1 Prologue: an atomistic view of electrical resistance

Let me start with a brief explanation since this is not a typical "prologue." For one it is too long, indeed as long as the average chapter. The reason for this is that I have a very broad objective in mind, namely to review *all* the relevant concepts needed to understand current flow through a very small object that has only one energy level in the energy range of interest. Remarkably enough, this can be done without invoking any significant background in quantum mechanics. What requires serious quantum mechanics is to understand where the energy levels come from and to describe large conductors with multiple energy levels. Before we get lost in these details (and we have the whole book for it!) it is useful to understand the factors that influence the current–voltage relation of a really small object.

This "bottom-up" view is different from the standard "top-down" approach to electrical resistance. We start in college by learning that the conductance G (inverse of the resistance) of a large macroscopic conductor is directly proportional to its cross-sectional area A and inversely proportional to its length L:

$$G = \sigma A/L \qquad \text{(Ohm's law)}$$

where the conductivity σ is a material property of the conductor. Years later in graduate school we learn about the factors that determine the conductivity and if we stick around long enough we eventually talk about what happens when the conductor is so small that one cannot define its conductivity. I believe the reason for this "top-down" approach is historical. Till recently, no one was sure how to describe the conductance of a really small object, or if it even made sense to talk about the conductance of something really small. To measure the conductance of anything we need to attach two large contact pads to it, across which a battery can be connected. No one knew how to attach contact pads to a small molecule till the late twentieth century, and so no one knew what the conductance of a really small object was. But now that we are able to do so, the answers look fairly simple, except for unusual things like the Kondo effect that are seen only for a special range of parameters. Of course, it is quite likely that many new effects will be discovered as we experiment more on small conductors and the description presented here is certainly not intended to be the last word. But I think it should be the "first

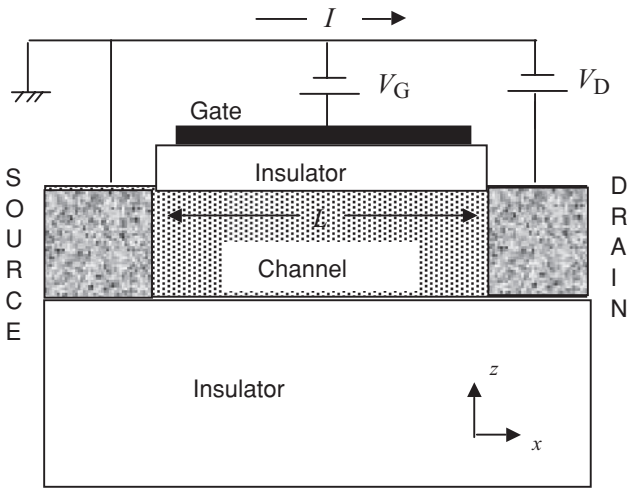


Fig. 1.1 Sketch of a nanoscale field effect transistor. The insulator should be thick enough to ensure that no current flows into the gate terminal, but thin enough to ensure that the gate voltage can control the electron density in the channel.

word" since the traditional top-down approach tends to obscure the simple physics of very small conductors.

The generic structure I will often use is a simple version of a "nanotransistor" consisting of a semiconducting channel separated by an insulator layer (typically silicon dioxide) from the metallic gate (Fig. 1.1). The regions marked source and drain are the two contact pads, which are assumed to be highly conducting. The resistance of the channel determines the current that flows from the source to the drain when a voltage V_D is applied between them. The voltage V_G on the gate is used to control the electron density in the channel and hence its resistance. Such a voltage-controlled resistor is the essence of any field effect transistor (FET) although the details differ from one version to another. The channel length L has been progressively reduced from $\sim$10 μm in 1960 to $\sim$0.1 μm in 2000, allowing circuit designers to pack $(100)^2 = 10\,000$ times more transistors (and hence that much more computing power) into a chip of given surface area. This increase in packing density is at the heart of the computer revolution. How much longer can the downscaling continue? No one really knows. However, one thing seems certain. Regardless of what form future electronic devices take, *we will have to learn how to model and describe the electronic properties of device structures that are engineered on an atomic scale.* The examples I will use in this book may or may not be important twenty years from now. But the problem of current flow touches on some of the deepest issues of physics related to the nature of "friction" on a microscopic scale and the emergence of irreversibility from reversible laws. The concepts we will discuss represent key fundamental concepts of quantum mechanics and non-equilibrium

statistical mechanics that should be relevant to the analysis and design of nanoscale devices for many years into the future.

Outline: To model the flow of current, the first step is to draw an equilibrium energy level diagram and locate the electrochemical potential μ (also called the Fermi level or Fermi energy) set by the source and drain contacts (Section 1.1). Current flows when an external device such as a battery maintains the two contacts at different electrochemical potentials μ_1 and μ_2, driving the channel into a non-equilibrium state (Section 1.2). The current through a really small device with only one energy level in the range of interest is easily calculated and, as we might expect, depends on the quality of the contacts. But what is not obvious (and was not appreciated before the late 1980s) is that there is a maximum conductance for a channel with one level (in the energy range of interest), which is a fundamental constant related to the charge on an electron and Planck's constant:

$$G_0 \equiv q^2/h = 38.7\,\mu\text{S} = (25.8\,\text{k}\Omega)^{-1} \tag{1.1}$$

Actually small channels typically have two levels (one for up spin and one for down spin) at the same energy ("degenerate" levels) making the maximum conductance equal to $2G_0$. We can always measure conductances lower than this, if the contacts are bad. But the point is that there is an upper limit to the conductance that can be achieved even with the most perfect of contacts (Section 1.3). In Section 1.4, I will explain the important role played by charging and electrostatics in determining the shape of the current–voltage (I–V) characteristics, and how this aspect is coupled with the equations for quantum transport. Once this aspect has been incorporated we have all the basic physics needed to describe a one-level channel that is coupled "well" to the contacts. But if the channel is weakly coupled, there is some additional physics that I will discuss in Section 1.5. Finally, in Section 1.6, I will explain how the one-level description is extended to larger devices with multiple energy levels, eventually leading to Ohm's law. It is this extension to larger devices that requires the advanced concepts of quantum statistical mechanics that constitute the subject matter of the rest of this book.

1.1 Energy level diagram

Figure 1.1.1 shows the typical current–voltage characteristics for a well-designed transistor of the type shown in Fig. 1.1 having a width of 1 μm in the y-direction perpendicular to the plane of the paper. At low gate voltages, the transistor is in its off state, and very little current flows in response to a drain voltage V_D. Beyond a certain gate voltage, called the threshold voltage V_T, the transistor is turned on and the ON-current increases with increasing gate voltage V_G. For a fixed gate voltage, the current I increases at first with drain voltage, but it then tends to level off and saturate at a value referred to as the

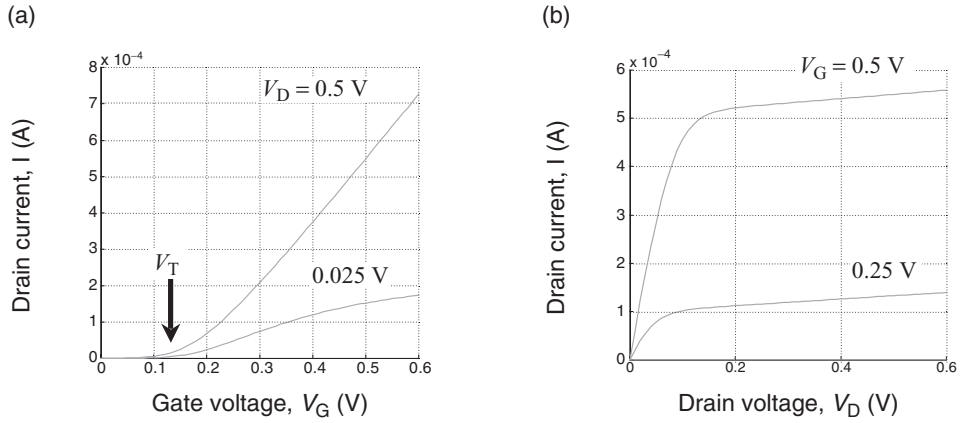


Fig. 1.1.1 (a) Drain current I as a function of the gate voltage V_G for different values of the drain voltage V_D. (b) Drain current as a function of the drain voltage for different values of the gate voltage.

ON-current. Let us start by trying to understand why the current increases when the gate voltage exceeds V_T (Fig. 1.1.1a).

The first step in understanding the operation of any inhomogeneous device structure (like the generic one shown in Fig. 1.1) is to draw an *equilibrium* energy level diagram (sometimes called a "band diagram") assuming that there is no voltage applied between the source and the drain. Electrons in a semiconductor occupy a set of energy levels that form bands as sketched in Fig. 1.1.2. Experimentally, one way to measure the occupied energy levels is to find the minimum energy of a photon required to knock an electron out into vacuum (photoemission (PE) experiments). We can describe the process symbolically as

$$S + h\nu \rightarrow S^+ + e^-$$

where "S" stands for the semiconductor device (or any material for that matter!).

The empty levels, of course, cannot be measured the same way since there is no electron to knock out. We need an inverse photoemission (IPE) experiment where an incident electron is absorbed with the emission of photons:

$$S + e^- \rightarrow S^- + h\nu$$

Other experiments like optical absorption also provide information regarding energy levels. All these experiments would be equivalent if electrons did not interact with each other and we could knock one electron around without affecting everything else around it. But in the real world subtle considerations are needed to relate the measured energies to those we use and we will discuss some of these issues in Chapter 2.

We will assume that the large contact regions (labeled source and drain in Fig. 1.1) have a continuous distribution of states. This is true if the contacts are metallic, but not

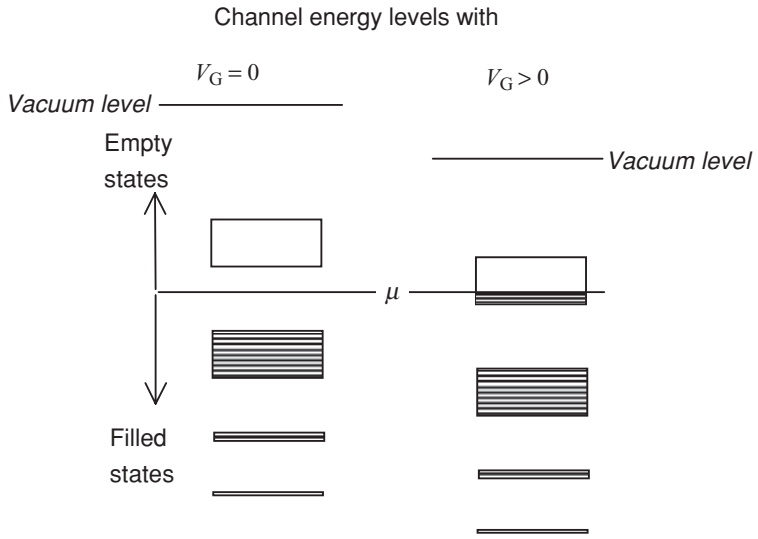


Fig. 1.1.2 Allowed energy levels that can be occupied by electrons in the active region of a device like the channel in Fig. 1.1. A positive gate voltage V_G moves the energy levels down while the electrochemical potential μ is fixed by the source and drain contacts, which are assumed to be in equilibrium with each other ($V_D = 0$).

exactly true of semiconducting contacts, and interesting effects like a decrease in the current with an increase in the voltage (sometimes referred to as negative differential resistance (NDR)) can arise as a result (see Exercise E.1.4); however, we will ignore this possibility in our discussion. The allowed states are occupied up to some energy μ (called the electrochemical potential) which too can be located using photoemission measurements. The work function is defined as the minimum energy of a photon needed to knock a photoelectron out of the metal and it tells us how far below the vacuum level μ is located.

Fermi function: If the source and drain regions are coupled to the channel (with V_D held at zero), then electrons will flow in and out of the device bringing them all in equilibrium with a common electrochemical potential, μ, just as two materials in equilibrium acquire a common temperature, T. In this equilibrium state, the average (over time) number of electrons in any energy level is typically not an integer, but is given by the Fermi function:

$$f_0(E-\mu) = \frac{1}{1+\exp[(E-\mu)/k_B T]} \tag{1.1.1}$$

Energy levels far below μ are always full so that $f_0 = 1$, while energy levels far above μ are always empty with $f_0 = 0$. Energy levels within a few $k_B T$ of μ are occasionally empty and occasionally full so that the average number of electrons lies

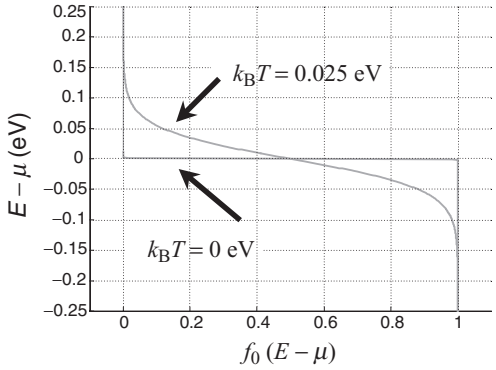


Fig. 1.1.3 The Fermi function (Eq. (1.1.1)) describing the number of electrons occupying a state with an energy E if it is in equilibrium with a large contact ("reservoir") having an electrochemical potential μ.

between 0 and 1: $0 \leq f_0 \leq 1$ (Fig. 1.1.3). Note that this number cannot exceed one because the exclusion principle forbids more than one electron per level.

n-type operation: A positive gate voltage V_G applied to the gate lowers the energy levels in the channel. However, the energy levels in the source and drain contacts are unchanged and hence the electrochemical potential μ (which must be the same everywhere) remains unaffected. As a result the energy levels move with respect to μ, driving μ into the empty band as shown in Fig. 1.1.2. This makes the channel more conductive and turns the transistor ON, since, as we will see in the next section, the current flow under bias depends on the number of energy levels available around $E = \mu$. The threshold gate voltage V_T needed to turn the transistor ON is thus determined by the energy difference between the equilibrium electrochemical potential μ and the lowest available empty state (Fig. 1.1.2) or what is called the conduction band edge.

p-type operation: Note that the number of electrons in the channel is not what determines the current flow. A negative gate voltage ($V_G < 0$), for example, reduces the number of electrons in the channel. Nevertheless the channel will become more conductive once the electrochemical potential is driven into the filled band as shown in Fig. 1.1.4, due to the availability of states (filled or otherwise) around $E = \mu$. This is an example of p-type or "hole" conduction as opposed to the example of n-type or electron conduction shown in Fig. 1.1.2. The point is that for current flow to occur, states are needed near $E = \mu$, but they need not be empty states. Filled states are just as good and it is not possible to tell from this experiment whether conduction is n-type (Fig. 1.1.2) or p-type (Fig. 1.1.4). This point should become clearer in Section 1.2 when we discuss why current flows in response to a voltage applied across the source and drain contacts.

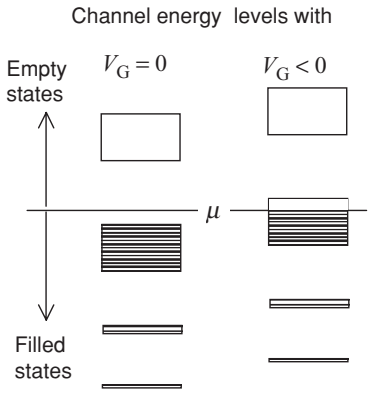


Fig. 1.1.4 Example of p-type or hole conduction. A negative gate voltage ($V_G < 0$) reduces the number of electrons in the channel. Nevertheless the channel will become more conductive once the electrochemical potential μ is driven into the filled band since conduction depends on the availability of states around $E = \mu$ and not on the total number of electrons.

Figures 1.1.2 and 1.1.4 suggest that the same device can be operated as an n-type or a p-type device simply by reversing the polarity of the gate voltage. This is true for short devices if the contacts have a continuous distribution of states as we have assumed. But in general this need not be so: for example, long devices can build up "depletion layers" near the contacts whose shape can be different for n- and p-type devices.

1.2 What makes electrons flow?

We have stated that conduction depends on the availability of states around $E = \mu$; it does not matter if they are empty or filled. To understand why, let us consider what makes electrons flow from the source to the drain. The battery lowers the energy levels in the drain contact with respect to the source contact (assuming V_D to be positive) and maintains them at distinct electrochemical potentials separated by qV_D

$$\mu_1 - \mu_2 = qV_D \tag{1.2.1}$$

giving rise to two different Fermi functions:

$$f_1(E) \equiv \frac{1}{1 + \exp[(E - \mu_1)/k_B T]} = f_0(E - \mu_1) \tag{1.2.2a}$$

$$f_2(E) \equiv \frac{1}{1 + \exp[(E - \mu_2)/k_B T]} = f_0(E - \mu_2) \tag{1.2.2b}$$

Each contact seeks to bring the channel into equilibrium with itself. The source keeps pumping electrons into it, hoping to establish equilibrium. But equilibrium is never

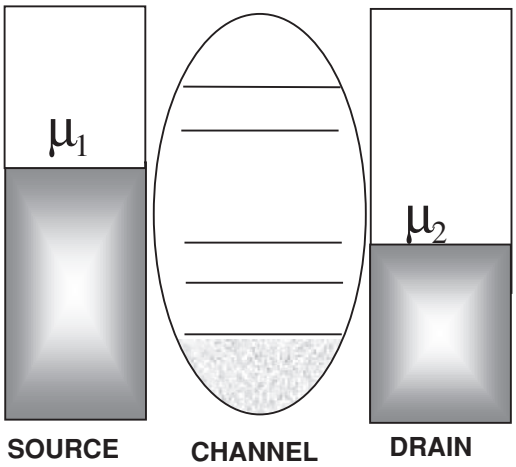


Fig. 1.2.1 A positive voltage V_D applied to the drain with respect to the source lowers the electrochemical potential at the drain: $\mu_2 = \mu_1 - qV_D$. Source and drain contacts now attempt to impose different Fermi distributions as shown, and the channel goes into a state intermediate between the two.

achieved as the drain keeps pulling electrons out in its bid to establish equilibrium with itself. The channel is thus forced into a balancing act between two reservoirs with different agendas and this sends it into a non-equilibrium state intermediate between what the source would like to see and what the drain would like to see (Fig. 1.2.1).

Rate equations for a one-level model: This balancing act is easy to see if we consider a simple one-level system, biased such that its energy ε lies between the electrochemical potentials in the two contacts (Fig. 1.2.2). Contact 1 would like to see $f_1(\varepsilon)$ electrons, while contact 2 would like to see $f_2(\varepsilon)$ electrons occupying the state where f_1 and f_2 are the source and drain Fermi functions defined in Eq. (1.2.2). The average number of electrons N at steady state will be something intermediate between $f_1(\varepsilon)$ and $f_2(\varepsilon)$. There is a net flux I_1 across the left junction that is proportional to $(f_1 - N)$, *dropping the argument* ε for clarity:

$$I_1 = \frac{q\,\gamma_1}{\hbar}(f_1 - N) \tag{1.2.3a}$$

where $-q$ is the charge per electron. Similarly the net flux I_2 across the right junction is proportional to $(f_2 - N)$ and can be written as

$$I_2 = \frac{q\,\gamma_2}{\hbar}(f_2 - N) \tag{1.2.3b}$$

We can interpret the rate constants $\gamma_1/\hbar$ and $\gamma_2/\hbar$ as the rates at which an electron placed initially in the level ε will escape into the source and drain contacts respectively. In principle, we could experimentally measure these quantities, which have the

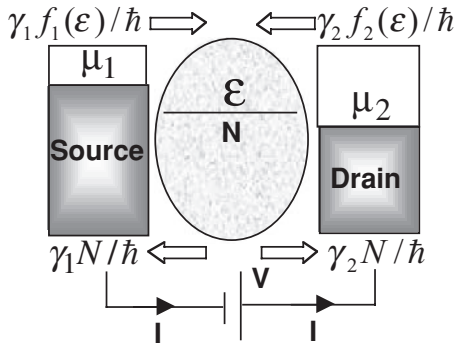


Fig. 1.2.2 Flux of electrons into and out of a one-level channel at the source and drain ends: simple rate equation picture.

dimension per second, so that γ_1 and γ_2 have the dimension of energy. At the end of this section I will say a few more words about the physics behind these equations. But for the moment, let us work out the consequences.

Current in a one-level model: At steady state there is no net flux into or out of the channel, $I_1 + I_2 = 0$, so that from Eqs. (1.2.3a, b) we obtain the reasonable result

$$N = \frac{\gamma_1 f_1 + \gamma_2 f_2}{\gamma_1 + \gamma_2} \tag{1.2.4}$$

that is, the occupation N is a weighted average of what contacts 1 and 2 would like to see. Substituting this result into Eq. (1.2.3a) or (1.2.3b) we obtain an expression for the steady-state current:

$$I = I_1 = -I_2 = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}[f_1(\varepsilon) - f_2(\varepsilon)] \tag{1.2.5}$$

This is the current per spin. We should multiply it by two if there are two spin states with the same energy.

This simple result serves to illustrate certain basic facts about the process of current flow. Firstly, no current will flow if $f_1(\varepsilon) = f_2(\varepsilon)$. A level that is way below both electrochemical potentials μ_1 and μ_2 will have $f_1(\varepsilon) = f_2(\varepsilon) = 1$ and will not contribute to the current, just like a level that is way above both potentials μ_1 and μ_2 and has $f_1(\varepsilon) = f_2(\varepsilon) = 0$. It is only when the level lies within a few k_BT of the potentials μ_1 and μ_2 that we have $f_1(\varepsilon) \neq f_2(\varepsilon)$ and a current flows. Current flow is thus the result of the "*difference in agenda*" between the contacts. Contact 1 keeps pumping in electrons striving to bring the number up from N to f_1, while contact 2 keeps pulling them out striving to bring it down to f_2. The net effect is a continuous transfer of electrons from contact 1 to 2 corresponding to a current I in the external circuit (Fig. 1.2.2). Note that the current is in a direction opposite to that of the flux of electrons, since electrons have negative charge.

It should now be clear why the process of conduction requires the presence of states around $E = \mu$. It does not matter if the states are empty (n-type, Fig. 1.1.2) or filled (p-type, Fig. 1.1.4) in equilibrium, before a drain voltage is applied. With empty states, electrons are first injected by the negative contact and subsequently collected by the positive contact. With filled states, electrons are first collected by the positive contact and subsequently refilled by the negative contact. Either way, we have current flowing in the external circuit in the same direction.

Inflow/outflow: Eqs. (1.2.3a, b) look elementary and I seldom hear anyone question them. But they hide many subtle issues that could bother more advanced readers and so I feel obliged to mention these issues briefly. I realize that I run the risk of confusing "satisfied" readers who may want to skip the rest of this section.

The right-hand sides of Eqs. (1.2.3a, b) can be interpreted as the difference between the influx and the outflux from the source and drain respectively (see Fig. 1.2.2). For example, consider the source. The outflux of $\gamma_1 N/\hbar$ is easy to justify since $\gamma_1/\hbar$ represents the rate at which an electron placed initially in the level ε will escape into the source contact. But the influx $\gamma_1 f_1/\hbar$ is harder to justify since there are many electrons in many states in the contacts, all seeking to fill up one state inside the channel and it is not obvious how to sum up the inflow from all these states. A convenient approach is to use a thermodynamic argument as follows. If the channel were in equilibrium with the source, there would be no net flux, so that the influx would equal the outflux. But the outflux under equilibrium conditions would equal $\gamma_1 f_1/\hbar$ since N would equal f_1. Under non-equilibrium conditions, N differs from f_1 but the influx remains unchanged since it depends only on the condition in the contacts which remains unchanged (note that the outflux does change giving a net current that we have calculated above).

"Pauli blocking"? Advanced readers may disagree with the statement I just made, namely that the influx "depends only on the condition in the contacts." Shouldn't the influx be reduced by the presence of electrons in the channel due to the exclusion principle ("Pauli blocking")? Specifically one could argue that the inflow and outflow (at the source contact) be identified respectively as

$$\gamma_1 f_1(1 - N) \quad \text{and} \quad \gamma_1 N(1 - f_1)$$

instead of

$$\gamma_1 f_1 \quad \text{and} \quad \gamma_1 N$$

as we have indicated in Fig. 1.2.2. It is easy to see that the net current given by the difference between inflow and outflow is the same in either case, so that the argument might appear "academic." What is not academic, however, is the level broadening that accompanies the process of coupling to the contacts, something we need to include in order to get quantitatively correct results (as we will see in the next section). I have chosen to define inflow and outflow in such a way that the outflow per electron

($\gamma_1 = \gamma_1 N/N$) is equal to the broadening (in addition to their difference being equal to the net current). Whether this broadening (due to the source) is γ_1 or $\gamma_1(1 - f_1)$ or something else is not an academic question. It can be shown that as long as energy relaxing or inelastic interactions are not involved in the inflow/outflow process, the broadening is γ_1, independent of the occupation factor f_1 in the contact. We will discuss this point a little further in Chapters 9 and 10, but a proper treatment requires advanced formalism as described in the appendix.

1.3 The quantum of conductance

Consider a device with a small voltage applied across it causing a splitting of the source and drain electrochemical potentials (Fig. 1.3.1a). We can write the current through this device from Eq. (1.2.5) and simplify it by assuming $\mu_1 > \varepsilon > \mu_2$ and the temperature is low enough that $f_1(\varepsilon) \equiv f_0(\varepsilon - \mu_1) \approx 1$ and $f_2(\varepsilon) \equiv f_0(\varepsilon - \mu_2) \approx 0$:

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2} = \frac{q\gamma_1}{2\hbar} \quad \text{if} \quad \gamma_2 = \gamma_1 \tag{1.3.1a}$$

This suggests that we could pump unlimited current through this one-level device by increasing γ_1 ($= \gamma_2$), that is by coupling it more and more strongly to the contacts. However, one of the seminal results of mesoscopic physics is that the maximum conductance of a one-level device is equal to G_0 (see Eq. (1.1)). What have we missed?

What we have missed is the broadening of the level that inevitably accompanies any process of coupling to it. This causes part of the energy level to spread outside the energy range between μ_1 and μ_2 where current flows. The actual current is then reduced below what we expect from Eq. (1.3.1a) by a factor $(\mu_1 - \mu_2)/C\gamma_1$ representing the fraction

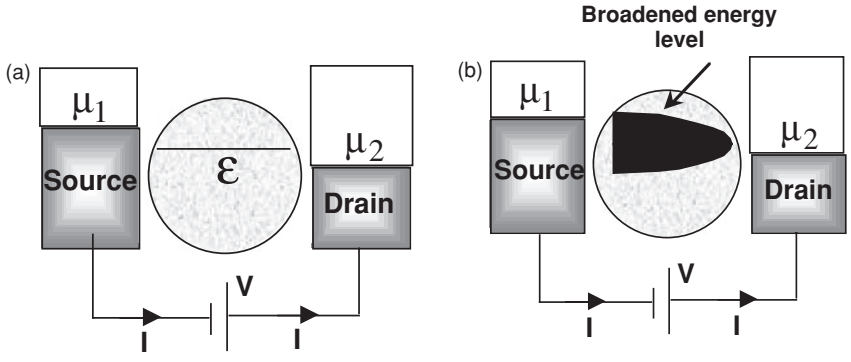


Fig. 1.3.1 (a) A channel with a small voltage applied across it causing a splitting of the source and drain electrochemical potentials $\mu_1 > \varepsilon > \mu_2$. (b) The process of coupling to the channel inevitably broadens the level, thereby spreading part of the energy level outside the energy range between μ_1 and μ_2 where current flows.

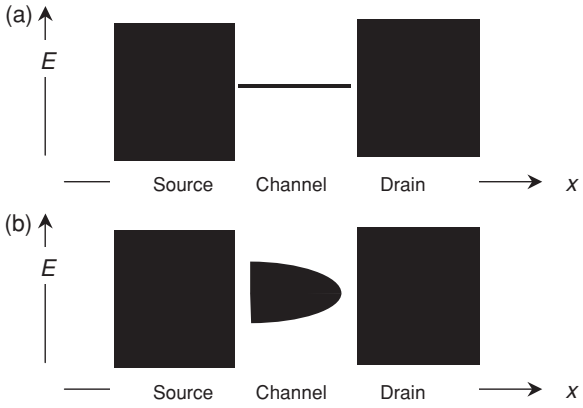


Fig. 1.3.2

of the level that lies in the window between μ_1 and μ_2, where $C\gamma_1$ is the effective width of the level, C being a numerical constant. Since $\mu_1 - \mu_2 = qV_D$, we see from Eq. (1.3.1b)

$$I = \frac{q\gamma_1}{2\hbar}\frac{qV_D}{C\gamma_1} \rightarrow G = \frac{I}{V_D} = \frac{q^2}{2C\hbar} \tag{1.3.1b}$$

that the conductance indeed approaches a constant value independent of the strength of the coupling ($\gamma_1 = \gamma_2$) to the contacts. We will now carry out this calculation a little more quantitatively so as to obtain a better estimate for C.

One way to understand this broadening is to note that, *before* we couple the channel to the source and the drain, the density of states (DOS) $D(E)$ looks something like Fig. 1.3.2a (dark indicates a high DOS). We have one sharp level in the channel and a continuous distribution of states in the source and drain contacts. On coupling, these states "spill over": the channel "loses" part of its state as it spreads into the contacts, but it also "gains" part of the contact states that spread into the channel. Since the loss occurs at a fixed energy while the gain is spread out over a range of energies, the overall effect is to broaden the channel DOS from its initial sharp structure (Fig. 1.3.2a) into a more diffuse structure (Fig. 1.3.2b). In Chapter 8 we will see that there is a "sum rule" that requires the loss to be exactly offset by the gain. Integrated over all energy, the level can still hold only one electron. The broadened DOS could in principle have any shape, but in the simplest situation it is described by a Lorentzian function centered around $E = \varepsilon$ (whose integral over all energy is equal to one):

$$D_\varepsilon(E) = \frac{\gamma/2\pi}{(E-\varepsilon)^2 + (\gamma/2)^2} \tag{1.3.2}$$

The initial delta function can be represented as the limiting case of $D_\varepsilon(E)$ as the broadening tends to zero: $\gamma \rightarrow 0$ (Fig. 1.3.3). The broadening γ is proportional to the strength of the coupling as we might expect. Indeed it turns out that $\gamma = \gamma_1 + \gamma_2$, where

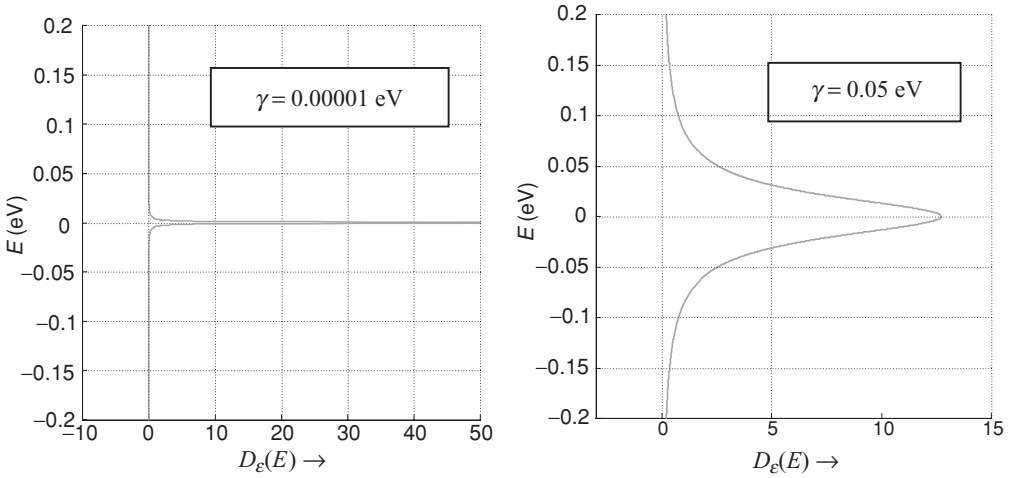


Fig. 1.3.3 An energy level at $E = \varepsilon$ is broadened into a continuous density of states $D_\varepsilon(E)$ by the process of coupling to the contacts. $D_\varepsilon(E)$ curves for two different values of coupling γ with $\varepsilon = 0\,\text{eV}$ are shown.

$\gamma_1/\hbar$ and $\gamma_2/\hbar$ are the escape rates introduced in Section 1.2. This will come out of our quantum mechanical treatment in Chapter 8, but at this stage we could rationalize it as a consequence of the "uncertainty principle" that requires the product of the lifetime ($= \hbar/\gamma$) of a state and its spread in energy (γ) to equal $\hbar$. Note that in general the lineshape need not be Lorentzian and this is usually reflected in an energy-dependent broadening $\gamma(E)$.

Anyway, the bottom line is that the coupling to the contacts broadens a single discrete energy level into a continuous density of states given by Eq. (1.3.2) and we can include this effect by modifying our expression for the current (Eq. (1.2.5))

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\,[f_1(\varepsilon) - f_2(\varepsilon)]$$

to integrate (that is, sum) over a distribution of states, $D_\varepsilon(E)\,\mathrm{d}E$:

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\,D_\varepsilon(E)\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\,[f_1(E) - f_2(E)] \tag{1.3.3}$$

At low temperatures, we can write

$$\begin{aligned} f_1(E) - f_2(E) &= 1 \quad \text{if} \quad \mu_1 > E > \mu_2 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

so that the current is given by

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\int_{\mu_2}^{\mu_1} \mathrm{d}E\,D_\varepsilon(E)$$

If the bias is small enough that we can assume the DOS to be constant over the range $\mu_1 > E > \mu_2$, we can use Eq. (1.3.2) to write

$$I = \frac{q}{\hbar}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}(\mu_1-\mu_2)\frac{(\gamma_1+\gamma_2)/2\pi}{(\mu-\varepsilon)^2+(\gamma_1+\gamma_2)^2}$$

The maximum current is obtained if the energy level ε coincides with μ, the average of μ_1 and μ_2. Noting that $\mu_1 - \mu_2 = qV_D$, we can write the maximum conductance as

$$G \equiv \frac{I}{V_D} = \frac{q^2}{h}\frac{4\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2} = \frac{q^2}{h} \quad \text{if} \quad \gamma_1 = \gamma_2$$

Equation (1.3.3) for the current extends our earlier result in Eq. (1.2.5) to include the effect of broadening. Similarly, we can extend the expression for the number of electrons N (see Eq. (1.2.4))

$$N = \frac{\gamma_1 f_1(\varepsilon) + \gamma_2 f_2(\varepsilon)}{\gamma_1+\gamma_2}$$

to account for the broadened DOS:

$$N = \int_{-\infty}^{+\infty} \mathrm{d}E\, D_\varepsilon(E)\frac{\gamma_1 f_1(E) + \gamma_2 f_2(E)}{\gamma_1+\gamma_2} \tag{1.3.4}$$

1.4 Potential profile

Physicists often focus on the low-bias conductance ("linear response"), which is determined solely by the properties of the energy levels around the equilibrium electrochemical potential μ. What is not widely appreciated is that this is not enough if we are interested in the full current–voltage characteristics. It is then important to pay attention to the actual potential inside the channel in response to the voltages applied to the external electrodes (source, drain, and gate). To see this, consider a one-level channel with an equilibrium electrochemical potential μ located slightly above the energy level ε as shown in Fig. 1.4.1. When we apply a voltage between the source and drain, the electrochemical potentials separate by qV: $\mu_1 - \mu_2 = qV$. We know that a current flows (at low temperatures) only if the level ε lies between μ_1 and μ_2. Depending on how the energy level ε is affected by the applied voltage, we have different possibilities.

If we ignore the gate we might expect the potential in the channel to lie halfway between the source and the drain, $\varepsilon \to \varepsilon - (V/2)$, leading to Fig. 1.4.2 for positive and negative voltages (note that we are assuming the source potential to be held constant, relative to which the other potentials are changing). It is apparent that the energy level

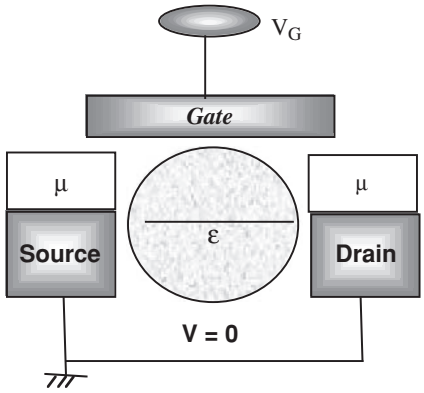


Fig. 1.4.1

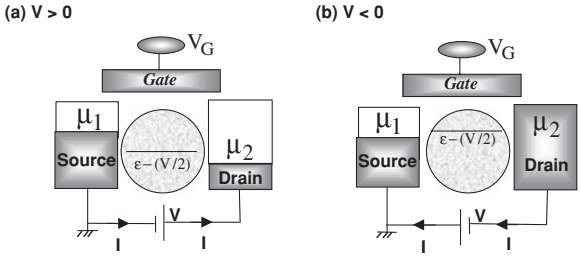


Fig. 1.4.2 Energy level diagram under (a) forward ($V > 0$) and (b) reverse ($V < 0$) bias, assuming that the channel potential lies halfway between the source and the drain.

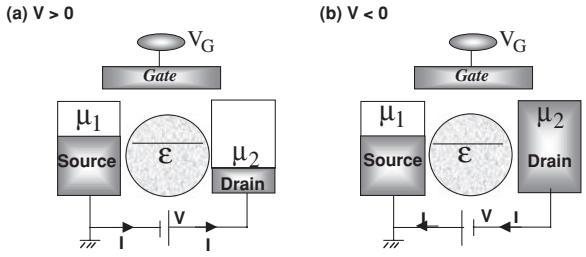


Fig. 1.4.3 Energy level diagram under (a) forward ($V > 0$) and (b) reverse ($V < 0$) bias assuming that the channel potential remains fixed with respect to the source.

lies halfway between μ_1 and μ_2 for either bias polarity ($V > 0$ or $V < 0$), leading to equal magnitudes for $+V$ and $-V$.

A different picture emerges if we assume that the gate is so closely coupled to the channel that the energy level follows the gate potential and is unaffected by the drain voltage or, in other words, ε remains fixed with respect to the source (Fig. 1.4.3). In this case the energy level lies between μ_1 and μ_2 for positive bias ($V > 0$) but

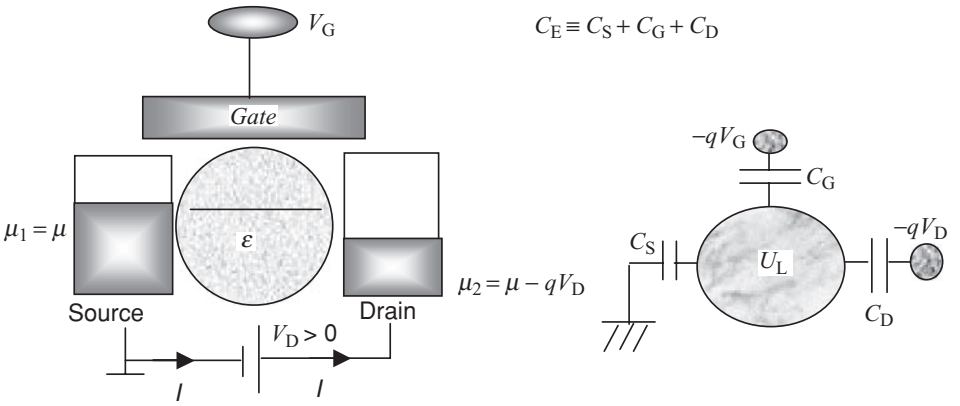


Fig. 1.4.4 A simple capacitive circuit model for the "Laplace" potential U_L of the active region in response to the external gate and drain voltages, V_G and V_D. The total capacitance is denoted C_E, where E stands for electrostatic. The actual potential U can be different from U_L if there is a significant density of electronic states in the energy range around μ_1 and μ_2.

not for negative bias ($V < 0$), leading to a current–voltage characteristic that can be very *asymmetric* in V. Clearly the shape of the current–voltage characteristic is affected strongly by the potential profile and even the simplest model needs to account for it.

So how do we calculate the potential inside the channel? If the channel were an insulator, we could solve Laplace's equation (ε_r is the relative permittivity, which could be spatially varying)

$$\vec{\nabla} \cdot (\varepsilon_r \vec{\nabla} V) = 0$$

subject to the boundary conditions that $V = 0$ (source electrode), $V = V_G$ (gate electrode), and $V = V_D$ (drain electrode). We could visualize the solution to this equation in terms of the capacitive circuit model shown in Fig. 1.4.4, if we treat the channel as a single point ignoring any spatial variation of the potential inside it.

The potential energy in the channel is obtained by multiplying the electrostatic potential V by the electronic charge $-q$:

$$U_L = \frac{C_G}{C_E}(-qV_G) + \frac{C_D}{C_E}(-qV_D) \tag{1.4.1a}$$

Here we have labeled the potential energy with a subscript L as a reminder that it is calculated from the Laplace equation ignoring any change in the electronic charge, which is justified if there are very few electronic states in the energy range around μ_1 and μ_2. Otherwise there is a change $\Delta\rho$ in the electron density in the channel and we need to solve the Poisson equation

$$\vec{\nabla} \cdot (\varepsilon_r \vec{\nabla} V) = -\Delta\rho/\varepsilon_0$$

for the potential. In terms of our capacitive circuit model, we could write the change in the charge as a sum of the charges on the three capacitors:

$$-q\Delta N = C_S V + C_G(V - V_G) + C_D(V - V_D)$$

so that the potential energy $U = -qV$ is given by the sum of the Laplace potential and an additional term proportional to the change in the number of electrons:

$$U = U_L + \frac{q^2}{C_E}\Delta N \tag{1.4.1b}$$

The constant $q^2/C_E \equiv U_0$ tells us the change in the potential energy due to *one* extra electron and is called the single-electron charging energy, whose significance we will discuss further in the next section. The *change* ΔN in the number of electrons is calculated with respect to the reference number of electrons, originally in the channel, N_0, corresponding to which the energy level is believed to be located at ε.

Iterative procedure for self-consistent solution: For a small device, the effect of the potential U is to raise the DOS in energy and can be included in our expressions for the number of electrons N (Eq. (1.3.4)) and the current I (Eq. (1.3.3)) in a straightforward manner:

$$N = \int_{-\infty}^{+\infty} dE\, D_\varepsilon(E - U)\frac{\gamma_1 f_1(E) + \gamma_2 f_2(E)}{\gamma_1 + \gamma_2} \tag{1.4.2}$$

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} dE\, D_\varepsilon(E - U)\frac{\gamma_1\gamma_2}{\gamma_1 + \gamma_2}[f_1(E) - f_2(E)] \tag{1.4.3}$$

Equation (1.4.2) has a U appearing on its right-hand side, which in turn is a function of N through the electrostatic relation (Eq. (1.4.1b)). This requires a simultaneous or "self-consistent" solution of the two equations which is usually carried out using the iterative procedure depicted in Fig. 1.4.5.

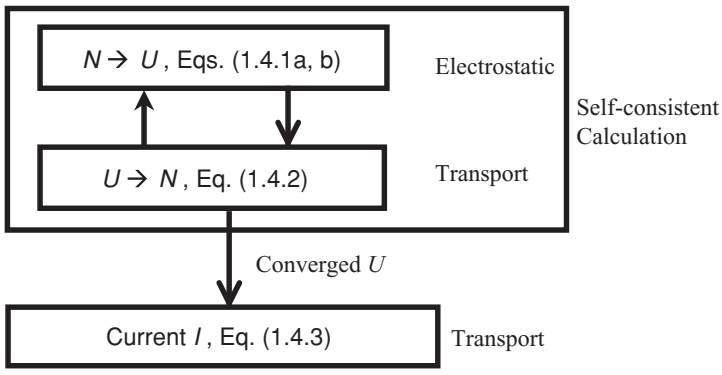


Fig. 1.4.5 Iterative procedure for calculating N and U self-consistently.

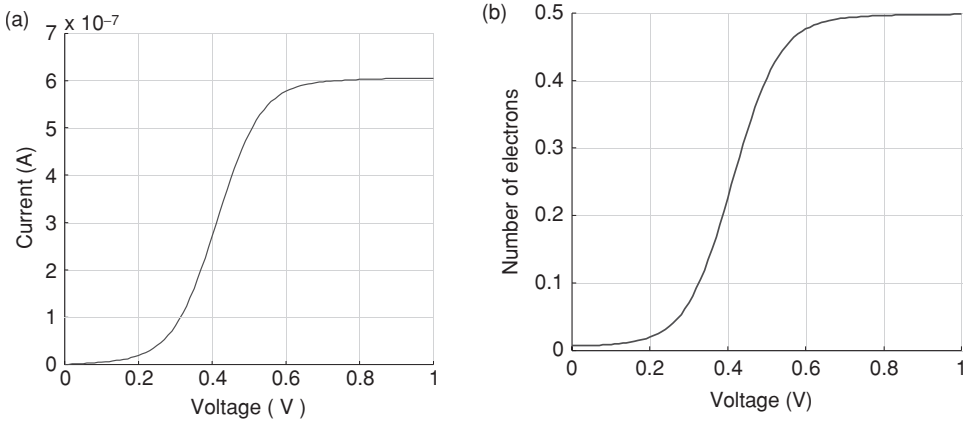


Fig. 1.4.6 (a) Current vs. voltage calculated using the SCF method (Fig. 1.4.5) with $\mu = 0$, $\varepsilon = 0.2\,\text{eV}$, $V_G = 0$, $k_BT = 0.025\,\text{eV}$, $U_0 = 0.025\,\text{eV}$, $C_D/C_E = 0.5$, and $\gamma_1 = \gamma_2 = 0.005\,\text{eV}$. (b) Number of electrons vs. voltage calculated using the SCF method (Fig. 1.4.5) with same parameters as in (a).

We start with an initial guess for U, calculate N from Eq. (1.4.2) with $D_\varepsilon(E)$ given by Eq. (1.3.2), calculate an appropriate U from Eq. (1.4.1b), with U_L given by Eq. (1.4.1a) and compare with our starting guess for U. If this new U is not sufficiently close to our original guess, we revise our guess using a suitable algorithm, say something like

$$U_{\text{new}} = U_{\text{old}} + \alpha(U_{\text{calc}} - U_{\text{old}}) \tag{1.4.4}$$

New guess Old guess Calculated

where α is a positive number (typically < 1) that is adjusted to be as large as possible without causing the solution to diverge (which is manifested as an increase in $U_{\text{calc}} - U_{\text{old}}$ from one iteration to the next). The iterative process has to be repeated till we find a U that yields an N that leads to a new U which is sufficiently close (say within a fraction of k_BT) to the original value. Once a converged U has been found, the current can be calculated from Eq. (1.4.3).

Figure 1.4.6 shows the current I and the number of electrons N calculated as a function of the applied drain voltage using the self-consistent field (SCF) method shown in Fig. 1.4.5.

1.5 Coulomb blockade

The charging model based on the Poisson equation represents a good zero-order approximation (sometimes called the Hartree approximation) to the problem of electron–electron interactions, but it is generally recognized that it tends to overestimate the

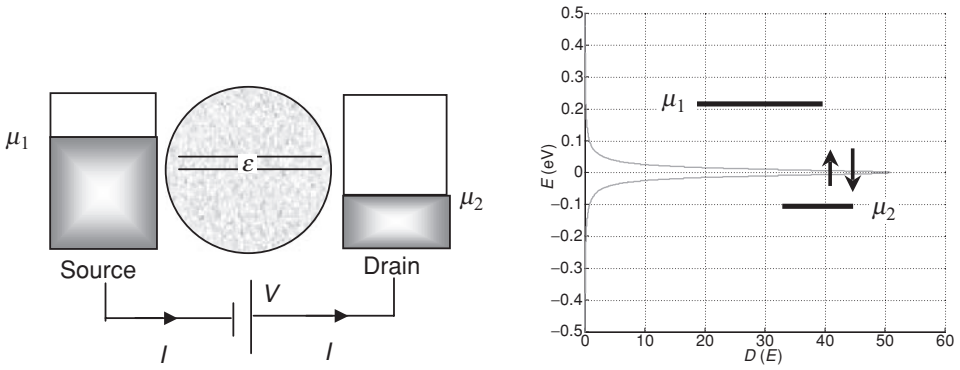


Fig. 1.5.1 A channel with two spin-degenerate levels containing one electron is expected to have an equilibrium electrochemical potential that lies in the center of its broadened density of states, so that current should flow easily under bias ($\gamma = 0.05\,\mathrm{eV}$).

effect and may need to be corrected (the so-called exchange and correlation effects). Discovering an appropriate function $U(N)$ (if there is one!) to replace our simple result (cf. Eq. (1.4.1b))

$$U(N) = q^2(N - N_0)/C_\mathrm{E}$$

is arguably one of the central topics in many-electron physics and can in some cases give rise to profound effects like magnetism, which are largely outside the scope of this book. However, there is one aspect that I would like to mention right away, since it can affect our picture of current flow even for a simple one-level device and put it in the so-called Coulomb blockade or single-electron charging regime. Let me explain what this means.

Energy levels come in pairs, one up-spin and one down-spin, which are degenerate, that is they have the same energy. Usually this simply means that all our results have to be multiplied by two. Even the smallest device has two levels rather than one, and its maximum conductance will be twice the conductance quantum $G_0 \equiv q^2/h$ discussed earlier. The expressions for the number of electrons and the current should all be multiplied by two. However, there is a less trivial consequence that I would like to explain.

Consider a channel with two spin-degenerate levels (Fig. 1.5.1), containing one electron when neutral ($N_0 = 1$). We expect the broadened DOS to be twice our previous result (see Eq. (1.3.2))

$$D_\varepsilon(E) = 2\ (\text{for spin}) \times \frac{\gamma/2\pi}{(E-\varepsilon)^2 + (\gamma/2)^2} \tag{1.5.1}$$

where the total broadening is the sum of those due to each of the two contacts individually: $\gamma = \gamma_1 + \gamma_2$, as before. Since the available states are only half filled for a neutral

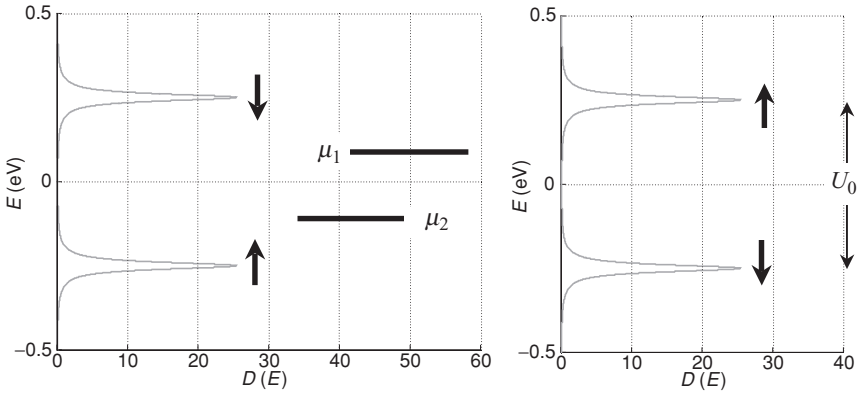


Fig. 1.5.2 Under certain conditions, the up-spin and down-spin density of states splits into two parts separated by the single-electron charging energy, U_0, instead of one single degenerate peak as shown in Fig. 1.5.1 ($\gamma = 0.05\,\text{eV}$, $U_0 = 0.25\,\text{eV}$).

channel, the electrochemical potential will lie exactly in the middle of the broadened DOS, so that we would expect a lot of current to flow when a bias is applied to split the electrochemical potentials in the source and drain as shown.

However, under certain conditions the DOS looks like one of the two possibilities shown in Fig. 1.5.2. The up-spin and the down-spin density of states splits into two parts separated by the single-electron charging energy

$$U_0 \equiv q^2/C_\text{E} \tag{1.5.2}$$

Very little current flows when we apply a small bias since there are hardly any states between μ_1 and μ_2 and this "Coulomb blockade" has been experimentally observed for systems where the charging energy U_0 exceeds the broadening γ.

It is hard to understand why the two peaks should separate based on the simple SCF picture. Two peaks with the same energy ("degenerate") should always remain degenerate as long as they feel the same self-consistent potential U. The point is that *no electron feels any potential due to itself*. Suppose the up-spin level gets filled first, causing the down-spin level to float up by U_0. But the up-spin level does not float up because it does not feel any self-interaction, leading to the picture shown on the left in Fig. 1.5.2. Of course, it is just as likely that the down-spin will fill up first leading to the picture on the right. In either case the DOS near μ is suppressed relative to the SCF picture (Fig. 1.5.1).

Describing the flow of current in this Coulomb blockade regime requires a very different point of view that we will not discuss in this book, except briefly in Section 3.4. But when do we have to worry about Coulomb blockade effects? Answer: only if U_0 exceeds both $k_\text{B}T$ and γ $(= \gamma_1 + \gamma_2)$. Otherwise, the SCF method will give results that look much like those obtained from the correct treatment (see Fig. 3.4.3).

So what determines U_0? Answer: the extent of the electronic wavefunction. If we smear out one electron over the surface of a sphere of radius R, then we know from freshman physics that the potential of the sphere will be $q/4\pi\varepsilon_r\varepsilon_0 R$, so that the energy needed to put another electron on the sphere will be $q^2/4\pi\varepsilon_r\varepsilon_0 R \cong U_0$, which is ~0.025 eV if $R = 5$ nm and $\varepsilon_r = 10$. Levels with well-delocalized wavefunctions (large R) have a very small U_0 and the SCF method provides an acceptable description even at the lowest temperatures of interest. But if R is small, then the charging energy U_0 can exceed k_BT and one could be in a regime dominated by single-electron charging effects that is not described well by the SCF method.

1.6 Towards Ohm's law

Now that we have discussed the basic physics of electrical conduction through small conductors, let us talk about the new factors that arise when we have large conductors. In describing electronic conduction through small conductors we can identify the following three regimes.

- *Self-consistent field (SCF) regime.* If k_BT and/or γ is comparable to U_0, we can use the SCF method described in Section 1.4.
- *Coulomb blockade (CB) regime.* If U_0 is well in excess of both k_BT and γ, the SCF method is not adequate. More correctly, one could use (if practicable) the multi-electron master equation that we will discuss in Section 3.4.
- *Intermediate regime*. If U_0 is comparable to the larger of k_BT and γ, there is no simple approach: the SCF method does not do justice to the charging, while the master equation does not do justice to the broadening.

It is generally recognized that the intermediate regime can lead to novel physics that requires advanced concepts, even for the small conductors that we have been discussing. For example, experimentalists have seen evidence for the Kondo effect, which is reflected as an extra peak in the density of states around $E = \mu$ in addition to the two peaks (separated by U_0) that are shown in Fig. 1.5.2.

With large conductors too we can envision three regimes of transport that evolve out of these three regimes. We could view a large conductor as an array of unit cells as shown in Fig. 1.6.1. The inter-unit coupling energy t has an effect somewhat (but not exactly) similar to the broadening γ that we have associated with the contacts. If $t \geq U_0$, the overall conduction will be in the SCF regime and can be treated using an extension of the SCF method from Section 1.4. If $t \ll U_0$, it will be in the CB regime and can in principle be treated using the multi-electron master equation (to be discussed in Section 3.4), under certain conditions (specifically if t is much less than the level broadening γ_s introduced by phase-breaking processes of the type to be discussed in Chapter 10). On the other hand, large conductors with $\gamma_s \ll t \leq U_0$ belong to an intermediate regime that presents major theoretical challenges, giving rise to intriguing

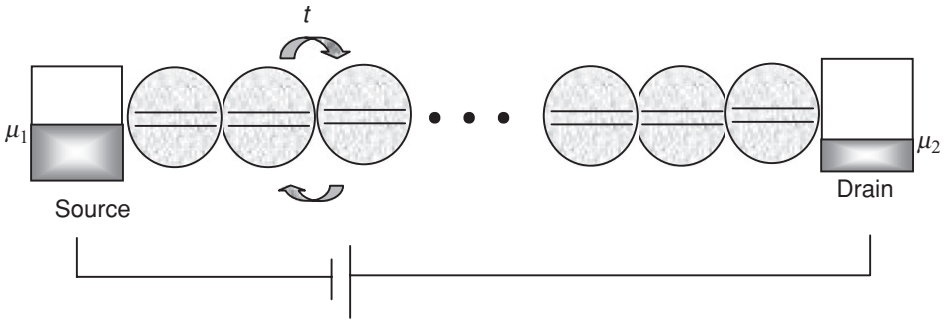


Fig. 1.6.1 A large conductor can be viewed as an array of unit cells. If the conductor is extended in the transverse plane, we should view each unit cell as representing an array of unit cells in the transverse direction.

possibilities. Indeed many believe that high-T_c superconductors (whose microscopic theory is still controversial) consist of unit cells whose coupling is delicately balanced on the borderline of the SCF and the CB regimes.

The more delocalized the electronic wavefunctions (large t), the more accurate the SCF description becomes and in this book I will focus on this regime. Basically I will try to explain how the simple one-level description from Section 1.4 is extended to larger conductors all the way to a nanotransistor, within the SCF picture that accounts for electron–electron interactions through an average potential $U(r)$ that one electron feels due to the other electrons.

Summary of results for one-level conductors: We have developed a model for current flow through a one-level device, starting with a simple discrete level (ε) in Section 1.2 and then extending it to include the broadening of the level into a Lorentzian density of states in Section 1.3

$$D_\varepsilon(E) = 2\,(\text{for spin}) \times \frac{\gamma/2\pi}{(E-\varepsilon)^2 + (\gamma/2)^2} \qquad \gamma \equiv \gamma_1 + \gamma_2 \tag{1.6.1}$$

and the self-consistent potential in Section 1.4

$$U = U_L + U_0\,(N - N_0) \tag{1.6.2}$$

$$U_L = \frac{C_G}{C_E}(-qV_G) + \frac{C_D}{C_E}(-qV_D)$$

$$U_0 = q^2/C_E \qquad C_E = C_G + C_S + C_D \tag{1.6.3}$$

The number of electrons N is given by (restricted SCF method)

$$N = \int_{-\infty}^{+\infty} \mathrm{d}E\, n(E)$$

where

$$n(E) = D(E-U)\left(\frac{\gamma_1}{\gamma} f_1(E) + \frac{\gamma_2}{\gamma} f_2(E)\right) \tag{1.6.4}$$

while the currents at the two terminals are given by

$$I_1 = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, \gamma_1[D(E-U)f_1(E) - n(E)] \tag{1.6.5a}$$

$$I_2 = \frac{q}{\hbar}\int_{-\infty}^{+\infty} \mathrm{d}E\, \gamma_2[D(E-U)f_2(E) - n(E)] \tag{1.6.5b}$$

At steady state, the sum of the two currents is equated to zero to eliminate $n(E)$:

$$I = \frac{q}{h}\int_{-\infty}^{+\infty} \mathrm{d}E\, \overline{T}(E)[f_1(E) - f_2(E)]$$

where

$$\overline{T}(E) = D(E-U)2\pi\gamma_1\gamma_2/\gamma \tag{1.6.6}$$

is called the *transmission* – a concept that plays a central role in the transmission formalism widely used in mesoscopic physics (see Section 9.4). Note that the Fermi functions f_1 and f_2 are given by

$$\begin{aligned} f_1(E) &= f_0(E-\mu_1) \\ f_2(E) &= f_0(E-\mu_2) \end{aligned} \tag{1.6.7}$$

where $f_0(E) \equiv [1+\exp(E/k_\mathrm{B}T)]^{-1}$ and the electrochemical potentials in the source and drain contacts are given by

$$\begin{aligned} \mu_1 &= \mu \\ \mu_2 &= \mu - qV_\mathrm{D} \end{aligned} \tag{1.6.8}$$

μ being the equilibrium electrochemical potential.

Note that in Eqs. (1.6.4) through (1.6.6) I have used $D(E)$ instead of $D_\varepsilon(E)$ to denote the DOS. Let me explain why.

Large conductors – a heuristic approach: $D_\varepsilon(E)$ (see Eq. (1.6.1)) is intended to denote the DOS obtained by broadening a single discrete level ε, while $D(E)$ denotes the DOS in general for a multi-level conductor with many energy levels (Fig. 1.6.2).

If we make the rather cavalier assumption that all levels conduct independently, then we could use exactly the same equations as for the one-level device, replacing the

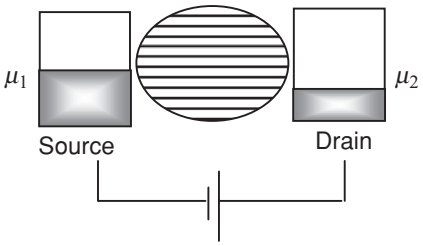


Fig. 1.6.2

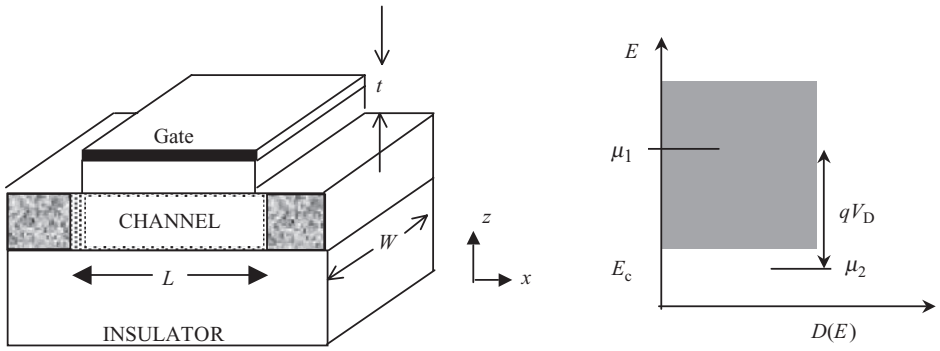


Fig. 1.6.3

one-level DOS $D_\varepsilon(E)$ in Eq. (1.6.1) with the total DOS $D(E)$. With this in mind, I will refer to Eqs. (1.6.4)–(1.6.6) as the *independent level model* for the current through a channel.

Nanotransistor – a simple model: As an example of this independent level model, let us model the nanotransistor shown in Fig. 1.1 by writing the DOS as (see Fig. 1.6.3, W is the width in the y-direction)

$$D(E) = m_c W L/\pi\hbar^2 \vartheta(E - E_c) \tag{1.6.9}$$

making use of a result that we will discuss in Chapter 6, namely that the DOS per unit area in a large two-dimensional (2D) conductor described by an effective mass m_c is equal to $m_c/\pi\hbar^2$, for energies greater than the energy E_c of the conduction band edge. The escape rates can be written down assuming that electrons are removed by the contact with a velocity v_R (somewhat like a "surface recombination velocity"):

$$\gamma_1 = \gamma_2 = \hbar v_R/L \tag{1.6.10}$$

The current–voltage relations shown in Fig. 1.1.1 were obtained using these model parameters: $m_c = 0.25m$, $C_G = 2\varepsilon_r\varepsilon_0 WL/t$, $C_S = C_D = 0.05C_G$, $W = 1\,\mu$m,

$L = 10\,\text{nm}$, insulator thickness $t = 1.5\,\text{nm}$, $v_R = 10^7$ cm/s. At high drain voltages V_D the current saturates when μ_2 drops below E_c since there are no additional states to contribute to the current. Note that the gate capacitance C_G is much larger than the other capacitances, which helps to hold the channel potential fixed relative to the source as the drain voltage is increased (see Eq. (1.6.3)). Otherwise, the bottom of the channel density of states, E_c will "slip down" with respect to μ_1 when the drain voltage is applied, so that the current will not saturate. The essential feature of a well-designed transistor is that the gate is much closer to the channel than L, allowing it to hold the channel potential constant despite the voltage V_D on the drain.

I should mention that our present model ignores the profile of the potential along the length of the channel, treating it as a little box with a single potential U given by Eq. (1.6.2). Nonetheless the results (Fig. 1.1.1) are surprisingly close to experiments/realistic models, because the current in well-designed nanotransistors is controlled by a small region in the channel near the source whose length can be a small fraction of the actual length L. Luckily we do not need to pin down the precise value of this fraction, since the present model gives the same current independent of L.

Ohm's law? Would this independent level model lead to Ohm's law if we were to calculate the low-bias conductance of a large conductor of length L and cross-sectional area S? Since the current is proportional to the DOS, $D(E)$ (see Eq. (1.6.5)), which is proportional to the volume SL of the conductor, it might seem that the conductance $G \sim SL$. However, the coupling to the contacts decreases inversely with the length L of the conductor, since the longer a conductor is, the smaller is its coupling to the contact. While the DOS goes up with the volume, the coupling to the contact goes down as $1/L$, so that the conductance

$$G \sim SL/L = S$$

However, Ohm's law tells us that the conductance should scale as S/L; we are predicting that it should scale as S. The reason is that we are really modeling a *ballistic* conductor, where electrons propagate freely, the only resistance arising from the contacts. The conductance of such a conductor is indeed independent of its length. The ohmic length dependence of the conductance comes from scattering processes within the conductor that are not yet included in our thinking.

For example, in a uniform channel the electronic wavefunction is spread out uniformly. But a scatterer in the middle of the channel could split up the wavefunctions into two, one on the left and one on the right with different energies. One has a small γ_2 while the other has a small γ_1, and so neither conducts very well. This *localization* of wavefunctions would seem to explain why the presence of a scatterer contributes to the resistance, but to get the story quantitatively correct it is in general

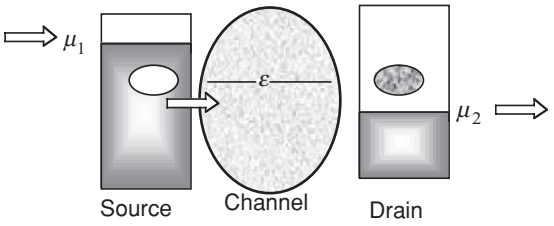


Fig. 1.6.4 When an electron goes from the source to the drain, an empty state or hole is left behind in the source, while an electron appears in the drain. Energy dissipating processes quickly take the electron down to μ_2 inside the drain and the hole up to μ_1 in the source. In our model we do not explicitly discuss these processes; we simply legislate that the contacts are maintained at equilibrium with the assumed electrochemical potentials.

necessary to go beyond the independent level model to account for interference between multiple paths. This requires a model that treats γ as a *matrix* rather than as simple numbers.

Such "coherent" scatterers, however, do not really lead to a resistance $R \sim 1/L$ (Ohm's law). The full story requires us to include phase-breaking scattering processes that cause a change in the state of an external object. For example, if an electron gets deflected by a rigid (that is unchangeable) defect in the lattice, the scattering is said to be coherent. But if the electron transfers some energy to the atomic lattice causing it to start vibrating, that would constitute a phase-breaking or incoherent process.

Such incoherent scatterers are also needed to remove energy from the electrons and cause dissipation. For example, in this chapter we have developed a simple model that allows us to calculate the resistance R, but none of the associated Joule heat I^2R is dissipated in the channel; it is all dissipated in the contacts. This is evident if we consider what happens when an electron goes from the source to the drain (Fig. 1.6.4). An empty state or hole is left behind in the source at an energy lower than μ_1 while an electron appears in the drain at an energy higher than μ_2. Energy dissipating processes quickly take the electron down to μ_2 inside the drain and the hole up to μ_1 in the source. The overall effect is to take an electron from μ_1 in the source to μ_2 in the drain, and in our model the energy $(\mu_1 - \mu_2)$ is dissipated partly in the source and partly in the drain, but none in the channel. In the real world too there is experimental evidence that in nanoscale conductors, most of the heating occurs in the contacts outside the channel, allowing experimentalists to pump a lot more current through a small conductor without burning it up. But long conductors have significant incoherent scattering inside the channel and it is important to include it in our model.

The point is that the transition from ballistic conductors to Ohm's law has many subtleties that require a much deeper model for the flow of current than the independent level model (Eqs. (1.6.4)–(1.6.6)), although the latter can often provide an adequate

description of short conductors. Let me now try to outline briefly the nature of this "deeper model" that we will develop in this book and illustrate with examples in Chapter 11.

Multi-level conductors – from numbers to matrices: The independent level model that we have developed in this chapter serves to identify the important concepts underlying the flow of current through a conductor, namely the location of the equilibrium *electrochemical potential* μ relative to the *density of states* $D(E)$, the *broadening* of the level $\gamma_{1,2}$ due to the coupling to contacts 1 and 2, the *self-consistent potential* U describing the effect of the external electrodes, and the change in the *number* of electrons N. In the general model for a multi-level conductor with n energy levels, each of these quantities is replaced by a corresponding matrix of size $(n \times n)$:

$\varepsilon \rightarrow [H]$	*Hamiltonian matrix*
$\gamma_{1,2} \rightarrow [\Gamma_{1,2}(E)]$	*Broadening matrix*
$2\pi D(E) \rightarrow [A(E)]$	*Spectral function*
$2\pi n(E) \rightarrow [G^{\mathrm{n}}(E)]$	*Correlation function*
$U \rightarrow [U]$	*Self-consistent potential matrix*
$N \rightarrow [\rho] = \int (\mathrm{d}E/2\pi)[G^{\mathrm{n}}(E)]$	*Density matrix*

Actually, the effect of the contacts is described by a "*self-energy*" *matrix,* $[\Sigma_{1,2}(E)]$, whose anti-Hermitian part is the broadening matrix: $\Gamma_{1,2} = \mathrm{i}[\Sigma_{1,2} - \Sigma_{1,2}^{+}]$. All quantities of interest can be calculated from these matrices. For example, in Section 1.2 we discussed the inflow/outflow of electrons from a one-level device. Figure 1.6.5 illustrates how these concepts are generalized in terms of these matrices. I should mention that in order to emphasize its similarity to the familiar concept of electron density, I have used $G^{\mathrm{n}}(E)$ to denote what is usually written in the literature as $-\mathrm{i}G^{<}(E)$ following the non-equilibrium Green's function (NEGF) formalism pioneered by the works of Martin and Schwinger (1959), Kadanoff and Baym (1962) and Keldysh (1965).

Note that in the matrix model (Fig. 1.6.5b), I have added a third "contact" labeled "s-contact" representing scattering processes, without which we cannot make the transition to Ohm's law. Indeed it is only with the advent of mesoscopic physics in the 1980s that the importance of the contacts (Γ_1 andΓ_2) in interpreting experiments became widely recognized. Prior to that, it was common to ignore the contacts as minor experimental distractions and try to understand the physics of conduction in terms of the s-contact, though no one (to my knowledge) thought of scattering as a "contact" till Büttiker introduced the idea phenomenologically in the mid 1980s (see Büttiker, 1988; Datta, 1995). Subsequently, Datta (1989) showed from a microscopic model that

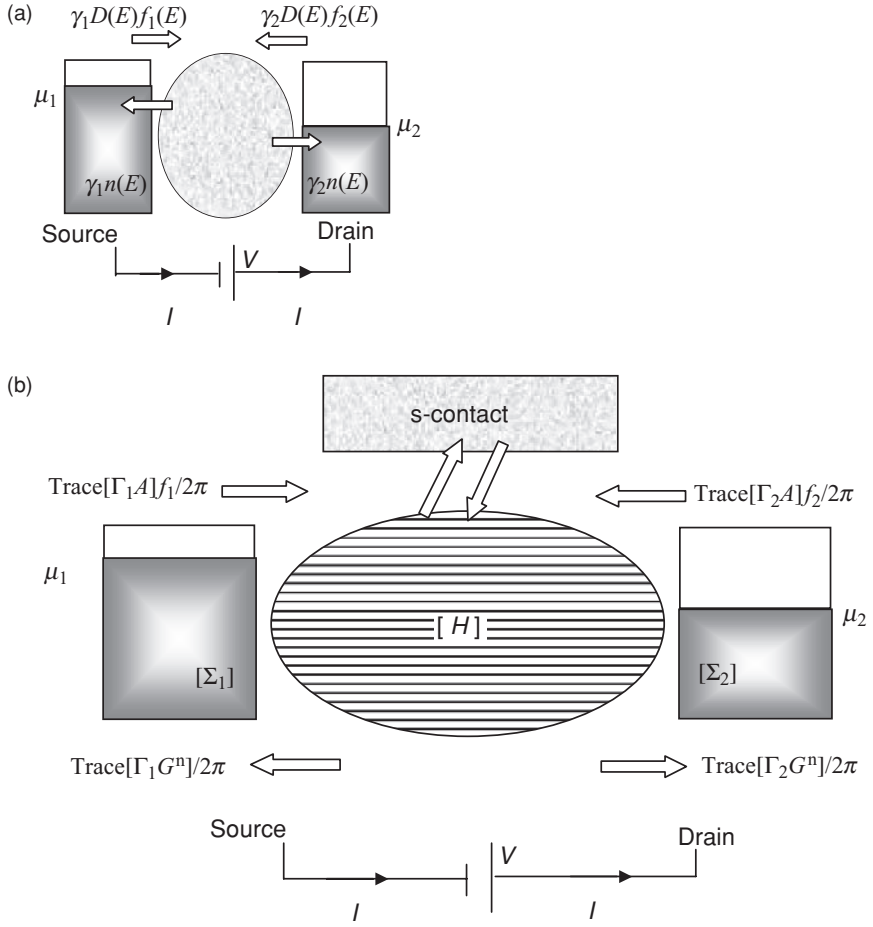


Fig. 1.6.5 From numbers to matrices: flux of electrons into and out of a device at the source and drain ends. (a) Simple result for independent level model, see Eqs. (1.6.4)–(1.6.6). (b) General matrix model, to be developed in this book. Without the "s-contact" this model is equivalent to Eq. (6) of Meir and Wingreen (1992). The "s-contact" distributed throughout the channel describes incoherent scattering processes (Datta, 1989). In general this "contact" cannot be described by a Fermi function, unlike the real contacts.

incoherent scattering processes act like a fictitious "contact" distributed throughout the channel that extracts and reinjects electrons. Like the real contacts, coupling to this "contact" too can be described by a broadening matrix Γ_s. However, unlike the real contacts, the scattering contact in general cannot be described by a Fermi function so that although the outflow is given by $\mathrm{Trace}[\Gamma_s G^n/2\pi]$, the inflow requires separate considerations that we will discuss in Chapter 10. The complete set of equations is summarized in Chapter 11.

The reader might wonder why the numbers become matrices, rather than just column vectors. For example, with one unit cell, we have an energy level ε. It seems reasonable

that with many unit cells, we should talk about an energy level $\varepsilon(n)$ in each cell "n". But why do we need a matrix $H(m, n)$? This is a question that goes to the heart of quantum mechanics whereby all physical quantities are represented by matrices. We can find a representation that diagonalizes $[H]$ and in this representation we could write the energy eigenvalues as a column vector $\varepsilon(n)$. If all the other matrices were also approximately diagonal in this representation, then we could indeed work with column vectors with n elements rather than matrices with n^2 elements and that is what "semi-classical" methods commonly do. In general, no single representation will diagonalize all the matrices and a full quantum treatment is needed.

Figure 1.6.5b without the s-contact is often used to analyze small devices and in this form it is identical to the result obtained by Meir and Wingreen (1992, their Eq. (6)) following the method of Caroli *et al*. (1972) based on the NEGF formalism. In order to make this approach accessible to readers unfamiliar with advanced many-body physics, I will derive these results using elementary arguments. What we have derived in this chapter (Fig. 1.6.5a) could be viewed as a special case of this general formalism with all the matrices being (1×1) in size. Indeed if there is a representation that diagonalizes all the matrices, then the matrix model without the s-contact would follow quite simply from Fig. 1.6.5a. We could write down separate equations for the current through each diagonal element (or level) for this special representation, add them up and write the sum as a trace. The resulting equations would then be valid in any representation, since the trace is invariant under a change in basis. In general, however, the matrix model cannot be derived quite so simply since no single representation will diagonalize all the matrices. In Chapters 8–10, I have derived the full matrix model (Fig. 1.6.5b) using elementary quantum mechanics. In the appendix, I have provided a brief derivation of the same results using the language of second quantization, but here too I have tried to keep the discussion less "advanced" than the standard treatments available in the literature.

I should mention that the picture in Fig. 1.6.5 is not enough to calculate the current: additional equations are needed to determine the "density of states" $[A(E)]$ and the "electron density" $[G^{\mathrm{n}}(E)]$. In our elementary model (Fig. 1.6.5a) we wrote down the density of states by "ansatz" (see Eq. (1.6.1)), but no separate equation was needed for the electron density which was evaluated by equating the currents (see derivation of Eq. (1.2.3) for a discrete level that was extended to obtain Eq. (1.6.4) for a broadened level). In the matrix model (Fig. 1.6.5b) too (without the s-contact), it was argued in Meir and Wingreen (1992) that $[G^{\mathrm{n}}(\mathrm{E})]$ can be similarly eliminated if $[\Gamma_1]$ is equal to a constant times $[\Gamma_2]$. However, this can be true only for very short channels. Otherwise, the source end is distinct from the drain end, making $[\Gamma_1]$ a very different matrix from $[\Gamma_2]$ since they couple to different ends. We then need additional equations to determine both $[A(E)]$ and $[G^{\mathrm{n}}(E)]$.

There is an enormous amount of physics behind all these matrices (both the diagonal and the off-diagonal elements) and we will introduce and discuss them in course of this

book: the next five chapters are about $[H]$, Chapter 7 is about $[\rho]$, Chapter 8 is about $[\Sigma]$, Chapter 9 combines these concepts to obtain the inflow/outflow diagram shown in Fig. 1.6.5b, and Chapter 10 introduces the matrix Γ_s describing scattering to complete the model for dissipative quantum transport. Finally, in Chapter 11, we illustrate the full "machinery" using a series of examples chosen to depict the transition from ballistic transport to Ohm's law, or in other words, from the atom to the transistor.

After that rather long introduction, we are now ready to get on with the "details." We will start with the question of how we can write down the Hamiltonian $[H]$ for a given device, whose eigenvalues will tell us the energy levels. We will work our way from the hydrogen atom in Chapter 2 "up" to solids in Chapter 5 and then "down" to nanostructures in Chapter 6. Let us now start where quantum mechanics started, namely, with the hydrogen atom.

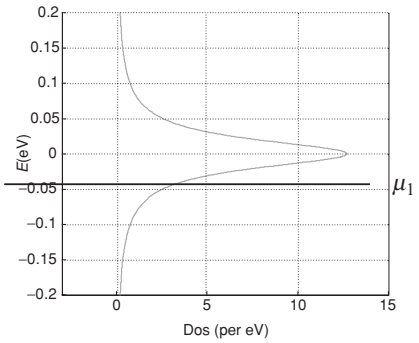
EXERCISES

E.1.1. Consider a channel with one spin-degenerate level assuming the following parameters: $\mu = 0, \varepsilon = 0.2\,\text{eV}, k_B T = 0.025\,\text{eV}, \gamma_1 = \gamma_2 = 0.005\,\text{eV}$. Calculate the current vs. drain voltage V_D assuming $V_G = 0$ with $U_L = -qV_D/2$ and $U_0 = 0.1\,\text{eV}$, 0.25 eV, using the SCF approach and compare with Fig. 1.4.6.

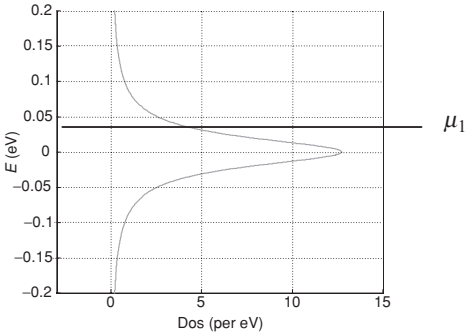
E.1.2. Calculate the current vs. gate and drain voltages for a nanotransistor as shown in Fig. 1.1.1 using the SCF equations summarized in Eqs. (1.6.2)–(1.6.7) with $D(E) = m_c WL/\pi\hbar^2$ and $\gamma_1 = \gamma_2 = \hbar v_R/L$ and the following parameters: $m_c = 0.25m$, $C_G = 2\varepsilon_r\varepsilon_0 WL/t$, $C_S = C_D = 0.05C_G$, $W = 1\,\mu m$, $L = 10\,\text{nm}$, insulator thickness, $t = 1.5\,\text{nm}$, $v_R = 10^7\,\text{cm/s}$.

E.1.3. Thermoelectric effect: In this chapter we have discussed the current that flows when a voltage is applied between the two contacts. In this case the current depends on the DOS near the Fermi energy and it does not matter whether the equilibrium Fermi energy μ_1 lies on (a) the lower end or (b) the upper end of the DOS:

(a) "n-type"



(b) "p-type"



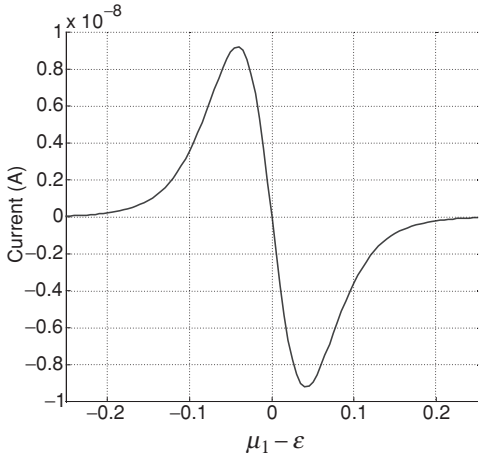


Fig. E.1.3 You should get a plot like this showing the reversal in the direction of the current from p-type ($\mu_1 < \varepsilon$) to n-type ($\mu_1 > \varepsilon$) samples.

However, if we simply heat up one contact relative to the other so that $T_1 > T_2$ (with no applied voltage) a thermoelectric current will flow whose direction will be different in case (a) and in case (b).

To see this, calculate the current from Eq. (1.6.6) with $U = 0$ (there is no need to perform a self-consistent solution), $V_D = 0$ and $V_G = 0$, and with $k_B T_1 = 0.026\,\text{eV}$ and $k_B T_2 = 0.025\,\text{eV}$:

$$f_1(E) \equiv \left[1 + \exp\left(\frac{E - \mu_1}{k_B T_1}\right)\right]^{-1} \quad \text{and} \quad f_2(E) \equiv \left[1 + \exp\left(\frac{E - \mu_1}{k_B T_2}\right)\right]^{-1}$$

and plot it as a function of $(\mu_1 - \varepsilon)$ as the latter changes from $-0.25\,\text{eV}$ to $+0.25\,\text{eV}$ assuming $\gamma_1 = \gamma_2 = 0.05\,\text{eV}$ (Fig. E.1.3). This problem is motivated by Paulsson and Datta (2003).

E.1.4. Negative differential resistance: Figure 1.4.6a shows the current–voltage (I–V_D) characteristics calculated from a self-consistent solution of Eqs. (1.6.2)–(1.6.5) assuming

$$\varepsilon = 0.2\,\text{eV}, k_B T = 0.025\,\text{eV}, U_0 = 0.025\,\text{eV}, V_G = 0,$$
$$\mu_1 = 0, \mu_2 = \mu_1 - q V_D, U_L = -q V_D/2$$

The broadening due to the two contacts γ_1 and γ_2 is assumed to be constant, equal to 0.005 eV.

Now suppose γ_1 is equal to 0.005 eV for $E > 0$, but is *zero* for $E < 0$ (γ_2 is still independent of energy and equal to 0.005 eV). Show that the current–voltage characteristics

will now show negative differential resistance (NDR), that is, a drop in the current with an increase in the voltage, in one direction of applied voltage but not the other as shown in Fig. E.1.4.

This problem is motivated by Rakshit *et al*. (2004).

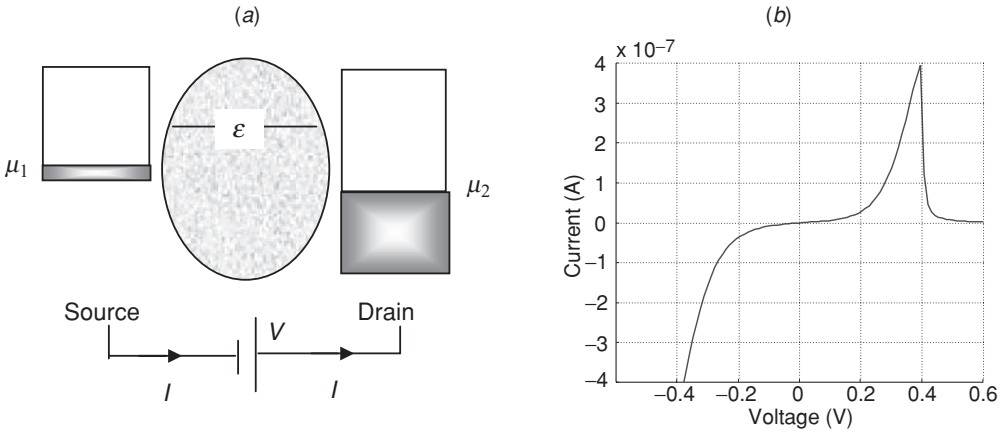


Fig. E.1.4

2 Schrödinger equation

Our objective for the next few chapters is to learn how the Hamiltonian matrix [H] for a given device structure (see Fig. 1.6.5) is written down. We start in this chapter with (1) the hydrogen atom (Section 2.1) and how it led scientists to the Schrödinger equation, (2) a simple approach called the finite difference method (Section 2.2) that can be used to convert this differential equation into a matrix equation, and (3) a few numerical examples (Section 2.3) showing how energy levels are calculated by diagonalizing the resulting Hamiltonian matrix.

2.1 Hydrogen atom

Early in the twentieth century scientists were trying to build a model for atoms which were known to consist of negative particles called electrons surrounding a positive nucleus. A simple model pictures the electron (of mass m and charge $-q$) as orbiting the nucleus (with charge Zq) at a radius r (Fig. 2.1.1) kept in place by electrostatic attraction, in much the same way that gravitational attraction keeps the planets in orbit around the Sun.

$$\frac{Zq^2}{4\pi\varepsilon_0 r^2} = \frac{mv^2}{r} \Rightarrow v = \sqrt{\frac{Zq^2}{4\pi\varepsilon_0 mr}} \tag{2.1.1}$$

Electrostatic force = Centripetal force

A faster electron describes an orbit with a smaller radius. The total energy of the electron is related to the radius of its orbit by the relation

$$E = -\frac{Zq^2}{4\pi\varepsilon_0 r} + \frac{mv^2}{2} = -\frac{Zq^2}{8\pi\varepsilon_0 r} \tag{2.1.2}$$

Potential energy + Kinetic energy = Total energy

However, it was soon realized that this simple viewpoint was inadequate since, according to classical electrodynamics, an orbiting electron should radiate electromagnetic waves like an antenna, lose energy continuously and spiral into the nucleus. Classically it is impossible to come up with a stable structure for such a system except with the

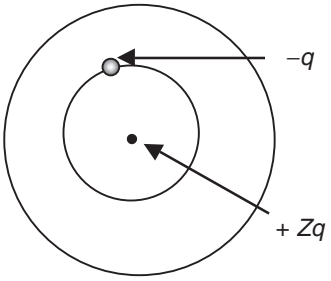


Fig. 2.1.1 Stationary orbits of an electron around a nucleus can be obtained by requiring their circumferences to be integer multiples of the de Broglie wavelength.

electron sitting right on top of the nucleus, in contradiction with experiment. It was apparent that a radical departure from classical physics was called for.

Bohr postulated that electrons could be described by stable orbits around the nucleus at specific distances from the nucleus corresponding to specific values of angular momenta. It was later realized that these distances could be determined by endowing the electrons with a wavelike character having a de Broglie wavelength equal to (h/mv), h being the Planck constant. One could then argue that the circumference of an orbit had to be an integer multiple of wavelengths in order to be stable:

$$2\pi r = n(h/mv) \tag{2.1.3}$$

Combining Eq. (2.1.3) with Eqs. (2.1.1) and (2.1.2) we obtain the radius and energy of stable orbits respectively:

$$r_n = (n^2/Z)a_0 \quad \text{(Bohr radius)} \tag{2.1.4}$$

$$\text{where } a_0 = 4\pi\varepsilon_0\hbar^2/mq^2 = 0.0529\,\text{nm} \tag{2.1.5}$$

$$E_n = -(Z^2/n^2)E_0 \tag{2.1.6a}$$

$$\text{where } E_0 = q^2/8\pi\varepsilon_0 a_0 = 13.6\,\text{eV} \quad \text{(1 Rydberg)} \tag{2.1.6b}$$

Once the electron is in its lowest energy orbit ($n = 1$) it cannot lose any more energy because there are no stationary orbits having lower energies available (Fig. 2.1.2a). If we heat up the atom, the electron is excited to higher stationary orbits (Fig. 2.1.2b). When it subsequently jumps down to lower energy states, it emits photons whose energy $h\nu$ corresponds to the energy difference between orbits m and n:

$$h\nu = E_m - E_n = E_0 Z^2 \left(\frac{1}{n^2} - \frac{1}{m^2}\right) \tag{2.1.7}$$

Experimentally it had been observed that the light emitted by a hydrogen atom indeed consisted of discrete frequencies that were described by this relation with integer values of n and m. This striking agreement with experiment suggested that there was some truth to this simple picture, generally known as the Bohr model.

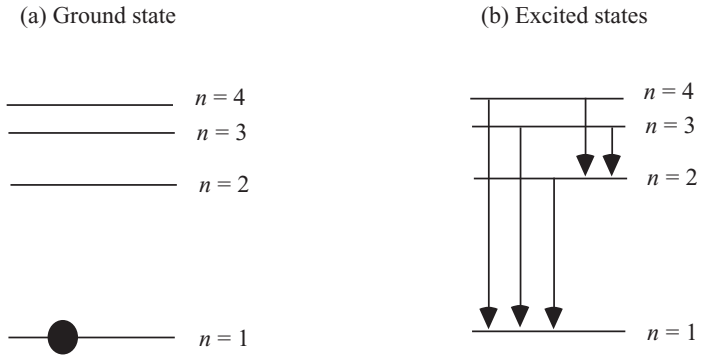


Fig. 2.1.2 (a) Left to itself, the electron relaxes to its lowest energy orbit ($n = 1$). (b) If we heat up the atom, the electron is excited to higher stationary orbits. When it subsequently jumps down to lower energy states, it emits photons whose energy $h\nu$ corresponds to the energy difference between the initial and final orbits.

The *Schrödinger equation* put this heuristic insight on a formal quantitative basis allowing one to calculate the energy levels for any confining potential $U(\vec{r})$.

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = \left(-\frac{\hbar^2}{2m}\nabla^2 + U(\vec{r})\right)\Psi \tag{2.1.8}$$

How does this equation lead to discrete energy levels? Mathematically, one can show that if we assume a potential $U(\vec{r}) = -Zq^2/4\pi\varepsilon_0 r$ appropriate for a nucleus of charge $+Zq$, then the solutions to this equation can be labeled with three indices n, l and m

$$\Psi(\vec{r}, t) = \phi_{nlm}(\vec{r})\exp(-\mathrm{i}E_n t/\hbar) \tag{2.1.9}$$

where the energy E_n depends only on the index n and is given by $E_n = -(Z^2/n^2)E_0$ in agreement with the heuristic result obtained earlier (see Eq. (2.1.6a)). The Schrödinger equation provides a formal wave equation for the electron not unlike the equation that describes, for example, an acoustic wave in a sound box . The energy E of the electron plays a role similar to that played by the frequency of the acoustic wave. It is well-known that a sound box resonates at specific frequencies determined by the size and shape of the box. Similarly an electron wave in an atomic box "resonates" at specific energies determined by the size and shape of the box as defined by the potential energy $U(\vec{r})$. Let us elaborate on this point a little further.

Waves in a box: To keep things simple let us consider the vibrations $u(x, t)$ of a one-dimensional (1D) string described by the 1D wave equation:

$$\frac{\partial^2 u}{\partial t^2} = v^2\frac{\partial^2 u}{\partial x^2} \tag{2.1.10}$$

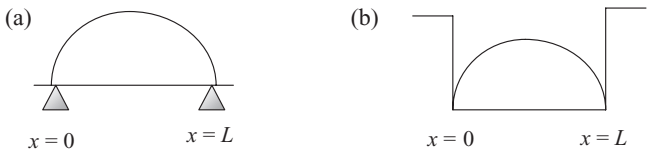


Fig. 2.1.3 Standing waves. (a) Acoustic waves in a "guitar" string with the displacement clamped to zero at either end. (b) Electron waves in a one-dimensional box with the wavefunction clamped to zero at both ends by an infinite potential.

The solutions to this equation can be written in the form of plane waves with a linear dispersion $\omega = \pm vk$:

$$u = u_0 \exp(\mathrm{i}kx)\exp(-\mathrm{i}\omega t) \Rightarrow \omega^2 = v^2k^2 \tag{2.1.11}$$

What happens if we clamp the two ends so that the displacement there is forced to be zero (Fig. 2.1.3)? We have to superpose solutions with $+k$ and $-k$ to obtain standing wave solutions. The allowed values of k are quantized leading to discrete resonant frequencies:

$$u = u_0 \sin(kx)\exp(-\mathrm{i}\omega t) \Rightarrow k = n\pi/L \Rightarrow \omega = n\pi v/L \tag{2.1.12}$$

Well, it's the same way with the Schrödinger equation. If there is no confining potential ($U = 0$), we can write the solutions to the 1D Schrödinger equation:

$$\mathrm{i}\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi}{\partial x^2} \tag{2.1.13}$$

in the form of plane waves with a parabolic dispersion law $E = \hbar^2k^2/2m$:

$$\Psi = \Psi_0 \exp(\mathrm{i}\,kx)\exp(-\mathrm{i}Et/\hbar) \Rightarrow E = \hbar^2k^2/2m \tag{2.1.14}$$

If we fix the two ends we get standing waves with quantized k and resonant frequency:

$$\begin{aligned}\Psi = \Psi_0 \sin(kx)\exp(-\mathrm{i}Et/\hbar) &\Rightarrow k = n\pi/L \\ &\Rightarrow E = \hbar^2\pi^2 n^2/2mL^2\end{aligned} \tag{2.1.15}$$

Atomic "boxes" are of course defined by potentials $U(\vec{r})$ that are more complicated than the simple rectangular 1D potential shown in Fig. 2.1.2b, but the essential point is the same: anytime we confine a wave to a box, the frequency or energy is discretized because of the need for the wave to "fit" inside the box.

"Periodic" box: Another kind of box that we will often use is a ring (Fig. 2.1.4) where the end point at $x = L$ is connected back to the first point at $x = 0$ and there are no ends. Real boxes are seldom in this form but this idealization is often used since it simplifies the mathematics. The justification for this assumption is that if we are interested in the properties in the interior of the box, then what we assume at the ends (or surfaces) should

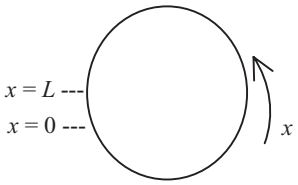


Fig. 2.1.4 Standing waves in a ring.

make no real difference and we could assume anything that makes our calculations simpler. However, this may not be a valid argument for "nanostructures" where the actual surface conditions can and do affect what an experimentalist measures.

Anyway, for a periodic box the eigenfunctions are given by (cf. Eq. (2.1.15))

$$\begin{aligned}&\Psi = \Psi_0 \sin(kx)\exp(-\mathrm{i}Et/\hbar)\\ &\text{and } \Psi = \Psi_0 \cos(kx)\exp(-\mathrm{i}Et/\hbar)\\ &\text{with } k = 2n\pi/L \Rightarrow E = 2\hbar^2\pi^2 n^2/mL^2\end{aligned} \tag{2.1.16}$$

The values of k are spaced by $2\pi/L$ instead of π/L, so that there are half as many allowed values. But for each value of k there is a sine and a cosine function which have the same eigenvalue, so that the eigenvalues now come in pairs.

An important point to note is that whenever we have degenerate eigenstates, that is, two or more eigenfunctions with the same eigenvalue, any linear combination of these eigenfunctions is also an eigenfunction with the same eigenvalue. So we could just as well write the eigenstates as

$$\begin{aligned}&\Psi = \Psi_0 \exp(+\mathrm{i}kx)\exp(-\mathrm{i}Et/\hbar)\\ &\text{and } \Psi = \Psi_0 \exp(+\mathrm{i}kx)\exp(-\mathrm{i}Et/\hbar)\\ &\text{with } k = 2n\pi/L \Rightarrow E = 2\hbar^2\pi^2 n^2/mL^2\end{aligned} \tag{2.1.17}$$

This is done quite commonly in analytical calculations and the first of these is viewed as the $+k$ state traveling in the positive x-direction while the second is viewed as the $-k$ state traveling in the negative x-direction.

Electron density and probability current density: An electron with a wavefunction $\Psi(x, t)$ has a probability of $\Psi^*\Psi\,\mathrm{d}V$ of being found in a volume $\mathrm{d}V$. When a number of electrons are present we could add up $\Psi^*\Psi$ for all the electrons to obtain the average electron density $n(x, t)$. What is the corresponding quantity we should sum to obtain the probability current density $J(x, t)$?

The appropriate expression for the probability current density

$$J = \frac{\mathrm{i}\hbar}{2m}\left(\Psi\frac{\partial\Psi^*}{\partial x} - \Psi^*\frac{\partial\Psi}{\partial x}\right) \tag{2.1.18}$$

is motivated by the observation that as long as the wavefunction $\Psi(x,t)$ obeys the Schrödinger equation, it can be shown that

$$\frac{\partial J}{\partial x} + \frac{\partial n}{\partial t} = 0 \tag{2.1.19}$$

if J is given by Eq. (2.1.18) and $n = \Psi^*\Psi$. This ensures that the continuity equation is satisfied regardless of the detailed dynamics of the wavefunction. The electrical current density is obtained by multiplying J by the charge $(-q)$ of an electron.

It is straightforward to check that the "$+k$" and "$-k$" states in Eq. (2.1.17) carry equal and opposite non-zero currents proportional to the electron density

$$J = (\hbar k/m)\,\Psi\Psi^* \tag{2.1.20}$$

suggesting that we associate $(\hbar k/m)$ with the velocity v of the electron (since we expect J to equal nv). However, this is true only for the plane wave functions in Eq. (2.1.17). The cosine and sine states in Eq. (2.1.16), for example, carry zero current. Indeed Eq. (2.1.18) will predict zero current for any *real* wavefunction.

2.2 Method of finite differences

The Schrödinger equation for a hydrogen atom can be solved analytically, but most other practical problems require a numerical solution. In this section I will describe one way of obtaining a numerical solution to the Schrödinger equation. Most numerical methods have one thing in common – they use some trick to convert the

	wavefunction $\Psi(\vec{r},t)$	*into a*	column vector $\{\psi(t)\}$
and the	differential operator H_{op}	*into a*	matrix $[H]$

so that the Schrödinger equation is converted from a

partial differential equation	*into a*	matrix equation
$\mathrm{i}\hbar\dfrac{\partial}{\partial t}\Psi(\vec{r},t) = H_{\text{op}}\Psi(\vec{r},t)$		$\mathrm{i}\hbar\dfrac{\mathrm{d}}{\mathrm{d}t}\{\psi(t)\} = [H]\{\psi(t)\}$

This conversion can be done in many ways, but the simplest one is to choose a discrete lattice. To see how this is done let us for simplicity consider just one dimension and discretize the position variable x into a lattice as shown in Fig. 2.2.1: $x_n = na$.

We can represent the wavefunction $\Psi(x,t)$ by a column vector $\{\psi_1(t)\ \psi_2(t)\ \cdots\ \cdots\}^{\mathrm{T}}$ ("T" denotes transpose) containing its values around each of the lattice points at time t. Suppressing the time variable t for clarity, we can write

$$\{\psi_1 \quad \psi_2 \quad \cdots \quad \cdots\} = \{\Psi(x_1) \quad \Psi(x_2) \quad \cdots \quad \cdots\}$$

This representation becomes exact only in the limit $a \to 0$, but as long as a is smaller than the spatial scale on which Ψ varies, we can expect it to be reasonably accurate.

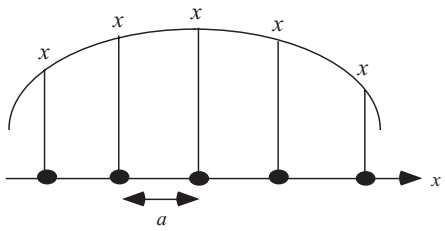


Fig. 2.2.1 A continuous function can be represented by its values at a set of points on a discrete lattice.

The next step is to obtain the matrix representing the Hamiltonian operator

$$H_{\text{op}} \equiv -\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + U(x)$$

Basically what we are doing is to turn a *differential* equation into a *difference* equation. There is a standard procedure for doing this – the finite difference technique:

$$\left(\frac{\partial^2 \Psi}{\partial x^2}\right)_{x=x_n} \rightarrow \frac{1}{a^2}\left[\Psi(x_{n+1}) - 2\Psi(x_n) + \Psi(x_{n-1})\right]$$

and

$$U(x)\,\Psi(x) \rightarrow U(x_n)\Psi(x_n)$$

This allows us to write (note: $t_0 \equiv \hbar^2/2ma^2$ and $U_n \equiv U(x_n)$)

$$\begin{aligned} \mathrm{i}\hbar\frac{\mathrm{d}\psi_n}{\mathrm{d}t} &= \left[H_{\text{op}}\,\psi\right]_{x=x_n} = (U_n + 2t_0)\,\psi_n - t_0\psi_{n-1} - t_0\psi_{n+1} \\ &= \sum_m \left[(U_n + 2t_0)\,\delta_{n,m} - t_0\delta_{n,m+1} - t_0\delta_{n,m-1}\right]\psi_m \end{aligned} \tag{2.2.1}$$

where $\delta_{n,m}$ is the Kronecker delta, which is one if $n = m$ and zero if $n \neq m$. We can write Eq. (2.2.1) as a matrix equation:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\{\psi(t)\} = [H]\{\psi(t)\} \tag{2.2.2}$$

The elements of the Hamiltonian matrix are given by

$$H_{n,m} = [U_n + 2t_0]\,\delta_{n,m} - t_0\delta_{n,m+1} - t_0\delta_{n,m-1} \tag{2.2.3}$$

where $t_0 \equiv \hbar^2/2ma^2$ and $U_n \equiv U(x_n)$. This means that the matrix representing H looks like this

$$
\begin{array}{lccccccc}
H = & 1 & & 2 & \dots & N-1 & & N \\
1 & 2t_0 + U_1 & & -t_0 & & 0 & & 0 \\
2 & -t_0 & & 2t_0 + U_2 & & 0 & & 0 \\
 & & \dots & & \dots & & \dots & \\
N-1 & 0 & & 0 & & 2t_0 + U_{N-1} & & -t_0 \\
N & 0 & & 0 & & -t_0 & & 2t_0 + U_N
\end{array}
\tag{2.2.4}
$$

For a given potential function $U(x)$ it is straightforward to set up this matrix, once we have chosen an appropriate lattice spacing a.

Eigenvalues and eigenvectors: Now that we have converted the Schrödinger equation into a matrix equation (Eq. (2.2.2))

$$
\mathrm{i}\hbar \frac{\mathrm{d}}{\mathrm{d}t} \{\psi(t)\} = [H]\,\{\psi(t)\}
$$

how do we calculate $\{\psi(t)\}$ given some initial state $\{\psi(0)\}$? The standard procedure is to find the eigenvalues E_α and eigenvectors $\{\alpha\}$ of the matrix $[H]$:

$$
[H]\,\{\alpha\} = E_\alpha\,\{\alpha\} \tag{2.2.5}
$$

Making use of Eq. (2.2.5) it is easy to show that the wavefunction $\{\psi(t)\} = \mathrm{e}^{-\mathrm{i}E_\alpha t/\hbar}\,\{\alpha\}$ satisfies Eq. (2.2.2). Since Eq. (2.2.2) is linear, any superposition of such solutions

$$
\{\psi(t)\} = \sum_\alpha C_\alpha \mathrm{e}^{-\mathrm{i}E_\alpha t/\hbar}\,\{\alpha\} \tag{2.2.6}
$$

is also a solution. It can be shown that this form, Eq. (2.2.6), is "complete," that is, any solution to Eq. (2.2.2) can be written in this form. Given an initial state we can figure out the coefficients C_α. The wavefunction at subsequent times t is then given by Eq. (2.2.6). Later we will discuss how we can figure out the coefficients. For the moment we are just trying to make the point that the dynamics of the system are easy to visualize or describe in terms of the eigenvalues (which are the energy levels that we talked about earlier) and the corresponding eigenvectors (which are the wavefunctions associated with those levels) of $[H]$. That is why the first step in discussing any system is to write down the matrix $[H]$ and to find its eigenvalues and eigenvectors. This is easily done using any standard mathematical package like MATLAB as we will discuss in the next section.

2.3 Examples

Let us now look at a few examples to make sure we understand how to find the eigenenergies and eigenvectors numerically using the method of finite differences described in the last section. These examples are all simple enough to permit analytical solutions that we can use to compare and evaluate our numerical solutions. The advantage of the numerical procedure is that it can handle more complicated problems just as easily, even when no analytical solutions are available.

2.3.1 Particle in a box

Consider, first the "particle in a box" problem that we mentioned in Section 2.1. The potential is constant inside the box which is bounded by infinitely high walls at $x = 0$ and at $x = L$ (Fig. 2.3.1). The eigenstates $\phi_\alpha(x)$ are given by

$$\phi_\alpha(x) \sim \sin(k_\alpha\, x) \qquad \text{where } k_\alpha = \alpha\pi/L,\ \alpha = 1, 2, \ldots$$

and their energies are given by $E_\alpha = \hbar^2 k_\alpha^2/2m$.

We could solve this problem numerically by selecting a discrete lattice with 100 points and writing down a 100×100 matrix $[H]$ using Eq. (2.2.4) with all $U_n = 0$:

$$H = \begin{array}{c|cccccc} & 1 & & 2 & \ldots & 99 & 100 \\ 1 & 2t_0 & & -t_0 & & 0 & 0 \\ 2 & -t_0 & & 2t_0 & & 0 & 0 \\ & & \ldots & & \ldots & \ldots & \\ 99 & 0 & & 0 & & 2t_0 & -t_0 \\ 100 & 0 & & 0 & & -t_0 & 2t_0 \end{array} \tag{2.3.1}$$

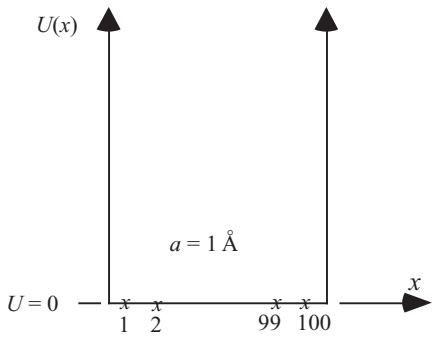


Fig. 2.3.1 Energy levels for a "particle in a box" are calculated using a discrete lattice of 100 points spaced by $a = 1$ Å.

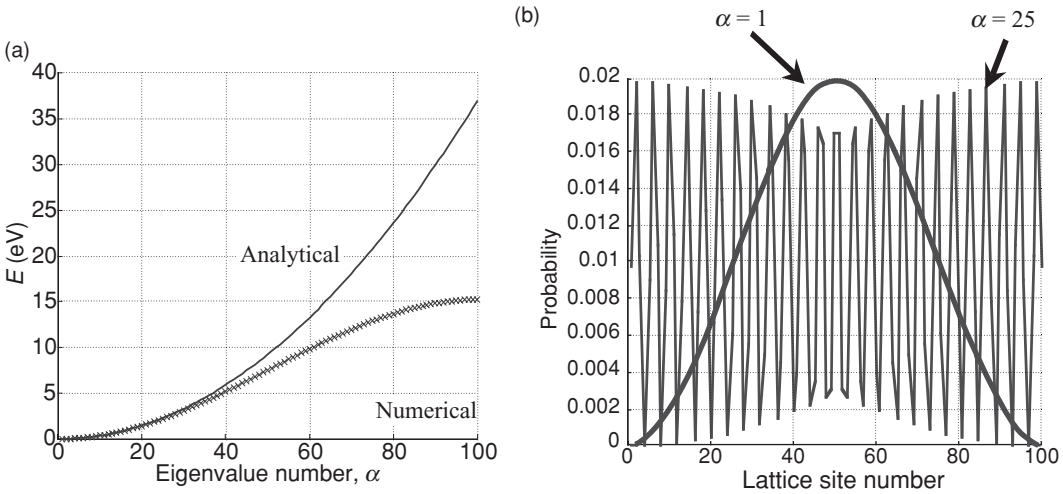


Fig. 2.3.2 (a) Numerical evaluation (see Fig. 2.3.1) yields 100 eigenvalues that follow the analytical result well for low energies but deviate at higher energies because the wavefunctions oscillate too rapidly. (b) Probability distribution (squared eigenfunction) for eigenvalues $\alpha = 1$ and $\alpha = 25$.

It is straightforward to set up this matrix and use any standard mathematical package like MATLAB to find the eigenvalues and the corresponding eigenvectors. We obtained 100 eigenvalues, which are plotted in Fig. 2.3.2a. They follow the analytical result $E_\alpha = \hbar^2\pi^2\alpha^2/2mL^2$, with $L = 101a$, fairly well at low energy, but deviate at higher energies because of the rapid oscillations in the wavefunction. Our finite difference approximation to the second derivative operator (note that $t_0 \equiv \hbar^2/2ma^2$)

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\Psi}{\partial x^2}\right)_{x=x_n} \rightarrow t_0\left[\Psi(x_{n+1}) - 2\Psi(x_n) + \Psi(x_{n-1})\right]$$

is accurate only if Ψ varies slowly enough on a length scale of a. Indeed if we put $\Psi \sim \sin(k_\alpha x)$ it is straightforward to show that

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\Psi}{\partial x^2}\right)_{x=x_n} = t_0\,(k_\alpha a)^2\,\Psi(x_n)$$

while

$$t_0\left[\Psi(x_{n+1}) - 2\Psi(x_n) + \Psi(x_{n-1})\right] = 2t_0(1 - \cos k_\alpha a)\Psi(x_n)$$

Since $k_\alpha = \alpha\pi/L$, the analytical eigenvalues follow a parabolic function while the numerical eigenvalues follow a cosine function:

$E_\alpha = t_0(\pi\alpha a/L)^2$ $\qquad$ $E_\alpha = 2t_0[1 - \cos(\alpha\pi a/L)]$
Analytical result $\qquad$ Numerical result

The two are equivalent only if $k_\alpha a = \alpha\pi a/L \ll 1$ so that $\cos(k_\alpha a) \approx 1 - (k_\alpha^2 a^2/2)$.

Normalization: Figure 2.3.2b shows the eigenfunction squared corresponding to the eigenvalues $\alpha = 1$ and $\alpha = 25$. A word about the normalization of the wavefunctions: In analytical treatments, it is common to normalize wavefunctions such that

$$\int_{-\infty}^{+\infty} \mathrm{d}x \, |\phi_\alpha(x)|^2 = 1$$

Numerically, a normalized eigenvector satisfies the condition

$$\sum_{n=1}^{N} |\phi_\alpha(x_n)|^2 = 1$$

So when we compare numerical results with analytical results we should expect

$$\underset{\text{Numerical}}{|\phi_\alpha(x_n)|^2} \quad = \underset{\text{Analytical}}{|\phi_\alpha(x)|^2 \, a} \tag{2.3.2}$$

where a is the lattice constant (see Fig. 2.3.1). For example, in the present case

$$\underset{\text{Analytical}}{|\phi_\alpha(x)|^2 = (2/L)\sin^2(k_\alpha \, x)} \longrightarrow \underset{\text{Numerical}}{|\phi_\alpha(x_n)|^2 = (2a/L)\sin^2(k_\alpha \, x_n)}$$

Since we used $a = 1$ Å and $L = 101$ Å, the numerical probability distribution should have a peak value of $2a/L \approx 0.02$ as shown in Fig. 2.3.2b.

Boundary conditions: One more point: Strictly speaking, the matrix $[H]$ is infinitely large, but in practice we always truncate it to a finite number, say N, of points. This means that at the two ends we are replacing (see Eq. (2.2.1))

$$-t_0\psi_0 + (2t_0 + U_1)\psi_1 - t_0\psi_2 \quad \textit{with} \quad (2t_0 + U_1)\psi_1 - t_0\psi_2$$

and

$$-t_0\psi_{N-1} + (2t_0 + U_N)\psi_N - t_0\psi_{N+1} \quad \textit{with} \quad -t_0\psi_{N-1} + (2t_0 + U_N)\psi_N$$

In effect we are setting ψ_0 and ψ_{N+1} equal to zero. This boundary condition is appropriate if the potential is infinitely large at point 0 and at point $N + 1$ as shown in Fig. 2.3.3. The actual value of the potential at the end points will not affect the results as long as the wavefunctions are essentially zero at these points anyway.

Another boundary condition that is often used is the *periodic boundary condition* where we assume that the last point is connected back to the first point so that there are no ends (Fig. 2.3.4). As we mentioned earlier (Fig. 2.1.4), the justification for this assumption is that if we are interested in the properties in the interior of a structure, then what we assume at the boundaries should make no real difference and we could assume anything to make our calculations simpler.

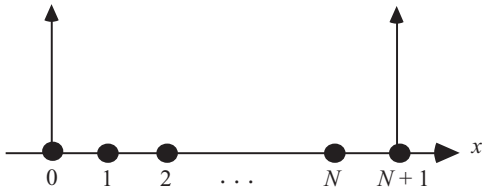


Fig. 2.3.3 The boundary condition $\psi_0 = 0$ and $\psi_{N+1} = 0$ can be used if we assume an infinitely large potential at points 0 and $N+1$.

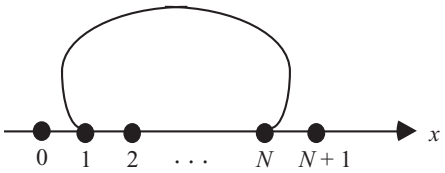


Fig. 2.3.4 Periodic boundary conditions assume that there are no "ends." Point N is connected back to point 1 as if the structure were in the form of a ring making $(N+1)$ equivalent to 1.

Mathematically, periodic boundary conditions are implemented by modifying the Hamiltonian to

$$
\begin{array}{rccccccc}
H = & 1 & & 2 & \ldots & 99 & & 100 \\
1 & 2t_0 & & -t_0 & & 0 & & -t_0 \\
2 & -t_0 & & 2t_0 & & 0 & & 0 \\
 & & \ldots & & \ldots & & \ldots & \\
99 & 0 & & 0 & & 2t_0 & & -t_0 \\
100 & -t_0 & & 0 & & -t_0 & & 2t_0
\end{array}
\tag{2.3.3}
$$

Note that compared to the infinite wall boundary conditions (cf. Eq. (2.3.1)) the only change is in the elements $H(1, 100)$ and $H(100, 1)$. This does change the resulting eigenvalues and eigenvectors, but the change is imperceptible if the number of points is large. The eigenfunctions are now given by

$$\phi_\alpha(x) \sim \sin(k_\alpha x) \quad \text{and} \quad \cos(k_\alpha x)$$

where $k_\alpha = \alpha\, 2\pi/L, \alpha = 1, 2, \ldots$ instead of

$$\phi_\alpha(x) \sim \sin(k_\alpha x)$$

where $k_\alpha = \alpha\, \pi/L, \alpha = 1, 2, \ldots$

The values of $k\alpha$ are spaced by $2\pi/L$ instead of π/L, so that there are half as many allowed values. But for each value of k_α there is a sine and a cosine function which have the same eigenvalue, so that the eigenvalues now come in pairs as evident from Fig. 2.3.5.

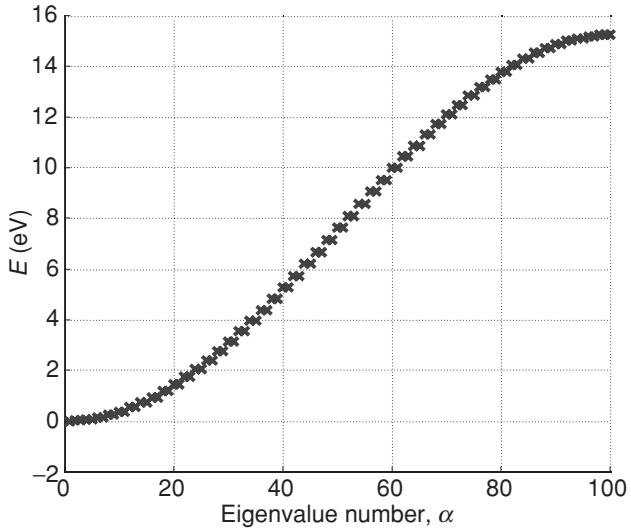


Fig. 2.3.5 Energy eigenvalues for a box of length 101 Å (same as Fig. 2.3.1) with periodic boundary conditions: the eigenvalues now come in pairs.

As we discussed earlier, instead of writing the eigenstates as

$$\cos(k_\alpha x) \quad \text{and} \quad \sin(k_\alpha x)$$

we could just as well write them as

$$e^{ik_\alpha x} = \cos(k_\alpha x) + i\sin(k_\alpha x) \quad \text{and} \quad e^{-ik_\alpha x} = \cos(k_\alpha x) - i\sin(k_\alpha x)$$

This is done quite commonly in analytical calculations, but numerical calculations will typically give the eigenvectors as $\cos(k_\alpha x)$ and $\sin(k_\alpha x)$. Both forms are equally correct though one may be more convenient than the other for certain calculations.

Number of eigenvalues: Another important point to note about the numerical solution is that it yields a finite number of eigenvalues (unlike the analytical solution for which the number is infinite). This is expected since a finite matrix can have only a finite number of eigenvalues, but one might wonder why we do not have an infinite number of E_α corresponding to an infinite number of $k_\alpha a = \alpha 2\pi a/L$, just as we have for the analytical result. The reason is that for a discrete lattice, the wavefunctions

$$\sin(k_\alpha\, x) \quad \text{and} \quad \sin([k_\alpha + (2\pi/a)]x)$$

represent the same state because at any lattice point $x_n = na$,

$$\sin(k_\alpha\, x_n) = \sin([k_\alpha + (2\pi/a)]x_n)$$

They are NOT equal between two lattice points and thus represent distinct states in a non-discrete representation. But once we adopt a discrete lattice, values of k_α differing

by $2\pi/a$ represent identical states and only the values of $k_\alpha a$ within a range of 2π yield independent solutions. Since $k_\alpha a = \alpha\pi a/L = \alpha\pi/N$, this means that there are only N values of α that need to be considered. It is common to restrict the values of $k_\alpha a$ to the range (sometimes called the first Brillouin zone)

$-\pi < k_\alpha a \leq +\pi$ for periodic boundary conditions

and

$0 < k_\alpha a \leq +\pi$ for infinite wall boundary conditions

2.3.2 Particle in a 3D "box"

For simplicity we have limited our discussion of the method of finite differences to one dimension, but the basic idea carries over in principle to two or three dimensions. The diagonal elements of $[H]$ are equal to t_0 times the number of nearest neighbors (two in one dimension, four in two dimensions and six in three dimensions) plus the potential $U(\vec{r})$ evaluated at the lattice site, while the off-diagonal elements are equal to $-t_0$ for neighboring sites on the lattice. That is, (ν is the number of nearest neighbors)

$$\begin{aligned} H_{nm} &= \nu t_0 && n = m \\ &= -t_0 && n, m \text{ are nearest neighbors} \\ &= 0 && \text{otherwise} \end{aligned} \tag{2.3.4}$$

However, we run into a practical difficulty in two or three dimensions. If we have lattice points spaced by 1 Å, then a one-dimensional problem with $L = 101$ Å requires a matrix $[H]$ 100×100 in size. But in three dimensions this would require a matrix $10^6 \times 10^6$ in size. This means that in practice we are limited to very small problems. However, if the coordinates are separable then we can deal with three separate one-dimensional problems as opposed to one giant three-dimensional problem. This is possible if the potential can be separated into an x-, a y-, and a z-dependent part:

$$U(\vec{r}) = U_x(x) + U_y(y) + U_z(z) \tag{2.3.5}$$

The wavefunction can then be written in product form:

$$\Psi(\vec{r}) = X(x)Y(y)Z(z)$$

where each of the functions $X(x)$, $Y(y)$, and $Z(z)$ is obtained by solving a separate one-dimensional Schrödinger equation:

$$\begin{aligned} E_x X(x) &= \left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2} + U_x(x)\right) X(x) \\ E_y Y(y) &= \left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}y^2} + U_y(y)\right) Y(y) \\ E_z Z(z) &= \left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}z^2} + U_z(z)\right) Z(z) \end{aligned} \tag{2.3.6}$$

The total energy E is equal to the sum of the energies associated with each of the three dimensions: $E = E_x + E_y + E_z$.

Spherically symmetric potential: Some problems may not be separable in Cartesian coordinates but could be separable in cylindrical or spherical coordinates. For example, the potential in a hydrogen atom $U(\vec{r}) = -q^2/4\pi\varepsilon_0 r$ cannot be separated in (x, y, z). But it is separable in (r, θ, ϕ) and the wavefunction may be written in the form

$$\Psi(r, \theta, \phi) = [f(r)/r]\, Y_l^m(\theta,\phi) \tag{2.3.7}$$

where the radial wavefunction $f(r)$ is obtained by solving the radial Schrödinger equation:

$$Ef(r) = \left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}r^2} + U(r) + \frac{l(l+1)\hbar^2}{2mr^2}\right) f(r) \tag{2.3.8}$$

Here $l = 0$ for s levels, $l = 1$ for p levels and so on. $Y_l^m(\theta, \phi)$ are the spherical harmonics given by

$$Y_0^0(\theta, \phi) = \sqrt{1/4\pi}$$
$$Y_1^0(\theta, \phi) = \sqrt{3/4\pi}\cos\theta$$
$$Y_1^{\pm1}(\theta, \phi) = \pm\sqrt{3/8\pi}\sin\theta \mathrm{e}^{\pm\mathrm{i}\phi}$$

etc. Equation (2.3.8) can be solved numerically using the method of finite differences that we have described.

Normalization: Note that the overall wavefunctions are normalized such that

$$\int_0^\infty \mathrm{d}r\, r^2 \int_0^\pi \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi\, |\Psi|^2 = 1$$

Since, from Eq. (2.3.7)

$$\Psi(r, \theta, \phi) = [f(r)/r]Y_l^m(\theta,\phi)$$

and the spherical harmonics are normalized such that

$$\int_0^\pi \mathrm{d}\theta \sin\theta \int_0^{2\pi} \mathrm{d}\phi \left|Y_l^m\right|^2 = 1$$

it is easy to see that the radial function $f(r)$ obeys the normalization condition

$$\int_0^\infty \mathrm{d}r|f(r)|^2 = 1 \tag{2.3.9}$$

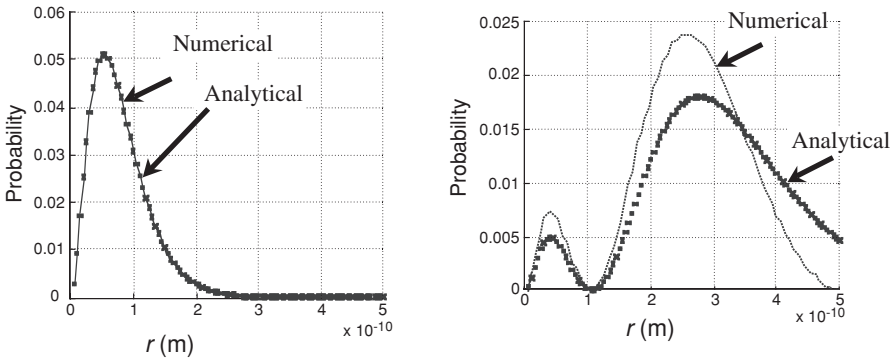


Fig. 2.3.6 Radial probability distribution $|f(r)|^2$ corresponding to the two lowest eigenvalues (-13.56 eV and -2.96 eV) for $l = 0$ (which correspond to the 1s and 2s levels). The dots show the analytical result (Eqs. (2.1.10a, b)) while the solid curve denotes the numerical result obtained using a lattice with 100 points spaced by $a = 0.05$ Å.

suggesting that we view $|f(r)|^2$ as a radial probability distribution function such that $|f(r)|^2 \Delta r$ tells us the probability of finding the electron in the volume between r and $(r + \Delta r)$. Numerical results with a lattice spacing of a should be compared with the analytical values of $|f(r)|^2 a$. For example, for the 1s and 2s levels,

$$|f_{1s}|^2 a = \left(4ar^2/a_0^3\right) \mathrm{e}^{-2r/a_0} \tag{2.3.10a}$$

$$|f_{2s}|^2 a = \left(ar^2/8a_0^3\right) \left(2 - \frac{r}{a_0}\right)^2 \mathrm{e}^{-2r/2a_0} \tag{2.3.10b}$$

Numerical results: If we use a lattice with 100 points spaced by $a = 0.05$ Å then the two lowest eigenvalues with $l = 0$ (which correspond to the 1s and 2s levels) are

$$E_{1s} = -13.56 \text{ eV and } E_{2s} = -2.96 \text{ eV}$$

as compared with the analytical values (see Eq. (2.2.6)) $E_{1s} = -13.59$ eV and $E_{2s} = -3.4$ eV. The 1s level agrees well, but the 2s level is considerably off. The reason is easy to see if we plot the corresponding $|f(r)|^2$ and compare with the analytical results. It is evident from Fig. 2.3.6 that the 1s wavefunction matches well, but it is apparent that we do not have enough range for the 2s function. This can be fixed by choosing a larger lattice spacing, namely $a = 0.1$ Å. Figure 2.3.7 shows that the wavefunction now matches the analytical result quite well and the 2s eigenvalue is -3.39 eV, in good agreement with the analytical result. However, the 1s eigenvalue degrades slightly to -13.47 eV, because the wavefunction is not sampled frequently enough. We could improve the agreement for both 1s and 2s levels by using 200 points spaced by $a = 0.05$ Å, so that we would have both fine sampling and large range. But the calculation would then take longer since we would have to calculate the eigenvalues of a (200×200) matrix instead of a (100×100) matrix.

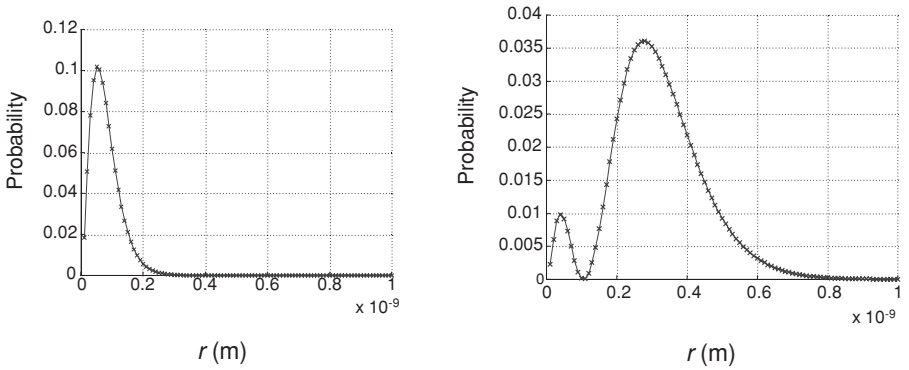


Fig. 2.3.7 Radial probability distribution $|f(r)|^2$ corresponding to the two lowest eigenvalues (−13.47 eV and −3.39 eV) for $l = 0$ (which correspond to the 1s and 2s levels). Solid line shows the analytical result (Eqs. (2.3.10a, b)) while the ×'s denote the numerical result obtained using a lattice with 100 points spaced by $a = 0.1$ Å.

This simple example illustrates the essential issues one has to consider in setting up the lattice for a numerical calculation. The lattice constant a has to be small enough to provide adequate sampling of the wavefunction while the size of the lattice has to be big enough to cover the entire range of the wavefunction. If it were essential to describe all the eigenstates accurately, our problem would be a hopeless one. Luckily, however, we usually need an accurate description of the eigenstates that lie within a certain range of energies and it is possible to optimize our matrix $[H]$ so as to provide an accurate description over a desired range.

EXERCISES

E.2.1. (a) Use a discrete lattice with 100 points spaced by 1 Å to calculate the eigenenergies for a particle in a box with infinite walls and compare with $E_\alpha = \hbar^2\pi^2\alpha^2/2mL^2$ (cf. Fig. 2.3.2a). Plot the probability distribution (eigenfunction squared) for the eigenvalues $\alpha = 1$ and $\alpha = 50$ (cf. Fig. 2.3.2b). (b) Find the eigenvalues using periodic boundary conditions and compare with Fig. 2.3.5.

E.2.2. (a) Obtain the radial equation given in Eq. (2.3.8) by (1) writing the operator ∇^2 in the Schrödinger equation in spherical coordinates:

$$\nabla^2 \equiv \left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right)$$

(2) noting that the spherical harmonics $Y_l^m(\theta,\phi)$ are eigenfunctions of the angular part:

$$\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right)Y_l^m = -l(l+1)Y_l^m$$

(3) writing the wavefunction $\Psi(r) = \Psi(r)\, Y_l^m(\theta, \phi)$ and noting that

$$\nabla^2\Psi = \left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} - \frac{l(l+1)}{r^2}\right)\Psi$$

(4) simplifying the Schrödinger equation to write for the radial part

$$E\psi = \left(-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\right) + U(r) + \frac{\hbar^2 l(l+1)}{2mr^2}\right)\psi$$

and finally (5) writing $\psi(r) = f(r)/r$, to obtain Eq. (2.3.8) for $f(r)$.
(b) Use a discrete lattice with 100 points spaced by a to solve Eq. (2.3.8)

$$Ef(r) = \left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}r^2} - \frac{q^2}{4\pi\varepsilon_0 r} + \frac{l(l+1)\hbar^2}{2mr^2}\right) f(r)$$

for the 1s and 2s energy levels of a hydrogen atom. Plot the corresponding radial probability distributions $|f(r)|^2$ and compare with the analytical results for (a) $a =$ 0.05 Å (cf. Fig. 2.3.6) and (b) $a = 0.1$ Å (cf. Fig. 2.3.7).

Strictly speaking one should replace the electron mass with the reduced mass to account for nuclear motion, but this is a small correction compared to our level of accuracy.

E.2.3. Use Eq. (2.1.18) to evaluate the current density associated with an electron having the wavefunction

$$\Psi(x, t) = (e^{+\gamma x} + ae^{-\gamma x})e^{-\mathrm{i}Et/\hbar}$$

assuming γ is (a) purely imaginary ($= \mathrm{i}\beta$) and (b) purely real.

3 Self-consistent field

As we move from the hydrogen atom (one electron only) to multi-electron atoms, we are immediately faced with the issue of electron–electron interactions, which is at the heart of almost all the unsolved problems in our field. In this chapter I will explain (1) the self-consistent field (SCF) procedure (Section 3.1), which provides an approximate way to include electron–electron interactions into the Schrödinger equation, (2) the interpretation of the energy levels obtained from this so-called "one-electron" Schrödinger equation (Section 3.2), and (3) the energetic considerations underlying the process by which atoms "bond" to form molecules (Section 3.3). Finally, a supplementary section elaborates on the concepts of Section 3.2 for interested readers (Section 3.4).

3.1 The self-consistent field (SCF) procedure

One of the first successes of quantum theory after the interpretation of the hydrogen atom was to explain the periodic table of atoms by combining the energy levels obtained from the Schrödinger equation with the Pauli exclusion principle requiring that each level be occupied by no more than one electron. The energy eigenvalues of the Schrödinger equation for each value of l starting from $l = 0$ (see Eq. (2.3.8)) are numbered with integer values of n starting from $n = l + 1$. For any (n, l) there are $(2l + 1)$ levels with distinct angular wavefunctions (labeled with another index m), all of which have the same energy. For each (n, l, m) there a is an up-spin and a down-spin level making the number of degenerate levels equal to $2(2l + 1)$ for a given (n, l). The energy levels look something like Fig. 3.1.1.

The elements of the periodic table are arranged in order as the number of electrons increases by one from one atom to the next. Their electronic structure can be written as: hydrogen, $1s^1$; helium, $1s^2$; lithium, $1s^2 2s^1$; beryllium, $1s^2 2s^2$; boron, $1s^2 2s^2\ 2p^1$, etc., where the superscript indicates the number of electrons occupying a particular orbital.

How do we calculate the energy levels for a multi-electron atom? The time-independent Schrödinger equation

$$E_\alpha \Phi_\alpha(\vec{r}) = H_{\text{op}} \Phi_\alpha(\vec{r}) \quad \text{where} \quad H_{\text{op}} \equiv -\frac{\hbar^2}{2m}\nabla^2 + U(\vec{r})$$

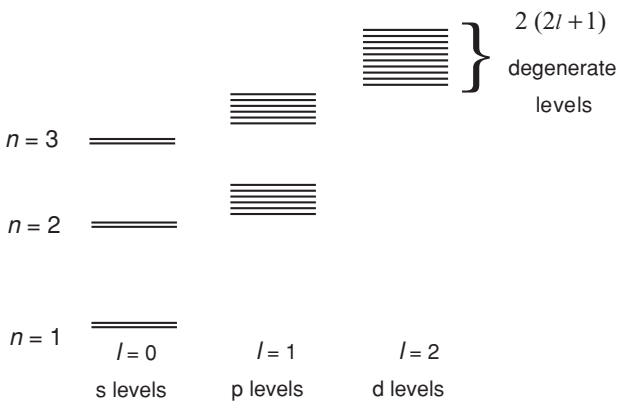


Fig. 3.1.1

provides a fairly accurate description of the observed spectra of all atoms, not just the hydrogen atom. However, multi-electron atoms involve electron–electron interactions that are included by adding a "self-consistent field (SCF)," $U_{\mathrm{SCF}}(\vec{r})$, to the nuclear potential $U_{\mathrm{nuc}}(\vec{r})$: $U(\vec{r}) = U_{\mathrm{nuc}}(\vec{r}) + U_{\mathrm{SCF}}(\vec{r})$, just as in Section 1.4 we added an extra potential to the Laplace potential U_{L} (see Eq. (1.4.1b)). The nuclear potential U_{nuc}, like U_{L}, is fixed, while U_{SCF} depends on the electronic wavefunctions and has to be calculated from a self-consistent iterative procedure. In this chapter we will describe this procedure and the associated conceptual issues.

Consider a helium atom consisting of two electrons bound to a nucleus with two positive charges $+2q$. What will the energy levels looks like? Our first guess would be simply to treat it just like a hydrogen atom except that the potential is

$$U(\vec{r}) = -2q^2/4\pi\varepsilon_0 r$$

instead of

$$U(\vec{r}) = -q^2/4\pi\varepsilon_0 r$$

If we solve the Schrödinger equation with $U(\vec{r}) = -Zq^2/4\pi\varepsilon_0 r$ we will obtain energy levels given by

$$E_n = -(Z^2/n^2)\,E_0 = -54.4\ \mathrm{eV}/n^2 \quad (Z = 2)$$

just as predicted by the simple Bohr model (see Eqs. (2.1.6a, b)). However, this does not compare well with experiment at all. For example, the ionization potential of helium is ~25 eV, which means that it takes a photon with an energy of at least 25 eV to ionize a helium atom:

$$\mathrm{He} + h\nu \rightarrow \mathrm{He}^+ + \mathrm{e}^- \tag{3.1.1a}$$

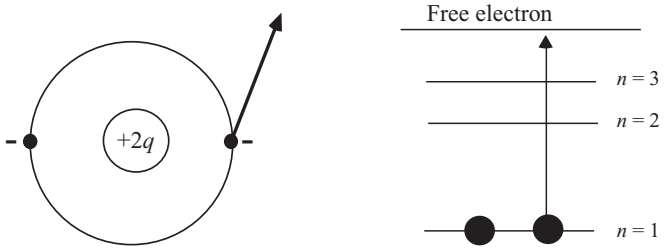


Fig. 3.1.2 Ionization of a neutral helium atom takes approximately 25 eV of energy, suggesting that the $n = 1$ level has an energy of -25 eV.

This suggests that the 1s level of a helium atom has an energy of -25 eV and not -54.4 eV as the simple argument would suggest. How could we be off by over 30 eV? It is because we did not account for the other electron in helium. If we were to measure the energy that it takes to remove the second electron from He^+

$$He^+ + h\nu \rightarrow He^{++} + e^- \tag{3.1.1b}$$

the result (known as the second ionization potential) is indeed close to 54.4 eV. But the (first) ionization potential is about 30 eV less, indicating that it takes 30 eV less energy to pull an electron out of a neutral helium atom than it takes to pull an electron out of a helium ion (He^+) that has already lost one electron. The reason is that an electron in a helium atom feels a repulsive force from the other electron, which effectively raises its energy by 30 eV and makes it easier for it to escape (Fig. 3.1.2).

In general, the ionization levels for multielectron atoms can be calculated approximately from the Schrödinger equation by adding to the nuclear potential $U_{\text{nuc}}(\vec{r})$, a self-consistent field $U_{\text{SCF}}(\vec{r})$ due to the other electrons (Fig. 3.1.3):

$$U(\vec{r}) = U_{\text{nuc}}(\vec{r}) + U_{\text{SCF}}(\vec{r}) \tag{3.1.2}$$

For all atoms, the nuclear potential arises from the nuclear charge of $+Zq$ located at the origin and is given by $U_{\text{nuc}}(\vec{r}) = -Zq^2/4\pi\varepsilon_0 r$. The self-consistent field arises from the other $(Z-1)$ electrons, since an electron does not feel any potential due to itself. In order to calculate the potential $U_{\text{SCF}}(\vec{r})$ we need the electronic charge which depends on the wavefunctions of the electron which in turn has to be calculated from the Schrödinger equation containing $U_{\text{SCF}}(\vec{r})$. This means that the calculation has to be done self-consistently as follows.

Step 1. Guess electronic potential $U_{\text{SCF}}(\vec{r})$.
Step 2. Find eigenfunctions and eigenvalues from Schrödinger equation.
Step 3. Calculate the electron density $n(\vec{r})$.
Step 4. Calculate the electronic potential $U_{\text{SCF}}(\vec{r})$.

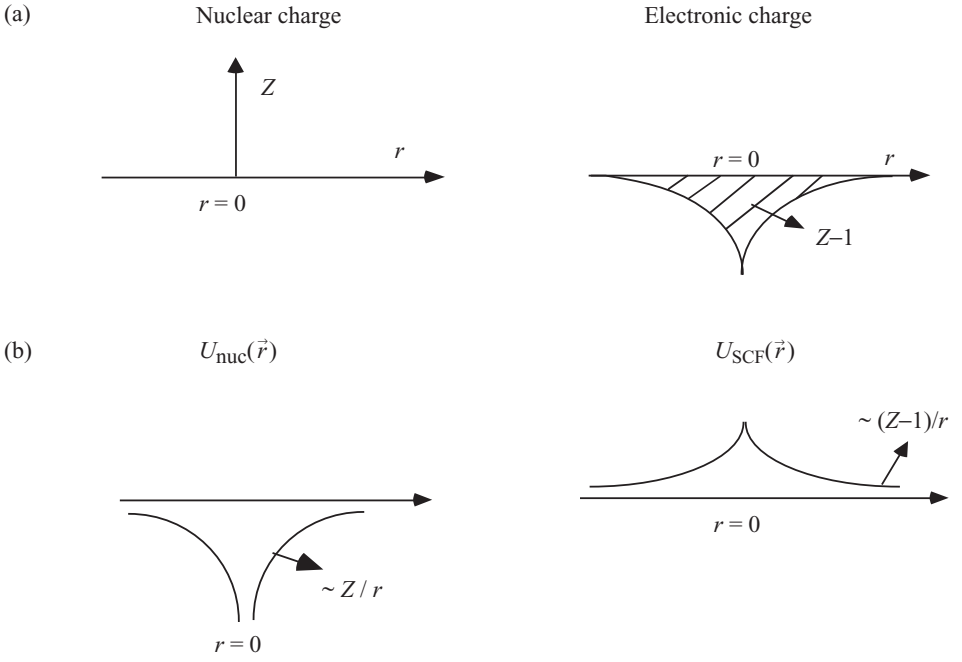


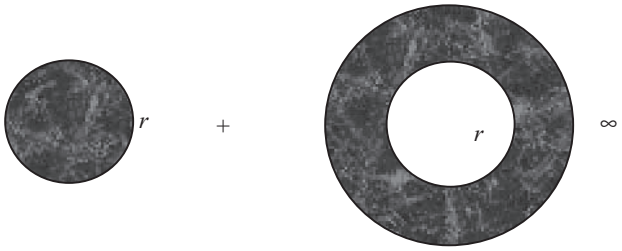
Fig. 3.1.3 Sketch of (a) the nuclear charge density and the electronic charge density; (b) potential energy felt by an additional electron due to the nucleus, $U_{nuc}(r)$, and the other electrons, $U_{SCF}(r)$. The latter has to be calculated self-consistently.

Step 5. If the new $U_{SCF}(\vec{r})$ is significantly different from last guess, update $U_{SCF}(\vec{r})$ and go back to Step 2. If the new $U_{SCF}(\vec{r})$ is within say 10 meV of the last guess, the result has converged and the calculation is complete.

For *Step 2* we can use essentially the same method as we used for the hydrogen atom, although an analytical solution is usually not possible. The potential $U_{SCF}(\vec{r})$ is in general not isotropic (which means independent of θ, ϕ) but for atoms it can be assumed to be isotropic without incurring any significant error. However, the dependence on r is quite complicated so that no analytical solution is possible. Numerically, however, it is just as easy to solve the Schrödinger equation with any $U(r)$ as it is to solve the hydrogen atom problem with $U(r) \sim 1/r$.

For *Step 3* we have to sum up the probability distributions for all the occupied eigenstates:

$$n(\vec{r}) = \sum_{\text{occ } \alpha} |\Phi_\alpha(\vec{r})|^2 = \sum_{\text{occ } n,l,m} \left| \frac{f_n(r)}{r} \right|^2 |Y_{lm}(\theta, \phi)|^2 \tag{3.1.3}$$



First term in Eq. (3.1.5)

Second term in Eq. (3.1.5)

Fig. 3.1.4

If we assume the charge distribution to be isotropic, we can write

$$\sigma(r) \equiv \int r^2 \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \, n(\vec{r}) = \sum_{\text{occ } n,l,m} |f_n(r)|^2 \tag{3.1.4}$$

For *Step 4* we can use straightforward electrostatics to show that

$$U_{\text{SCF}}(r) = \frac{Z-1}{Z}\left[\frac{q^2}{4\pi\varepsilon_0 r}\int_0^r \mathrm{d}r'\,\sigma(r') + \frac{q^2}{4\pi\varepsilon_0}\int_r^\infty \mathrm{d}r'\,\frac{\sigma(r')}{r'}\right] \tag{3.1.5}$$

The two terms in Eq. (3.1.5) arise from the contributions due to the charge within a sphere of radius r and that due to the charge outside of this sphere as shown in Fig. 3.1.4. The first term is the potential at r outside a sphere of charge that can be shown to be the same as if the entire charge were concentrated at the center of the sphere:

$$\frac{q^2}{4\pi\varepsilon_0 r}\int_0^r \mathrm{d}r'\,\sigma(r')$$

The second term is the potential at r inside a sphere of charge and can be shown to be the same as the potential at the center of the sphere (the potential is the same at all points inside the sphere since the electric field is zero)

$$\frac{q^2}{4\pi\varepsilon_0}\int_r^\infty \mathrm{d}r'\,\frac{\sigma(r')}{r'}$$

We obtain the total potential by adding the two components.

To understand the reason for the factor $(Z-1)/Z$ in Eq. (3.1.5), we note that the appropriate charge density for each eigenstate should exclude the eigenstate under consideration, since no electron feels any repulsion due to itself. For example, silicon has 14 electrons $1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^2$ and the self-consistent field includes all but one of

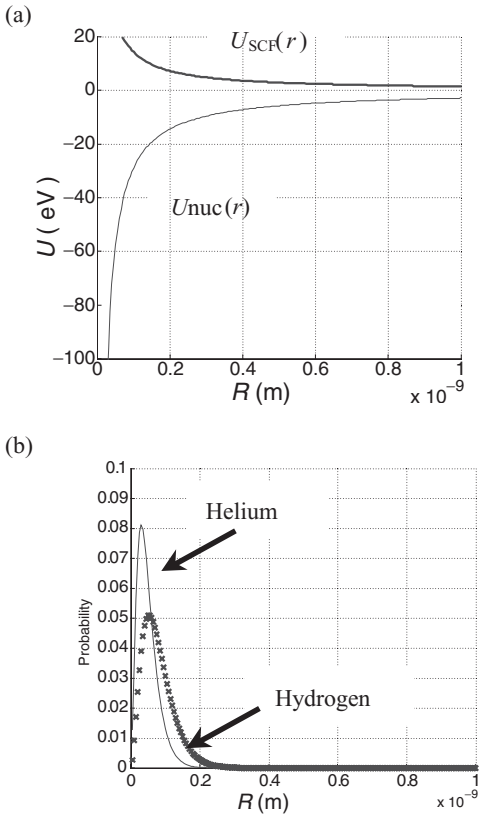


Fig. 3.1.5 Self-consistent field method applied to the helium atom. (a) Nuclear potential $U_{\text{nuc}}(r)$ and the self-consistent electronic potential $U_{\text{SCF}}(r)$. (b) Radial probability distribution for the 1s state in helium and hydrogen.

these electrons – for the 3p level we exclude the 3p electron, for the 3s level we exclude the 3s electron etc. However, it is more convenient to simply take the total charge density and scale it by the factor $(Z-1)/Z$. This preserves the spherical symmetry of the charge distribution and the difference is usually not significant. Note that the total electronic charge is equal to Z:

$$\int_0^\infty \mathrm{d}r\ \sigma(r) = \sum_{\text{occ}\, n,l,m} 1 = Z \tag{3.1.6}$$

since the radial eigenfunctions are normalized: $\int_0^\infty \mathrm{d}r |f_n(r)|^2 = 1$.

Helium atom: Figure 3.1.5 shows the potential profile and the probability distribution for the 1s state of helium obtained using the SCF method we have just described.

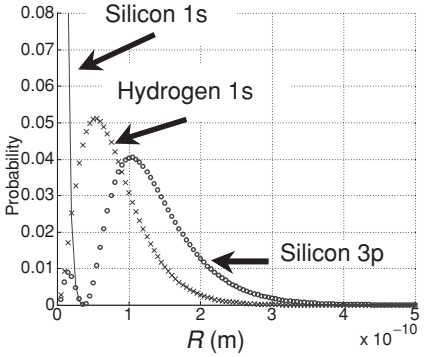


Fig. 3.1.6 Self-consistent field method applied to the silicon atom. The radial probability distributions for hydrogen 1s level and silicon 1s level and 3p level are shown.

Also shown for comparison is the 1s level of the hydrogen atom, discussed in the last chapter.

Silicon atom: Figure 3.1.6 shows the probability distribution for the 1s and 3p states of silicon obtained using the SCF method. Also shown for comparison is the 1s level of the hydrogen atom. Note that the silicon 1s state is very tightly confined relative to the 3p state or the hydrogen 1s state. This is typical of core states and explains why such states remain well-localized in solids, while the outer electrons (like 3p) are delocalized.

3.2 Relation to the multi-electron picture

Multi-electron Schrödinger equation: It is important to recognize that the SCF method is really an approximation that is widely used only because the correct method is virtually impossible to implement. For example, if we wish to calculate the eigen-states of a helium atom with two electrons we need to solve a two-electron Schrödinger equation of the form

$$E\Psi(\vec{r}_1,\ \vec{r}_2) = \left(-\frac{\hbar^2}{2m}\nabla^2 + U(\vec{r}_1) + U(\vec{r}_2) + U_{\text{ee}}(\vec{r}_1,\ \vec{r}_2)\right)\Psi(\vec{r}_1,\ \vec{r}_2) \tag{3.2.1}$$

where $\vec{r}_1$ and $\vec{r}_2$ are the coordinates of the two electrons and U_{ee} is the potential energy due to their mutual repulsion: $U_{\text{ee}}(\vec{r}_1, \vec{r}_2) = e^2/4\pi\varepsilon_0|\vec{r}_1 - \vec{r}_2|$. This is more difficult to solve than the "one-electron" Schrödinger equation that we have been talking about, but it is not impossible. However, this approach quickly gets out of hand as we go to bigger atoms with many electrons and so is seldom implemented directly. But suppose we could actually calculate the energy levels of multi-electron atoms. How would we use our results (in principle, if not in practice) to construct a *one-electron energy level*

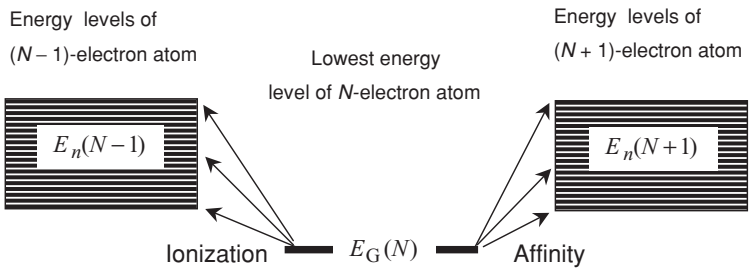


Fig. 3.2.1 One-electron energy levels represent energy differences between the energy levels of the N-electron atom and the $(N-1)$- or the $(N+1)$-electron atom. The former (called the ionization levels) are the filled states from which an electron can be removed while the latter (the affinity levels) are the empty states to which an electron can be added.

diagram like the ones we have been drawing? The answer depends on what we want our one-electron energy levels to tell us.

Ionization levels and affinity levels: Our interest is primarily in describing the flow of current, which involves inserting an electron and then taking it out or vice versa, as we discussed in Chapter 1. So we would want the one-electron energy levels to represent either the energies needed to take an electron out of the atom (ionization levels) or the energies involved in inserting an electron into the atom (affinity levels) (Fig. 3.2.1).

For the ionization levels, the one-electron energies ε_n represent the difference between the ground state energy $E_G(N)$ of the neutral N-electron atom and the nth energy level $E_n(N-1)$ of the positively ionized $(N-1)$-electron atom:

$$\varepsilon_n = E_G(N) - E_n(N-1) \tag{3.2.2a}$$

These ionization energy levels are measured by looking at the photon energy needed to ionize an electron in a particular level. Such photoemission experiments are very useful for probing the occupied energy levels of atoms, molecules, and solids. However, they only provide information about the occupied levels, like the 1s level of a helium atom or the valence band of a semiconductor. To probe unoccupied levels such as the 2s level of a helium atom or the conduction band of a semiconductor we need an inverse photoemission (IPE) experiment (see Fig. 3.2.2):

$$\mathrm{He} + \mathrm{e}^- \rightarrow \mathrm{He}^- + h\nu$$

with which to measure the affinity of the atom for acquiring additional electrons. To calculate the affinity levels we should look at the difference between the ground state energy $E_G(N)$ and the nth energy level $E_n(N+1)$ of the negatively ionized $(N+1)$-electron atom:

$$\varepsilon_n = E_n(N+1) - E_G(N) \tag{3.2.2b}$$

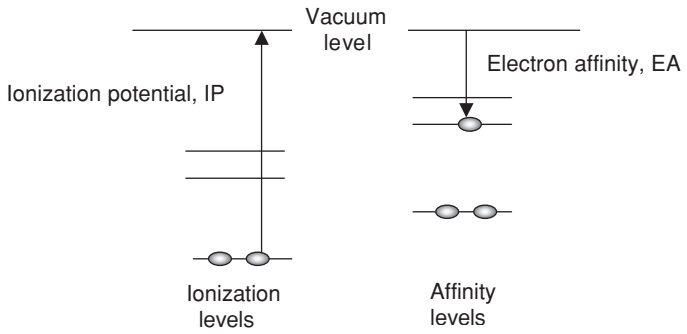


Fig. 3.2.2 The ionization levels include the repulsive potential from $Z-1$ electrons while the affinity levels include that of Z electrons, so that the latter is higher in energy by the single-electron charging energy U_0.

Note that if we want the energy levels to correspond to optical transitions then we should look at the difference between the ground state energy $E_{\mathrm{G}}(N)$ and the nth energy level $E_n(N)$ of the N-electron atom, since visible light does not change the total number of electrons in the atom, just excites them to a higher energy:

$$\varepsilon_n = E_n(N) - E_{\mathrm{G}}(N)$$

There is no *a priori* reason why the energy gap obtained from this calculation should correspond to the energy gap obtained from either the ionization or the affinity levels. In large solids (without significant excitonic effects) we are accustomed to assuming that the optical gap is equal to the gap between the valence and conduction bands, but this need not be true for small nanostructures.

Single-electron charging energy: As we have explained above, the straightforward approach for calculating the energy levels would be to calculate the energies $E_{\mathrm{G}}(N)$ and $E_n(N \pm 1)$ from an N-electron and an $(N \pm 1)$-electron Schrödinger equation (cf. Eq. (3.2.1) which is a two-electron Schrödinger equation) respectively. This, however, is usually impossible and the only practical approach for large atoms, molecules, or solids is to include an effective potential $U_{\mathrm{SCF}}(\vec{r})$ in the Schrödinger equation as we have been discussing.

How do we choose this effective potential? If we use $U_{\mathrm{ee}}(N)$ to denote the total electron–electron interaction energy of an N-electron system then the appropriate U_{SCF} for the ionization levels is equal to the change in the interaction energy as we go from an N-electron to an $(N-1)$-electron atom:

$$[U_{\mathrm{SCF}}]_{\mathrm{ionization}} = U_{\mathrm{ee}}(N) - U_{\mathrm{ee}}(N-1) \tag{3.2.3a}$$

Similarly the appropriate U_{SCF} for the affinity levels is equal to the change in the

interaction energy between an N-electron and an $(N+1)$-electron atom:

$$[U_{\mathrm{SCF}}]_{\mathrm{affinity}} = U_{\mathrm{ee}}(N+1) - U_{\mathrm{ee}}(N) \tag{3.2.3b}$$

The electron–electron interaction energy of a collection of N electrons is proportional to the number of distinct pairs:

$$U_{\mathrm{ee}}(N) = U_0 N(N-1)/2 \tag{3.2.4}$$

where U_0 is the average interaction energy per pair, similar to the single-electron charging energy introduced in Section 1.4. From Eqs. (3.2.3a, b) and (3.2.4) it is easy to see that

$$[U_{\mathrm{SCF}}]_{\mathrm{ionization}} = U_0(N-1) \quad \text{while} \quad [U_{\mathrm{SCF}}]_{\mathrm{affinity}} = U_0 N \tag{3.2.5}$$

This means that to calculate the ionization levels of a Z-electron atom, we should use the potential due to $(Z-1)$ electrons (one electron for helium) as we did in the last section. But to calculate the affinity levels we should use the potential due to Z electrons (two electrons for helium). The energy levels we obtain from the first calculation are lower in energy than those obtained from the second calculation by the single-electron charging energy U_0.

As we discussed in Section 1.5, the single-electron charging energy U_0 depends on the degree of localization of the electronic wavefunction and can be several electron-volts in atoms. Even in nanostructures that are say 10 nm or less in dimension, it can be quite significant (that is, comparable to $k_{\mathrm{B}}T$).

Typically one uses a single self-consistent potential

$$U_{\mathrm{SCF}} = \partial U_{\mathrm{ee}}/\partial N = U_0 N - (U_0/2) \tag{3.2.6}$$

for all levels so that the ionization levels are $(U_0/2)$ lower while the affinity levels are $(U_0/2)$ higher than the energy levels we calculate. One important consequence of this is that even if an SCF calculation gives energy levels that are very closely spaced compared to $k_{\mathrm{B}}T$ (see Fig. 3.2.3a), a structure may not conduct well, because the one-electron charging effects will create a "Coulomb gap" between the occupied and unoccupied levels (Fig. 3.2.3b). Of course, this is a significant effect only if the single-electron charging energy U_0 is larger than $k_{\mathrm{B}}T$.

Hartree approximation: In large conductors (large R) U_0 is negligible and the distinction between Z and $(Z-1)$ can be ignored. The self-consistent potential for both ionization and affinity levels is essentially the same and the expression

$$U_{\mathrm{SCF}} = \partial U_{\mathrm{ee}}/\partial N \tag{3.2.7}$$

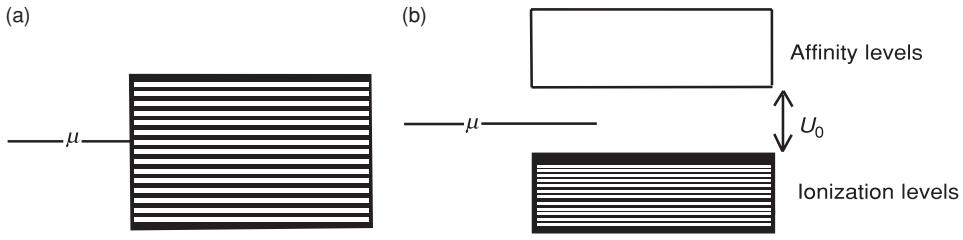


Fig. 3.2.3

can be generalized to obtain the standard expression used in density functional theory (DFT):

$$U_{\mathrm{SCF}}(\vec{r}) = \frac{\partial U_{\mathrm{ee}}}{\partial [n(\vec{r})]} \tag{3.2.8}$$

which tells us that the self-consistent potential at any point $\vec{r}$ is equal to the change in the electron–electron interaction energy due to an infinitesimal change in the number of electrons at the same point. If we use the standard expression for U_{ee} from classical electrostatics

$$U_{\mathrm{ee}} = \frac{1}{2} \int \mathrm{d}\vec{r} \int \mathrm{d}\vec{r}' \, \frac{q^2 n(\vec{r}) n(\vec{r}')}{4\pi\varepsilon \, |\vec{r} - \vec{r}'|} \tag{3.2.9}$$

Equation (3.2.8) yields the Hartree approximation, $U_{\mathrm{H}}(\vec{r})$ for the self-consistent potential:

$$U_{\mathrm{H}}(\vec{r}) = \int \mathrm{d}\vec{r}' \, \frac{q^2 n(\vec{r}')}{4\pi\varepsilon \, |\vec{r} - \vec{r}'|} \tag{3.2.10}$$

which is a solution of the Poisson equation $-\nabla^2 U_{\mathrm{H}} = -q^2 n/\varepsilon$ in a homogeneous medium. Device problems often require us to incorporate complicated boundary conditions including different materials with different dielectric constants. It is then more convenient to solve a modified form of the Poisson equation that allows a spatially varying relative permittivity:

$$-\vec{\nabla} \cdot (\varepsilon_r \nabla U_{\mathrm{H})} \; = q^2 n/\varepsilon_0 \tag{3.2.11}$$

But for atoms, there is no complicated inhomogeneity to account for and it is more convenient to work with Eq. (3.2.10).

Correlation energy: The actual interaction energy is less than that predicted by Eq. (3.2.9) because electrons can correlate their motion so as to avoid each other – this correlation would be included in a many-electron picture but is missed in the one-particle picture. One way to include it is to write

$$U_{\mathrm{ee}} = \frac{1}{2} \int \mathrm{d}\vec{r} \int \mathrm{d}\vec{r}' \, \frac{e^2 n(\vec{r}) n(\vec{r}') \, [1 - g(\vec{r}, \vec{r}')]}{4\pi\varepsilon \, |\vec{r} - \vec{r}'|}$$

where g is a correlation function that accounts for the fact that the probability of finding two electrons simultaneously at $\vec{r}$ and $\vec{r}'$ is not just proportional to $n(\vec{r})n(\vec{r}')$, but is somewhat reduced because electrons try to avoid each other (actually this correlation factor is spin-dependent, but we are ignoring such details). The corresponding self-consistent potential is also reduced (cf. Eq. (3.2.10)):

$$U_{\mathrm{SCF}} = \int \mathrm{d}\vec{r}' \; \frac{e^2 n(\vec{r}') \, [1 - g(\vec{r}, \vec{r}')]}{4\pi\varepsilon \, |\vec{r} - \vec{r}'|} \tag{3.2.12}$$

Much research has gone into estimating the function $g(\vec{r}, \vec{r}')$ (generally referred to as the exchange-correlation "hole").

The basic effect of the correlation energy is to add a *negative* term $U_{\mathrm{xc}}(\vec{r})$ to the Hartree term $U_{\mathrm{H}}(\vec{r})$ discussed above (cf. Eq. (3.2.10)):

$$U_{\mathrm{SCF}}(\vec{r}) = U_{\mathrm{H}}(\vec{r}) + U_{\mathrm{xc}}(\vec{r}) \tag{3.2.13}$$

One simple approximation, called the local density approximation (LDA) expresses U_{xc} at a point in terms of the electron density at that point:

$$U_{\mathrm{xc}}(\vec{r}) \; = -\frac{q^2}{4\pi\varepsilon_0} \, C[n(\vec{r})]^{\,1/3} \tag{3.2.14}$$

Here, C is a constant of order one. The physical basis for this approximation is that an individual electron introduced into a medium with a background electron density $n(r)$ will push other electrons in its neighborhood, creating a positive correlation "hole" around it. If we model this hole as a positive sphere of radius r_0 then we can estimate r_0 by requiring that the total charge within the sphere be equal in magnitude to that of an electron:

$$n(r) \, 4\pi r_0^3/3 = 1 \rightarrow r_0 = \frac{1}{C} \, [n(r)]^{-1/3}$$

C being a constant of order one. The potential in Eq. (3.2.14) can be viewed as the potential at the center of this positive charge contained in a sphere of radius r_0:

$$U_{\mathrm{xc}}(\vec{r}) = -\frac{q^2}{4\pi\varepsilon_0 \, r_0}$$

Much work has gone into the SCF theory and many sophisticated versions of Eq. (3.2.14) have been developed over the years. But it is really quite surprising that the one-electron picture with a suitable SCF often provides a reasonably accurate description of multi-electron systems. The fact that it works so well is not something that can be proved mathematically in any convincing way. Our confidence in the SCF method stems from the excellent agreement that has been obtained with experiment for virtually every atom in the periodic table (see Fig. 3.2.4). Almost all the work on the theory of electronic structure of atoms, molecules, and solids is based on this method and that is what we will be using.

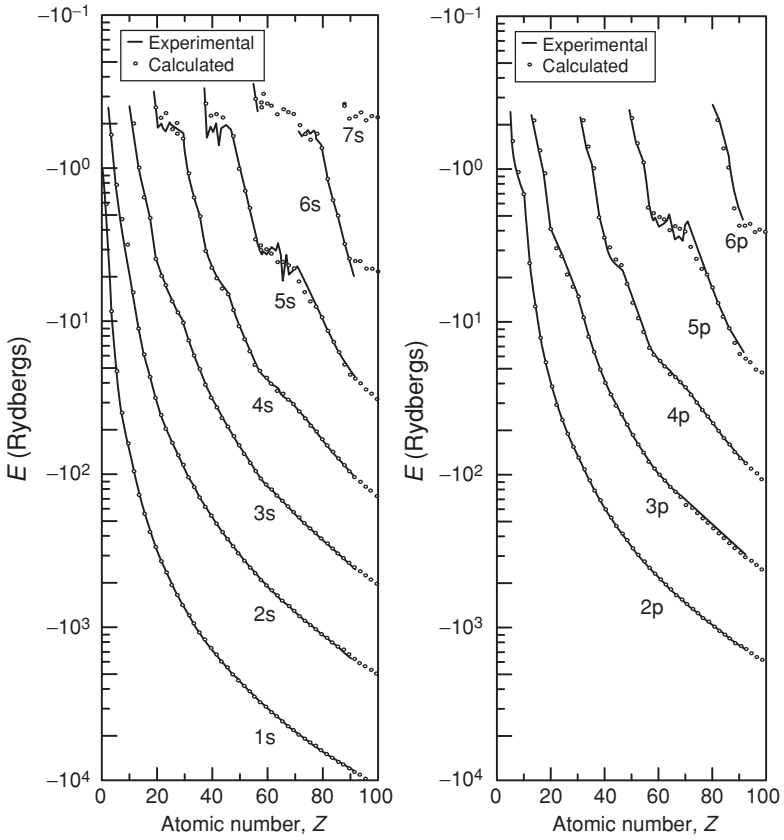


Fig. 3.2.4 Energy levels as a function of the atomic number calculated theoretically using a self-consistent field method. The results are in excellent agreement with experiment (adapted from Herman and Skillman (1963)). For a hydrogen atom, the s and p levels are degenerate (that is, they have the same energy). This is a consequence of the $\sim 1/r$ dependence of the nuclear potential. But this is not true of the self-consistent potential due to the electrons and, for multi-electron atoms, the s state has a lower energy than the p state.

3.3 Bonding

One of the first successes of quantum theory was to explain the structure of the periodic table of atoms by combining the energy levels obtained from the Schrödinger equation with the Pauli exclusion principle requiring that each level be occupied by no more than one electron. In Section 3.3.1 we will discuss the general trends, especially the periodic character of the energy levels of individual atoms. We will then discuss two bonding mechanisms (ionic (Section 3.3.2) and covalent (Section 3.3.3)) whereby a pair of atoms, A and B, can lower their overall energy by forming a molecule AB: $E(\mathrm{AB}) < E(\mathrm{A}) + E(\mathrm{B})$.

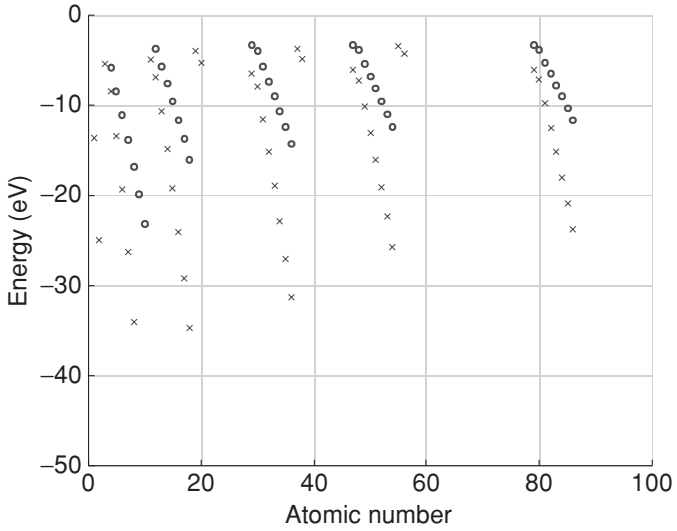


Fig. 3.3.1 Energy of the outermost s (×) and p levels (○) of the first 86 elements of the periodic table excluding the d- and f-shell transition metals ($Z = 21$–28, 39–46, and 57–78). The numbers are taken from Harrison (1999) and Mann (1967).

3.3.1 Valence electrons

It is important to note that only the electrons in the outermost shell, referred to as the valence electrons, participate in the bonding process. The energies of these valence electrons exhibit a periodic variation as shown in Fig. 3.3.1 for the first 86 atoms of the periodic table from hydrogen (atomic number $Z = 1$) to radon ($Z = 86$), excluding the d- and f-shell transition metals (see Table 3.3.1). The main point to notice is that the energies tend to go down as we go across a row of the periodic table from lithium (Li) to neon (Ne), increase abruptly as we step into the next row with sodium (Na) and then decrease as we go down the row to argon (Ar). This trend is shown by both the s and p levels and continues onto the higher rows. Indeed this periodic variation in the energy levels is at the heart of the periodic table of the elements.

3.3.2 Ionic bonds

Ionic bonds are typically formed between an atom to the left of the periodic table (like sodium, Na) and one on the right of the periodic table (like chlorine, Cl). The energy levels of Na and Cl look roughly as shown in Fig. 3.3.2. It seems natural for the 3s electron from Na to "spill over" into the 3p levels of Cl, thereby lowering the overall energy as shown. Indeed it seems "obvious" that the binding energy, E_{bin}, of NaCl would be

$$E_{\text{bin}} = E(\text{Na}) + E(\text{Cl}) - E(\text{Na}^{+}\text{Cl}^{-}) = 12.3 - 5.1 = 7.2\,\text{eV}.$$

Table 3.3.1 *First 86 atoms of the periodic table from hydrogen (atomic number Z = 1) to radon (Z = 86), excluding the d- and f-shell transition metals (Z = 21–28, 39–46, and 57–78)*

H										He
(Z=1)										(Z=2)
Li	Be				B	C	N	O	F	Ne
(Z=3)	(Z=4)				(Z=5)	(Z=6)	(Z=7)	(Z=8)	(Z=9)	(Z=10)
Na	Mg				Al	Si	P	S	Cl	Ar
(Z=11)	(Z=12)				(Z=13)	(Z=14)	(Z=15)	(Z=16)	(Z=17)	(Z=18)
K	Ca	...	Cu	Zn	Ga	Ge	As	Se	Br	Kr
(Z=19)	(Z=20)	...	(Z=29)	(Z=30)	(Z=31)	(Z=32)	(Z=33)	(Z=34)	(Z=35)	(Z=36)
Rb	Sr	...	Ag	Zn	In	Sn	Sb	Te	I	Xe
(Z=37)	(Z=38)	...	(Z=47)	(Z=48)	(Z=49)	(Z=50)	(Z=51)	(Z=52)	(Z=53)	(Z=54)
Cs	Ba	...	Au	Hg	Tl	Pb	B	Po	At	Rn
(Z=55)	(Z=56)	...	(Z=79)	(Z=80)	(Z=81)	(Z=82)	(Z=83)	(Z=84)	(Z=85)	(Z=86)

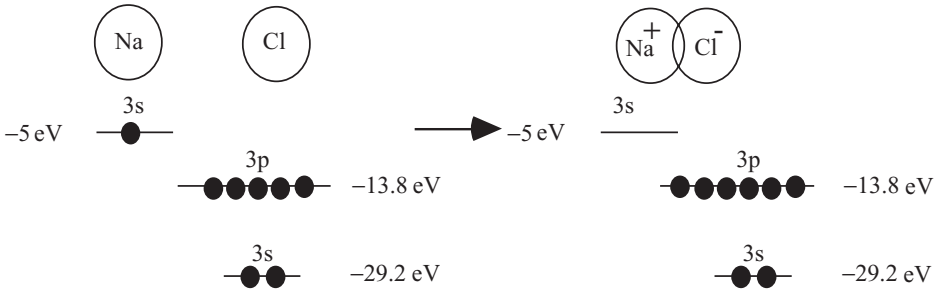


Fig. 3.3.2 Formation of Na^+Cl^- from individual Na and Cl atoms with a 3s electron from Na "spilling over" into the 3p levels of Cl thereby lowering the overall energy. This is only part of the story, since the overall energetics also includes the electrostatic energy stored in the microscopic capacitor formed by the two ions as explained in the text.

But this argument is incomplete because we also need to consider the change in the electrostatic energy due to the bonding. The correct binding energy is more like 4 eV.

The point is that the energy levels we have drawn here are all *ionization* levels. The energy needed to create a sodium ion is given by its ionization potential (IP)

$$E(\mathrm{Na}^+) - E(\mathrm{Na}) = \mathrm{IP(Na)} = 5\,\mathrm{eV} \tag{3.3.1a}$$

But the energy needed to create a chlorine ion is given by the *electron affinity* (EA) of Cl and this includes an extra charging energy U_0:

$$E(\mathrm{Cl}) - E(\mathrm{Cl}^-) = \mathrm{EA(Cl)} = \mathrm{IP(Cl)} - U_0 = 13.8\,\mathrm{eV} - U_0 \tag{3.3.1b}$$

Combining Eqs. (3.3.1a) and (3.3.1b) we obtain

$$E(\mathrm{Na}) + E(\mathrm{Cl}) - E(\mathrm{Na}^+) - E(\mathrm{Cl}^-) = 8.8\,\mathrm{eV} - U_0 \tag{3.3.2}$$

However, this is not the binding energy of NaCl. It gives us the energy gained in converting neutral Na and neutral Cl into a Na^+ and a Cl^- ion completely separated from each other. If we let a Na^+ and a Cl^- ion that are infinitely far apart come together to form a sodium chloride molecule, Na^+Cl^-, it will gain an energy U_0' in the process.

$$E(\mathrm{Na}^+) + E(\mathrm{Cl}^-) - E(\mathrm{Na}^+\mathrm{Cl}^-) = U_0'$$

so that the binding energy is given by

$$E_{\mathrm{bin}} = E(\mathrm{Na}) + E(\mathrm{Cl}) - E(\mathrm{Na}^+\mathrm{Cl}^-) = 8.8\,\mathrm{eV} - U_0 + U_0' \tag{3.3.3}$$

$U_0 - U_0'$ is approximately 5 eV, giving a binding energy of around 4 eV. The numerical details of this specific problem are not particularly important or even accurate. The main point I wish to make is that although the process of bonding by electron transfer may seem like a simple one where one electron "drops" off an atom into another with

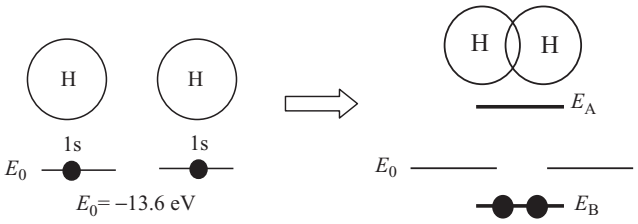


Fig. 3.3.3 Formation of H_2 from individual H atoms with a bonding level E_B and an anti-bonding level E_A.

a lower energy level, the detailed energetics of the process require a more careful discussion. In general, care is needed when using one-electron energy level diagrams to discuss electron transfer on an atomic scale.

3.3.3 Covalent bonds

We have just seen how a lowering of energy comes about when we bring together an atom from the left of the periodic table (like sodium) and one from the right (like chlorine). The atoms on the right of the periodic table have lower electronic energy levels and are said to be more electronegative than those on the left. We would expect electrons to transfer from the higher energy levels in the former to the lower energy levels in the latter to form an ionic bond.

However, this argument does not explain covalent bonds which involve atoms with roughly the same electronegativity. The process is a little more subtle. For example, it is hard to see why two identical hydrogen atoms would want to form a H_2 molecule, since no lowering of energy is achieved by transferring an electron from one atom to the other. What happens is that when the two atoms come close together the resulting energy levels split into a bonding level (E_B) and an anti-bonding level (E_A) as shown in Fig. 3.3.3. Both electrons occupy the bonding level which has an energy lower than that of an isolated hydrogen atom: $E_B < E_0$.

How do we calculate E_B? By solving the Schrödinger equation:

$$E_\alpha \Phi_\alpha(\vec{r}) = \left(-\frac{\hbar^2}{2m} \nabla^2 + U_N(\vec{r}) + U_{N'}(\vec{r}) + U_{SCF}(\vec{r}) \right) \Phi_\alpha(\vec{r}) \tag{3.3.4}$$

where $U_N(r)$ and $U_{N'}(r)$ are the potentials due to the left and the right nuclei respectively and $U_{SCF}(r)$ is the potential that one electron feels due to the other. To keep things simple let us ignore $U_{SCF}(r)$ and calculate the electronic energy levels due to the nuclear potentials alone:

$$E_{\alpha 0} \Phi_{\alpha 0}(\vec{r}) = \left(-\frac{\hbar^2}{2m} \nabla^2 + U_N(\vec{r}) + U_{N'}(\vec{r}) \right) \Phi_{\alpha 0}(\vec{r}) \tag{3.3.5}$$

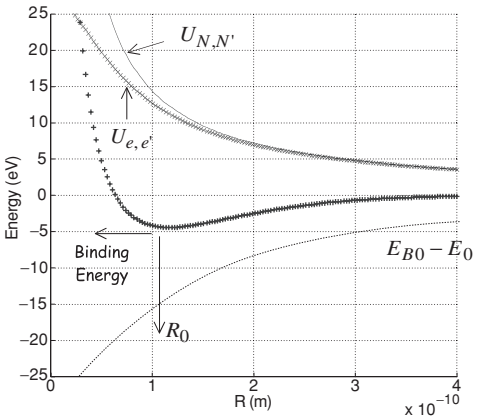


Fig. 3.3.4 Various energies as a function of the nuclear distance R. $\times\times\times$, approximate electron–electron repulsive energy ($U_{e,e'}$). Solid curve, nucleus–nucleus repulsive energy ($U_{N,N'}$). Dashed curve, $E_{B0}-E_0$; energy of the bonding level in a H_2 molecule relative to the 1s level in a hydrogen atom calculated approximately from the Schrödinger equation without any self-consistent potential. ++++, binding energy of a H_2 molecule relative to two hydrogen atoms estimated from $2(E_{B0}-E_0) + U_{N,N'} + U_{e,e'}$.

The lowest energy solution to Eq. (3.3.5) can be written approximately as

$$E_{B0} = E_0 + \frac{a+b}{1+s} \tag{3.3.6}$$

where

$$a = -2E_0 \frac{1-(1+\overline{R})e^{-2\overline{R}}}{\overline{R}}$$

$$b = -2E_0(1+\overline{R})\,e^{-\overline{R}}$$

$$s = e^{-\overline{R}}[1+\overline{R}+(\overline{R}^2/3)]$$

$$\overline{R} \equiv R/a_0$$

R being the center-to-center distance between the hydrogen atoms.

Let us now try to understand the competing forces that lead to covalent bonding. The dashed curve in Fig. 3.3.4 shows $E_{B0} - E_0$ versus the bond length R as given by Eq. (3.3.6). Experimentally, the bond length R for a H_2 molecule is 0.074 nm, indicating that the overall energy is a minimum for this value of R. Since the energy keeps decreasing as R is decreased, one might wonder why the two hydrogen atoms do not just sit on top of each other ($R = 0$). To answer this question we need to calculate the overall energy which should include the electron–electron repulsion (note that $U_{SCF}(r)$

was left out from Eq. (3.3.6)) as well as the nucleus–nucleus repulsion. To understand the overall energetics let us consider the difference in energy between a hydrogen molecule (H_2) and two isolated hydrogen atoms (2H).

The energy required to assemble two separate hydrogen atoms from two protons (N, N′) and two electrons (e, e′) can be written as

$$E(2\mathrm{H}) = U_{\mathrm{e,N}} + U_{\mathrm{e'},\mathrm{N'}} = 2E_0 \tag{3.3.7a}$$

The energy required to assemble an H_2 molecule from two protons (N, N′) and two electrons (e, e′) can be written as

$$E(\mathrm{H}_2) = U_{\mathrm{N,N'}} + U_{\mathrm{e,e'}} + U_{\mathrm{e,N}} + U_{\mathrm{e,N'}} + U_{\mathrm{e'},\mathrm{N}} + U_{\mathrm{e'},\mathrm{N'}} \tag{3.3.7b}$$

Equation (3.3.6) gives the quantum mechanical value of $(U_{\mathrm{e,N}} + U_{\mathrm{e,N'}})$ as well as $(U_{\mathrm{e'},\mathrm{N}} + U_{\mathrm{e'},\mathrm{N'}})$ as E_{B0}. Hence

$$E(\mathrm{H}_2) = U_{\mathrm{N,N'}} + U_{\mathrm{e,e'}} + 2E_{\mathrm{B0}} \tag{3.3.7c}$$

The binding energy is the energy it takes to make the hydrogen molecule dissociate into two hydrogen atoms and can be written as

$$E_{\mathrm{bin}} = E(\mathrm{H}_2) - E(2\mathrm{H}) = 2(E_{\mathrm{B0}} - E_0) + U_{\mathrm{N,N'}} + U_{\mathrm{e,e'}} \tag{3.3.8}$$

This is the quantity that ought to be a minimum at equilibrium and it consists of three separate terms. Eq. (3.3.6) gives us only the first term. The second term is easily written down since it is the electrostatic energy between the two nuclei, which are point charges:

$$U_{\mathrm{N,N'}} = q^2/4\pi\varepsilon_0 R \tag{3.3.9a}$$

The electrostatic interaction between the two electrons should also look like $q^2/4\pi\varepsilon_0 R$ for large R, but should saturate to $\sim q^2/4\pi\varepsilon_0 a_0$ at short distances since the electronic charges are diffused over distances $\sim a_0$. Let us approximate it as

$$U_{\mathrm{e,e'}} \cong q^2/4\pi\varepsilon_0\sqrt{R^2 + a_0^2} \tag{3.3.9b}$$

noting that this is just an oversimplified approximation to what is in general a very difficult quantum mechanical problem – indeed, electron–electron interactions represent the central outstanding problem in the quantum theory of matter.

The solid curve in Fig. 3.3.4 shows $U_{\mathrm{N,N'}}$ (Eq. (3.3.9a)), while the ××× curve shows $U_{\mathrm{e,e'}}$ (Eq. (3.3.9b)). The +++ curve shows the total binding energy estimated from Eq. (3.3.8). It has a minimum around 0.1 nm, which is not too far from the experimental bond length of 0.074 nm. Also the binding energy at this minimum is ~4.5 eV, very close to the actual experimental value. Despite the crudeness of the approximations used, the basic physics of bonding is illustrated fairly well by this example.

Fig. 3.3.5 A hydrogen molecule can be viewed as two masses connected by a spring.

Vibrational frequency: The shape of the binding energy vs. R curve suggests that we can visualize a hydrogen molecule as two masses connected by a spring (Fig. 3.3.5). An ideal spring with a spring constant K has a potential energy of the form $U(R) = K(R - R_0)^2/2$. The binding energy of the hydrogen molecule (see Fig. 3.3.4) can be approximated as $U(R) \cong U(R_0) + K(R - R_0)^2/2$, where the effective spring constant K is estimated from the curvature $[\mathrm{d}^2U/\mathrm{d}R^2]_{R=R_0}$. Indeed the vibrational frequency of the H–H bond can be estimated well from the resonant frequency $\sqrt{2K/M}$ of the mass and spring system where M is the mass of a hydrogen atom.

Ionization levels: As we have discussed, the energy levels of a multi-electron system usually denote the ionization levels, that is the energy it takes to strip an electron from the system. This means that in the present context the energy level E_B for a hydrogen molecule should represent

$$E_\mathrm{B} = E(\mathrm{H}_2) - E(\mathrm{H}_2^+)$$

Since $E(\mathrm{H}_2^+) = U_{\mathrm{N,N'}} + U_{\mathrm{e',N}} + U_{\mathrm{e',N'}}$, we can write using Eq. (3.3.7b),

$$E_\mathrm{B} = U_{\mathrm{e,e'}} + U_{\mathrm{e,N}} + U_{\mathrm{e,N'}} = U_{\mathrm{e,e'}} + E_{\mathrm{B0}} \tag{3.3.10}$$

It is easy to check that for our model calculation (see Fig. 3.3.4) E_{B0} is nearly 15 eV below E_0, but E_B lies only about 4 eV below E_0. If we were to include a self-consistent field $U_{\mathrm{SCF}}(r)$ in the Schrödinger equation, we would obtain the energy E_B which would be higher (less negative) than the non-interacting value of E_{B0} by the electron–electron interaction energy $U_{\mathrm{e,e'}}$.

Binding energy: It is tempting to think that the binding energy is given by

$$E_{\mathrm{bin}} = 2(E_\mathrm{B} - E_0) + U_{\mathrm{N,N'}}$$

since E_B includes the electron–electron interaction energy $U_{\mathrm{e,e'}}$. However, it is easy to see from Eqs. (3.3.8) and (3.3.10) that the correct expression is

$$E_{\mathrm{bin}} = 2(E_\mathrm{B} - E_0) + (U_{\mathrm{N,N'}} - U_{\mathrm{e,e'}})$$

The point I am trying to make is that if we include the electron–electron interaction in our calculation of the energy level E_B then the overall energy of two electrons is NOT $2E_\mathrm{B}$, for that would double-count the interaction energy between the two

electrons. The correct energy is obtained by subtracting off this double-counted part: $2E_B - U_{e,e'}$.

3.4 Supplementary notes: multi-electron picture

As I mentioned in Section 3.2, the SCF method is widely used because the exact method based on a multi-electron picture is usually impossible to implement. However, it is possible to solve the multi-electron problem exactly if we are dealing with a small channel weakly coupled to its surroundings, like the one-level system discussed in Section 1.4. It is instructive to recalculate this one-level problem in the multi-electron picture and compare with the results obtained from the SCF method.

One-electron vs. multi-electron energy levels: If we have one spin-degenerate level with energy ε, the one-electron and multi-electron energy levels would look as shown in Fig. 3.4.1. Since each one-electron energy level can either be empty (0) or occupied (1), multi-electron states can be labeled in the form of binary numbers with a number of digits equal to the number of one-particle states. N one-electron states thus give rise to 2^N multi-electron states, which quickly diverges as N increases, making a direct treatment impractical. That is why SCF methods are so widely used, even though they are only approximate.

Consider a system with two degenerate one-electron states (up-spin and down-spin) that can either be filled or empty. All other one-electron states are assumed not to change their occupation: those below remain filled while those above remain empty. Let us assume that the electron–electron interaction energy is given by

$$U_{ee}(N) = (U_0/2)N(N-1) \qquad \text{(same as Eq. (3.2.4))}$$

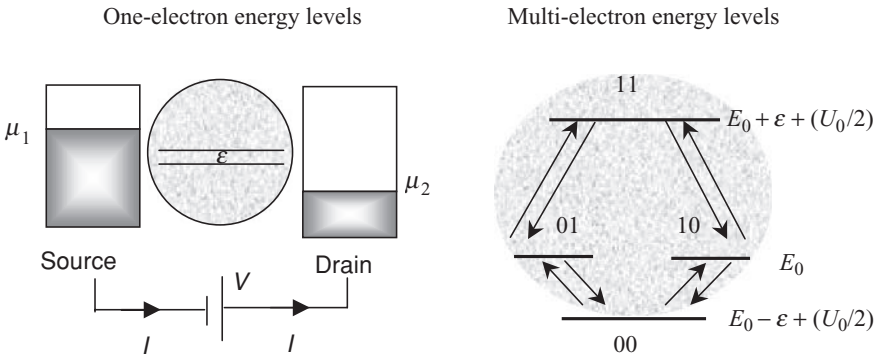


Fig. 3.4.1 One-electron vs. multi-electron energy levels in a channel with one spin-degenerate level having energy ε.

corresponding to a self-consistent potential (see Eq. (3.2.6))

$$\partial U_{ee}/\partial N = U_0 N - (U_0/2)$$

Suppose the number of electrons N_0 in the neutral state corresponds to having one of these states filled. The one-electron energy levels ε can be written as the sum of the "bare" levels $\tilde{\varepsilon}$ (obtained from a Schrödinger equation with just the nuclear potential, U_N) plus the self-consistent potential $[\partial U_{ee}/\partial N]_{N=N_0}$:

$$\varepsilon = \tilde{\varepsilon} \; + \; [\partial U_{ee}/\partial N]_{N=N_0} = \tilde{\varepsilon} + U_0 N_0 - (U_0/2)$$

Consider now the multi-electron picture. We have four available multi-electron states which we can designate as 00, 01, 10, and 11. In the neutral state, the system is in either the (10) or the (01) state whose total energy we denote as

$$E(10) = E(01) \equiv E_0$$

We can write the energies of the other multi-electron states as

$$\begin{aligned} E(11) &= E_0 + \tilde{\varepsilon} + U_{ee}(N_0 + 1) - U_{ee}(N_0) \\ &= E_0 + \tilde{\varepsilon} + U_0 N_0 = E_0 + \varepsilon + (U_0/2) \end{aligned}$$

and

$$\begin{aligned} E(00) &= E_0 - \tilde{\varepsilon} - U_{ee}(N_0) + U_{ee}(N_0 - 1) \\ &= E_0 - \tilde{\varepsilon} - U_0(N_0 - 1) = E_0 - \varepsilon + (U_0/2) \end{aligned}$$

Master equation: In the multi-electron picture, the overall system has different probabilities P_α of being in one of the 2^N possible states α and all the probabilities must add up to one:

$$\sum_\alpha P_\alpha = 1 \rightarrow P_{00} + P_{01} + P_{10} + P_{11} = 1 \tag{3.4.1}$$

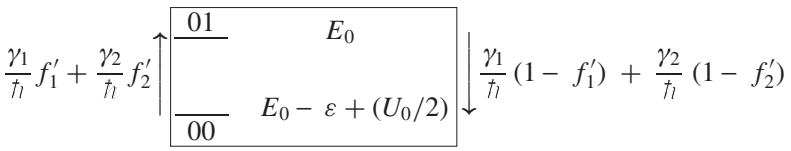
We can calculate the individual probabilities by noting that the system is continually shuffled among these states and under steady-state conditions there must be no net flow into or out of any state:

$$\sum_\beta R\,(\alpha \rightarrow \beta)\, P_\alpha = \sum_\beta R\,(\beta \rightarrow \alpha) P_\beta \tag{3.4.2}$$

Knowing the rate constants, we can calculate the probabilities by solving Eq. (3.4.2). Equations involving probabilities of different states are called master equations. We could call Eq. (3.4.2) a multi-electron master equation.

The rate constants $R(\alpha \rightarrow \beta)$ can be written down assuming a specific model for the interaction with the surroundings. For example, if we assume that the interaction only involves the entry and exit of individual electrons from the source and drain contacts

then for the 00 and 01 states the rate constants are given by

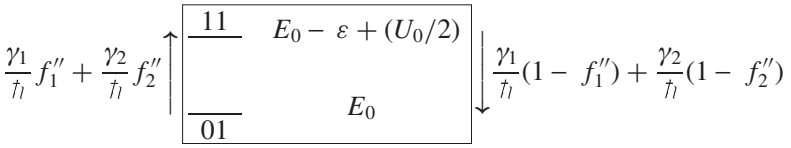


where

$$f_1' \equiv f_0(\varepsilon_1 - \mu_1) \quad \text{and} \quad f_2' \equiv f_0(\varepsilon_1 - \mu_2)$$

tell us the availability of electrons with energy $\varepsilon_1 = \varepsilon - (U_0/2)$ in the source and drain contacts respectively. The entry rate is proportional to the available electrons, while the exit rate is proportional to the available empty states. The same picture applies to the flow between the 00 and the 10 states, assuming that up- and down-spin states are described by the same Fermi function in the contacts, as we would expect if each contact is locally in equilibrium.

Similarly we can write the rate constants for the flow between the 01 and the 11 states



where

$$f_1'' \equiv f_0(\varepsilon_2 - \mu_1) \quad \text{and} \quad f_2'' \equiv f_0(\varepsilon_2 - \mu_2)$$

tell us the availability of electrons with energy $\varepsilon_2 = \varepsilon + (U_0/2)$ in the source and drain contacts corresponding to the energy difference between the 01 and 11 states. This is larger than the energy difference ε between the 00 and 01 states because it takes more energy to add an electron when one electron is already present due to the interaction energy U_0.

Using these rate constants it is straightforward to show from Eq. (3.4.2) that

$$\frac{P_{10}}{P_{00}} = \frac{P_{01}}{P_{00}} = \frac{\gamma_1 f_1' + \gamma_2 f_2'}{\gamma_1(1 - f_1') + \gamma_2(1 - f_2')} \tag{3.4.3a}$$

and

$$\frac{P_{11}}{P_{10}} = \frac{P_{11}}{P_{01}} = \frac{\gamma_1 f_1'' + \gamma_2 f_2''}{\gamma_1(1 - f_1'') + \gamma_2(1 - f_2'')} \tag{3.4.3b}$$

Together with Eq. (3.4.1), this gives us all the individual probabilities. Figure 3.4.2 shows the evolution of these probabilities as the gate voltage V_G is increased

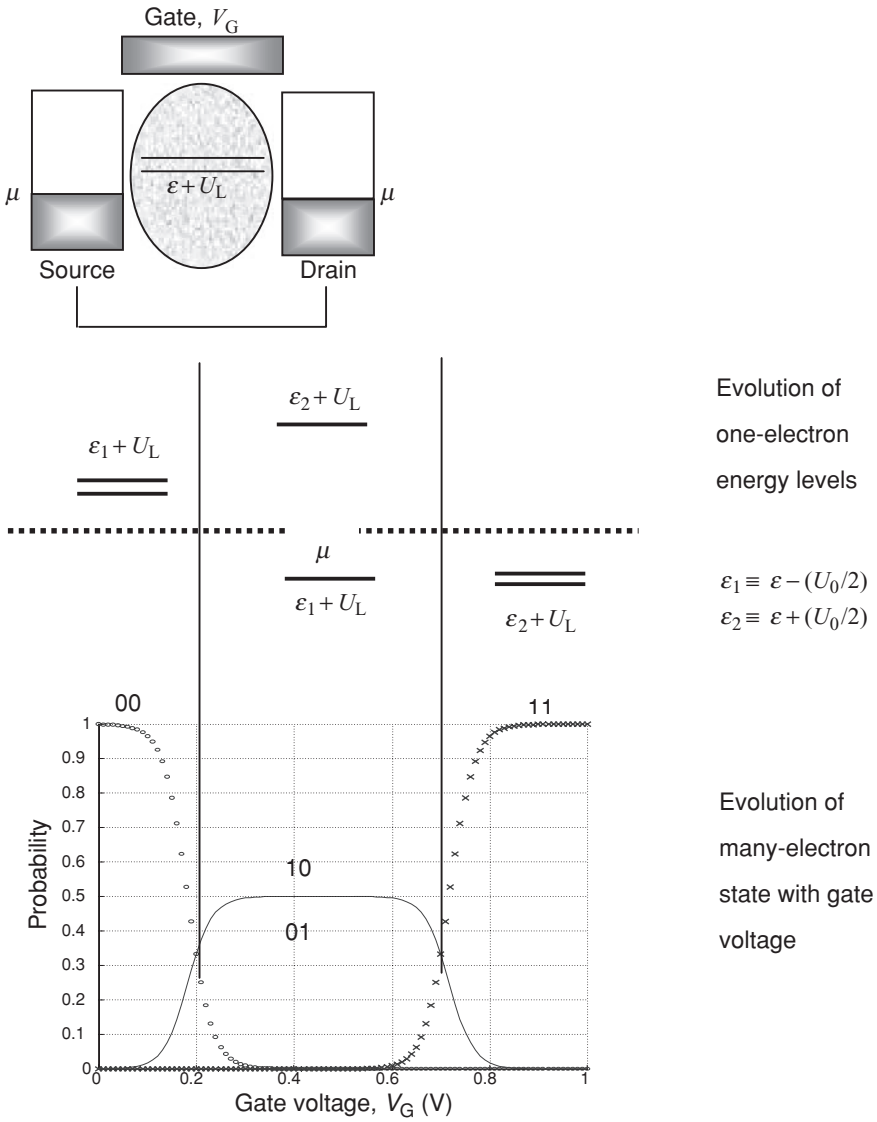


Fig. 3.4.2 Evolution of the energy levels of a channel with one spin-degenerate level as the gate voltage V_G is made more positive, holding the drain voltage V_D equal to zero. $\mu = 0$, $\varepsilon = 0.2$ eV, $k_B T = 0.025$ eV, $U_0 = 0.25$ eV, $U_L = -qV_G$. Lower plot shows the probabilities of finding the channel in one of its four states: P_{00} ($\circ$), $P_{01} = P_{10}$ (solid) and P_{11} ($\times$).

holding the drain voltage V_D equal to zero. The gate voltage shifts the one-electron level $\varepsilon \to \varepsilon + U_L$ (we have assumed $U_L = -qV_G$) and the probabilities are calculated from Eqs. (3.4.3a, b) and (3.4.1) noting that the Fermi functions are given by

$$f_1' = f_0(\varepsilon_1 + U_L - \mu_1),\ f_2' = f_0(\varepsilon_1 + U_L - \mu_2) \tag{3.4.4a}$$

$$f_1'' = f_0(\varepsilon_2 + U_L - \mu_1),\ f_2'' = f_0(\varepsilon_2 + U_L - \mu_2) \tag{3.4.4b}$$

The system starts out in the 00 state ($P_{00} = 1$), shifts to the 01 and 10 states ($P_{01} = P_{10} = 0.5$) once $\varepsilon_1 + U_L$ drops below μ, and finally goes into the 11 state ($P_{11} = 1$) when $\varepsilon_2 + U_L$ drops below μ.

Relation between the multi-electron picture and the one-electron levels: As I have emphasized in Section 3.2, one-electron energy levels represent *differences* between energy levels in the multi-electron picture corresponding to states that differ by *one electron*. Transitions involving the addition of one electron are called *affinity* levels while those corresponding to the removal of one electron are called *ionization* levels. For example (see Fig. 3.4.2), if the system is in the 00 state then there are two degenerate one-electron levels $\varepsilon_1 + U_L$ representing

$$\varepsilon_1 + U_L = E(10) - E(00) = E(01) - E(00) \qquad \textit{Affinity levels}$$

Once it is in the 10 state there are two one-electron levels

$$\varepsilon_1 + U_L = E(10) - E(00) \qquad \textit{Ionization level}$$

and $$\varepsilon_2 + U_L = E(11) - E(10) \qquad \textit{Affinity level}$$

In the 11 state there are two degenerate one-electron levels

$$\varepsilon_2 + U_L = E(11) - E(10) = E(11) - E(01) \qquad \textit{Ionization levels}$$

Affinity levels lie above μ, while ionization levels lie below μ as shown in Fig. 3.4.2. This is a very important general concept regarding the interpretation of the one-electron energy levels when dealing with complicated interacting objects. The occupied (or ionization) levels tell us the energy levels for removing an electron while the unoccupied (or affinity) levels tell us the energy levels for adding an extra electron. Indeed that is exactly how these levels are measured experimentally, the occupied levels by photoemission (PE) and the unoccupied levels by inverse photoemission (IPE) as mentioned in Section 1.1.

Law of equilibrium: Figure 3.4.2 represents an equilibrium calculation with both source and drain contacts having the same Fermi function: $f_1 = f_2$. Equilibrium problems do not really require the use of a master equation like Eq. (3.4.2). We can use the general principle of equilibrium statistical mechanics which states that the probability P_α that the system is in a multi-electron state α with energy E_α and N_α electrons is given by

$$P_\alpha = \frac{1}{Z}\exp[-(E_\alpha - \mu N_\alpha)/k_B T] \tag{3.4.5}$$

where the constant Z (called the partition function) is determined so as to ensure that the probabilities given by Eq. (3.4.5) for all states α add up to one:

$$Z = \sum_\alpha \exp[-(E_\alpha - \mu N_\alpha)/k_B T] \tag{3.4.6}$$

This is the central law of equilibrium statistical mechanics that is applicable to any system of particles (electrons, photons, atoms, etc.), interacting or otherwise (see for example, Chapter 1 of Feynman, 1972). The Fermi function is just a special case of this general relation that can be obtained by applying it to a system with just a single one-electron energy level, corresponding to two multi-electron states:

α	N_α	E_α	P_α
0	0	0	$1/Z$
1	1	ε	$(1/Z)\exp[(\mu-\varepsilon)/k_\mathrm{B}T]$

so that $Z = 1 + \exp[(\mu - \varepsilon)/k_\mathrm{B}T]$ and it is straightforward to show that the average number of electrons is equal to the Fermi function (Eq. (1.1.1)):

$$N = \sum_\alpha N_\alpha P_\alpha = P_1 = \frac{\exp[(\mu-\varepsilon)/k_\mathrm{B}T]}{1+\exp[(\mu-\varepsilon)/k_\mathrm{B}T]} = f_0\,[\varepsilon - \mu]$$

For multi-electron systems, we can use the Fermi function only if the electrons are not interacting. It is then justifiable to single out one level and treat it independently, ignoring the occupation of the other levels. The SCF method uses the Fermi function assuming that the energy of each level depends on the occupation of the other levels. But this is only approximate. The exact method is to abandon the Fermi function altogether and use Eq. (3.4.5) instead to calculate the probabilities of the different multi-particle states.

One well-known example of this is the fact that localized donor or acceptor levels (which have large charging energies U_0) in semiconductors at equilibrium are occupied according to a modified Fermi function (ν is the level degeneracy)

$$f = \frac{1}{1+(1/\nu)\exp\,[(\varepsilon-\mu)/k_\mathrm{B}T]} \tag{3.4.7}$$

rather than the standard Fermi function (cf. Eq. (1.1.1)). We can easily derive this relation for two spin-degenerate levels ($\nu = 2$) if we assume that the charging energy U_0 is so large that the 11 state has zero probability. We can then write for the remaining states

α	N_α	E_α	P_α
00	0	0	$1/Z$
01	1	ε	$(1/Z)\exp[(\mu-\varepsilon)/k_\mathrm{B}T]$
10	1	ε	$(1/Z)\exp[(\mu-\varepsilon)/k_\mathrm{B}T]$

so that $Z = 1 + 2\exp[(\mu - \varepsilon)/k_\mathrm{B}T]$ and the average number of electrons is given by

$$N = \sum_\alpha N_\alpha P_\alpha = P_{01} + P_{10} = \frac{2\exp[(\mu-\varepsilon)/k_\mathrm{B}T]}{1+2\exp[(\mu-\varepsilon)/k_\mathrm{B}T]}$$
$$= \frac{1}{1+(1/2)\exp[(\varepsilon-\mu)/k_\mathrm{B}T]}$$

in agreement with Eq. (3.4.7). This result, known to every device engineer, could thus be viewed as a special case of the general result in Eq. (3.4.5).

Equation (3.4.5), however, can only be used to treat equilibrium problems. Our primary interest is in calculating the current under non-equilibrium conditions and that is one reason we have emphasized the master equation approach based on Eq. (3.4.2). For equilibrium problems, it gives the same answer. However, it also helps to bring out an important conceptual point. One often hears concerns that the law of equilibrium is a statistical one that can only be applied to large systems. But it is apparent from the master equation approach that the law of equilibrium (Eq. (3.4.5)) is not a property of the system. It is a property of the contacts or the "reservoir." The only assumptions we have made relate to the energy distribution of the electrons that come in from the contacts. As long as these "reservoirs" are simple, it does not matter how complicated or how small the "system" is.

Current calculation: Getting back to non-equilibrium problems, once we have solved the master equation for the individual probabilities, the source current can be obtained from

$$I_1 = -q \sum_{\beta} (\pm) R_1(\alpha \to \beta)\, P_\alpha$$

$+$ if β has one more electron than α
$-$ if β has one less electron than α

where R_1 represents the part of the total transition rate R associated with the source contact. In our present problem this reduces to evaluating the expression

$$\begin{aligned} I_1 = (-q/\hbar)\,(&2\gamma_1 f_1' P_{00} - \gamma_1(1-f_1')(P_{01}+P_{10}) \\ &+ \gamma_1 f_1''(P_{01}+P_{10}) - 2\gamma_1(1-f_1'')P_{11}) \end{aligned} \tag{3.4.8}$$

Figure 3.4.3 shows the current–drain voltage (I–V_D) characteristics calculated from the approach just described. The result is compared with a calculation based on the restricted SCF method described in Section 1.4. The SCF current–voltage characteristics look different from Fig. 1.4.6a because the self-consistent potential $U_0(N - N_0)$ has $N_0 = 1$ rather than zero and we have now included two spins. The two approaches agree well for $U_0 = 0.025$ eV, but differ appreciably for $U_0 = 0.25$ eV, showing evidence for Coulomb blockade or single-electron charging (see Exercise E.3.6).

The multi-electron master equation provides a suitable framework for the analysis of current flow in the Coulomb blockade regime where the single-electron charging energy U_0 is well in excess of the level broadening $\gamma_{1,2}$ and/or the thermal energy $k_B T$. We cannot use this method more generally for two reasons. Firstly, the

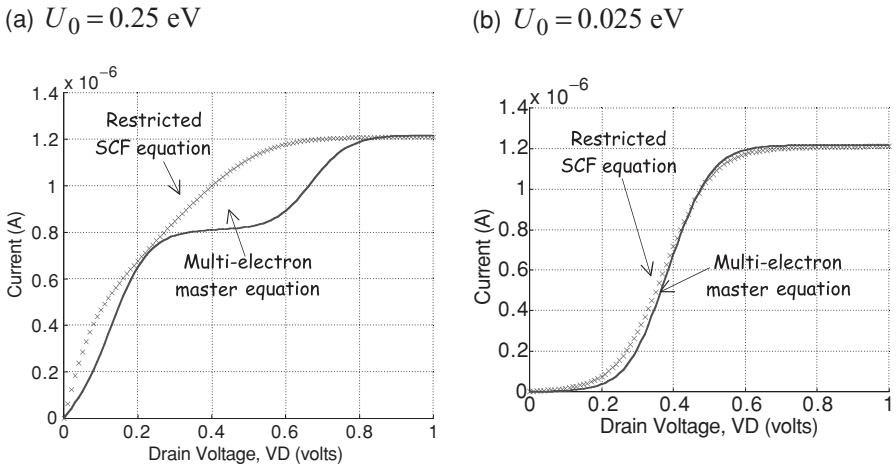


Fig. 3.4.3 Current vs. drain voltage V_D calculated assuming $V_G = 0$ with $\mu = 0$, $\varepsilon = 0.2$ eV, $k_B T = 0.025$ eV, $\gamma_1 = \gamma_2 = 0.005$ eV, $U_L = -qV_D / 2$. The two approaches (the SCF and the multi-electron master equation) agree well for $U_0 = 0.1$ eV, but differ appreciably for $U_0 = 0.25$ eV, showing evidence for Coulomb blockade or single-electron charging.

size of the problem increases exponentially and becomes prohibitive. Secondly, it is not clear how to incorporate broadening into this picture and apply it to the transport regime where the broadening is comparable to the other energy scales. And so it remains a major challenge to provide a proper theoretical description of the intermediate transport regime $U_0 \sim \gamma_{1,2}, k_B T$: the regime where electronic motion is "strongly correlated" making a two-electron probability like $P(11)$ very different from the product of one-electron probabilities like $P(01)P(10)$. A lot of work has gone into trying to discover a suitable SCF within the one-electron picture that will capture the essential physics of correlation. For example, the self-consistent potential $U_{SCF} = U_0 \Delta N$ we have used is the same for all energy levels or orbitals. One could use an "unrestricted" self-consistent field that is orbital-dependent such that the potential felt by level j excludes any self-interaction due to the number of electrons n_j in that level:

$$U_{SCF}(j) = U_0(\Delta N - \Delta n_j) \tag{3.4.9}$$

Such approaches can lead to better agreement with the results from the multi-electron picture but must be carefully evaluated, especially for non-equilibrium problems.

EXERCISES

E.3.1. Use the SCF method (only the Hartree term) to calculate the energy of the 1s level in a helium atom. (a) Plot the nuclear potential $U_N(r)$ and the self-consistent electronic

potential $U_{\mathrm{SCF}}(r)$ (cf. Fig. 3.1.4a). (b) Plot the wavefunction for the 1s level in helium and compare with that for the 1s level in hydrogen (cf. Fig. 3.1.4b).

E.3.2. Use the SCF method (only the Hartree term) to calculate the energies of the 3s and 3p levels in a silicon atom. Plot the wavefunction for the 1s and 3p levels in silicon and compare with that for the 1s level in hydrogen (cf. Fig. 3.1.4b).

E.3.3. Plot the approximate binding energy for a hydrogen molecule as a function of the hydrogen–hydrogen bond length, making use of Eqs. (3.3.6) and (3.3.9a, b) and compare with Fig. 3.3.4.

E.3.4: In Section 1.2 we obtained the following expression for the current through a single level

$$I = \frac{q}{h}\frac{\gamma_1\gamma_2}{\gamma_1+\gamma_2}\left[f_1(\varepsilon) - f_2(\varepsilon)\right]$$

and for the average number of electrons

$$N = \frac{\gamma_1 f_1 + \gamma_2 f_2}{\gamma_1+\gamma_2}$$

by writing a set of rate equations for a single one-electron energy level (without spin degeneracy). In the multi-electron picture we have two levels "0" and "1" corresponding to the one-electron level being empty or full respectively. Write down the appropriate rate equations in this picture and re-derive the expressions for "N" and "I".

E.3.5: Consider a channel with two spin-degenerate levels assuming the following parameters: $\mu = 0$, $\varepsilon = 0.2\,\mathrm{eV}$, $k_{\mathrm{B}}T = 0.025\,\mathrm{eV}$, $\gamma_1 = \gamma_2 = 0.005\,\mathrm{eV}$.

(a) Calculate the number of electrons vs. gate voltage V_{G}, with $V_{\mathrm{D}} = 0$ and $U_{\mathrm{L}} = -qV_{\mathrm{G}}$, using (1) the multi-electron master equation and (2) a restricted SCF potential given by $U_{\mathrm{SCF}} = U_0(N - N_0)$ with $N_0 = 1$. Use two different values of $U_0 = 0.025\,\mathrm{eV}, 0.25\,\mathrm{eV}$.

(b) Calculate the current vs. drain voltage V_{D} assuming $V_{\mathrm{G}} = 0$ with $U_{\mathrm{L}} = -qV_{\mathrm{D}}/2$, using (1) the multi-electron master equation and (2) the restricted SCF potential given in (a). Use two different values of $U_0 = 0.025\,\mathrm{eV}, 0.25\,\mathrm{eV}$ and compare with Fig. 3.4.3.

(c) Repeat (a) and (b) with an unrestricted SCF potential (Eq. (3.4.9)) that excludes the self-interaction:

$$U_{\mathrm{SCF}}(\uparrow) = U_0(\Delta N - \Delta n_\uparrow) = U_0(\Delta n_\downarrow) = U_0(n_\downarrow - 0.5)$$
$$U_{\mathrm{SCF}}(\downarrow) = U_0(n_\uparrow - 0.5)$$

Note: The result may be different depending on whether the initial guess is symmetric, $U_{\mathrm{SCF}}(\uparrow) = U_{\mathrm{SCF}}(\downarrow)$ or not, $U_{\mathrm{SCF}}(\uparrow) \neq U_{\mathrm{SCF}}(\downarrow)$.

E.3.6: In Fig. 3.4.3a ($U_0 = 0.25\,\mathrm{eV}$) the multi-electron approach yields two current plateaus: a lower one with $\varepsilon_2 + U_\mathrm{L} > \mu_1 > \varepsilon_1 + U_\mathrm{L}$ such that $f_1' \simeq 1$, $f_1'' \simeq 0$ and an upper one with $\mu_1 > \varepsilon_2 + U_\mathrm{L} > \varepsilon_1 + U_\mathrm{L}$, such that $f_1' \simeq 1$, $f_1'' \simeq 1$. In either case $f_2' \simeq 0$, $f_2'' \simeq 0$. Show from Eqs. (3.4.3) and (3.4.8) that the current at these plateaus is given by

$$\frac{2\gamma_1\,\gamma_2}{2\gamma_1 + \gamma_2} \quad \text{and} \quad \frac{2\gamma_1\,\gamma_2}{\gamma_1 + \gamma_2}$$

respectively.