points are, however, illustrated by the results for the rectangular lattice.

As we saw in Chapter 13, these electrons can have energies only in a certain energy band—called the *conduction band.* Within this band the energy is related to the wave-number k of the probability amplitude C (see Eq. 13.24) by

$$E = E_0 - 2A_x \cos k_x a - 2A_y \cos k_y b - 2A \cos k_z c. \qquad (14.1)$$

The A's are the amplitudes for jumping in the x-, y-, and z-directions, and a, b, and c are the lattice spacings in these directions.

For energies near the bottom of the band, we can approximate Eq. (14.1) by

$$E = E_{\min} + A_x a^2 k_x^2 + A_y b^2 k_y^2 + A_z c^2 k_z^2 \qquad (14.2)$$

(see Section 13–4).

If we think of electron motion in some particular direction, so that the components of k are always in the same ratio, the energy is a quadratic function of the wave number—and as we have seen of the momentum of the electron. We can write

$$E = E_{\min} + \alpha k^2, \qquad (14.3)$$

where α is some constant, and we can make a graph of E versus k as in Fig. 14–1. We'll call such a graph an "energy diagram." An electron in a particular state of energy and momentum can be indicated by a point such as S in the figure.

Reference: C. Kittel, *Introduction to Solid State Physics*, Chapters 13, 14, and 18.

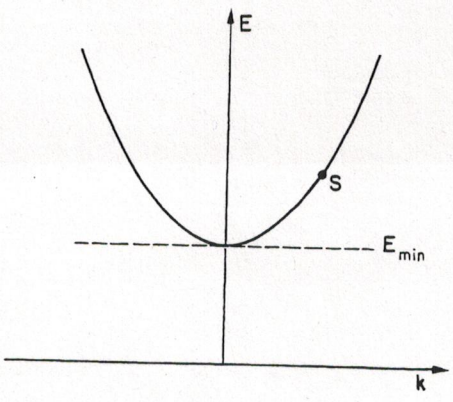

Fig. 14–1. The energy diagram for an electron in an insulating crystal.

As we also mentioned in Chapter 13, we can have a similar situation if we *remove* an electron from a neutral insulator. Then, an electron can jump over from a nearby atom and fill the "hole," but leaving another "hole" at the atom it started from. We can describe this behavior by writing an amplitude to find the *hole* at any particular atom, and by saying that the *hole* can jump from one atom to the next. (Clearly, the amplitudes A that the hole jumps from atom a to atom b is just the same as the amplitude that an electron on atom b jumps into the hole at atom a.) The mathematics is just the same for the *hole* as it was for the extra electron, and we get again that the energy of the hole is related to its wave number by an equation just like Eq. (14.1) or (14.2), except, of course, with different numerical values for the amplitudes A_x, A_y, and A_z. The hole has an energy related to the wave number of its probability amplitudes. Its energy lies in a restricted band, and near the bottom of the band its energy varies quadratically with the wave number—or momentum—just as in Fig. 14–1. Following the arguments of Section 13–3, we would find that *the hole also behaves like a classical particle* with a certain effective mass—except that in noncubic crystals the mass depends on the direction of motion. So the hole behaves like a *positive particle* moving through the crystal. The charge of the hole-particle is positive, because it is located at the site of a missing electron; and when it moves in one direction there are actually electrons moving in the opposite direction.

If we put several electrons into a neutral crystal, they will move around much like the atoms of a low-pressure gas. If there are not too many, their interactions will not be very important. If we then put an electric field across the crystal, the electrons will start to move and an electric current will flow. Eventually they would all be drawn to one edge of the crystal, and, if there is a metal electrode there, they would be collected, leaving the crystal neutral.

Similarly we could put many holes into a crystal. They would roam around at random unless there is an electric field. With a field they would flow toward the negative terminal, and would be "collected"—what actually happens is that they are neutralized by electrons from the metal terminal.

One can also have both holes and electrons together. If there are not too many, they will all go their way independently. With an electric field, they will all contribute to the current. For obvious reasons, electrons are called the *negative carriers* and the holes are called the *positive carriers*.

We have so far considered that electrons are put into the crystal from the outside, or are removed to make a hole. It is also possible to "create" an electron-hole pair by taking a bound electron away from one neutral atom and putting it some distance away in the same crystal. We then have a free electron and a free hole, and the two can move about as we have described.

The energy required to put an electron *into* a state S—we say to "create" the state S—is the energy E^- shown in Fig. 14–2. It is some energy above E^-_{min}. The energy required to "create" a hole in some state S' is the energy E^+ of Fig. 14–3, which is some energy greater than E^+_{min}. Now if we create a pair in the states S and S', the energy required is just $E^- + E^+$.

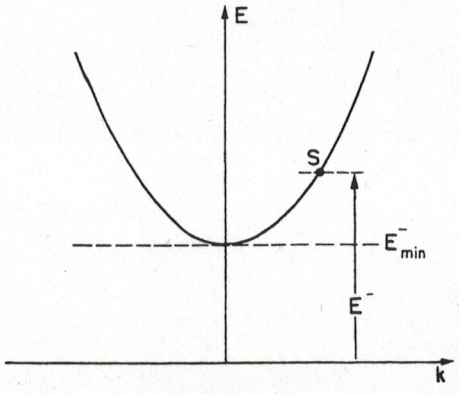

Fig. 14–2. The energy E^- is required to "create" a free electron.

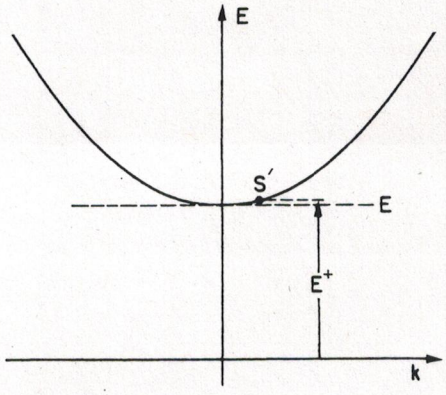

Fig. 14–3. The energy E^+ is required to "create" a hole in the state S'.

The creation of pairs is a common process (as we will see later), so many people like to put Fig. 14–2 and Fig. 14–3 together on the same graph—with the *hole* energy plotted *downward*, although it is, of course a *positive* energy. We have combined our two graphs in this way in Fig. 14–4. The advantage of such a graph is that the energy $E_{\text{pair}} = E^- + E^+$ required to create a pair with the electron in S and the hole in S' is just the vertical distance between S and S' as shown in Fig. 14–4. The minimum energy required to create a pair is called the "gap" energy and is equal to $E^-_{\text{min}} + E^+_{\text{min}}$.

Sometimes you will see a simpler diagram called an energy level diagram which is drawn when people are not interested in the k variable. Such a diagram—shown in Fig. 14–5—just shows the possible energies for the electrons and holes.†

How can electron-hole pairs be created? There are several ways. For example, photons of light (or x-rays) can be absorbed and create a pair if the photon energy is above the energy of the gap. The rate at which pairs are produced is proportional to the light intensity. If two electrodes are plated on a wafer of the crystal and a "bias" voltage is applied, the electrons and holes will be drawn to the electrodes. The circuit current will be proportional to the intensity of the light. This mechanism is responsible for the phenomenon of photoconductivity and the operation of photoconductive cells.

Electron hole pairs can also be produced by high-energy particles. When a fast-moving charged particle—for instance, a proton or a pion with an energy of tens or hundreds of Mev—goes through a crystal, its electric field will knock electrons out of their bound states creating electron-hole pairs. Such events occur hundreds of thousands of times per millimeter of track. After the passage of the particle, the carriers can be collected and in doing so will give an electrical pulse. This is the mechanism at play in the semiconductor counters recently put to use for experiments in nuclear physics. Such counters do not require semiconductors, they can also be made with crystalline insulators. In fact, the first of such counters was made using a diamond crystal which is an insulator at room temperature. Very pure crystals are required if the holes and electrons are to be able to move freely to the electrodes without being trapped. The semiconductors silicon and germanium are used because they can be produced with high purity in reasonable large sizes (centimeter dimensions).

So far we have been concerned with semiconductor crystals at temperatures near absolute zero. At any finite temperature there is still another mechanism by which electron-hole pairs can be created. The pair energy can be provided from the thermal energy of the crystal. The thermal vibrations of the crystal can transfer their energy to a pair—giving rise to "spontaneous" creation.

The probability per unit time that the energy as large as the gap energy E_{gap} will be concentrated at one atomic site is proportional to $e^{-E_{\text{gap}}/\kappa T}$, where T is the temperature and κ is Boltzmann's constant (see Chapter 40, Vol. I). Near absolute zero there is no appreciable probability, but as the temperature rises there is an increasing probability of producing such pairs. At any finite temperature the production should continue forever at a constant rate giving more and more negative and positive carriers. Of course that does not happen because after awhile the electrons and holes accidentally find each other—the electron drops into the hole and the excess energy is given to the lattice. We say that the electron and hole "annihilate." There is a certain probability per second that a hole meets an electron and the two things annihilate each other.

If the number of electrons per unit volume is N_n (n for negative carriers) and the density of positive carriers is N_p, the chance per unit time that an electron and a hole will find each other and annihilate is proportional to the product N_nN_p. In equilibrium this rate must equal the rate that pairs are created. You see that in

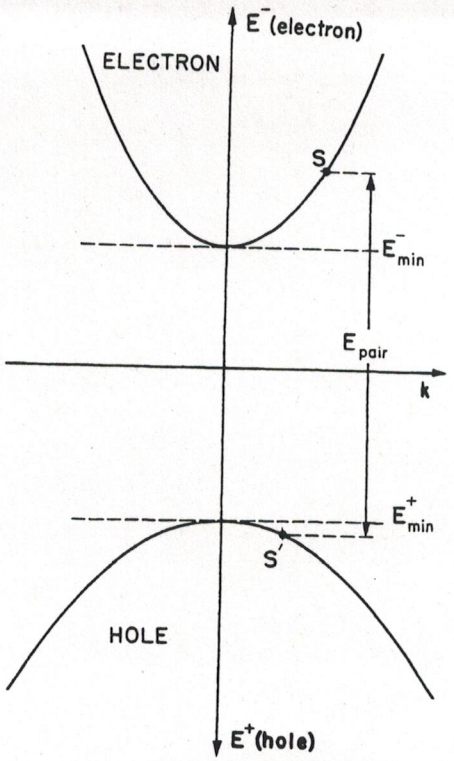

Fig. 14–4. Energy diagrams for an electron and a hole drawn together.

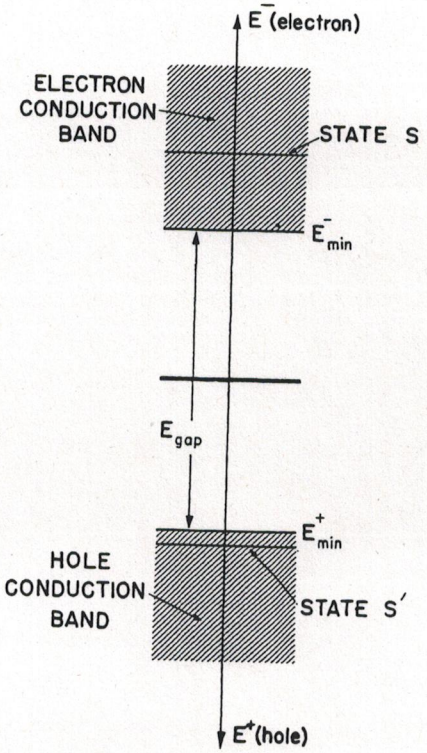

Fig. 14–5. Energy level diagram for electrons and holes.

† In many books this same energy diagram is interpreted in a different way. The energy scale refers only to *electrons*. Instead of thinking of the energy of the hole, they think of the energy an electron *would* have if it filled the hole. This energy is *lower* than the free-electron energy—in fact, just the amount lower that you see in Fig. 14–5. With this interpretation of the energy scale, the gap energy is the minimum energy which must be given *to an electron* to move it from its bound state to the conduction band.

equilibrium the product of N_n and N_p should be given by some constant times the Boltzmann factor:

$$N_n N_p = \text{const}\, e^{-E_{\text{gap}}/\kappa T}. \tag{14.4}$$

When we say constant, we mean nearly constant. A more complete theory—which includes more details about how holes and electrons "find" each other—shows that the "constant" is slightly dependent upon temperature, but the major dependence on temperature is in the exponential.

Let's consider, as an example, a pure material which is originally neutral. At a finite temperature you would expect the number of positive and negative carriers to be equal, $N_n = N_p$. Then each of them should vary with temperature as $e^{-E_{\text{gap}}/2\kappa T}$. The variation of many of the properties of a superconductor—the conductivity for example—is mainly determined by the exponential factor because all the other factors vary much more slowly with temperature. The gap energy for germanium is about 0.72 ev and for silicon 1.1 ev.

At room temperature κT is about 1/40 of an electron volt. At these temperatures there are enough holes and electrons to give a significant conductivity, while at, say, 30°K—one-tenth of room temperature—the conductivity is imperceptible. The gap energy of diamond is 6 or 7 ev and diamond is a good insulator at room temperature.

14–2 Impure semiconductors

So far we have talked about two ways that extra electrons can be put into an otherwise ideally perfect crystal lattice. One way was to inject the electron from an outside source; the other way, was to knock a bound electron off a neutral atom creating simultaneously an electron and a hole. It is possible to put electrons into the conduction band of a crystal in still another way. Suppose we imagine a crystal of germanium in which one of the germanium atoms is replaced by an arsenic atom. The germanium atoms have a valence of 4 and the crystal structure is controlled by the four valence electrons. Arsenic, on the other hand, has a valence of 5. It turns out that a single arsenic atom can sit in the germanium lattice (because it has approximately the correct size), but in doing so it must act as a valence 4 atom—using four of its valence electrons to form the crystal bonds and having one electron left over. This extra electron is very loosely attached—the binding energy is less than 1/10 of a volt. At room temperature the electron easily picks up that much energy from the thermal energy of the crystal, and then takes off on its own—moving about in the lattice as a free electron. An impurity atom such as the arsenic is called a *donor site* because it can give up a negative carrier to the crystal. If a crystal of germanium is grown from a melt to which a very small amount of arsenic has been added, the arsenic donor sites will be distributed throughout the crystal and the crystal will have a certain density of negative carriers built in.

You might think that these carriers would get swept away as soon as any small electric field was put across the crystal. This will not happen, however, because the arsenic atoms in the body of the crystal each have a positive charge. If the body of the crystal is to remain neutral, the average density of negative carrier electrons must be equal to the density of donor sites. If you put two electrodes on the edges of such a crystal and connect them to a battery, a current will flow; but as the carrier electrons are swept out at one end, new conduction electrons must be introduced from the electrode on the other end so that the average density of conduction electrons is left very nearly equal to the density of donor sites.

Since the donor sites are positively charged, there will be some tendency for them to capture some of the conduction electrons as they diffuse around inside the crystal. A donor site can, therefore, act as a trap such as those we discussed in the last section. But if the trapping energy is sufficiently small—as it is for arsenic—the number of carriers which are trapped at any one time is a small fraction of the total. For a complete understanding of the behavior of semiconductors

one must take into account this trapping. For the rest of our discussion, however, we will assume that the trapping energy is sufficiently low and the temperature is sufficiently high, that all of the donor sites have given up their electrons. This is, of course, just an approximation.

It is also possible to build into a germanium crystal some impurity atom whose valence is 3, such as aluminum. The aluminum atom tries to act as a valence 4 object by stealing an extra electron. It can steal an electron from some nearby germanium atom and end up as a negatively charged atom with an effective valence of 4. Of course, when it steals the electron from a germanium atom, it leaves a hole there; and this hole can wander around in the crystal as a positive carrier. An impurity atom which can produce a hole in this way is called an *acceptor* because it "accepts" an electron. If a germanium or a silicon crystal is grown from a melt to which a small amount of aluminum impurity has been added, the crystal will have built-in a certain density of holes which can act as positive carriers.

When a donor or an acceptor impurity is added to a semiconductor, we say that the material has been "doped."

When a germanium crystal with some built-in donor impurities is at room temperature, some conduction electrons are contributed by the thermally induced electron-hole pair creation as well as by the donor sites. The electrons from both sources are, naturally, equivalent, and it is the total number N_n which comes into play in the statistical processes that lead to equilibrium. If the temperature is not too low, the number of negative carriers contributed by the donor impurity atoms is roughly equal to the number of impurity atoms present. In equilibrium Eq. (14.4) must still be valid; at a given temperature the product N_nN_p is determined. This means that if we add some donor impurity which increases N_n, the number N_p of positive carriers will have to decrease by such an amount that N_nN_p is unchanged. If the impurity concentration is high enough, the number N_n of negative carriers is determined by the number of donor sites and is nearly independent of temperature—all of the variation in the exponential factor is supplied by N_p, even though it is much less than N_n. An otherwise pure crystal with a small concentration of donor impurity will have a majority of negative carriers; such a material is called an "n-type" semiconductor.

If an acceptor-type impurity is added to the crystal lattice, some of the new holes will drift around and annihilate some of the free electrons produced by thermal fluctuation. This process will go on until Eq. (14.4) is satisfied. Under equilibrium conditions the number of positive carriers will be increased and the number of negative carriers will be decreased, leaving the product a constant. A material with an excess of positive carriers is called a "p-type" semiconductor.

If we put two electrodes on a piece of semiconductor crystal and connect them to a source of potential difference, there will be an electric field inside the crystal. The electric field will cause the positive and the negative carriers to move, and an electric current will flow. Let's consider first what will happen in an n-type material in which there is a large majority of negative carriers. For such material we can disregard the holes; they will contribute very little to the current because there are so few of them. In an ideal crystal the carriers would move across without any impediment. In a real crystal at a finite temperature, however,—especially in a crystal with some impurities—the electrons do not move completely freely. They are continually making collisions which knock them out of their original trajectories, that is, changing their momentum. These collisions are just exactly the scatterings we talked about in the last chapter and occur at any irregularity in the crystal lattice. In an n-type material the main causes of scattering are the very donor sites that are producing the carriers. Since the conduction electrons have a very slightly different energy at the donor sites, the probability waves are scattered from that point. Even in a perfectly pure crystal, however, there are (at any finite temperature) irregularities in the lattice due to thermal vibrations. From the classical point of view we can say that the atoms aren't lined up exactly on a regular lattice, but are, at any instant, slightly out of place due to their thermal

vibrations. The energy E_0 associated with each lattice point in the theory we described in Chapter 13 varies a little bit from place to place so that the waves of probability amplitude are not transmitted perfectly but are scattered in an irregular fashion. At very high temperatures or for very pure materials this scattering may become important, but in most doped materials used in practical devices the impurity atoms contribute most of the scattering. We would like now to make an estimate of the electrical conductivity of such a material.

When an electric field is applied to an n-type semiconductor, each negative carrier will be accelerated in this field, picking up velocity until it is scattered from one of the donor sites. This means that the carriers which are ordinarily moving about in a random fashion with their thermal energies will pick up an average drift velocity along the lines of the electric field and give rise to a current through the crystal. The drift velocity is in general rather small compared with the typical thermal velocities so that we can estimate the current by assuming that the average time that the carrier travels between scatterings is a constant. Let's say that the negative carrier has an effective electric charge q_n. In an electric field $\mathcal{E}$, the force on the carrier will be $q_n\mathcal{E}$. In Section 43–3 of Volume I we calculated the average drift velocity under such circumstances and found that it is given by $F\tau/m$, where F is the force on the charge, τ is the mean free time between collisions, and m is the mass. We should use the effective mass we calculated in the last chapter but since we want to make a rough calculation we will suppose that this effective mass is the same in all directions. Here we will call it m_n. With this approximation the average drift velocity will be

$$v_{\text{drift}} = \frac{q_n\mathcal{E}\tau_n}{m_n}. \tag{14.5}$$

Knowing the drift velocity we can find the current. Electric current density j is just the number of carriers per unit volume, N_n, multiplied by the average drift velocity, and by the charge on each carrier. The current density is therefore

$$j = N_n v_{\text{drift}} q_n = \frac{N_n q_n^2 \tau_n}{m_n}\mathcal{E}. \tag{14.6}$$

We see that the current density is proportional to the electric field; such a semiconductor material obeys Ohm's law. The coefficient of proportionality between j and $\mathcal{E}$, the conductivity σ, is

$$\sigma = \frac{N_n q_n^2 \tau_n}{m_n}. \tag{14.7}$$

For an n-type material the conductivity is relatively independent of temperature. First, the number of majority carriers N_n is determined primarily by the density of donors in the crystal (so long as the temperature is not so low that too many of the carriers are trapped). Second, the mean time between collisions τ_n is mainly controlled by the density of impurity atoms, which is, of course, independent of the temperature.

We can apply all the same arguments to a p-type material, changing only the values of the parameters which appear in Eq. (14.7). If there are comparable numbers of both negative and positive carriers present at the same time, we must add the contributions from each kind of carrier. The total conductivity will be given by

$$\sigma = \frac{N_n q_n^2 \tau_n}{m_n} + \frac{N_p q_p^2 \tau_p}{m_p}. \tag{14.8}$$

For very pure materials, N_p and N_n will be nearly equal. They will be smaller than in a doped material, so the conductivity will be less. Also they will vary rapidly with temperature (like $e^{-E_{\text{gap}}/\kappa T}$, as we have seen), so the conductivity may change extremely fast with temperature.

14-3 The Hall effect

It is certainly a peculiar thing that in a substance where the only relatively free objects are electrons, there should be an electrical current carried by holes that behave like positive particles. We would like, therefore, to describe an experiment that shows in a rather clear way that the sign of the carrier of electric current is quite definitely positive. Suppose we have a block made of semiconductor material—it could also be a metal—and we put an electric field on it so as to draw a current in some direction, say the horizontal direction as drawn in Fig. 14–6. Now suppose we put a magnetic field on the block pointing at a right angle to the current, say *into* the plane of the figure. The moving carriers will feel a magnetic force $q(v \times B)$. And since the average drift velocity is either right or left—depending on the sign of the charge on the carrier—the average magnetic force on the carriers will be either up or down. No, that is not right! For the directions we have assumed for the current and the magnetic field the magnetic force on the moving charges will always be *up*. Positive charges moving in the direction of j (to the right) will feel an upward force. If the current is carried by negative charges, they will be moving left (for the same sign of the conduction current) and they will also feel an upward force. Under steady conditions, however, there is no upward motion of the carriers because the current can flow only from left to right. What happens is that a few of the charges initially flow upward, producing a surface charge density along the upper surface of semiconductor—leaving an equal and opposite surface charge density along the bottom surface of the crystal. The charges pile up on the top and bottom surfaces until the electric forces they produce on the moving charges just exactly cancel the magnetic force (on the average) so that the steady current flows horizontally. The charges on the top and bottom surfaces will produce a potential difference vertically across the crystal which can be measured with a high-resistance voltmeter, as shown in Fig. 14–7. The sign of the potential difference registered by the voltmeter will depend on the sign of the carrier charges responsible for the current.

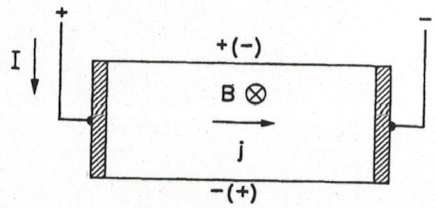

Fig. 14–6. The Hall effect comes from the magnetic forces on the carriers.

When such experiments were first done it was expected that the sign of the potential difference would be negative as one would expect for negative conduction electrons. People were, therefore, quite surprised to find that for some materials the sign of the potential difference was in the opposite direction. It appeared that the current carrier was a particle with a positive charge. From our discussion of doped semiconductors it is understandable that an n-type semiconductor should produce the sign of potential difference appropriate to negative carriers, and that a p-type semiconductor should give an opposite potential difference, since the current is carried by the positively charged holes.

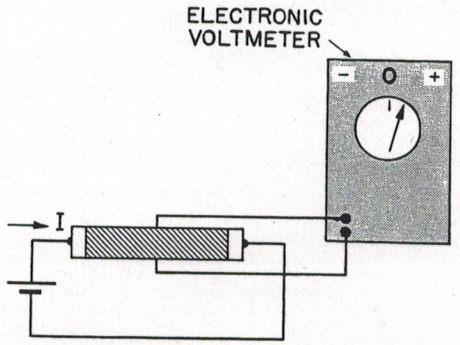

Fig. 14–7. Measuring the Hall effect.

The original discovery of the anomalous sign of the potential difference in the Hall effect was made in a metal rather than a semiconductor. It had been assumed that in metals the conduction was always by electron; however, it was found out that for berylium the potential difference had the wrong sign. It is now understood that in metals as well as in semiconductors it is possible, in certain circumstances, that the "objects" responsible for the conduction are holes. Although it is ultimately the electrons in the crystal which do the moving, nevertheless, the relationship of the momentum and the energy, and the response to external fields is exactly what one would expect for an electric current carried by positive particles.

Let's see if we can make a quantitative estimate of the magnitude of the voltage difference expected from the Hall effect. If the voltmeter in Fig. 14–7 draws a negligible current, then the charges inside the semiconductor must be moving from left to right and the vertical magnetic force must be precisely cancelled by a vertical electric field which we will call $\mathcal{E}_{tr}$ (the "tr" is for "transverse"). If this electric field is to cancel the magnetic forces, we must have

$$\mathcal{E}_{tr} = -v_{drift} \times B. \tag{14.9}$$

Using the relation between the drift velocity and the electric current density given

in Eq. (14.6), we get

$$\mathcal{E}_{tr} = -\frac{1}{qN} jB.$$

The potential difference between the top and the bottom of the crystal is, of course, this electric field strength multiplied by the height of the crystal. The electric field strength $\mathcal{E}_{tr}$ in the crystal is proportional to the current density and to the magnetic field strength. The constant of proportionality $1/qN$ is called the Hall coefficient and is usually represented by the symbol R_H. The Hall coefficient depends just on the density of carriers—provided that carriers of one sign are in a large majority. Measurement of the Hall effect is, therefore, one convenient way of determining experimentally the density of carriers in a semiconductor.

14–4 Semiconductor junctions

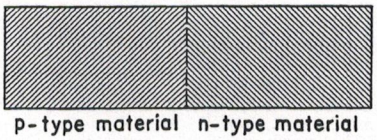

Fig. 14–8. A *p-n* junction.

We would like to discuss now what happens if we take two pieces of germanium or silicon with different internal characteristics—say different kinds or amounts of doping—and put them together to make a "junction." Let's start out with what is called a *p-n* junction in which we have *p*-type germanium on one side of the boundary and *n*-type germanium on the other side of the boundary—as sketched in Fig. 14–8. Actually, it is not practical to put together two separate pieces of crystal and have them in uniform contact on an atomic scale. Instead, junctions are made out of a single crystal which has been modified in the two separate regions. One way is to add some suitable doping impurity to the "melt" after only half of the crystal has grown. Another way is to paint a little of the impurity element on the surface and then heat the crystal causing some impurity atoms to diffuse into the body of the crystal. Junctions made in these ways do not have a sharp boundary, although the boundaries can be made as thin as 10^{-4} centimeters or so. For our discussions we will imagine an ideal situation in which these two regions of the crystal with different properties meeting at a sharp boundary.

On the *n*-type side of *p-n* junction there are free electrons which can move about, as well as the fixed donor sites which balance the overall electric charge. On the *p*-type side there are free holes moving about and an equal number of negative acceptor sites keeping the charge balanced. Actually, that describes the situation before we put the two materials in contact. Once they are connected together the situation will change near the boundary. When the electrons in the *n*-type material arrive at the boundary they will not be reflected back as they would at a free surface, but are able to go right on into the *p*-type material. Some of the electrons of the *n*-type material will, therefore, tend to diffuse over into the *p*-type material where there are fewer electrons. This cannot go on forever because as we lose electrons from the *n*-side the net positive charge there increases until finally an electric voltage is built up which retards the diffusion of electrons into the *p*-side. In a similar way, the positive carriers of the *p*-type material can diffuse across the junction into the *n*-type material. When they do this they leave behind an excess of negative charge. Under equilibrium conditions the net diffusion current must be zero. This brought about by the electric fields which are established in such a way as to draw the positive carriers back toward the *p*-type material.

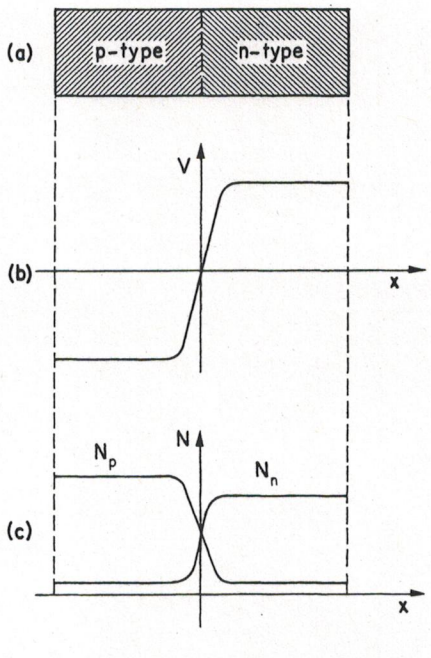

Fig. 14–9. The electric potential and the carrier densities in an unbiased semiconductor junction.

The two diffusion processes we have been describing go on simultaneously and, you will notice, both act in the direction which will charge up the *n*-type material in a positive sense and the *p*-type material in a negative sense. Because of the finite conductivity of the semiconductor material, the change in potential from the *p*-side to the *n*-side will occur in a relatively narrow region near the boundary; the main body of each block of material will have a uniform potential. Let's imagine an *x*-axis in a direction perpendicular to the boundary surface. Then the electric potential will vary with *x*, as shown in Fig. 14–9(b). We have also shown in part (c) of the figure the expected variation of the density N_n of *n*-carriers and the density N_p of *p*-carriers. Far away from the junction the carrier densities N_p and N_n should be just the equilibrium density we would expect for individual blocks of materials at the same temperature. (We have drawn the figure for a

junction in which the p-type material is more heavily doped than the n-type material.) Because of the potential gradient at the junction, the positive carriers have to climb up a potential hill to get to the n-type side. This means that under equilibrium conditions there can be fewer positive carriers in the n-type material than there are in the p-type material. Remembering the laws of statistical mechanics, we expect that the ratio of p-type carriers on the two sides to be given by the following equation:

$$\frac{N_p(n\text{-side})}{N_p(p\text{-side})} = e^{-q_pV/\kappa T}. \tag{14.10}$$

The product q_pV in the numerator of the exponential is just the energy required to carry a charge of q_p through a potential difference V.

We have a precisely similar equation for the densities of the n-type carriers:

$$\frac{N_n(n\text{-side})}{N_n(p\text{-side})} = e^{-q_nV/\kappa T}. \tag{14.11}$$

If we know the equilibrium densities in each of the two materials, we can use either of the two equations above to determine the potential difference across the junction.

Notice that if Eqs. (14.10) and (14.11) are to give the same value for the potential difference V, the product N_pN_n must be the same for the p-side as for the n-side. (Remember that $q_n = -q_p$.) We have seen earlier, however, that this product depends only on the temperature and the gap energy of the crystal. Provided both sides of the crystal are at the same temperature, the two equations are consistent with the same value of the potential difference.

Since there is a potential difference from one side of the junction to the other, it looks something like a battery. Perhaps if we connect a wire from the n-type side to the p-type side we will get an electrical current. That would be nice because then the current would flow forever without using up any material and we would have an infinite source of energy in violation of the second law of thermodynamics! There is, however, no current if you connect a wire from the p-side to the n-side. And the reason is easy to see. Suppose we imagine first a wire made out of a piece of undoped material. When we connect this wire to the n-type side, we have a junction. There will be a potential difference across this junction. Let's say that it is just one-half the potential difference from the p-type material to the n-type material. When we connect our undoped wire to the p-type side of the junction, there is also a potential difference at this junction—again, one-half the potential drop across the p-n junction. At all the junctions the potential differences adjust themselves so that there is no net current flow in the circuit. Whatever kind of wire you use to connect together the two sides of the n-p junction, you are producing two new junctions, and so long as all the junctions are at the same temperature, the potential jumps at the junctions all compensate each other and no current will flow in the circuit. It does turn out, however—if you work out the details—that if some of the junctions are at a different temperature than the other junctions, currents will flow. Some of the junctions will be heated and others will be cooled by this current and thermal energy will be converted into electrical energy. This effect is responsible for the operation of thermocouples which are used for measuring temperatures, and of thermoelectric generators. The same effect is also used to make small refrigerators.

If we cannot measure the potential difference between the two sides of an n-p junction, how can we really be sure that the potential gradient shown in Fig. 14–9 really exists? One way is to shine light on the junction. When the light photons are absorbed they can produce an electron-hole pair. In the strong electric field that exists at the junction (equal to the slope of the potential curve of Fig. 14–9) the hole will be driven into the p-type region and the electron will be driven into the n-type region. If the two sides of the junction are now connected to an external circuit, these extra charges will provide a current. The energy of the light will be converted into electrical energy in the junction. The solar cells which generate electrical power for the operation of some of our satellites operate on this principle.

In our discussion of the operation of a semiconductor junction we have been assuming that the holes and the electrons act more-or-less independently—except that they somehow get into proper statistical equilibrium. When we were describing the current produced by light shining on the junction, we were assuming that an electron or a hole produced in the junction region would get into the main body of the crystal before being annihilated by a carrier of the opposite polarity. In the immediate vicinity of the junction, where the density of carriers of both signs is approximately equal, the effect of electron-hole annihilation (or as it is often called, "recombination") is an important effect, and in a detailed analysis of a semiconductor junction must be properly taken into account. We have been assuming that a hole or an electron produced in a junction region has a good chance of getting into the main body of the crystal before recombining. The typical time for an electron or a hole to find an opposite partner and annihilate it is for typical semiconductor materials in the range between 10^{-3} and 10^{-7} seconds. This time is, incidentally, much longer than the mean free time τ between collisions with scattering sites in the crystal which we used in the analysis of conductivity. In a typical *n-p* junction, the time for an electron or hole formed in the junction region to be swept away into the body of the crystal is generally much shorter than the recombination time. Most of the pairs will, therefore, contribute to an external current.

14–5 Rectification at a semiconductor junction

We would like to show next how it is that a *p-n* junction can act like a rectifier. If we put a voltage across the junction, a large current will flow if the polarity is in one direction, but a very small current will flow if the same voltage is applied in the opposite direction. If an alternating voltage is applied across the junction, a net current will flow in one direction—the current is "rectified." Let's look again at what is going on in the equilibrium condition described by the graphs of Fig. 14–9. In the *p*-type material there is a large concentration N_p of positive carriers. These carriers are diffusing around and a certain number of them each second approach the junction. This current of positive carriers which approaches the junction is proportional to N_p. Most of them, however, are turned back by the high potential hill at the junction and only the fraction $e^{-qV/\kappa T}$ gets through. There is also a current of positive carriers approaching the junction from the other side. This current is also proportional to the density of positive carriers in the *n*-type region, but the carrier density here is much smaller than the density on the *p*-type side. When the positive carriers approach the junction from the *n*-type side, they find a hill with a negative slope and immediately slide downhill to the *p*-type side of the junction. Let's call this current I_0. Under equilibrium the currents from the two directions are equal. We expect then the following relation:

$$I_0 \sim N_p(n\text{-side}) = N_p(p\text{-side})e^{-qV/\kappa T}. \tag{14.12}$$

You will notice that this equation is really just the same as Eq. (14–10). We have just derived it in a different way.

Suppose, however, that we lower the voltage on the *n*-side of the junction by an amount ΔV—which we can do by applying an external potential difference to the junction. Now the difference in potential across the potential hill is no longer V but $V - \Delta V$. The current of positive carriers from the *p*-side to the *n*-side will now have this potential difference in its exponential factor. Calling this current I_1, we have

$$I_1 \sim N_p(p\text{-side})e^{-q(V-\Delta V)/\kappa T}.$$

This current is larger than I_0 by just the factor $e^{q\Delta V/\kappa T}$. So we have the following relation between I_1 and I_0:

$$I_1 = I_0 e^{+q\Delta V/\kappa T}. \tag{14.13}$$

The current from the *p*-side increases exponentially with the externally applied voltage ΔV. The current of positive carriers from the *n*-side, however, remains

constant so long as ΔV is not too large. When they approach the barrier, these carriers will still find a downhill potential and will all fall down to the p-side. (If ΔV is larger than the natural potential difference V, the situation would change, but we will not consider what happens at such high voltages.) The net current I of positive carriers which flows across the junction is then the difference between the currents from the two sides:

$$I = I_0(e^{+q\Delta V/\kappa T} - 1). \tag{14.14}$$

The net current I of holes flows into the n-type region. There the holes diffuse into the body of the n-region, where they are eventually annihilated by the majority n-type carriers—the electrons. The electrons which are lost in this annihilation will be made up by a current of electrons from the external terminal of the n-type material.

When ΔV is zero, the net current in Eq. (14.14) is zero. For positive ΔV the current increases rapidly with the applied voltage. For negative ΔV the current reverses in sign, but the exponential term soon becomes negligible and the negative current never exceeds I_0—which under our assumptions is rather small. This back current I_0 is limited by the small density of the minority carriers on the n-side of the junction.

If you go through exactly the same analysis for the current of negative carriers which flows across the junction, first with no potential difference and then with a small externally applied potential difference ΔV, you get again an equation just like (14.14) for the net electron current. Since the total current is the sum of the currents contributed by the two carriers, Eq. (14.14) still applies for the total current provided we identify I_0 as the maximum current which can flow for a reversed voltage.

The voltage-current characteristic of Eq. (14.14) is shown in Fig. 14–10. It shows the typical behavior of solid state diodes—such as those used in modern computers. We should remark that Eq. (14.14) is true only for small voltages. For voltages comparable to or larger than the natural internal voltage difference V, other effects come into play and the current no longer obeys the simple equation.

You may remember, incidentally, that we got exactly the same equation we have found here in Eq. (14.14) when we discussed the "mechanical rectifier"—the ratchet and pawl—in Chapter 46 of Volume I. We get the same equations in the two situations because the basic physical processes are quite similar.

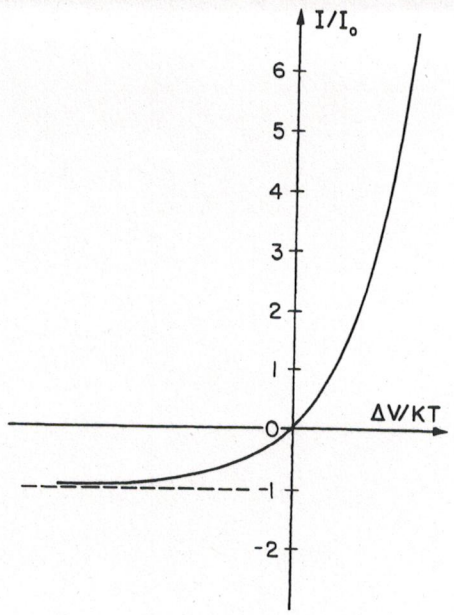

Fig. 14–10. The current through a junction as a function of the voltage across it.

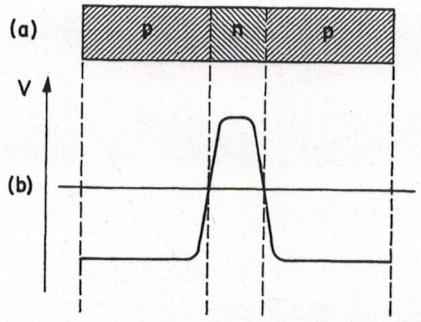

Fig. 14–11. The potential distribution in a transistor with no applied voltages.

14–6 The transistor

Perhaps the most important application of semiconductors is in the transistor. The transistor consists of two semiconductor junctions very close together. Its operation is based in part on the same principles that we just described for the semiconductor diode—the rectifying junction. Suppose we make a little bar of germanium with three distinct regions, a p-type region, an n-type region, and another p-type region, as shown in Fig. 14–11(a). This combination is called a p-n-p transistor. Each of the two junctions in the transistor will behave much in the way we have described in the last section. In particular, there will be a potential gradient at each junction having a certain potential drop from the n-type region to each p-type region. If the two p-type regions have the same internal properties, the variation in potential as we go across the crystal will be as shown in the graph of Fig. 14–11(b).

Now let's imagine that we connect each of the three regions to external voltage sources as shown in part (a) of Fig. 14–12. We will refer all voltages to the terminal connected to the left-hand p-region so it will be, by definition, at zero potential. We will call this terminal the *emitter*. The n-type region is called the *base* and it is connected to a slightly negative potential. The right-hand p-type region is called the *collector*, and is connected to a somewhat larger negative potential. Under these circumstances the variation of potential across the crystal will be as shown in the graph of Fig. 14–12(b).

Let's first see what happens to the positive carriers, since it is primarily their behavior which controls the operation of the p-n-p transistor. Since the emitter is

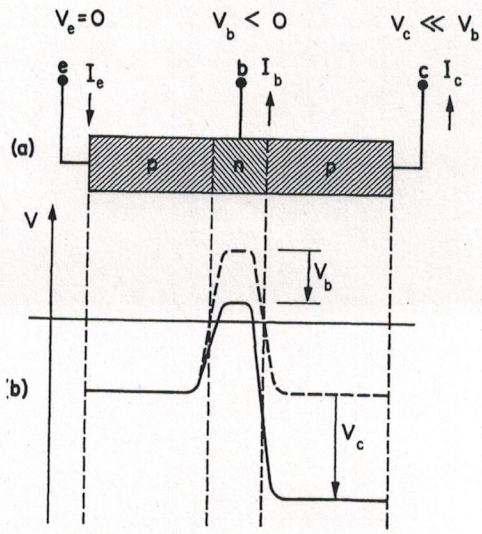

Fig. 14–12. The potential distribution in an operating transistor.

at a relatively more positive potential than the base, a current of positive carriers will flow from the emitter region into the base region. A relatively large current flows, since we have a junction operating with a "forward voltage"—corresponding to the right-hand half of the graph in Fig. 14–10. With these conditions, positive carriers or holes are being "emitted" from the *p*-type region into the *n*-type region. You might think that this current would flow out of the *n*-type region through the base terminal *b*. Now, however, comes the secret of the transistor. The *n*-type region is made very thin—typically 10^{-3} cm or less, much narrower than its transverse dimensions. This means that as the holes enter the *n*-type region they have a very good chance of diffusing across to the other junction before they are annihilated by the electrons in the *n*-type region. When they get to the right-hand boundary of the *n*-type region they find a steep downward potential hill and immediately fall into the right-hand *p*-type region. This side of the crystal is called the collector because it "collects" the holes after they have diffused across the *n*-type region. In a typical transistor, all but a fraction of a percent of the hole current which leaves the emitter and enters the base is collected in the collector region, and only the small remainder contributes to the net base current. The sum of the base and collector currents is, of course, equal to the emitter current.

Now imagine what happens if we vary slightly the potential V_b on the base terminal. Since we are on a relatively steep part of the curve of Fig. 14–10, a small variation of the potential V_b will cause a rather large change in the emitter current I_e. Since the collector voltage V_c is much more negative than the base voltage, these slight variations in potential will not effect appreciably the steep potential hill between the base and the collector. Most of the positive carriers emitted into the *n*-region will still be caught by the collector. Thus as we vary the potential of the base electrode, there will be a corresponding variation in the collector current I_c. The essential point, however, is that the base current I_b always remains a small fraction of the collector current. The transistor is an amplifier; a small current I_b introduced into the base electrode gives a large current —100 or so times higher—at the collector electrode.

What about the electrons—the negative carriers that we have been neglecting so far? First, note that we do not expect any significant electron current to flow between the base and the collector. With a large negative voltage on the collector, the electrons in the base would have to climb a very high potential energy hill and the probability of doing that is very small. There is a very small current of electrons to the collector.

On the other hand, the electrons in the base *can* go into the emitter region. In fact, you might expect the electron current in this direction to be comparable to the hole current from the emitter into the base. Such an electron current isn't useful, and, on the contrary, is bad because it increases the total base current required for a given current of holes to the collector. The transistor is, therefore, designed to minimize the electron current to the emitter. The electron current is proportional to N_n(base), the density of negative carriers in the base material while the hole current from the emitter depends on N_p(emitter), the density of positive carriers in the emitter region. By using relatively little doping in the *n*-type material N_n(base) can be made much smaller than N_p(emitter). (The very thin base region also helps a great deal because the sweeping out of the holes in this region by the collector increases significantly the average hole current from the emitter into the base, while leaving the electron current unchanged.) The net result is that the electron current across the emitter-base junction can be made much less than the hole current, so that the electrons do not play any significant role in operation of the *p-n-p* transistor. The currents are dominated by motion of the holes, and the transistor performs as an amplifier as we have described above.

It is also possible to make a transistor by interchanging the *p*-type and *n*-type materials in Fig. 14–11. Then we have what is called an *n-p-n* transistor. In the *n-p-n* transistor the main currents are carried by the electrons which flow from the emitter into the base and from there to the collector. Obviously, all the arguments we have made for the *p-n-p* transistor also apply to the *n-p-n* transistor if the potentials of the electrodes are chosen with the opposite signs.