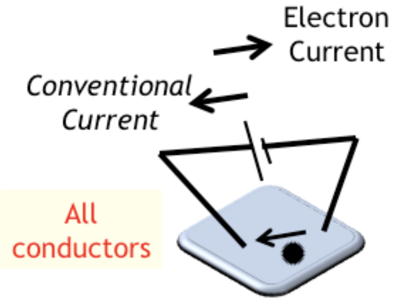
II. Old Topics in New Light

Lecture 10

Thermoelectricity



Conductance measurements ordinarily do not tell us anything about the nature of the conduction process inside the conductor. If we connect the terminals of a battery across any conductor, electron current flows out of the negative terminal back to its positive terminal. Since this is true of all conductors, it clearly does not tell us anything about the conductor itself.

On the other hand, thermoelectricity, that is, electricity driven by a temperature difference, is an example of an effect that does. A very simple experiment is to look at the current between a hot probe and a cold probe (Fig.10.1).

For an n-type conductor (see Fig.5.1) the direction of the external current will be consistent with what we expect if electrons travel from the hot to the cold probe inside the conductor, but for a p-type conductor (see Fig.5.2) the direction is reversed, consistent with electrons traveling from the cold to the hot probe. Why ?

It is often said that p-type conductors show the opposite effect because the carriers have the opposite sign. As we discussed in Lecture 5, p-type conductors involve the flow of electrons near the top of a band of energies and it is convenient to keep track of the empty states above μ rather than the filled states below μ. These empty states are called holes and since they represent the absence of an electron behave like positively charged entities.

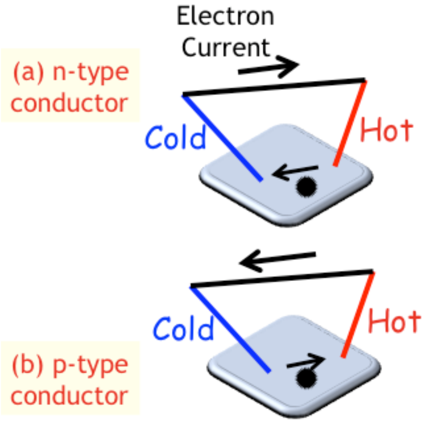


Fig.10.1.
Thermoelectric currents driven by temperature differences flow in opposite directions for n- and p-type conductors

However, this is not quite satisfactory since what moves is really an electron with a negative charge. "Holes" are at best a conceptual convenience and effects observed in a laboratory should not depend on subjective conveniences.

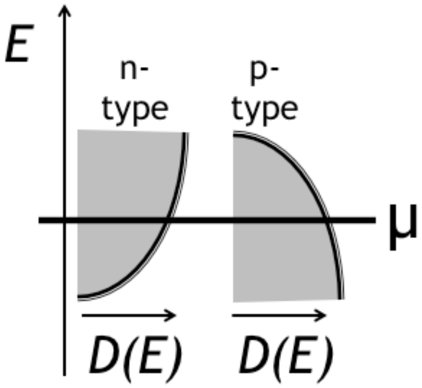


Fig.10.2. In n-type conductors the electrochemical potential is located near the bottom of a band of energies, while in p-type conductors it is located near the top. In n-conductors *D(E)* increases with increasing *E*, while in p-conductors it decreases with increasing *E*.

As we will see in this Lecture the difference between n- and p-conductors requires no new principles or assumptions beyond what we have already discussed, namely that the current is driven by the difference between f_1 and f_2. The essential difference between n- and p-conductors is that while one has a density of states *D(E)* that increases with energy E, the other has a *D(E)* decreasing with *E*.

Later in Lecture 13 we will discuss another important phenomenon called the Hall effect which changes sign for n-type and p-type conductors and this too is commonly blamed on negative and positive charges. This effect, however, has a totally different origin related to the negative mass ($m=p/v$) associated with *E(p)* relations in p-conductors that point downwards. By contrast the thermoelectric effect does not require a conductor to even have a *E(p)* relation. Even small molecules show sensible thermoelectric effects (Baheti et al. 2008).

The basic idea is easy to see starting from our old expression for the current obtained in Lecture 3:

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\left(f_1(E) - f_2(E) \right)$$

(10.1, same as Eq.(3.3))

So far the difference in f_1 and f_2 has been driven by difference in electrochemical potentials μ_1 and μ_2. But it could just as well be driven by a temperature difference, since in general

$$f_1(E) = \frac{1}{\exp\left(\dfrac{E-\mu_1}{kT_1}\right)+1}$$

and

$$f_2(E) = \frac{1}{\exp\left(\dfrac{E-\mu_2}{kT_2}\right)+1}$$

(10.2)

But why would such a current reverse directions for an n-type and a p-type conductor?

To see this, consider two contacts with the same electrochemical potential μ, but with different temperatures as shown in Fig. 10.3.

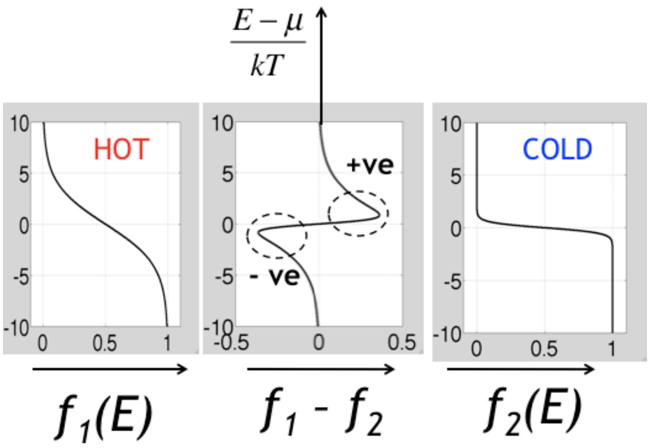


Fig.10.3. Two contacts with the same μ, but different temperatures: $f_1 - f_2$ is positive for $E>\mu$, and negative for $E<\mu$.

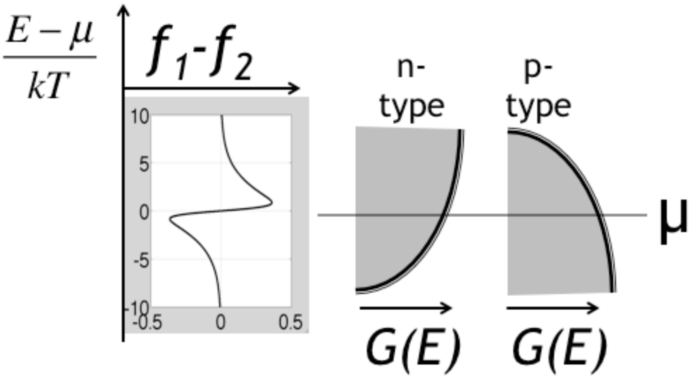
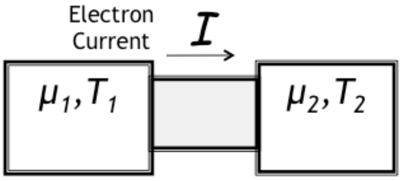


Fig.10.4 For n-type channels, the current for $E>\mu$ dominates that for $E<\mu$, while for p-type channels the current for $E<\mu$ dominates that for $E>\mu$. Consequently, electrons flow from hot to cold across an n-type channel, but from cold to hot in a p-type channel.

The key point is that the difference between $f_1(E)$ and $f_2(E)$ has a different sign for energies E greater than μ and for energies less than μ (see Fig.10.3).

In an n-type channel, the conductance $G(E)$ is an increasing function of energy, so that the net current is dominated by states with energy $E>\mu$ and thus flows from 1 to 2, that is from hot to cold (Fig.10.4). But in a p-type channel it is the opposite. The conductance $G(E)$ is a decreasing function of energy, so that the net current is dominated by states with energy $E<\mu$ and thus flows from 2 to 1, that is from cold to hot.

10.1.Seebeck Coefficient



We can use Eq.(10.1) directly to calculate currents without making any approximations. But it is often convenient to use a Taylor series expansion like we did earlier (Eq.(2.5)) to obtain results that are reasonably accurate for low "bias".

We could write approximately from Eq.(10.1)

$$I = G\,(V_1 - V_2) + G_S\,(T_1 - T_2) \tag{10.3}$$

where we have defined V_1, V_2 as μ_1/q and μ_2/q. The conductance is given by

$$G = \int_{-\infty}^{+\infty} dE\, G(E) \left(\frac{\partial f_0}{\partial \mu} \right)$$

$$= \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) G(E) \tag{10.4a}$$

as we have seen before in Section 2.4.

The new coefficient G_S that we have introduced is given by

$$G_S = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\left(\frac{\partial f_0}{\partial T}\right)$$

$$= \int_{-\infty}^{+\infty} dE\left(-\frac{\partial f_0}{\partial E}\right)\frac{E-\mu_0}{qT}\,G(E) \tag{10.4b}$$

This last step, relating the derivatives with respect to T and with respect to E, requires a little algebra (see Appendix A).

One point regarding the notation: I should really use a different symbol for the averaged conductance G (which we have not used elsewhere in these lectures) to distinguish it from the energy-dependent conductance *G(E)*. To avoid confusion, in this Lecture I will try to write *G(E)* whenever I mean the latter.

Eq.(10.4b) expresses mathematically the basic point we just discussed. Energies E greater and less than μ_0, contribute with opposite signs to the thermoelectric coefficient, G_S. It is clear that if we wanted to design a material with the best Seebeck coefficient, S we would try to choose a material with all its density of states on one side of μ_0 since anything on the other side contributes with an opposite sign and brings it down.

We can visualize Eq.(10.3) as shown in Fig.10.5, where the short circuit current is given by

$$I_{sc} = G_S\,(T_1 - T_2) \tag{10.5}$$

Experimentally what is often measured is the open circuit voltage

$$V_{oc} = -\frac{I_{sc}}{G} = -\frac{G_S}{G}\,(T_1 - T_2) \tag{10.6}$$

Note that we are using I and V for electron current and electron voltage μ/q whose sign is opposite that of the conventional current and voltage.

For n-type conductors, for example, G_S is positive, so that Eq.(10.6) tells us that V_{oc} is negative if $T_1 > T_2$. This means that the contact with the higher temperature has a negative electron voltage and hence a positive conventional voltage. By convention this is defined as a negative Seebeck coefficient.

$$S \equiv \frac{V_{oc}}{T_1 - T_2} = -\frac{G_S}{G} \tag{10.7}$$

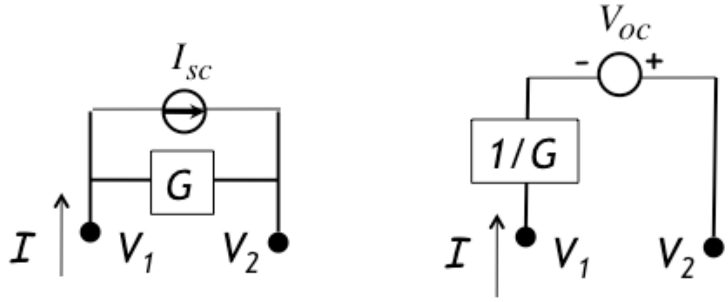


Fig.10.5. Circuit representations of Eq.(10.3).

10.2. Thermoelectric Figures of Merit

The practical importance of thermoelectric effects arise from the possibility of converting waste heat into electricity and from this point of view the important figure of merit is the amount of power that could be generated from a given T_1 - T_2. What load resistor R_L will maximize the power delivered to it (Fig.10.6)? A standard theorem in circuit theory says (this is not too hard to prove for yourself) that the answer is a "matched load" for which R_L equal to $1/G$:

$$P_{\max} = V_{oc}^2 G/4 = S^2 G \ (T_1 - T_2)^2/4 \tag{10.8}$$

The quantity S^2G is known as the power factor and is one of the standard figures of merit for thermoelectric materials.

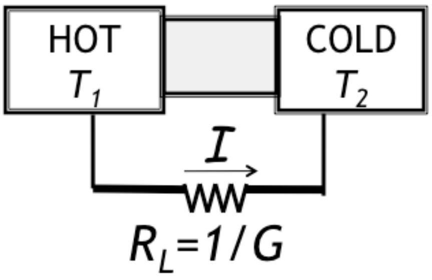
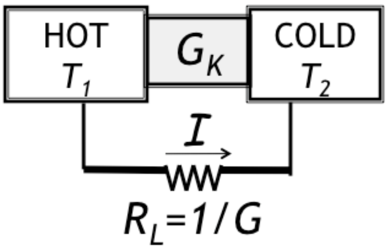


Fig.10.6. A thermoelectric generator can convert a temperature difference into an electrical output.



However, there is a second figure of merit that is more commonly used. To see where this comes from, we first note that when the contacts at different temperatures, we expect a constant flow of heat through the conductor due to its *heat conductance* G_K

$$G_K\left(T_1-T_2\right)$$

which has to be supplied by the source that maintains the temperature difference. Actually this is not quite right, it only gives the heat flow under open circuit conditions and ignores a component that depends on *I*. But this is good enough for our purpose which is simply to provide an intuitive feeling for where the standard thermoelectric figure of merit comes from.

The ratio of the maximum generated power to the power that is supplied by the external source is a good measure of the efficiency of the thermoelectric material in converting heat to electricity and can be written as

$$\frac{P_{\max}}{G_K(T_1-T_2)} = \underbrace{\frac{S^2GT}{G_K}}_{\equiv ZT} \frac{T_1-T_2}{4T} \tag{10.9}$$

where *T* is the average temperature *$(T_1+T_2)/2$*. The standard figure of merit for thermoelectric materials, called its *ZT product,* is proportional to the ratio of S^2G to G_K:

$$ZT \equiv \frac{S^2GT}{G_K} = \frac{S^2\sigma T}{\kappa} \tag{10.10}$$

where κ is the thermal conductivity related to the thermal conductance G_K by the same geometric factor *A/L* connecting the corresponding electrical quantities *G* and σ . Indeed the Ohm's law for heat conduction (known as Fourier's law) also needs the same correction for interface resistance namely the replacement of *L* with $L+\lambda$.

However, while the electrical conductivity arises solely from charged particles like electrons, the thermal conductivity also includes a contribution from phonons which describes the vibrations of the atoms comprising the solid lattice. Ordinarily it is the phonon component that dominates the thermal conductivity and we will discuss it briefly in the next Lecture. For the moment let us talk about the heat carried by electrons, something we have not discussed so far at all.

10.3. Heat Current

We have discussed the thermoelectric currents in a material with any arbitrary conductance function *G(E)*. The nice thing about the elastic resistor is that channels at different energies all conduct in parallel, so that we can think of one energy at a time and add them up at the end. Consider a small energy range located between *E* and *E+dE*, either above or below the electrochemical potentials $\mu_{1,2}$ as shown in Fig.10.7.

As we discussed in Section 10.1, these two channels will make contributions with opposite signs to the Seebeck effect.

It has been known for a long time that the Seebeck effect is associated with a Peltier effect. The connection can be easily understood as follows. Earlier in Lecture 3 we saw that for an elastic resistor the associated Joule heat I^2R is dissipated in the contacts (see Fig.3.3). But if we consider the n-type or p-type channels in Fig.10.7 apparent that unlike Fig.3.3, both contacts do not get heated.

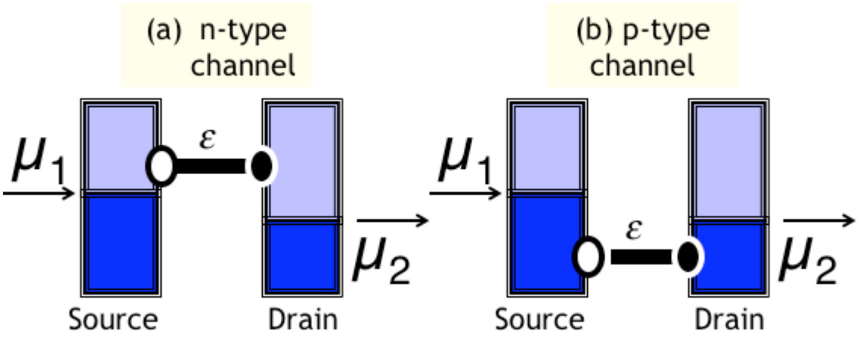


Fig.10.7. A one-level elastic resistor having just one level with E= ε , (a) above or (b) below the electrochemical potentials $\mu_{1,2}$,

Fig.10.8 is essentially the same as Fig.3.3 except that we have shown the heat absorbed from the surroundings rather than the heat dissipated. For n-type conductors the heat absorbed is positive at the source, negative at the drain, indicating that the source is cooled and the drain is heated. For p-type conductors it is exactly the opposite.

This is the essence of the Peltier effect that forms the basis for practical thermoelectric refrigerators. Note that the sign of the Peltier coefficient like that of the Seebeck coefficient is related to the sign of E-μ and not the sign of q.

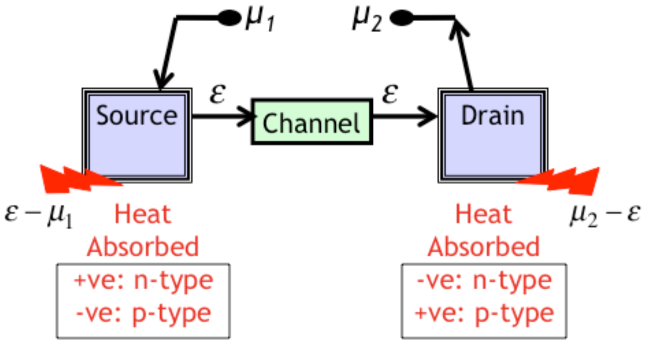


Fig.10.8. Same as Fig.3.3 but showing the heat absorbed at each contact. For n-type conductors the heat absorbed is positive at the source, negative at the drain showing that the electrons COOL the source and HEAT the drain. For p-type conductors it is exactly the opposite.

To write the heat current carried by electrons, we can simply extend what we wrote for the ordinary current earlier:

$$I = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, G(E)\left(f_1(E) - f_2(E) \right) \qquad \text{(same as Eq.(3.3))}$$

Noting that an electron with energy E carrying a charge $-q$ also extracts an energy $E-\mu_1$ from the source and dumps an energy $E-\mu_2$ in the drain, we can write the heat currents I_{Q1} and I_{Q2} *extracted* from the source and drain respectively as

$$I_{Q1} = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, \frac{E-\mu_1}{q}\, G(E)\left(f_1(E) - f_2(E)\right) \qquad (10.11a)$$

$$I_{Q2} = \frac{1}{q}\int_{-\infty}^{+\infty} dE\, \frac{\mu_2 - E}{q}\, G(E)\left(f_1(E) - f_2(E)\right) \qquad (10.11b)$$

The energy extracted from the external source per unit time is given by

$$I_E \quad = \quad \frac{\mu_1 - \mu_2}{q} I \quad = \quad V I \tag{10.11c}$$

so that the sum of all three energy currents is zero:

$$I_{Q1} + I_{Q2} + I_E \quad = \quad 0$$

as we would expect due to overall energy conservation.

10.3.1. Linear response

Just as we linearized the current equation (Eq.(3.3)) to obtain an expression for the current in terms of voltage and temperature differences (Eqs.(10.4)), we can linearize the heat current equation to obtain

$$I_Q \quad = \quad G_P \ (V_1 - V_2) \quad + \quad G_Q \ (T_1 - T_2) \tag{10.12}$$

where

$$G_P \quad = \quad \int\limits_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) \frac{E - \mu_0}{q} G(E) \tag{10.13a}$$

$$G_Q \quad = \quad \int\limits_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E} \right) \frac{(E - \mu_0)^2}{q^2 \, T} G(E) \tag{10.13b}$$

These are the standard expressions for the thermoelectric coefficients due to electrons which are usually obtained from the Boltzmann equation.

I should mention that the quantity G_Q we have obtained is not the thermal conductance G_K that is normally used in the *ZT* expression cited earlier (Eq.(10.10)). One reason is what we have stated earlier, namely that G_K also has a phonon component that we have not yet discussed. But there is another totally different reason.

The quantity G_K is defined as the heat conductance under electrical open circuit conditions (*I=0*):

$$G_K = \left(\frac{\partial I_Q}{\partial (T_1 - T_2)} \right)_{I=0}$$

while it can be seen from Eq.(10.12) that G_Q is the heat conductance under electrical short circuit conditions (*V=0*):

$$G_Q = \left(\frac{\partial I_Q}{\partial (T_1 - T_2)} \right)_{V_1=V_2}$$

However, we can rewrite Eqs.(10.3) and (10.12) in a form that gives us the open circuit coefficients (as noted earlier, *V* and *I* represent the electron voltage *μ/q* and the electron current, which are opposite in sign to the conventional voltage and current)

$$(V_1 - V_2) = \frac{1}{G} I - \overbrace{\frac{G_S}{G}}^{\substack{S, \\ Seebeck}} (T_1 - T_2) \tag{10.14a}$$

$$I_Q = \underbrace{\frac{G_P}{G}}_{\substack{Peltier \\ -\Pi}} I + \underbrace{\left(G_Q - \frac{G_P G_S}{G} \right)}_{\substack{Heat\ conductance \\ G_K}} (T_1 - T_2) \tag{10.14b}$$

We have indicated the coefficients that are normally measured experimentally and are named after the experimentalists who discovered them. Eqs.(10.3) and (10.12), on the other hand, come more naturally in theoretical models because of our Taylor's series expansion and it is important to be aware of the difference.

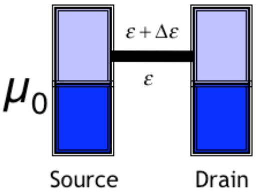
Incidentally, if we use the expressions in Eqs.(10.13a) and (10.4b), the Peltier and Seebeck coefficients in Eq.(10.14) obey the Kelvin relation

$$\Pi = T S \tag{10.15}$$

which is a special case of the fundamental Onsager relations that the linear coefficients are required to obey (Lecture 15).

10.4. “Delta Function” Thermoelectric

It is instructive to look at a so-called “delta function” thermoelectric, which is a hypothetical material with a narrow conductance function located at energy ε with a width $\Delta\varepsilon$ that is much less than *kT*.



It is straightforward to obtain the thermoelectric coefficients of this delta function thermoelectric formally starting from the general relations we have obtained in this Lecture, reproduced below for convenience:

$$G = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) G(E) \qquad \text{(same as 10.4a)}$$

$$G_S = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) \frac{E-\mu_0}{qT}\, G(E) \qquad \text{(same as 10.4b)}$$

$$G_P = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) \frac{E-\mu_0}{q}\, G(E) \qquad \text{(same as 10.13a)}$$

$$G_Q = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) \frac{\left(E-\mu_0\right)^2}{q^2 T}\, G(E) \qquad \text{(same as 10.13b)}$$

We argue that factors like E-μ_0 can be pulled out of the integrals assuming they are almost constant over the very narrow energy range where $G(E)$ is non-zero. This gives

$$G = G(\varepsilon)\,\Delta\varepsilon \left(-\frac{\partial f_0}{\partial E}\right)_{E=\varepsilon} \tag{10.16a}$$

$$G_S = \frac{\varepsilon-\mu_0}{qT}\,G \tag{10.16b}$$

$$G_P = \frac{\varepsilon-\mu_0}{q}\,G \tag{10.16c}$$

$$G_Q = \frac{\left(\varepsilon-\mu_0\right)^2}{q^2 T}\,G \tag{10.16d}$$

From Eq.(10.14) we obtain the coefficients for the ***delta function thermoelectric***:

$$S = -\frac{G_S}{G} = -\frac{\varepsilon-\mu_0}{qT} \tag{10.17a}$$

$$\Pi = -\frac{G_P}{G} = -\frac{\varepsilon-\mu_0}{q} \tag{10.17b}$$

$$G_K = G_Q - \frac{G_P G_S}{G} = 0 \tag{10.17c}$$

Let us now see how we can understand these results from intuitive arguments without any formal calculations. The Seebeck coefficient in Eq.(10.17a) is the open circuit voltage required to maintain zero current. Since the channel conducts only at a single energy $E=\varepsilon$, in order for no current to flow, the Fermi functions at this energy must be equal:

$$f_1(\varepsilon) = f_2(\varepsilon) \rightarrow \frac{\varepsilon-\mu_1}{kT_1} = \frac{\varepsilon-\mu_2}{kT_2}$$

Hence

$$\frac{\varepsilon-\mu_1}{kT_1} = \frac{\varepsilon-\mu_2}{kT_2} = \frac{(\varepsilon-\mu_1)-(\varepsilon-\mu_2)}{k(T_1-T_2)} = -\frac{\mu_1-\mu_2}{k(T_1-T_2)}$$

Noting that the Seebeck coefficient is defined as

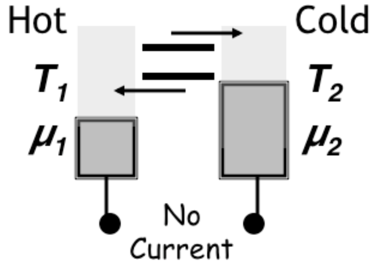
$$S \equiv \frac{(\mu_1-\mu_2)/q}{T_1-T_2} \qquad \text{(with I = 0)}$$

we obtain $S = -\frac{\varepsilon-\mu_1}{qT_1} = -\frac{\varepsilon-\mu_2}{qT_2} \approx -\frac{\varepsilon-\mu_0}{qT}$

in agreement with the result in Eq.(10.17a).

The expression in Eq.(10.17b) for the Peltier coefficient too can be understood in simple terms by arguing that every electron carries a charge -q and a heat $\varepsilon-\mu_0$, and hence the ratio of the heat current to the charge current must be $(\varepsilon-\mu_0)/-q$.

That brings us to the zero heat conductance in Eq.(10.17c) which tells us that the heat current is zero under open circuit conditions, that is when the regular charge current is zero. This seems quite reasonable. After all if there is no electrical current, how can there be a heat current? But if this were the whole story, then no thermoelectric would have any heat conductance, and not just delta function thermoelectrics.



The full story can be understood by considering a two-channel thermoelectric with a temperature difference. Under open circuit conditions, there is a voltage between the two contacts with $\mu_1 < \mu_2$. Although the total current is zero, the individual currents in each level are non-zero. They are equal and opposite, thereby canceling each other out. But the corresponding energy currents do not cancel, since the channel with higher energy carries more energy. Zero charge current thus does not guarantee zero heat current, except for a delta function thermoelectric with its sharply peaked *G(E)*.

Since the delta function thermoelectric has zero heat conductance, the *ZT* product (see Eq.(10.10)) should be very large and it would be seem that is what an ideal thermoelectric should look like. However, as we mentioned earlier, even if the electronic heat conductance were zero, we would still have the phonon contribution which would prevent the *ZT*-product from getting too large. We will talk briefly about this aspect in the next Lecture.

10.4.1. Optimizing Power Factor

Let us end this Lecture by discussing what factors might maximize the power factor S^2G (see Eq.(10.8)) for a thermoelectric. If getting the highest Seebeck coefficient S were our sole objective then it is apparent from Eq.(10.17b) that we should choose our energy ε as far from μ_0 as possible. But that would make the conductance *G* from Eq.(10.17a) unacceptably low, because the factor $(-\partial f_0 / \partial E)$ dies out quickly as the energy *E* moves away from μ_0.

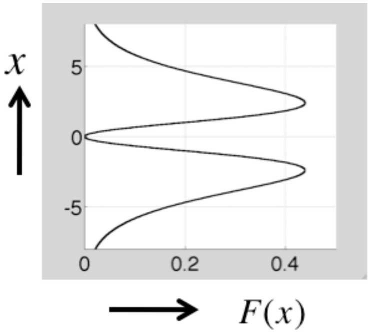
From Eq.(10.17a) and (10.16a) we have

$$S^2G \;=\; G(\varepsilon)\,\Delta\varepsilon\left(\frac{\varepsilon-\mu_0}{qT}\right)^2\left(-\frac{\partial f_0}{\partial E}\right)_{E=\varepsilon}$$

$$=\; G(\varepsilon)\frac{\Delta\varepsilon}{kT}\left(\frac{k}{q}\right)^2 \underbrace{x^2\,\frac{e^x}{(e^x+1)^2}}_{\equiv F(x)}, \quad where \quad x\equiv\frac{\varepsilon-\mu_0}{kT} \tag{10.18}$$

Fig.10.9. Plot of

$$F(x) \equiv \frac{x^2 e^x}{(e^x+1)^2}$$



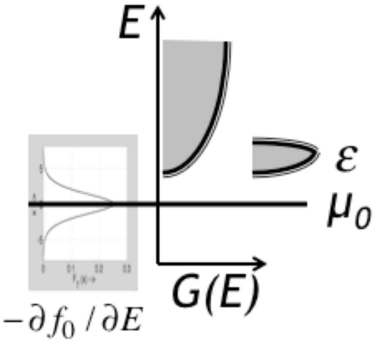
It is apparent from Fig.10.9 that the function *F(x)* has a maximum around x ~ 2, suggesting that ideally we should place our level approximately 2kT above or below the electrochemical potential μ_0.

The corresponding Seebeck coefficient is approximately

$$S \approx 2\frac{k}{q} \tag{10.19a}$$

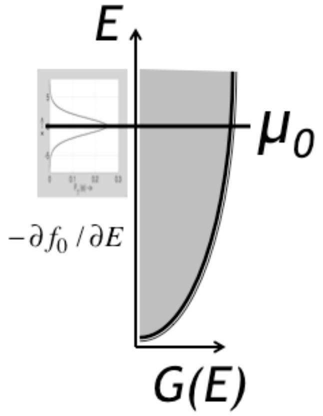
$$S^2 G \approx 0.5\left(\frac{k}{q}\right)^2 G(\varepsilon)\,\frac{\Delta\varepsilon}{kT} \tag{10.19b}$$

The best thermoelectrics typically have Seebeck coefficients that are not too far from the *2 (k/q) = 170 μV/K* expected from Eq.(10.19a). They are usually designed to place μ_0 a little below the bottom of the band so that the product of *G(E)* and $(-\partial f_0/\partial E)$ looks like a "delta function" around ε a little above the bottom of the band as shown in the sketch.



The problem is that the corresponding values of conductance are not as large as they could possibly be if μ_0 were located higher up in the band as sketched here. This would be characteristic of metals.

But metals show little promise as thermoelectric materials, because their Seebeck coefficients are far less than *k/q,* since the electrochemical potential lies in the middle of a band of states and there are nearly as many states above μ_0 as there are below μ_0.



For this reason the field of thermoelectric materials is dominated by semiconductors which show the highest power factors. However, the power factor determines only the numerator of the *ZT* product in Eq.(10.10). As we mentioned earlier the heat conductance in the denominator is dominated by phonon transport involving a physics that is very different from the electronic transport properties that this book is largely about. In the next lecture we will digress briefly to talk about phonon transport.

Lecture 11

Phonon Transport

11.1. Phonon Heat Current
11.2. Thermal Conductivity

We have seen earlier that the electrical conductivity is given by (Eq.(4.14))

$$\sigma = \frac{q^2}{h}\left(\frac{M\lambda}{A}\right) \tag{11.1}$$

where the number of channels per unit area *M/A* and the mean free path λ are evaluated in an energy window ~ a few *kT* around μ_0. The degeneracy factor g (Section 5.4) due to spins and valleys is assumed to be included in *M*.

In this Lecture we will extend the transport theory for electrons to handle something totally different, namely phonons and obtain a similar expression for the thermal conductivity due to phonons

$$\kappa_{ph} = \frac{\pi^2}{3}\frac{k^2T}{h}\left(\frac{M\lambda}{A}\right)_{ph} \tag{11.2}$$

where the number of channels per unit area *M/A* and the mean free path λ are evaluated in a frequency window $\hbar\omega$ ~ a few *kT*. There is a degeneracy factor of $g=3$ due to three polarizations that is assumed to be included in *M*.

Our purpose in this Lecture is two-fold. The first is to provide an interesting perspective in the hunt for high-*ZT* thermoelectrics. We have seen in Lecture 10 that with careful design it is possible to achieve a Seebeck coefficient ~ *2k/q* while maximizing the numerator in Eq.(10.10). We can write

$$ZT \approx 4\frac{k^2T}{q^2}\frac{\sigma}{\kappa+\kappa_{ph}} \approx 4\frac{k^2T}{q^2}\frac{\sigma}{\kappa_{ph}}$$

if we assume that the thermoelectric has been designed to provide a Seebeck coefficient *S ~ 2k/q* and the heat conductivity is dominated by phonons. Using Eqs.(11.1) and (11.2) we have

$$ZT \approx \frac{M\lambda / A}{(M\lambda / A)_{ph}} \tag{11.3}$$

where we have dropped a factor of $12/\pi^2$ ~ 1 since it is just an approximate number anyway. This is an interesting expression suggesting that once a material has been optimized to provide a respectable Seebeck coefficient (*S*), the *ZT* product we obtain simply reflects the ratio of $M\lambda / A$ for electrons and phonons.

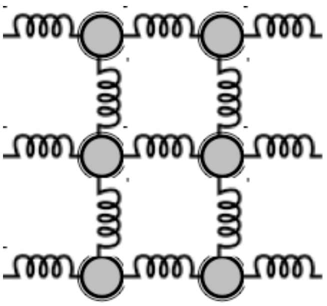
As we discussed at the end of the last Lecture, the process of achieving a high S usually puts us in a regime with a low *M/A* for electrons. *M/A* for phonons on the other hand is often much higher ~ $(1\text{ nm})^{-2}$ at room temperature, so that the ratio of *M/A*'s in Eq.(11.3) is ~ *0.1* or less. But electrons tend to have a longer mean free path, resulting in a *ZT ~ 1* for the best thermoelectrics. The most promising approach for improving *ZT* at this time seems to be to try to suppress the mean free path for phonons without hurting the electrons (the so-called "electron crystal, phonon glass").

The highest *ZT* is on the order of 1 to 3 and it has been that way for long time. Experts say that an increase of *ZT* to 4 - 10 would have a major impact on its practical applications and researchers hope that nanostructured materials might enable this increase. Whether they are right, only future experiments can tell, but it is clearly of interest to understand the principles that govern *ZT* in nanoscale materials and we hope this Lecture will contribute to this understanding.

But my real objective is to demonstrate the power of the elastic resistor approach that allows us not only to obtain linear transport coefficients for electrons easily, but also extend the results to a totally different entity (the phonons) with relative ease.

11.1. Phonon Heat Current

As we mentioned earlier the thermal conductance of solids has a significant phonon component in addition to the electronic component we just talked about. I will not go into this in any depth. My purpose is simply to show how easily our elastic transport model is extended to something totally different.



The atoms comprising the solid lattice are often pictured as an array of masses connected by springs as sketched here. The vibrational frequencies of such a system are described by a dispersion relation $\omega(\beta)$ analogous to the *E(k)* relation that describes electron waves, with β playing the role of *k,* and $\hbar\omega$ playing the role of *E*.

The key difference with electrons is that unlike electrons, there is no exclusion principle. Millions of phonons can be packed into a single channel creating a sound wave that we can even hear, if the frequency is low enough. One consequence of this lack of a exclusion principle is that the equilibrium distribution of phonons is given by a Bose function

$$n(\omega) \equiv \frac{1}{\exp\left(\dfrac{\hbar\omega}{kT}\right)-1} \tag{11.4}$$

instead of the Fermi function for electrons introduced in Lecture 2:

$$f(E) \equiv \frac{1}{\exp\left(\dfrac{E-\mu}{kT}\right)+1} \qquad \text{(same as Eq.(2.2))}$$

The difference with Eq.(11.4) is just the +1 instead of the -1 in the denominator, which restricts *f(E)* to values between 0 and 1, unlike the $n(\omega)$ in Eq.(11.4). The other difference is the absence of an electrochemical potential μ in Eq.(11.4) which is attributed to the lack of conservation of phonons. Unlike electrons, they can appear and disappear as long as other entities are around to take care of energy conservation.

These results are of course not meant to be obvious, but they represent basic results from equilibrium statistical mechanics that are discussed in standard texts. In Lecture 16 on the Second Law, we will try to say a little more about the basics of equilibrium statistical mechanics. We make use of these equilibrium results but we cannot really do justice to them without a major detour from our main objective of presenting a new approach to non-equilibrium problems.

Anyway, the bottom line is that our result for the charge current carried by electrons

$$I = \frac{q}{h}\int_{-\infty}^{+\infty} dE \left(\frac{M\lambda}{L+\lambda}\right)\left(f_1(E)-f_2(E)\right)$$

can be modified to represent the heat current due to phonons

$$I_Q = \frac{1}{h}\int_{0}^{+\infty} d(\hbar\omega)\left(\frac{M\lambda}{L+\lambda}\right)_{ph} \hbar\omega\,\left(n_1(\omega)-n_2(\omega)\right) \qquad (11.5)$$

simply by replacing the charge with the energy: $q \rightarrow \hbar\omega$

and the Fermi functions with the Bose functions:

$$n_1(\omega) = \frac{1}{\exp(\hbar\omega / kT_1) - 1}$$

$$n_2(\omega) = \frac{1}{\exp(\hbar\omega / kT_2) - 1}$$

and changing the lower integration limit to zero.

Again we can linearize Eq.(11.5) to write (see Appendix A)

$$I_Q \approx G_\kappa \, (T_1 - T_2) \tag{11.6}$$

where the thermal conductance due to phonons can be written as

$$G_K = \frac{k^2 T}{h} \int_0^{+\infty} dx \left(\frac{M\lambda}{L+\lambda}\right)_{ph} \frac{x^2 e^x}{(e^x - 1)^2} \, , \quad x \equiv \frac{\hbar\omega}{kT} \tag{11.7}$$

Note that just as the elastic resistor model for electrons ignores effects due to the inelastic scattering between energy channels, this model for phonons ignores effects due to the so-called "anharmonic interactions" that cause phonons to convert from one frequency to another. While ballistic electron devices have been widely studied for nearly two decades, much less is known about ballistic phonon devices.

11.1.1. Ballistic Phonon Current

Before moving on let us take a brief detour to point out that the ballistic conductance due to phonons is well-known though in a slightly different form, similar to the Stefan-Boltzmann law for photons.

From Eq.(11.7) we can write the ballistic heat conductance as

$$\left[G_\kappa\right]_{ballistic} = \frac{k^2 T}{h} \int_0^{\infty} dx \, M_{ph} \, \frac{x^2 e^x}{(e^x - 1)^2} \tag{11.8}$$

To evaluate this expression we need to evaluate the number of modes which is related to the number of wavelengths that fit into the cross-section, as we discussed for electrons (see Eq.(5.17))

$$M_{ph} = \frac{\pi A}{(wavelength)^2} \underbrace{3}_{\substack{number\ of\\ polarizations}}$$

but we have a degeneracy factor of 3 for the three allowed polarizations. Noting that for phonons (c_s : acoustic wave velocity)

$$wavelength = \frac{c_s}{\omega / 2\pi}$$

we have

$$M_{ph} = \frac{3\omega^2 A}{4\pi c_s^2} = \frac{3k^2 T^2 A}{4\pi \hbar^2 c_s^2} x^2$$

From Eq.(11.8),

$$\left[G_K\right]_{ballistic} = \frac{3k^4 T^3}{8\pi^2 \hbar^3 c_s^2} \underbrace{\int_0^\infty dx\, \frac{x^4 e^x}{(e^x - 1)^2}}_{= 4\pi^4/15} = \frac{\pi^2 k^4 T^3}{10\hbar^3 c_s^2}$$

Making use of this expression we can write the ballistic heat current from Eq.(11.6) as

$$\left[I_Q\right]_{ballistic} = \frac{\pi^2 k^4 T^3}{10\hbar^3 c_s^2} \Delta T$$

However, the ballistic current is usually written in a different form making use of the relation $T^3\ \Delta T = \Delta(T^4 / 4)$:

$$\left[I_Q\right]_{ballistic} = \frac{\pi^2 k^4}{40\hbar^3 c_s^2} \Delta T^4 = \frac{\pi^2 k^4}{40\hbar^3 c_s^2} \left(T_1^4 - T_2^4\right) \tag{11.9}$$

The corresponding result for photons is known as the Stefan-Boltzmann relation for which the numerical factor differs by a factor of *2/3* because the number of polarizations is *2* instead of *3*. But this is just a detour. Let us get back to diffusive phonon transport.

11.2. Thermal Conductivity

Returning to Eq.(11.7) for the thermal conductance due to phonons, we could define the thermal conductivity

$$\kappa \;=\; \frac{k^2 T}{h} \int_{0}^{+\infty} dx \left(\frac{M\lambda}{A}\right)_{ph} \frac{x^2\, e^x}{(e^x - 1)^2} \tag{11.10}$$

such that

$$G_\kappa = \frac{\kappa A}{\lambda + L}$$

Note the similarity with the electrical conductivity due to electrons:

$$\sigma \;=\; \frac{q^2}{h} \int_{-\infty}^{+\infty} dx \left(\frac{M\lambda}{L+\lambda}\right) \frac{e^x}{(e^x + 1)^2}$$

The function

$$F_T(x) \;\equiv\; \frac{e^x}{(e^x + 1)^2}$$

appearing in all electronic transport coefficients is different from the function

$$\frac{3}{\pi^2} \frac{x^2\, e^x}{(e^x - 1)^2}$$

appearing in Eq.(11.10) but they are of similar shape as shown. The factor $3/\pi^2$ is needed to make the area under the curve equal to one, as it is for the broadening function $F_T(x)$ for electrons (see Eq.(2.4)).

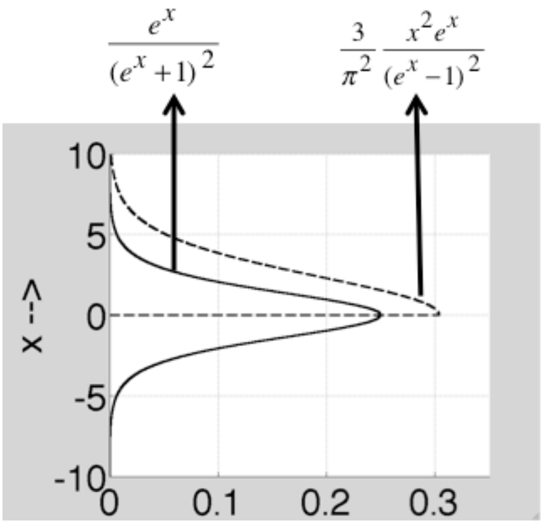


Fig.11.1. Broadening function for phonons compared to that for electrons, $F_T(x)$. These are the window functions defined by Jeong et al. (2011), see Eqs.(7e,f).

So we can think of electrical and thermal conductance at least qualitatively in the same way. Just as the electrical conductivity is given by the product

$$\underbrace{\frac{q^2}{h}}_{38\,\mu S}\left(\frac{M\lambda}{A}\right)$$

the thermal conductivity is given by

$$\underbrace{\frac{\pi^2}{3}\frac{k^2T}{h}}_{284\,pW/K}\left(\frac{M\lambda}{A}\right)_{ph}$$

The factor $\pi^2/3$ is just the inverse of the $3/\pi^2$ needed to normalize the phonon broadening function.

We mentioned at the end of the last Lecture that *M/A* for electrons tends to be rather small for good thermoelectric materials whose electrochemical potential μ lies within a *kT* of the bottom of the band. One way to get around this is to use materials where the entire electronic band of energies is a few *kT* wide, which is unusual. Unfortunately for the phonon band this condition is common, giving an average *M/A* close to the maximum value.

The most popular thermoelectric material Bi_2Te_3 appears to have a phonon bandwidth much less than *kT*, thus making the average value of *M/A* for phonons relatively small. The phonon mean free path is also relatively small, helping raise *ZT*. For example, $M/A \sim 4e17/m^2$, $\lambda \sim 15$ nm gives $\kappa \sim 2\ W/m/K$, numbers that are approximately representative of Bi_2Te_3.

The possible role of nanostructuring in engineering a better thermoelectric is still a developing story whose ending is not clear. At this time all we can do is to present a different viewpoint that may help us see some new options. And that is what we have tried to do here.

Lecture 12

Measuring Electrochemical Potentials

12.1. The Landauer Formulas
12.2. Büttiker Formula

Electrochemical potentials have played an important role in our discussion, starting from Lecture 2 where I stressed that electron flow is driven by the difference in the electrochemical potentials μ_1 and μ_2 in the two contacts. However, talking about electrochemical potentials inside the channel, as we did later in Chapter 6 when discussing the diffusion equation, often raises eyebrows. This is because an electrochemical potential of μ implies that the occupation of all available states are described by the corresponding Fermi function (Eq.(2.2))

$$f(E) \quad = \quad \frac{1}{1+\exp((E-\mu)/kT)}$$

This is approximately true of large contacts which always remain close to equilibrium, but not necessarily true of small conductors even for small applied voltages. As we saw in Lecture 6, it was important to introduce two separate electrochemical potentials μ^+ and μ^- in order to understand the interface resistance that is the key feature of the new Ohm's law (Eq.(4.2)).

Non-equilibrium electrochemical potentials of this type can be very useful in understanding current flow and is widely used by device engineers. It is common to use two different potentials (often called quasi-Fermi levels) for conduction and valence bands and in Lecture 14 we will talk about different potentials for upspin and downspin electrons.

Indeed in Lecture 9 we even argued that controlling such potentials with creatively designed "smart" contacts could lead to unique devices.

In spite of the obvious utility of the concept, many experts are uneasy about invoking non-equilibrium electrochemical potentials inside nanoscale devices which they view as ill-defined concepts that cannot be measured. Instead they feel conceptually on solid ground by sticking to terminal descriptions in terms of the electrochemical potentials at the contacts.

In this Lecture I would like to address some of these issues related to non-equilibrium potentials and their measurability using a simple example which will also allow us to connect our discussion to the Landauer formulas and the Büttiker formula that form the centerpiece of the transmission formalism widely used in mesoscopic physics.

So far we have talked about normal resistors with uniformly distributed scatterers characterized by a mean free path. Instead, following Landauer, let us consider an otherwise ballistic channel with a single localized defect that lets a fraction *T* of all the incident electrons proceed along the original direction, while the rest *1-T* get turned around. (see Fig.12.1).

We could follow our arguments from Lecture 6 to obtain the spatial variation of the potentials μ^+ and μ^- across the scatterer, and use it to deduce the resistance of the scatterer. But experts are often uneasy about non-equilibrium potentials and one way to bypass these questions is to consider a four-terminal measurement (Fig.12.2) using two additional voltage probes that draw negligible current, to measure the voltage drop across the defect.

We will show that if the voltage probes are identical and weakly coupled (non-invasive) then this four-terminal conductance G_{4t} is given by

$$G_{4t} \quad = \quad \frac{I}{(\mu_{1*} - \mu_{2*})/q} \quad = \quad M\frac{q^2}{h}\frac{T}{1-T} \tag{12.1}$$

M being the number of channels or modes in the conductor discussed in Lecture 4. But if we were to determine the conductance using the actual voltage applied to the current-carrying terminals we would obtain a lower conductance:

$$G_{2t} \quad = \quad \frac{I}{(\mu_1 - \mu_2)/q} \quad = \quad M\frac{q^2}{h}T \tag{12.2}$$

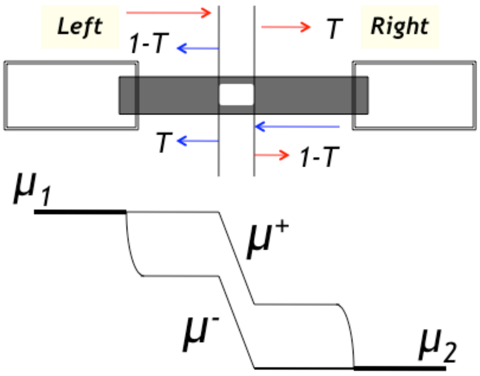


Fig.12.1.
Potential variation across a defect.

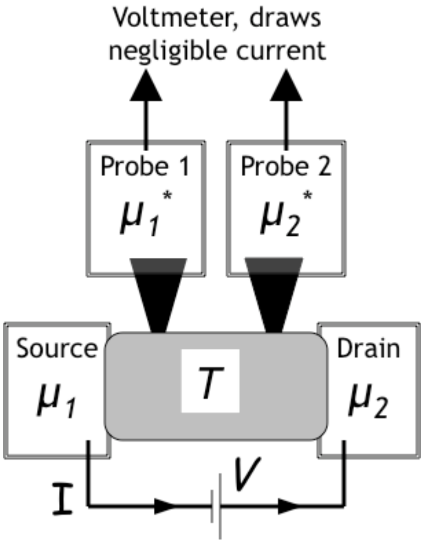


Fig.12.2: Four-terminal measurement of conductance of an otherwise ballistic one-dimensional conductor having a single “defect” in the middle, through which electrons have a probability *T* of transmitting.

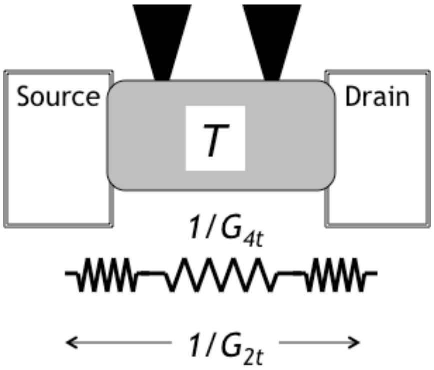


Fig.12.3: The two-terminal resistance can be viewed as the four-terminal resistance in series with the interface resistance.

The difference between the two-terminal (Eq.(12.2)) and four-terminal (Eq.(12.1)) resistances reflects the same *interface resistance*

$$\frac{1}{G_{2t}} - \frac{1}{G_{4t}} \quad = \quad \frac{h}{q^2 M}$$

introduced in Lecture 4 that differentiates the new Ohm's law (Eq.(1.4)) from the standard one (Eq.(1.1a)).

Although the interface resistance was recognized for metallic resistors in the late 1960's and is known as the *Sharvin resistance*, its ubiquitous role is not widely appreciated even today. In the early 1980's there was considerable confusion and discussion about the difference between the two conductance formulas in Eqs.(12.1) and (12.2) and *Imry* is credited with identifying the difference as a quantized Sharvin resistance related to the interfaces. With the rise of mesoscopic physics, Eq.(12.2) has come to be widely used and known as the Landauer formula while Landauer's original formula (Eq.(12.1)) is relatively forgotten, and not many recognize the difference.

The reader may wonder why the four-terminal Landauer formula came to be "forgotten." After all resistance measurements are commonly made in the four terminal configuration in order to exclude any contact resistance. Don't such measurements require Eq.(12.1) for their interpretation? Sort of, but not exactly. Let me explain.

One problem in the early days of mesoscopic physics was that the voltage probes were strongly coupled to the main conductor and behaved like "additional defects" whose effect could not simply be ignored. In order to interpret real experiments using four-terminal configurations, Büttiker (see Büttiker 1988) found an elegant solution by writing the current I_m at terminal m of a multi-terminal conductor in terms of the terminal potentials μ_n:

$$I_m = \ (1/q) \sum_n G_{mn} (\mu_m - \mu_n) \tag{12.3}$$

where $G_{m,n}$ is the conductance determined by the transmission $T_{m,n}$ between terminals m and n. With just two terminals, Büttiker's formula reduces to

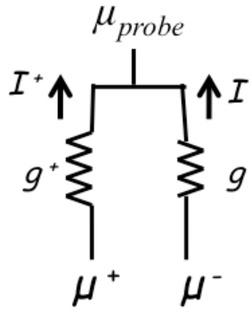
$$I_1 = \ (1/q)\ G_{12}(\mu_1 - \mu_2) \ \ = - I_2$$

which is the same as the two-terminal Landauer formula (Eq.(12.2)) if we identify G_{12} as $(q^2/h)M$. But the power of Eq.(12.3) lies in its ability to provide a quantitative basis for the analysis of multi-terminal structures like the one in Fig.12.2.

Knowing G_{mn}, if we knew all the potentials μ_m, we could use Eq.(12.3) to calculate the currents I_m at all the terminals. Of course for the voltage probes 1* and 2* we do not know the voltages they will float to and so we do not know μ_{1^*} or μ_{2^*}, to start with. But we do know the currents I_{1^*} and I_{2^*}, each of which must be zero, since the high impedance voltmeter draws negligible current. The point is that if we know either μ_m or I_m at each terminal m we can solve Eq.(12.3) to obtain whatever we do not know.

In this Lecture we will look at a specific problem, namely the voltage drop across a defect (Fig.12.1) and show that with weakly coupled non-invasive probes the Büttiker formula indeed gives the same answers as we get by looking directly at the electrochemical potentials μ^+ and μ^- inside the conductor.

This is reassuring because the approach due to Büttiker deals directly with measurable terminal quantities and so appears conceptually on more comfortable ground. The development of scanning probe microscopy (SPM) has made it possible to use nanoscale tunneling contacts as voltage probes whose effect is indeed negligible. Measurements using such "non-invasive" probes do provide experimental support for the four-terminal Landauer formula, but there is a subtlety involved.



What a voltage probe measures is some weighted average of the two potentials μ^+ and μ^- , the exact weighting depending on the construction of the probes. We could model it by associating conductances g^+ and g^- with the transmission of electrons from the + and the - streams into the probes respectively.

Setting the net probe current to zero we can write

$$g^+ (\mu^+ - \mu_{probe}) + g^- (\mu^- - \mu_{probe}) = 0$$

so that

$$\mu_{probe} = \underbrace{\frac{g^+}{g^+ + g^-}}_{\alpha} \mu^+ + \underbrace{\frac{g^-}{g^+ + g^-}}_{1-\alpha} \mu^- \tag{12.4}$$

For atomic scale probes that are much smaller than an electron wavelength we expect the two conductances to be similar so that the weighting factor $\alpha \sim 50\%$, so that the probe measures the average potential

$$\mu_{probe} = (\mu^+ + \mu^-)/2$$

For larger probes, however, it is possible for a voltage probe to have a pronounced bias for one stream or the other leading to a weighting factor α different from 50%. If this weighting happens to be different for the two probes 1* and 2*, it could change the measured resistance from that predicted by Eq.(12.1).

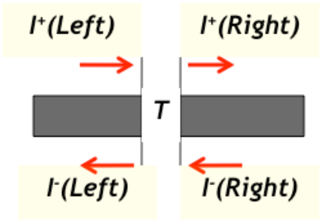
Indeed, experimental measurements have even shown *negative resistance*, something that cannot be understood in terms of Eq.(12.1). However, some of this is due to quantum interference effects that make the simple semiclassical description in terms of $\mu^{\pm}$ inadequate as we will see in Lecture 20. However, one could use a more sophisticated version of Eq.(12.4) (Lecture 22) or use the Büttiker formula, with the conductances $G_{m,n}$ calculated from an appropriate quantum transport model.

The bottom line is that if we know the correct internal state of the conductor in terms of a set of non-equilibrium electrochemical potentials, we can predict what a specific non-invasive voltage probe will measure and the result should match what the Büttiker formula predicts. The reverse, however, is not true. Knowing what a specific probe will measure, we cannot deduce the internal state of the conductor. But if we have measurements from multiple probes we could back out the internal state as we will see in Lecture 14 when we discuss non-equilibrium spin potentials.

With that rather long "introduction" let us now look at the two Landauer formulas (Eqs.(12.1), (12.2)) and the Büttiker formula (Eq.(12.3)) in a little more detail.

12.1.The Landauer formulas (Eqs.(12.1), (12.2))

Getting back to the problem of finding the potential variation across a defect in an otherwise ballistic conductor (Fig.12.1), we start by relating the outgoing currents to the incoming currents as follows



$$I^+\,(Right) \;=\; T\,I^+\,(Left) \;+\; (1-T)\,I^-\,(Right)$$

$$I^-\,(Left) \;=\; (1-T)I^+\,(Left) \;+\; T\,I^-\,(Right)$$

We can then convert the currents to occupation factors (see Eqs.(6.13))

$$f^+\,(Right) = \; T\,f^+(Left) \,+\, (1-T)\,f^-(Right)$$

$$f^-\,(Left) = (1-T)\,f^+\,(Left) \,+\, T\,f^-(Right)$$

and then to potentials using the same argument as in Lecture 6 (see discussion following Eq.(6.11b)):

$$\begin{aligned}\mu^+(Right) &= T\,\mu^+(Left) \;+\; (1-T)\,\mu^-(Right)\\ &= T\,\mu_1 \;+\; (1-T)\,\mu_2\end{aligned} \tag{12.5a}$$

$$\begin{aligned}\mu^-(Left) &= (1-T)\,\mu^+(Left) \;+\; T\,\mu^-(Right)\\ &= (1-T)\,\mu_1 \;+\; T\,\mu_2\end{aligned} \tag{12.5b}$$

The algebra can be simplified by choosing the potential for one of the contacts as zero and the other as one. The actual potential differences can then be obtained by multiplying by the actual $\mu_1 - \mu_2 = qV$.

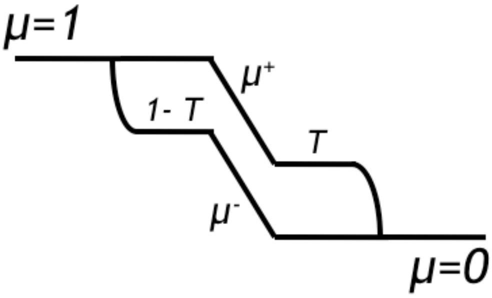


Fig.12.4. Spatial profile of μ^+ and μ^- across a scatterer normalized to an overall potential difference of one. The actual potential differences can be obtained by multiplying by the actual $\mu_1 - \mu_2 = qV$.

Eqs.(12.5a,b) then give us the picture shown in Fig.12.4 leading to

$$\mu^{+} - \mu^{-} = T(\mu_1 - \mu_2)$$

as long as both μ^+ and μ^- are evaluated at the same location on the left or on the right of the scatterer. Using

$$I = \frac{q}{h} M (\mu^{+} - \mu^{-})$$

for the current we obtain the standard Landauer formula (Eq.(12.2)).

To obtain the first Landauer formula we find the drop in either μ^+ or μ^- across the scatterer:

$$\mu^{+}(Left) - \mu^{+}(Right) = (1-T)(\mu_1 - \mu_2) \tag{12.6a}$$

$$\mu^{-}(Left) - \mu^{-}(Right) = (1-T)(\mu_1 - \mu_2) \tag{12.6b}$$

and then divide the current by it to obtain the result stated in Eq.(12.1):

$$G_{4t} = \frac{q^2}{h} M \frac{T}{1-T}$$

Note, however, that we are looking at the electrochemical potentials inside the conductor. How does this relate to the voltage measured by non-invasive voltage probes implemented using scanning tunneling probes (STP)? Assuming that the probe measures the average of μ^+ and μ^- we obtain the plot shown in Fig.12.5 from Fig.12.4.

What if the probe measures a weighted average of μ^+ and μ^- with some α (see Eq.(12.4)) other than *50%*? As long as α is the same for both probes, the drop across the scatterer would still be given by

$$\mu_{probe}(Left) - \mu_{probe}(Right) = (1-T)(\mu_1 - \mu_2) \tag{12.6c}$$

thus leading to the same Landauer formula (Eq.(12.1)). But if the weighting factor α were different for the two probes then the result would not match Eq.(12.1). As an extreme example if α were zero on the left and one on the right,

$$\mu_{probe}(Left) - \mu_{probe}(Right) = (1-2T)(\mu_1-\mu_2)$$

leading to a negative resistance for $T > 0.5$.

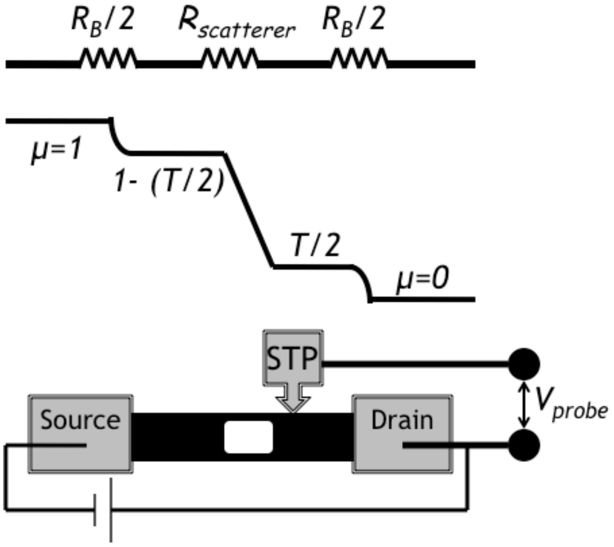


Fig.12.5. A scanning tunneling probe (STP) measures the average electrochemical potential $(\mu^+ + \mu^-)/2$.

Clearly the concept of non-equilibrium potentials μ^+ and μ^- should be used wisely with caution. But it does lead to intuitive understandable results. The potential drops across the defect but not across the ballistic regions, suggesting that the defect represents a resistance given by Eq.(12.1). Note, however, that we are still talking about elastic resistors. We have an IR drop in the voltage, but no corresponding I^2R in Joule dissipation. All dissipation is still in the contacts.

12.2. Büttiker Formula (Eq.(12.3))

Eq.(12.3) deals directly with the experimentally measured terminal quantities bypassing any questions regarding the internal variables. The point we wish to stress is the general applicability of this result irrespective of whether the resistor is elastic or not. Indeed, as we will see we can obtain it invoking very little beyond linear circuit theory.

We start by defining a multi-terminal conductance

$$G_{m,n} \equiv -\frac{\partial I_m}{\partial(\mu_n / q)}, \quad m \neq n \tag{12.7a}$$

$$G_{m,m} \equiv +\frac{\partial I_m}{\partial(\mu_m / q)} \tag{12.7b}$$

Why do we have a negative sign for $m \neq n$, but not for $m=n$? The motivation can be appreciated by looking at a representative multi-terminal structure (Fig.12.6).

An increase in μ_1 leads to an incoming or positive current at terminal 1, but leads to ***outgoing*** or negative currents at the other terminals. The signs in Eq.(12.7a,b) are included to make the coefficients come out positive as we intuitively expect a conductance to be.

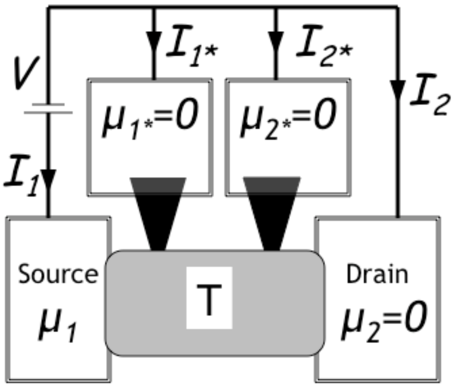


Fig.12.6. Thought experiment based on the four-terminal measurement set-up in Fig.12.1.

In terms of these conductance coefficients, we can write the current as

$$I_m = G_{m,m}\frac{\mu_m}{q} - \sum_{n\neq m} G_{m,n}\frac{\mu_n}{q} \tag{12.8}$$

The conductance coefficients must obey two important "sum rules" in order to meet two important conditions.

Firstly, the currents predicted by Eq.(12.8) must all be zero if all the μ's are equal, since there should be no external currents at equilibrium. This requires that

$$G_{m,m} = \sum_{n\neq m} G_{m,n} \tag{12.9a}$$

Secondly, for any choice of μ's, the currents I_m must add up to zero. This requires that

$$G_{m,m} = \sum_{n\neq m} G_{n,m} \tag{12.9b}$$

but it takes a little algebra to see this from Eq.(12.8). First we sum over all *m*

$$\sum_m I_m = 0 = \sum_m G_{m,m}\frac{\mu_m}{q} - \sum_m \sum_{n\neq m} G_{m,n}\frac{\mu_n}{q}$$

and interchange the indices *n* and *m* for the second term to write

$$0 = \sum_m G_{m,m}\frac{\mu_m}{q} - \sum_m \sum_{n\neq m} G_{n,m}\frac{\mu_m}{q}$$

which can be true for all choices of μ_m only if Eq.(12.9b) is satisfied. We can combine Eqs.(12.9a, b) to obtain the "sum rule" succinctly:

$$G_{m,m} \;=\; \sum_{n \neq m} G_{m,n} \;=\; \sum_{n \neq m} G_{n,m} \qquad (12.10)$$

Making use of the sum rule (Eq.(12.10)) we can re-write the first term in Eq.(12.8) to obtain Eq.(12.3):

$$I_m \;=\; (1/q)\sum_n G_{m,n}(\mu_m - \mu_n) \qquad \text{(same as Eq.(12.3))}$$

Note that it is not necessary to restrict the summation to *n≠m*, since the term with *n=m* is zero anyway. An alternate form that is sometimes useful is to write

$$I_m \;=\; \sum_n g_{m,n}\frac{\mu_n}{q} \qquad (12.11)$$

where the response coefficients defined as

$$g_{m,n} \;\equiv\; -G_{m,n}\,, \quad m \neq n \qquad (12.12a)$$

$$g_{m,m} \;\equiv\; G_{m,m} \qquad (12.12b)$$

The sum rule in Eq.(12.10) can be rewritten in term of this new response coefficient:

$$\sum_n g_{m,n} \;=\; \sum_n g_{n,m} \;=\; 0 \qquad (12.13)$$

12.2.1. Application

In Section 12.1 we analyzed the potential profile across a single scatterer with transmission probability *T*. Based on this discussion (Fig.12.5) we would expect that two non-invasive probes inserted before and after the scatterer should float to potentials *1-(T/2)* and *T/2* as indicated in Fig.12.7. But will Büttiker's approach get us the same result?

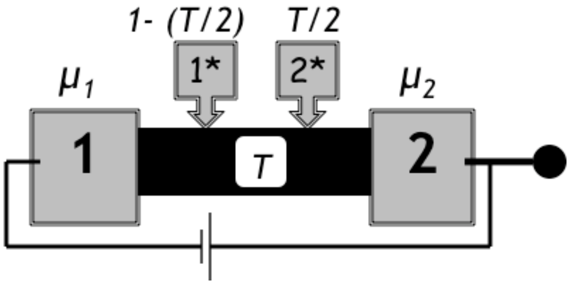


Fig.12.7. Based on Fig.12.5, we expect that two non-invasive probes inserted before and after a scatterer with transmission probability *T* to float to potentials *1-(T/2)* and *T/2* respectively.

We start from Eq.(12.11) noting that we have four currents and four potentials, labeled 1, 2, 1* and 2*:

$$\begin{Bmatrix} I_1 \\ I_2 \\ I_{1*} \\ I_{2*} \end{Bmatrix} = \frac{Mq}{h} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{Bmatrix} \mu_1 \\ \mu_2 \\ \mu_{1*} \\ \mu_{2*} \end{Bmatrix} \tag{12.14}$$

where A,B,C and D are each (2x2) matrices.

Since $$\begin{Bmatrix} I_{1*} \\ I_{2*} \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$$

we have $$\begin{Bmatrix} \mu_{1*} \\ \mu_{2*} \end{Bmatrix} = -D^{-1} C \begin{Bmatrix} \mu_1 \\ \mu_2 \end{Bmatrix} \tag{12.15}$$

Now we can write [C] and [D] in the form

$$\begin{bmatrix} C & D \end{bmatrix} = \begin{bmatrix} -t_1 & -t_2 & r & 0 \\ -t'_2 & -t'_1 & 0 & r' \end{bmatrix} \tag{12.16}$$

where the elements t_1, t_2, t_1' and t_2' of the matrix [C] can be visualized as the probability that an electron transmit from 1 to 1*, 2 to 1*, 2 to 2* and 1 to 2* respectively as sketched in Fig.12.8.

We have assumed that both probes 1* and 2* are weakly coupled so that any direct transmission between them can be ignored. The sum rule in Eq.(12.13) then requires that

$$r = t_1 + t_2$$

and

$$r' = t_2' + t_1' \qquad (12.17)$$

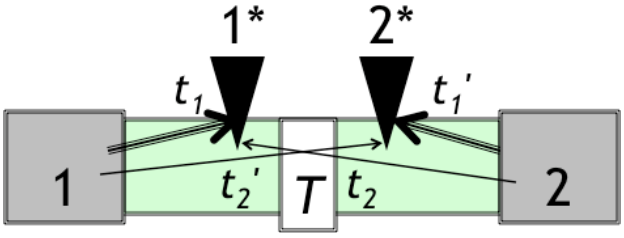


Fig.12.8. The elements t_1, t_2, t_1' and t_2' of the matrix [C] can be visualized as the probability that an electron transmit from 1 to 1*, 2 to 1*, 2 to 2* and 1 to 2* respectively.

Using Eqs.(12.15)-(12.17) we have

$$\mu_{1*} = \frac{t_1}{t_1 + t_2}\mu_1 + \frac{t_2}{t_1 + t_2}\mu_2 \qquad (12.18a)$$

and

$$\mu_{2*} = \frac{t_2'}{t_1' + t_2'}\mu_1 + \frac{t_1'}{t_1' + t_2'}\mu_2 \qquad (12.18b)$$

So far we have kept things general, making no assumptions other than that of weakly coupled probes. Now we note that for our problem (Fig.12.8), t_1 can be written as

$$t_1 \;=\; \tau \;+\; (1-T)\tau \qquad \text{(12.19a)}$$

since an electron from 1 has a probability of τ to get into probe 1* directly plus a probability of *1-T* times τ to get reflected from the scatterer and then get into probe 1*.

Similarly t_2 can be written as

$$t_2 \;=\; T\,\tau \qquad \text{(12.19b)}$$

since an electron from 2 has to cross the scatterer (probability *T*) and then enter the weakly coupled probe 1* (probability τ) . Similarly we can argue $t_1 = t_1'$, $t_2 = t_2'$.

Using Eqs.(12.19a,b) and setting $\mu_1=1$, $\mu_2=0$, we then obtain from Eqs.(12.18a,b),

$$\mu_{1*} \;=\; 1-(T/2)$$

$$\mu_{2*} \;=\; T/2$$

in agreement with what we expected from the last section (Fig.12.7). As mentioned earlier, this is reassuring since the Büttiker formula deals only with terminal quantities, bypassing the subtleties of non-equilibrium electrochemical potentials.

However, the real strength of Eq.(12.3) lies in its model-independent generality. It should be valid in the linear response regime for all conductors, simple and complex, large and small. The conductances G_{mn} in Eq.(12.3) can be calculated from a microscopic transport model like the Boltzmann equation introduced in Lecture 7 or the quantum transport model discussed in Part three of these lectures. Sometimes they can even be guessed and as long as we are careful about not violating the sum rules we should get reasonable results.

12.2.2. Is Eq.(12.3) obvious?

Some might argue that Eq.(12.3) is not really telling us much. After all, we can always view any structure as a network of effective resistors like the one shown in Fig.12.9 for three terminals? Wouldn't the standard circuit equations applied to this network give us Eq.(12.3)?

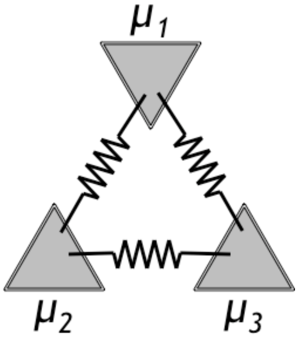


Fig.12.9. The Büttiker formula (Eq.(12.3)) can be visualized as a network of resistors, only if the conductances are reciprocal, that is, if $G_{mn} = G_{nm}$.

The answer is "yes" if we consider only normal circuits for which electrons transmit just as easily from *m* to *n* as from *n* to *m* so that

$$G_{m\leftarrow n} = G_{n\leftarrow m}$$

where we have added the arrows in the subscripts to denote the standard convention for the direction of electron transfer. Eq.(12.3), however, goes far beyond such normal circuits and was intended to handle conductors in the presence of magnetic fields for which

$$G_{m\leftarrow n} \neq G_{n\leftarrow m}$$

For such conductors, Eq.(12.3) is not so easy to justify. Indeed if we were to reverse the subscripts *m* and *n* in Eq.(12.3) to write

$$I_m = (1/q) \sum_n G_{nm} (\mu_m - \mu_n) \qquad \text{WRONG!}$$

it would not even be correct. Its predictions would be different from those of Eq.(12.3) for multi-terminal non-reciprocal circuits of the type we will discuss in the next Lecture.

Lecture 13

Hall effect

13.1. Why n- and p- Conductors Are Different
13.2. Spatial Profile of Electrochemical Potential
13.3. Measuring the Potential
13.4. Non –Reciprocal Circuits

Let me briefly explain what the Hall effect is about. Consider a two-dimensional conductor (see Fig.13.1) carrying current, subject to a perpendicular magnetic field along the *y*-direction which exerts a force on the electrons perpendicular to its velocity.

$$\vec{F} = \frac{d\vec{p}}{dt} = \quad -q\vec{v}\, x\, \vec{B} \tag{13.1}$$

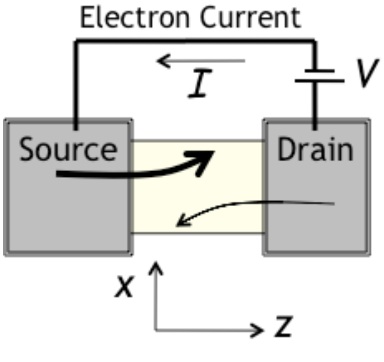


Fig.13.1.
A magnetic field *B* in the y-direction makes electrons from the source veer “up”wards.

This would cause an electron from the source to veer "up"wards and an electron from the drain to veer "down"wards as shown. Since there are more electrons headed from source to drain, we expect electrons to pile up on the top side causing a voltage V_H to develop in the *x*-direction transverse to current flow (see Fig.13.2).

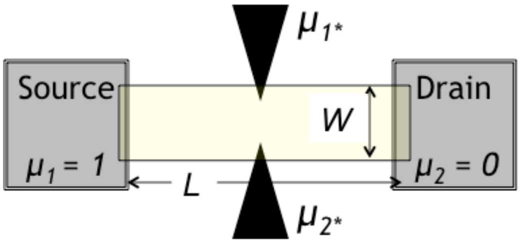


Fig.13.2. Basic structure with two voltage probes whose potential difference measure the Hall voltage, $qV_H = \mu_{1*} - \mu_{2*}$.

The Hall effect has always been important since its discovery around 1880, and has acquired a renewed importance since the discovery of the quantum Hall effect in 1980 at high magnetic fields. In these lectures we have seen the conductance quantum q^2/h appear repeatedly and it is very common in the context of nanoelectronics and mesoscopic physics. But the quantum Hall effect was probably the first experimental observation where it played a clear identifiable role and the degree of precision is so fantastic that the National Bureau of Standards uses it as the resistance standard. We will talk briefly about it later at the end of this lecture.

I will also talk about an interesting property that magnetic (B) fields introduce into any circuit, namely that of non-reciprocity, making it difficult to visualize in terms of ordinary resistances. This is particularly important, now that there is the possibility that a new class of materials called "topological insulators" might show such non-reciprocity even without B-fields. But I am getting ahead of myself.

For the moment let us focus on the conventional Hall effect at low magnetic fields. One reason it is particularly important is that it changes sign for n- and p-type materials, thus providing an experimental technique for telling the difference. Like the thermoelectric current discussed in Lecture 10, this too is commonly explained by invoking "holes" as the positive charge carriers in p-type materials. Once again, this is not satisfactory because it is really the electrons that move even in p-type conductors. Both n-type and p-type conductors have negative charge carriers.

For the thermoelectric effect we saw that its sign is determined by the slope of the density of states *D(E),* that is whether it is an increasing or a decreasing function of the energy E. By contrast, the sign of the Hall effect is determined by the sign of the effective mass defined as the ratio of the momentum *p* to the velocity *dE/dp* (see Lecture 5). As a result although the magnetic force (see Eq.(13.1)) is the same for both n- and p-type conductors, giving the same *dp/dt*, the resulting *dv/dt* has opposite signs. This makes electrons in p-conductors veer in the opposite direction giving rise to a Hall voltage of the opposite sign.

Clearly this requires the existence of an *E(p)* relation underlying the density of states function. Perhaps it is for this reason that amorphous semiconductors which lack a well-defined *E(p)* often show strange results for the Hall effect and yet show reasonable thermoelectric effect.

The simple theory of the Hall effect given in freshman physics texts goes like this. First the current is written as

$$I \;=\; q\,(N/L)\,v_d \tag{13.2a}$$

with the drift velocity given by the product of the mobility and the electric field in the *z*-direction:

$$v_d \;=\; \bar{\mu}\;(V/L) \tag{13.2b}$$

These two relations are normally combined to yield the Drude formula (see Eq.(5.1)) that we discussed in Lecture 5

$$\frac{I}{V} \;=\; \underbrace{q\frac{N}{WL}\bar{\mu}}_{\sigma}\,\frac{W}{L} \tag{13.2c}$$

For the Hall effect, it is argued that an electric field V_H/W must appear in the *x*-direction to offset the magnetic force

$$V_H / W = v_d B \tag{13.3}$$

Combining Eq.(13.3) with Eq.(13.2a) one obtains the standard expression for the Hall resistance

$$R_H = \frac{V_H}{I} = \frac{B}{q(N/LW)} \tag{13.4}$$

One reason the Hall effect is widely used is that Eq.(13.4) allows us to determine the electron density *N/LW* from the slope of the Hall resistance versus *B*-field curve.

This looks like a straightforward transparent theory for a well-established effect. What more could we add to it? The main concern we have about this derivation is the same concern that we voiced regarding the Drude formula, namely that if electric field were indeed what drives currents then all electrons should feel its effect.

Indeed Eq.(13.4) for the Hall resistance conveys the impression that the Hall effect depends on the total electron density *N/LW* over all energies. But we believe this is not correct. Like the other transport coefficients we have discussed, the Hall resistance too is a "Fermi surface property" that depends only on the electrons in an energy window ~ a few *kT* around $E=\mu_0$ and not on the total number of electrons obtained by integrating over energy.

We will show that the Hall resistance for a single energy channel of an elastic resistor is given by

$$R_H(E) = \frac{2B\,LW}{q\,D(E)v(E)p(E)} \tag{13.5a}$$

which can be averaged over an energy window of a few *kT* around $E=\mu_0$ using our standard broadening function:

$$\frac{1}{R_H} = \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) \frac{1}{R_H(E)} \tag{13.5b}$$

Note that in general we should integrate the conductance *1/R*$_H$ rather than the resistance R_H since different energy channels all have the same voltage so that they conduct "in parallel" as circuit theorists would put it.

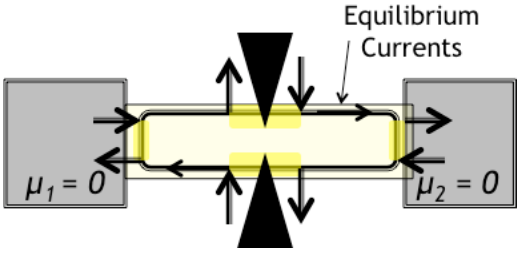
Eq.(13.5) can be reduced to the standard result (Eq.(13.4)) by making use of the single band relation obtained in Lecture 5

$$D(E)\nu(E)p(E) = N(E)\ .d \qquad \text{(same as Eq.(5.4))}$$

with $d= 2$ for a two-dimensional conductor and relating the average of *N(E)* to the total number of electrons as we did in Section 5.3. But if the single band relation (Eq.(5.7)) is not applicable one should use the expression in Eq.(13.5) rather than Eq.(13.4).

In any case, Eq.(13.5b) shows that the effect really does not involve electrons at all energies. One reason this point causes some confusion is the existence of equilibrium currents inside the sample in the presence of a magnetic field which involve all electrons at all energies.

Fig.13.3. Equilibrium currents can exist in any conductor in the presence of a magnetic field.



However, these are dissipationless currents of the type that exist even if we put a hydrogen atom in a magnetic field and have nothing to do with the transport coefficients we are talking about. In any transport model it is important to eliminate these non Fermi surface currents. A similar issue arises with respect to spin currents even without any magnetic fields (see Rashba 2003, Lecture 22).

Getting back to the problem of determining the Hall voltage, as we saw in the last Lecture there are two approaches: (a) calculate the non-equilibrium electrochemical potential inside the conductor or (b) treat it as a four terminal structure using the Büttiker equation. We will discuss both approaches sequentially in Sections 13.2 and 13.3 respectively after we have briefly discussed the dynamics of electrons in a *B*-field in Section 13.1.

13.1.Why n- and p- Conductors Are Different

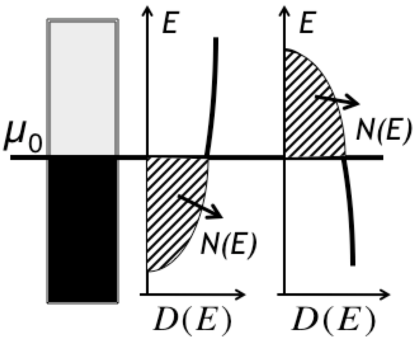


Fig.13.4
The Hall resistance changes sign for n- and p-type conductors and is inversely proportional to *N(E),* at $E=\mu_0$.

Why do n- and p-type conductors show opposite signs of the Hall effect? The basic difference is that in n-type conductors, the velocity is parallel to the momentum, while in p-type conductors, it is anti-parallel because $v = dE/dp$ and in p-type conductors, the energy decreases with increasing p (Fig.13.4).

To see why the relative sign of p and v matters, let us consider the magnetic force described by Eq.(13.1) in a little more detail.

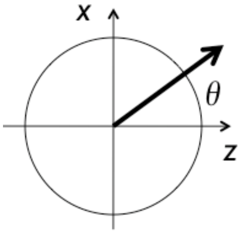
For any isotropic *E(p)* relation, the velocity and momentum are collinear (parallel or anti-parallel) pointing, say at an angle θ to the *z*-axis, so that

$$\vec{p} = \quad p\cos\theta\,\hat{z} + p\sin\theta\,\hat{x}$$

$$\vec{v} = \quad v\cos\theta\,\hat{z} + v\sin\theta\,\hat{x}$$

Inserting into Eq.(13.1) we obtain

$$\frac{d\theta}{dt} = \frac{qvB}{p} \qquad (13.6)$$



showing that the angle θ increases linearly with time so that the velocity and momentum vectors rotate uniformly in the *z-x* plane.

But the sense of rotation is opposite for n- and p-type conductors because the ratio *p/v* has opposite signs. This is the ratio we defined as mass (see Eq.(5.6)) and is constant for parabolic dispersion (Eq.(5.7a)).

$$\omega_c = \left|\frac{qvB}{p}\right|_{E=\mu_0} = \left|\frac{qB}{m}\right|_{E=\mu_0} \qquad (13.7)$$

But for linear dispersion (Eq.(5.7b)) the mass increases with energy, so that the cyclotron frequency decreases with increased carrier density, as is observed in graphene.

The magnetic field tries to make the electron would go round and round in a circle with an angular frequency ω_c. However, it gets scattered after a mean free time τ, so that if $\omega_c \tau << 1$ the electron never really gets to complete a full rotation. This is the low field regime where the Hall resistance in given by Eq.(13.5), while the high field regime characterized by $\omega_c \tau >> 1$ leads to the quantum Hall effect mentioned earlier.

Let us now discuss our first approach to the problem of determining the Hall resistance (Eq.(13.5)) based on looking at the non-equilibrium electrochemical potentials inside the conductor.

13.2.Spatial Profile of Electrochemical Potential

As I mentioned earlier, the textbook derivation of the Hall resistance (Eq.(13.4)) looks fairly straightforward, but we are attempting to provide a different expression (Eq.(13.5)) motivated by the same reasons that prompted us to describe an alternative expression for the conductivity back in Lecture 5.

In our elastic resistor model, the role of the drift velocity in the text-book derivation is played by the potential separation

$$\delta\mu = \mu^{+} - \mu^{-}$$

between drainbound and sourcebound states, so that instead of Eq.(13.2a)) we have from Lecture 6 (see Eq.(6.16))

$$I(E) \quad = \quad \frac{q}{h} M(E)\left(-\frac{\partial f_0}{\partial E}\right)\delta\mu \tag{13.8a}$$

with

$$\frac{M(E)}{h} \quad = \quad \frac{D(E)v(E)}{\pi L} \tag{13.8b}$$

where we have used the result for 2D conductors from Eq.(4.13).

Instead of Eq.(13.2b), we have the potential separation $\delta\mu$ related to the applied voltage by (see Eq.(9.1))

$$\delta\mu \quad = \quad \frac{qV\lambda}{L+\lambda} \quad \cong \quad q\lambda\frac{V}{L} \tag{13.8c}$$

Just as Eqs.(13.2a,b) yield the Drude expression for the conductivity, similarly Eqs.(13.8a,b) can be combined to yield the more general expression that we discussed in Lecture 5 (see Eq.(5.15)).

For the Hall effect we need a replacement for Eq.(13.3)

$$V_H / W \quad = \quad v_d B$$

which we will show is given by

$$\frac{V_H}{W} = \frac{2}{\pi}\frac{\delta\mu}{p}B \tag{13.9}$$

Eq.(13.9) together with Eqs.(13.8a,b) gives us the result for Hall resistance stated earlier in Eq.(13.5).

Unfortunately we do not have a one-line argument for Eq.(13.9) like the one used for Eq.(13.3). Instead I need to describe a two-page argument using the BTE discussed in Lecture 7.

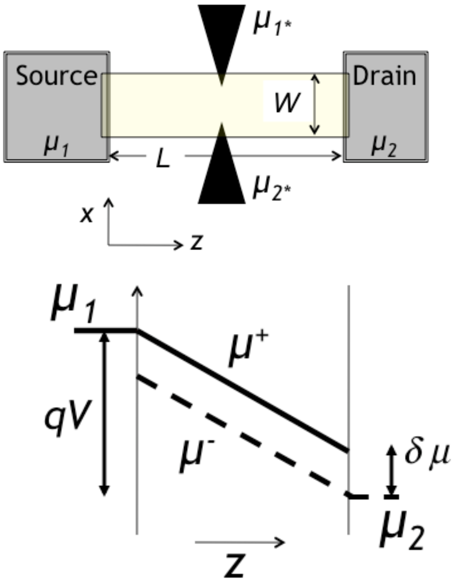


Fig.13.5.
Spatial variation of $\mu^{\pm}$ along z.

Back in Lecture 6 we obtained a solution for a subset of this problem based on a solution of

$$v_z\frac{\partial\mu}{\partial z} = -\frac{\mu-\mu_0}{\tau} \tag{13.10}$$

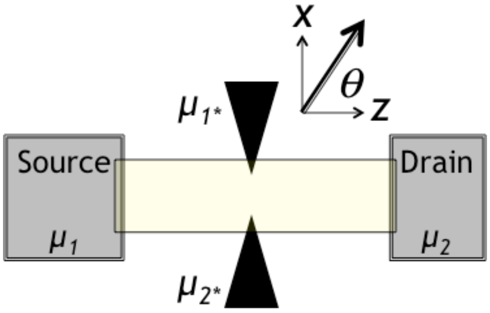
and obtained the solutions for the electrochemical potentials μ^+ and μ^- sketched above in Fig.13.5.

The solutions could be written in the form

$$\mu(z,\theta) \quad = \quad \bar{\mu}(z) + \frac{2}{\pi}\,\delta\mu\,\cos\theta \tag{13.11}$$

where we have separated out a z-dependent part $\bar{\mu}(z)$ from the momentum-dependent part at a specific location, z. The latter needs a little discussion.

Since we are discussing an elastic resistor for which electrons have a fixed energy E and hence a fixed momentum p, it is convenient to use cylindrical coordinates for the momentum (p, θ) instead of (p_x , p_z). Suppose we were dealing only with electrons at a fixed angle θ (or the exact opposite direction) then making use of Eq.(13.8c) we could write



$$\mu(z) \quad = \quad \bar{\mu}(z) + \frac{qV}{L}\,v\tau\cos\theta \tag{13.12a}$$

noting that the mean free path in this case is simply

$$2v_z\tau = 2v\tau\,\cos\theta\;.$$

Comparing with Eq.(13.11), we have

$$\delta\mu \quad \approx \quad \frac{qV}{L}\,\frac{\pi}{2}\,v\tau \tag{13.12b}$$

The question is how we expect the solution in Eq.(13.11) to change when we turn on the magnetic field so that it exerts a force on the electrons. For this we could use a linearized version of the BTE like Eq.(7.17), but retaining both z- and x- components since we have a two-dimensional problem

$$v_x \frac{\partial \mu}{\partial x} + v_z \frac{\partial \mu}{\partial z} + F_x \frac{\partial \mu}{\partial p_x} + F_z \frac{\partial \mu}{\partial p_z} = -\frac{\mu - \mu_0}{\tau} \tag{13.13}$$

Note that Eq.(13.10) is a “subset” of this equation which includes three extra terms. The last two coming from the magnetic force (Eq.(13.1)) can be written as

$$F_x \frac{\partial \mu}{\partial p_x} + F_z \frac{\partial \mu}{\partial p_z} = \vec{F}.\vec{\nabla}_p \mu = \frac{F_\theta}{p} \frac{\partial \mu}{\partial \theta} + F_r \frac{\partial \mu}{\partial p}$$

The force due to a magnetic field has no radial component, only a θ component:

$$F_r = 0 \ , \ F_\theta = -qvB$$

This is because the velocity is purely radial and so when we take a cross-product with a magnetic field in the z-direction, we get a vector that is purely in the θ-direction. This allows us to rewrite Eq.(13.13) in the form

$$v_x \frac{\partial \mu}{\partial x} + v_z \frac{\partial \mu}{\partial z} - \frac{qvB}{p} \frac{\partial \mu}{\partial \theta} = -\frac{\mu - \mu_0}{\tau} \tag{13.14}$$

Noting that our solution in Eq.(13.11) satisfies Eq.(13.10), it is easy to check that if we add an extra term varying only with x to it, the resulting expression

$$\mu(z,\theta,x) = \bar{\mu}(z) + \frac{2}{\pi} \delta\mu \cos\theta - \frac{2}{\pi} \frac{\delta\mu}{p} qBx \tag{13.15}$$

will satisfy Eq.(13.14). From this solution we obtain Eq.(13.9) by writing

$$-qV_H = \mu(x = W) - \mu(x = 0) = -\frac{2}{\pi} \frac{\delta\mu}{p} qBW$$

13.3.Measuring the Potential

Let us now look at how we could calculate the Hall voltage using the Büttiker equation for a four-terminal conductor with two current probes and two voltage probes (Fig.13.6) not unlike the one we discussed in the last Lecture. But the two probes are now on two sides of the conductor and would normally not register any potential difference. But when an applied potential causes electrons to flow from left to right, the applied *B*-field causes them to veer upwards or downwards making μ_{1*} different from μ_{2*} and we would like to calculate the resulting voltage for a given current.

The basic result from the last Lecture still holds:

$$\mu_{1*} \;=\; \frac{t_1}{t_1+t_2}\,\mu_1 \;+\; \frac{t_2}{t_1+t_2}\,\mu_2 \qquad \text{(same as (12.18a))}$$

and

$$\mu_{2*} \;=\; \frac{t_2\,'}{t_1\,'+t_2\,'}\,\mu_1 \;+\; \frac{t_1\,'}{t_1\,'+t_2\,'}\,\mu_2 \qquad \text{(same as (12.18b))}$$

as long as we interpret the various probabilities according to Fig.13.6.

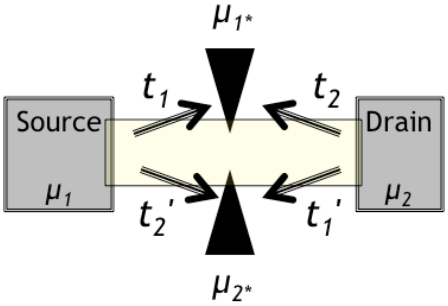


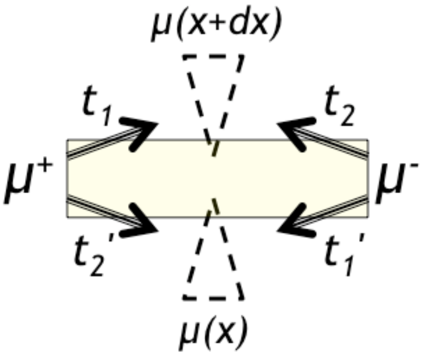
Fig.13.6. We can use the results from Eq.(12.18) as long as we visualize the different probabilities as shown.

These probabilities can be calculated numerically using either semiclassical or numerical models, but I do not have a simple analytical argument showing that this indeed yields a Hall voltage

$$-qV_H \quad = \quad \mu_{1*} - \mu_{2*}$$

in agreement with our basic result in Eq.(13.9).

To obtain Eq.(13.9) we could apply our basic result in Eqs.(12.18a,b) to a thin slice of the conductor less than a mean free path long somewhere in the middle of a long channel with electrons from the left having a potential μ^+ and those from the right having a potential μ^-.



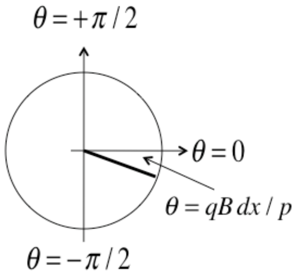
We can then use Eqs.(12.18a,b) to write down the potential on the upper and lower edges which we could treat as conceptual probes 1* and 2*:

$$\mu(x+dx) \quad = \quad \frac{t_1}{t_1+t_2}\,\mu^+ \;+\; \frac{t_2}{t_1+t_2}\,\mu^- \tag{13.16a}$$

and

$$\mu(x) \quad = \quad \frac{t_2\,'}{t_1\,'+t_2\,'}\,\mu^+ \;+\; \frac{t_1\,'}{t_1\,'+t_2\,'}\,\mu^- \tag{13.16b}$$

To write down the probabilities appearing in Eqs.(13.16a,b) we note that without a magnetic field all electrons with a velocity lying within the range of angles



$$0 \quad < \quad \theta \quad < \quad +\pi/2$$

will go upwards while those with a velocity in the range

$$-\pi/2 \quad < \quad \theta \quad < \quad 0$$

will go downwards.

The magnetic field causes electrons to bend upwards by an angle obtained by multiplying the angular rate given by Eq.(13.6) with the time dx/v it takes to cross a distance dx:

$$\frac{qvB}{p}\frac{dx}{v} = \frac{qB\,dx}{p}$$

With this picture in mind we can write

$$t_1 = t_1' \sim \frac{\pi}{2} + \frac{qB}{p}dx \qquad (13.17a)$$

$$t_2 = t_2' \sim \frac{\pi}{2} - \frac{qB}{p}dx \qquad (13.17b)$$

and use it in Eqs.(13.16a,b) to obtain

$$\mu(x+dx)-\mu(x) = \frac{2qB\,dx}{\pi p}\left(\mu^+ - \mu^-\right)$$

so that
$$\frac{d\mu}{dx} = \frac{2qB}{\pi p}\delta\mu$$

in agreement with Eq.(13.15) and hence with Eq.(13.9).

13.3.1. Edge states:

Some of the most illuminating use of the Büttiker approach is in the quantum Hall regime where the *B*-fields are so high that no electron from the source ever gets to probe 2*, and no electron from the drain gets to 1*. This makes

$$\frac{V_H}{(\mu_1-\mu_2)/q} = \frac{t_1 - t_2'}{t_1 + t_2'} = 1$$

since $t_2 = t_2' = 0$ so that the Hall voltage becomes equal to the applied voltage making the Hall resistance equal to the ordinary two-terminal resistance.

Interestingly, in this regime this resistance is given by

$$R = \frac{h}{q^2 i} \tag{13.18}$$

where i is an integer to a fantastic degree of precision, making this a resistance standard used by the National Bureau of Standards. It is as if we have an unbelievably perfect ballistic conductor whose only resistance is the interface resistance. Since these conductors are often hundreds of micrometers long, this perfect ballisticity is amazing and was recognized with a Nobel prize in 1985 (von Klitzing K. et al. 1980)

In these lectures we have seen the conductance quantum q^2/h appear repeatedly and it is very common in the context of nanoelectronics and mesoscopic physics. But the quantum Hall effect was probably the first experimental observation where it played a clear identifiable role.

The simplest picture that helps understand it is the so-called "skipping orbits" (Fig.13.7) that result in a "divided" electronic highway with drainbound electrons so well-separated from the sourcebound electrons that backscattering is extremely unlikely, resulting in an incredibly ballistic conductor.

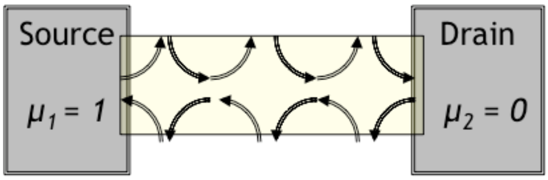


Fig.13.7. Skipping orbits in high B-fields leads to a "divided highway" with drainbound electrons on one side and sourcebound electrons on the other.

This simple picture, however, is a little too simple. It does not for example tell us the significance of the integer i in Eq.(13.18) which requires some input from quantum mechanics, as we will see in Lecture 21.

13.4.Non-Reciprocal Circuits

This may be a good place to raise an interesting property of conductors with non-reciprocal transmission of the type expected from edge states. Consider the structure shown in Fig.13.8 with a B-field that makes an electron coming in from contact 2 veer towards contact 1, but makes an electron coming from contact 1 veer away from contact 2.

Is $G_{1,2} \neq G_{2,1}$?

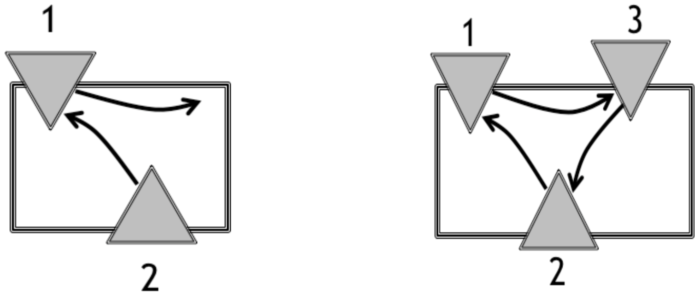


Fig.13.8. A magnetic field makes an electron coming in from contact 2 veer towards contact 1, but makes an electron coming from contact 1 veer away from contact 2. Is $G_{1,2} \neq G_{2,1}$? Yes, if there are more than two terminals, but not in a two-terminal circuit.

Not in the linear response regime as evident from the sum rule discussed in Lecture 12 (Eq.(12.10)) which for a structure with two terminals requires that

$$G_{1,1} = G_{1,2} = G_{2,1} .$$

However, there is no such requirement for a structure with more than two terminals. For example with three terminals, Eq.(12.10) tells us that

$$G_{1,1} \quad = \quad G_{1,2} + G_{1,3} \quad = \quad G_{2,1} + G_{3,1}$$

which does not require $G_{1,2}$ to equal $G_{2,1}$. The effects of such non-reciprocal transmission have been observed clearly with "edge states" in the quantum Hall regime.

This idea of "edge states" providing unidirectional ballistic channels over macroscopic distances is a very remarkable effect, but it has so far been restricted to low temperatures and high B-fields making it not too relevant from an applied point of view. That may change with the advent of new materials like "topological insulators" which show edge states even without B-fields.

But can we have non-reciprocal transmission without magnetic fields? In general the conductance matrix (which is proportional to the transmission matrix) obeys the Onsager reciprocity relation

$$G_{n,m}(+B) \quad = \quad G_{m,n}(-B) \tag{13.19}$$

requiring the current at n due to a voltage at m to equal the current at m due to a voltage at n with any magnetic field reversed. This is one of the deep principles of statistical mechanics which is usually proved for large conductors starting from the Kubo formula (Lecture 15).

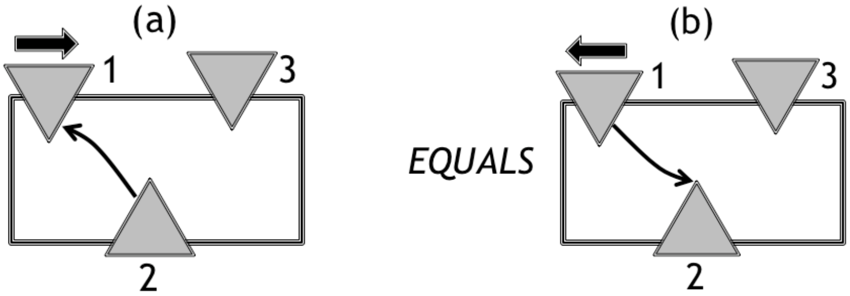
Doesn't this Onsager relation require the conductance to be reciprocal

$$G_{m,n} \quad = \quad G_{n,m}$$

when $B=0$? The answer is yes if the structure does not include magnetic materials. Otherwise we need to reverse not just the external magnetic field but the internal magnetization too.

$$G_{n,m}(+B,+M) \quad = \quad G_{m,n}(-B,-M) \tag{13.20}$$

For example if one contact is magnetic, Onsager relations would require the $G_{1,2}$ in structure (a) to equal $G_{2,1}$ in structure (b) with the contact magnetization reversed as sketched above. ***But that does not mean $G_{1,2}$ equals $G_{2,1}$ in the same structure,*** (a) or (b).



And so based on our current understanding a "topological insulator" which is a non-magnetic material could not show non-reciprocal conductances at zero magnetic field with ordinary contacts, but might do so if magnetic contacts were used. But this is an evolving story whose ending is not yet clear.

What has become very clear in the last twenty years is the operation of magnetic contacts, and that is what we will discuss next.

Lecture 14

Spin valve

One of the major developments in the last two decades is the spin valve, a device with two magnetic contacts (Fig.14.1) If they are magnetized in the same direction (parallel configuration, P) the resulting resistance is lower than if they are magnetized in opposite directions (anti-parallel configuration, AP). Since its first demonstration in 1988, it rapidly found application as a "reading" device to sense the information stored in a magnetic memory and the discovery was recognized with a Nobel prize in 2007.

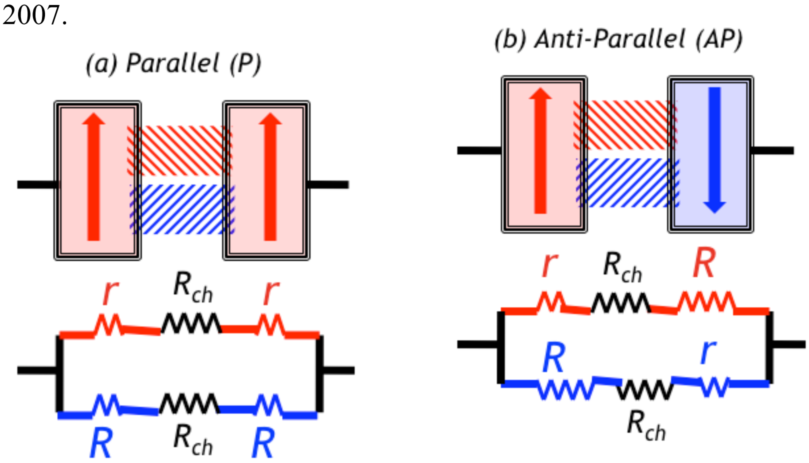


Fig.14.1. Spin valve: (a) Parallel (P) configuration. (b) Anti-Parallel (AP) configuration.

So far we have only mentioned spin as part of a "degeneracy factor, g" (Section 5.4), the idea being that electronic states always come in pairs, one corresponding to each spin. We could call these “up” and “down” or “left” and “right” or even “red” and “blue” as we have done in Fig.14.1. Note that the two spins are not spatially separated even though we have separated the red and the blue channel for clarity. Ordinarily the two channels are identical and we can calculate the conductance due to one and remember to multiply by two.

But in spin valve devices the contacts are magnets that treat the two spin channels differently and the operation of a spin valve can be understood in fairly simple terms if we postulate that each spin channel has a different interface resistance with the magnet depending on whether it is parallel (majority spin) or anti-parallel (minority spin) to the magnetization.

If we assume the interface resistance for majority spins to be r and for minority spins to be R $(r < R)$ we can draw simple circuit representations for the P and AP configurations as shown, with R_{ch} representing the channel resistance. Elementary circuit theory then gives us the resistance for the parallel configuration as

$$R_P = \left(\frac{1}{2r+R_{ch}} + \frac{1}{2R+R_{ch}} \right)^{-1}$$

and that for the anti-parallel configuration as

$$R_{AP} = \frac{r+R+R_{ch}}{2}$$

The essence of the spin valve device is the difference between R_P and R_{AP} and we would expect this to be most pronounced when the channel resistance is negligible and everything is dominated by the interfaces. Fig.14.2 shows the variation in the magnetoresistance (MR, defined below) as a function of the channel resistance (per spin) R_{ch} normalized to $r+R$ assuming $P=0.5$. Note how the MR dies out quickly once the normalized R_{ch} increases beyond say ~5.

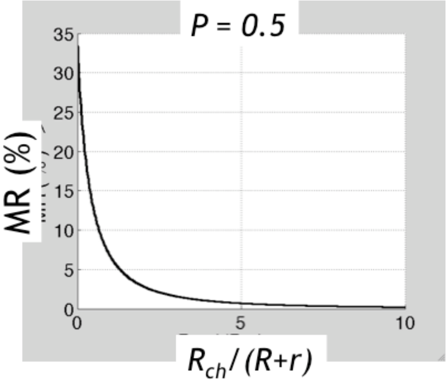


Fig.14.2. Variation in magnetoresistance (MR) as a function of the normalized channel resistance.

If we set $R_{ch} = 0$, we obtain a simple result for the maximum MR

$$MR \equiv \frac{R_{AP}}{R_P} - 1 \quad = \quad \frac{(R-r)^2}{4rR}, \quad \mathit{if}\ R_{ch} = 0 \tag{14.1}$$

which can be written in terms of the polarization:

$$P \quad \equiv \quad \frac{R-r}{R+r} \tag{14.2}$$

$$MR \quad = \quad \frac{P^2}{1-P^2}, \quad \mathit{if}\ R_{ch} = 0 \tag{14.3a}$$

I should mention here that the expression commonly seen in the literature has an extra factor of 2

$$MR \quad = \quad \frac{2P^2}{1-P^2} \tag{14.3b}$$

which is applicable to magnetic tunnel junctions (MTJ's) that use short tunnel junctions as channels instead of the metallic channels we have been discussing. We get this extra factor of 2, if we assume that two resistors R_1 and R_2 in series give a total resistance of KR_1R_2, K being a constant, instead of the standard result $R_1 + R_2$ expected of ordinary Ohmic resistors. The product dependence captures the physics of tunnel resistors.

The point to note is that the key to spin valve operation is the different interface resistance associated with each spin for magnetic contacts. In the ***Section 14.1*** I will try to elaborate on the origin of this difference based on the approach developed in this book. Next (***Section 14.2***) we will discuss how magnetic contacts lead to nonequilibrium spin potentials which ties in with the other important theme we have been trying to stress, namely that nanoscale device design of the future will need to emphasize the control of electrochemical potentials through creative design of contacts (Lecture 9).

While spin valves showed us how to use magnets to inject spins and control spin potentials, later researchers have shown how to use non-equilibrium spins to turn nanoscale magnets thus integrating spintronics and magnetics into a single and very active area of research with exciting possibilities that are yet to be explored.

In ***Section 14.3*** we will try to give the reader a feeling for this intriguing interplay of spins and magnets. Finally in ***Section 14.4*** we will try to illustrate the interesting dichotomy presented by spins, where some aspects can be understood in semiclassical terms, while others require a quantum viewpoint which we will take up in part three (Lecture 22).

14.1. Mode Mismatch and Interface Resistance

The original spin valve devices used metallic channels like copper and have developed rapidly since the original experiments in 1988. For many practical applications they have now been largely replaced by MTJ’s that use insulating oxides as channels, due to the much larger values of MR that have been achieved.

By contrast all efforts to use semiconductors as the channel material proved singularly unsuccessful till around the year 2000, when the cause for poor MR was identified as the high R_{ch} compared to $R+r$ and this led to the idea of deliberately increasing the interface resistances by inserting barrier layers as sketched in Fig.14.3 (See for example, review by Schmidt 2005).

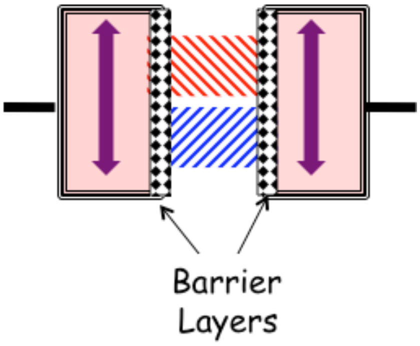


Fig.14.3. Barrier layers are deliberately inserted to increase the interface resistance when injecting spins into semiconductor channels.

Inserting a barrier layer is now a standard technique used by all experimentalists trying to inject spins into a semiconducting channel and so the “facts” seem quite clear. But why does it work?

The common explanation is that the barriers increase the interface resistances R , r thereby reducing the ratio $R_{ch}/(R+r)$ and increasing MR in accordance with Fig.14.2. However, it seems to us that if this were the whole story we should be able to increase the MR by reducing R_{ch} using a channel that is short enough to be ballistic (Length, L << mean free path, λ). But experimentally it seems clear that making the channel short does not help.

Fig.14.4. shows a sketch of the number of channels $M(E)$ (or the density of states, $D(E)$) for a magnetic contact and a non-magnetic channel plotted separately for upspin and downspin electrons to the right and to the left respectively.

In the non-magnetic channel, the two are identical, but in the magnetic contact the minority spin band is typically shifted up in energy making the number of modes at E=μ smaller for the minority spin M_{dn} than for the majority spin M_{up}. What are the interface resistances?

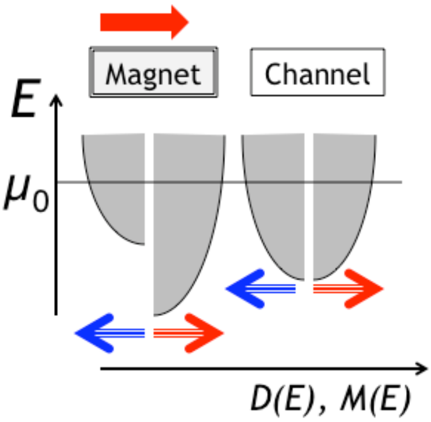


Fig.14.4. Sketch of D(E) or M(E) for magnetic contact and non-magnetic channel plotted separately for upspin and downspin electrons to the right and to the left respectively.

We will show shortly (Section 14.1.1) that the interface resistance at an interface between two materials with different numbers of modes ($M_1 > M_2$) is given by

$$R_{\text{int}} = \frac{h}{2q^2}\left(\frac{1}{M_2} - \frac{1}{M_1}\right) \tag{14.5}$$

If M_1 is much much greater than M_2, then

$$R_{\text{int}} \approx \frac{h}{2q^2 M_2}$$

which is the result we have discussed earlier (Fig.6.2) corresponding to ***good contacts*** ($M_1 >> M_2$).

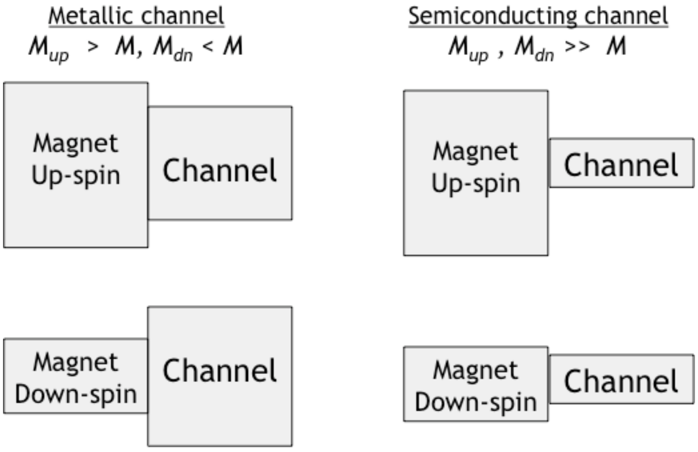
Typically for a metallic channel the number of modes in the channel has a value intermediate between the two. Ideally

$$M_{up} \;>>\; M \;>>\; M_{dn}$$

so that the magnet provides a good contact for the majority spin but not for the minority spin:

$$r \;=\; \frac{h}{2q^2}\left(\frac{1}{M}-\frac{1}{M_{up}}\right) \;\approx\; \frac{h}{2q^2 M}$$

$$R \;=\; \frac{h}{2q^2}\left(\frac{1}{M_{dn}}-\frac{1}{M}\right) \;\approx\; \frac{h}{2q^2 M_{dn}}$$



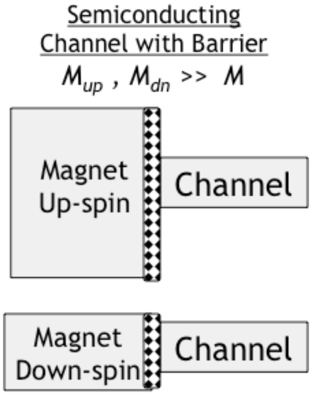
With semiconducting channels, on the other hand the number of modes in the channel is typically the smallest

$$M_{up} \;>\; M_{dn} \;>>\; M$$

so that $$r \;=\; R \;=\; \frac{h}{2q^2 M}$$

and the polarization P (see Eq.(14.2)) is zero.

In other words, it seems to us that the problem with spin injection into semiconducting channels is not just related to a high R_{ch} which could be eliminated with short ballistic channels, but is also related to the loss of distinction in the interface resistances R and r for the two spins. With 10 modes in a channel, it really does not matter whether the contact has 100 modes (minority spin) or 1000 modes (majority spin). Both are equally effective in keeping the channel well stocked with electrons.



Why does it help to insert a barrier? With a barrier present, the interface resistance is no longer given by Eq.(14.5) which applies only to good contacts where electrons can exit easily from the channel. But once a barrier is inserted it becomes harder and harder for the contact to keep the channel well-stocked and at some point the minority spin cannot do that any more.

For a tunneling barrier it is believed that the conductance is proportional to the product of the density of states and hence the number of modes on the two sides of the tunneling barrier so that instead of Eq.(14.5) for the interface resistance we can write

$$\frac{1}{R_{\text{int}}} = K M_1 M_2 \tag{14.6}$$

K being a constant. This seems reasonable if an electron from any mode on the left can transition into any mode on the right, but the exact conditions under which Eq.(14.5) changes to Eq.(14.6) need further discussion which we will not get into. We simply note that if we accept Eq.(14.6) we obtain

$$\frac{1}{r} = K M_{up} M \ , \quad \frac{1}{R} = K M_{dn} M$$

so that the polarization *P* (Eq.(14.2)) can now be sizeable irrespective of the number of modes *M* in the channel:

$$P = \frac{M_{up} - M_{dn}}{M_{up} + M_{dn}}$$

Of course the interface resistance values are larger than those for Ohmic interfaces described by Eq.(14.5).

14.1.1. Interface Resistance Due to Mode Mismatch

Let me briefly explain where Eq.(14.5) for the interface resistance comes from. Consider an interface between two channels with different mode numbers $M_1 > M_2$ with large contacts (effectively infinite number of modes) at both ends as shown in Fig.14.5.

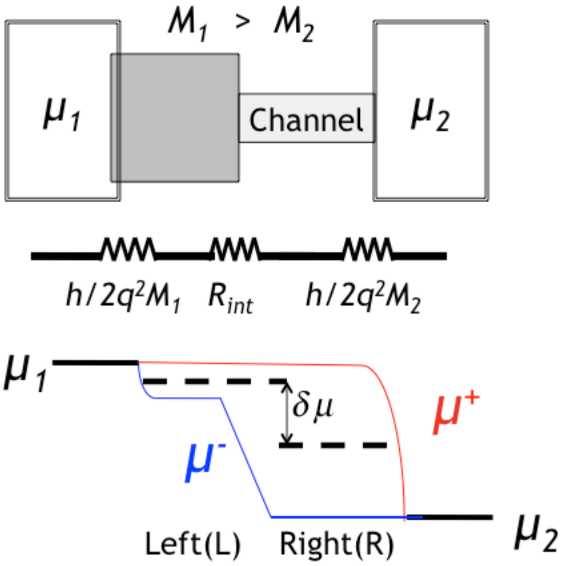


Fig.14.5. Interface between two channels with mode numbers M_1 and M_2 with large contacts (infinite number of modes) at either end.

Consider the electrochemical potentials μ^+, μ^- for right-moving and left-moving electrons respectively. As we discussed in Lecture 6, the contacts impose the boundary conditions (L: Left, R: Right)

$$\mu^+(L) \;=\; \mu_1 \;\text{ and }\; \mu^-(R) \;=\; \mu_2 \qquad (14.7)$$

The current I is the same on the left and on the right and is given by

$$I \;=\; \frac{q}{h} M_1 \left(\mu^+ - \mu^-\right)_L \;=\; \frac{q}{h} M_2 \left(\mu^+ - \mu^-\right)_R \qquad (14.8)$$

The electrons flow freely across the interface, so that the right-moving lanes in the narrow channel on the right remain in equilibrium with the right-moving electrons on the left:

$$\mu^+(R) \;=\; \mu_1 \qquad (14.9a)$$

The left-moving lanes in the wide channel on the left, however, cannot be populated adequately by the narrow channel on the right and we do not know a priori what the resulting potential is. We can determine it by noting from Eq.(14.8) that

$$\mu^+(L) - \mu^-(L) \;=\; \frac{M_2}{M_1} \left(\mu^+(R) - \mu^-(R)\right)$$

and making use of Eqs.(14.7) and (14.9a) to write

$$\mu^-(L) \;=\; \mu_1 - \frac{M_2}{M_1} \left(\mu_1 - \mu_2\right) \qquad (14.9b)$$

To evaluate the interface resistance we need the drop $\delta\mu$ in the average potential across the interface:

$$\delta\mu \;=\; \left(\frac{\mu^+ + \mu^-}{2}\right)_L \;-\; \left(\frac{\mu^+ + \mu^-}{2}\right)_R$$

Making use of Eqs.(14.7) and (14.9)

$$\delta\mu = \frac{1}{2}\left(1-\frac{M_2}{M_1}\right)(\mu_1-\mu_2)$$

$$I = \frac{q}{h}M_2\,(\mu_1-\mu_2)$$

so that we obtain the result stated earlier in Eq.(14.5):

$$R_{\text{int}} \equiv \frac{\delta\mu/q}{I} = \frac{h}{2q^2}\left(\frac{1}{M_2}-\frac{1}{M_1}\right)$$

14.2. Spin Potentials

The difference in the interface resistance between a magnet and the up and downspin channels allows us to create "spin potentials" electrically inside a non-magnetic conductor, a phenomenon that is at the heart of the field of spintronics. This has been possible in metallic conductors like copper ever since its first demonstration in 1985, but has now become possible in semiconductors as well, once the idea of deliberately introducing a barrier layer was demonstrated.

The concept of "spin potentials" is easy to appreciate considering a simple structure having just one magnetic contact, as shown in Fig.14.6a. If no spin was involved we would expect the electrochemical potential to vary spatially as sketched in Fig.14.6b. We could obtain a quantitative plot by solving the diffusion equation (See Eqs.(6.1a,b))

$$\frac{dI}{dz} = 0,$$

$$I = -\frac{\sigma A}{q}\frac{d\mu}{dz}$$

(same as Eqs.(6.1a,b))

subject to the appropriate boundary conditions on $\mu(z)$ at the contacts. Now because the interface resistance is different for the two spins we would expect different drops across the magnet-channel interface for them and so when we solve Eqs.(6.1a,b) for the upspins we will get a different profile from that for the downspins as sketched roughly in Fig.14.6c.

We expect the electrochemical potentials for the two spins to separate around the magnetic contact but they are eventually brought back down to a common level by spin-flip processes, that continually try to restore local equilibrium by equalizing μ_{up} and μ_{dn}.

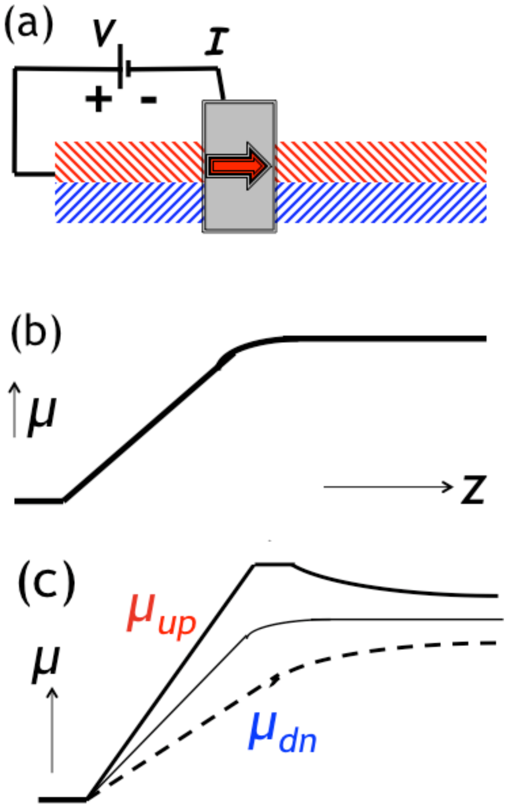


Fig.14.6. (a) Structure designed to cause separation of spin potentials in a channel using magnetic contacts. (b) Expected potential variation for non-magnetic contacts. (c) Magnetic contacts cause a separation of potentials for up and down spins.

Quantitative plots can be obtained by writing separate diffusion equations for up and down spins

$$I_{up} \;=\; -\frac{\sigma A}{2q}\frac{d\mu_{up}}{dz} \tag{14.10a}$$

$$I_{dn} \;=\; -\frac{\sigma A}{2q}\frac{d\mu_{dn}}{dz} \tag{14.10b}$$

We are using half the total conductivity σ for each of the up and down spin channels. It is the interface that discriminates between them, and this is reflected in a different interface resistance and hence a different boundary condition at the magnetic contact (Eqs.(6.4a,b)).

Noting that spin-flip processes convert upspin currents into downspin currents and viceversa, so that

$$\frac{dI_{up}}{dz} \;=\; -\frac{dI_{dn}}{dz} \;=\; -K\left(\mu_{up}-\mu_{dn}\right) \tag{14.11}$$

where K is a constant proportional to the strength of spin-flip processes that seek to equalize μ_{up} and μ_{dn}.

Combining Eqs.(14.11) with (14.10a,b) we obtain

$$\frac{d^2\mu_{up}}{dz^2} \;=\; \frac{\mu_{up}-\mu_{dn}}{2\lambda_{sf}^2} \;=\; -\frac{d^2\mu_{dn}}{dz^2} \tag{14.12}$$

where $$\lambda_{sf} = \frac{1}{2}\sqrt{\frac{\sigma A}{qK}}$$

can be interpreted as a characteristic length that provides a measure of the distance over which spin-related information is preserved. It varies widely from tens of nanometers to hundreds of micrometers depending on the material and the temperature.

Eq.(14.12), known as the Valet-Fert equation, was originally obtained from the Boltzmann equation discussed in Lecture 7 and is widely used to analyze spin diffusion problems. We could use the upspin and downspin potentials to define charge and spin potentials

$$\mu \equiv (\mu_{up} + \mu_{dn})/2 \quad ,$$

$$\mu_s \equiv \mu_{up} - \mu_{dn} \tag{14.13a}$$

Similarly the upspin and downspin currents in Eq.(14.10) can be used to define charge and spin currents:

$$I = I_{up} + I_{dn}$$

$$I_s = I_{up} - I_{dn} \tag{14.13b}$$

It is straightforward to show that the charge components obey the normal equations (Eqs.(6.1a,b)), while the spin component is affected by the spinflip length λ_{sf}.

$$\frac{d^2\mu_s}{dz^2} = \frac{\mu_s}{\lambda_{sf}^2} \tag{14.14}$$

Can we measure the spin voltage inside the channel? The answer is yes, not just along the current path, but also outside the path as shown in Fig.14.7. The latter is often referred to as a non-local spin voltage and is “routinely” measured in spin transport experiments.

The spin voltage is measured by looking at the change in the potential at the output probe when it is switched from parallel to anti-parallel relative to the injecting probe. We will show that this spin voltage is given by

$$V_s \equiv \frac{\mu_P - \mu_{AP}}{q} = P_1 P_2 \, I R_S \, e^{-L/\lambda_{sf}} \tag{14.15}$$

where P_1 and P_2 are the polarizations of the injecting and detecting contacts and

$$R_S = \lambda_{sf} / \sigma A \tag{14.16}$$

These results in a more general form are discussed in Takahashi and Maekawa (2003). Here I just want to give the reader a feeling for the physics by going through the case when both contacts have high resistance due to tunneling barriers. Unfortunately even this simplest case requires a relatively extended discussion.

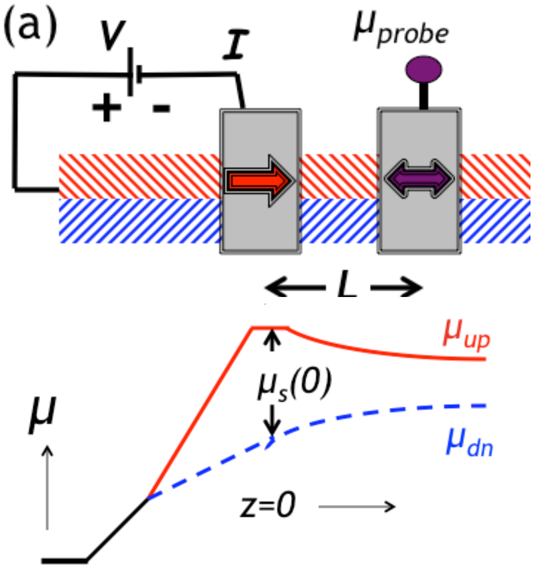


Fig.14.7. The spin voltage can be measured even outside the current path as shown.

14.2.1. *Non-local spin voltage, Eq.(14.15)*

Obtaining Eq.(14.15) involves two steps. Step 1 is to show that the spin potential at the injecting probe is given by

$$\mu_s(0) = P_1 qIR_s \tag{14.17a}$$

Step 2 is to show that the difference between the output potentials for the parallel and anti-parallel configurations is given by

$$\mu_P - \mu_{AP} \quad = \quad P_2\, \mu_s(0) e^{-L/\lambda_{sf}} \tag{14.17b}$$

Combining the two we obtain the stated result in Eq.(14.15).

Step 1 starts by noting that the spin potential obeys Eq.(14.14) which requires it to decay exponentially away from the injecting probe in either direction as sketched in Fig.14.8.

$$\mu_s \quad = \quad \mu_s(0)\, e^{-|z|/\lambda_{sf}} \tag{14.18}$$

We can then use the relation

$$I_s \quad = \quad -\frac{\sigma A}{2q} \frac{d\mu_s}{dz} \tag{14.19}$$

obtained from Eqs.(14.10) and (14.13) to calculate the spin current flowing in either direction. The net spin current drawn from the injecting contact is the sum of the two so that at injecting contact we should have

$$I_{up} - I_{dn} = \frac{\mu_s(0)}{qR_s} \tag{14.20}$$

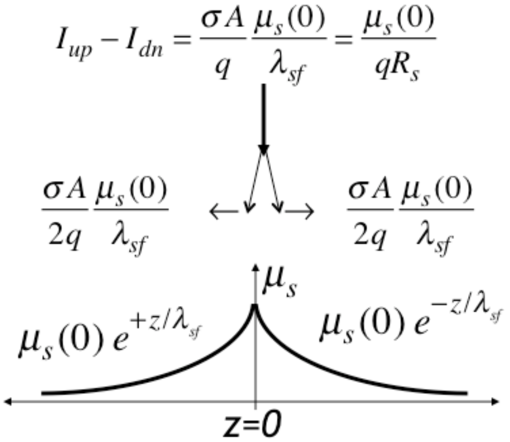
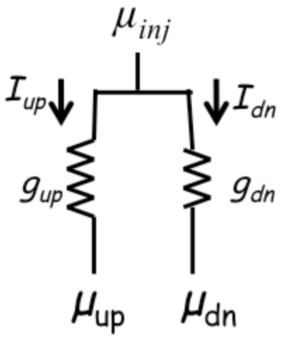


Fig.14.8. Calculating the net spin current at the injecting probe in the structure of Fig.14.7.

Consider now the flow of current at the injecting contact modeling it in terms of two interfacial conductances g_{up} , g_{dn} for up and down spins respectively.



Simple circuit theory yields

$$\frac{\mu_s(0)}{q} \equiv \frac{\mu_{up}-\mu_{dn}}{q} = \frac{I_{dn}}{g_{dn}} - \frac{I_{up}}{g_{up}}$$

which can be rewritten in the form

$$\frac{\mu_s(0)}{q} = \frac{g_{up}+g_{dn}}{2g_{up}g_{dn}}\Big(P_1 I - (I_{up}-I_{dn})\Big)$$

so that using Eq.(14.20)

$$\Big(I_{up}-I_{dn}\Big)R_s = \frac{g_{up}+g_{dn}}{2g_{up}g_{dn}}\Big(P_1 I - (I_{up}-I_{dn})\Big) \tag{14.21}$$

where P_1 is the polarization of the injecting probe defined as

$$P_1 \equiv \frac{g_{up}-g_{dn}}{g_{up}+g_{dn}} \tag{14.22}$$

Now the resistance R_s (Eq.(14.16)) represents the resistance of a section of the channel of length equal to the spin-flip length λ_{sf} which is usually much smaller than the interface resistances $1/g_{up}$ or $1/g_{dn}$ which are relatively large due to the use of barriers to enhance the polarization. Under these conditions we can set the right hand side to zero to obtain

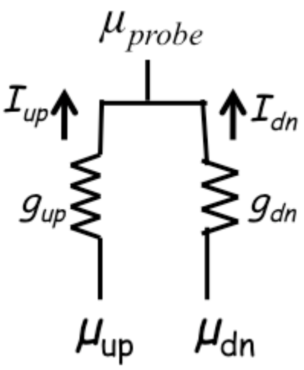
$$I_{up}-I_{dn} = P_1 I$$

which combined with Eq.(14.20) yields our desired result in Eq.(14.17a).

To obtain Eq.(14.17b) (Step 2), we start by noting that the potential at the detecting probe is given by

$$\mu_s(L) \quad = \quad \mu_s(0)\, e^{-L/\lambda_{sf}} \tag{14.22}$$

To find the potential registered by the detecting probe we adopt a circuit model similar to that for the injecting probe. Note also the similarity with the model used to obtain Eq.(12.4), but with the role positive and negative going electrons replaced by up and down spins.



Setting the current equal to zero as we did earlier in Lecture 12 in a different context (see Eq.(12.4)), we have from simple circuit theory

$$I \;=\; 0 \;=\; g_{up}(\mu_{up} - \mu_{probe}) \;+\; g_{dn}(\mu_{dn} - \mu_{probe})$$

so that $$\mu_{probe} \;=\; \frac{g_{up}\mu_{up} + g_{dn}\mu_{dn}}{g_{up} + g_{dn}}$$

We can use Eq.(14.13a) to write μ_{up} and μ_{dn} in terms of μ and μ_s

$$\mu_{up} = \quad \mu + \frac{\mu_s}{2}$$

$$\mu_{dn} = \quad \mu - \frac{\mu_s}{2}$$

and use these relations to write

$$\mu_{probe} = \quad \mu + \frac{P_2 \mu_s}{2}$$

where P_2 is the polarization of the detecting probe defined as before (Eq.(14.2)) in terms of the interface conductances.

With the probe in anti-parallel configuration, the potential is given by the same expression but with g_{up} and g_{dn} interchanged, that is with P_2 replaced by $-P_2$:

$$\mu_{AP} = \mu - \frac{P_2 \mu_s}{2}$$

Hence $$\mu_P - \mu_{AP} = P_2 \mu_s(L) \tag{14.23}$$

Combining Eqs.(14.23) with (14.22) we obtain the result stated in Eq.(14.15).

14.3. Spin-Torque

We have seen how the spin-specific interface resistances associated with magnetic contacts have led to the creation of spin potentials that have been measured experimentally and give rise to large magnetic field dependent resistances that are used routinely to read information stored in magnets.

Another important development has been the demonstration of "spin-torque" which allows spin currents to turn magnets provided they are not more than a few atomic layers in thickness. The basic experimental fact is summarized in Fig.14.9 showing a spin valve structure with a fixed magnet on the left pointing down and a nanomagnet on the right free to point up or down. A negative voltage on the fixed magnet creates a large down spin potential

$$\mu_s \equiv \mu_{up} - \mu_{dn} < 0$$

that exerts a "spin-torque" on the nanomagnet which if it is sufficiently large can make it turn from up to down. Next if we reverse the polarity of the voltage, the positive voltage on the fixed magnet pulls out down spins from the channel reversing the spin potential

$$\mu_s \equiv \mu_{up} - \mu_{dn} > 0$$

Again if this potential is large enough it can turn the nanomagnet back to the up configuration. This effect is now experimentally well-established and it seems likely that it will soon be used to ***write*** information into nanomagnets, just as magnetoresistance phenomena are used to ***read*** information from them.

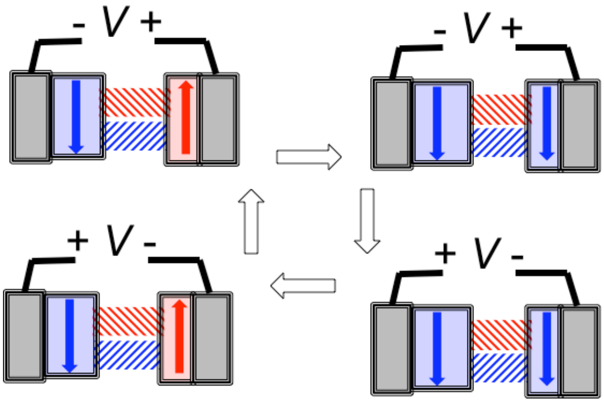
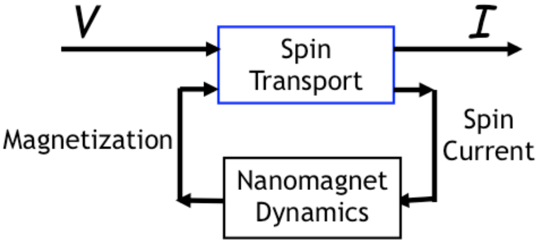


Fig.14.9. Spin valve structure with a fixed magnet on the left pointing down, and a nanomagnet on the right free to point up or down.

These two demonstrations, namely that magnets can create excess spins and excess spins can turn magnets, have combined spintronics and nanomagnetics into a single field that will require its practitioners to acquire an understanding of both spin transport and nanomagnet dynamics (Fig.14.10).

Fig.14.10. Theoretical models need to combine spin transport with nanomagnet dynamics.



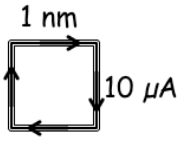
These Lectures are largely about transport, which in this context represents the top half, namely spin transport. We have discussed the spin diffusion equation in the last Section and will discuss the full quantum version later in Lecture 22. These spin transport models let us calculate the spin current given the magnetization.

To complete the story we need a model for the nanomagnet dynamics which will tell us the magnetization of a magnet given the spin current that is incident on it. This model is based on the Landau-Lifshitz-Gilbert (LLG) equation and in the rest of this Section let me try to say a few brief words about it.

The first point to note is that roughly speaking (we will discuss some subtleties in Part III) every electron is like an elementary magnet with a magnetic moment given by the Bohr magneton

$$\mu_B \equiv q\hbar/2m = 9.27e-24\ A-m^2 \tag{14.24}$$

roughly what we would get if a current of 10 μA were circulating in a square loop with dimensions 1 nm x 1 nm, or say a current of 1 mA in a 0.1 nm x 0.1 nm loop.



This was established back in the 1920's by the celebrated experiment due to Stern and Gerlach. More correctly the electron magnetic moment is given by

$$\mu_{el} = \frac{g_s}{2}\mu_B$$

g_s being the "g-factor" which is approximately equal to 2 for electrons in vacuum but could be significantly different in solids, just as the effective mass of electrons in solids can differ from that in vacuum. We will not worry about this "detail" and assume $g_s = 2$ for the following discussion.

If each electron is like a magnet then why are all materials not magnetic? Because usually the electrons are all paired with every up magnet balanced by a corresponding down magnet. It is only in magnetic materials like iron that internal interactions make a large number of electrons line up in the same direction giving rise to a macroscopic magnetization whose ***magnitude*** is given by

$$M_s \;=\; \mu_B \frac{N_s}{\Omega} \tag{14.25}$$

N_s being the number of spins in a volume Ω.

The magnitude of the magnetization of a magnet is fixed but its ***direction*** denoted by the unit vector $\hat{m}$ can change when a magnetic field $\vec{H}$ are applied. The dynamics of $\hat{m}$ is described by the LLG equation

$$(1+\alpha^2)\frac{d\hat{m}}{dt} \;=\; \underbrace{-\gamma\,\mu_0\left(\hat{m}\times\vec{H}\right)}_{Dynamics} - \underbrace{\alpha\gamma\,\mu_0\left(\hat{m}\times\hat{m}\times\vec{H}\right)}_{Dissipation} \tag{14.26}$$

where γ is the "gyromagnetic ratio" given by

$$\gamma \;\equiv\; \frac{q}{m} \;=\; \frac{2\mu_B}{\hbar}$$

and μ_0 is the permeability of vacuum which may not appear explicitly in much of the literature since it is common to use cgs units rather than the SI units we are using.

As indicated in Eq.(14.26) the first term on the right represents dynamical processes while the second term represents "frictional" processes, α being known as the damping coefficient, typically ~ 0.01.

To get some insight, let us see how we can use this equation to understand a basic experimental fact, namely that a magnet has an "easy axis" (assumed to be along z). An external magnetic field H_{ext} can be used to change its magnetization between -1 and + 1 if it exceeds a critical field H_K, as sketched in Fig.14.11. With the magnetic field pointing along the z-direction

$$\vec{H} \;=\; H\,\hat{z}$$

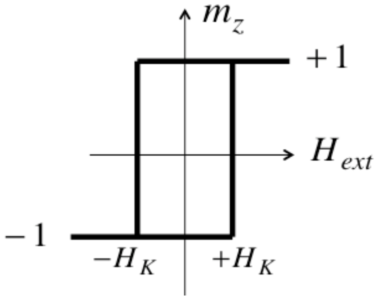
Eq.(14.26) has the form (dropping the term α^2 which is usually << 1)

$$\frac{d\hat{m}}{dt} \;=\; -\gamma\,\mu_0 H\,(\hat{m}\times\hat{z}) \;-\; \alpha\gamma\,\mu_0 H\,(\hat{m}\times\hat{m}\times\hat{z})$$

so that taking its dot product with the unit vector $\hat{z}$ we have

$$\frac{dm_z}{dt} \;=\; \left(1-m_z^2\right)\alpha\gamma\,\mu_0 H \qquad (14.27)$$

Fig.14.11. A magnet has an "easy axis" (assumed to be along *z*). An external magnetic field H_{ext} can be used to change its magnetization between -1 and + 1 if it exceeds a critical field H_K.



Clearly the two steady states predicted by this equation are

$$m_z = +1 \quad and \quad m_z = -1$$

since either choice makes $dm_z/dt = 0$. But are they stable? To answer this question let us assume a small deviation δ from +1

$$m_z = +1-\delta$$

so that Eq.(14.27) reduces to

$$-\frac{d}{dt}\delta \;\approx\; (2\alpha\gamma\,\mu_0 H)\,\delta$$

showing that such deviations will die out if H is positive. Similarly if we assume

$$m_z = -1 + \delta$$

so that Eq.(14.27) reduces to

$$\frac{d}{dt}\delta \approx (2\alpha\gamma\mu_0 H)\,\delta$$

showing that such deviations will die out if H is negative. In short,

$$m_z = +1 \quad \textit{is stable} \quad \textit{if } H > 0 \tag{14.28a}$$

$$m_z = -1 \quad \textit{is stable} \quad \textit{if } H < 0 \tag{14.28b}$$

How does this help us understand Fig.14.11?

First we note that in a magnet, the total field H consists of an external field H_{ext} and an internal field that each electron feels due to all the other electrons, whose sign is determined by m_z:

$$H = H_{ext} + \underbrace{H_K m_z}_{\substack{\text{int}\textit{ernal}\\ \textit{field}}} \tag{14.29}$$

Taking this into account the stability conditions in Eq.(14.28) translate into

$$m_z = +1 \quad \textit{is stable} \quad \textit{if } H_{ext} > -H_K \tag{14.30a}$$

$$m_z = -1 \quad \textit{is stable} \quad \textit{if } H_{ext} < H_K \tag{14.30b}$$

which is exactly what Fig.14.11 indicates.

How do we describe switching with a spin current, $\vec{I}_s$? We have to add an extra term to the right hand side of Eq.(14.26)

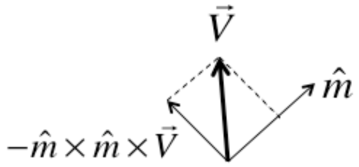
$$\frac{d\hat{m}}{dt} \;=\; \left(\cdots From\ Eq.(14.26)\cdots\right) \;-\; \left(\hat{m}\times\hat{m}\times\frac{\vec{I}_s}{qN_s}\right) \tag{14.31}$$

proportional to the spin current per spin (N_s is the number of spins comprising the magnet). Why is the additional term

$$\hat{m}\times\hat{m}\times\frac{\vec{I}_s}{qN_s} \quad and\ not\ just \quad \frac{\vec{I}_s}{qN_s}\,?$$

What the operation $\hat{m}\times\hat{m}$ does to any vector $\vec{V}$ is to subtract out any component of $\vec{V}$ along $\hat{m}$ as described by the following identity:

$$-\hat{m}\times\hat{m}\times\vec{V} \;=\; \vec{V}-(\hat{m}.\vec{V})\,\hat{m}$$



Hence

$$-\hat{m}\times\hat{m}\times\frac{\vec{I}_s}{qN_s} \;=\; component\ of\ \frac{\vec{I}_s}{qN_s}\ perpendicular\ to\ \hat{m}$$

which is justified by noting that the magnitude of the magnetization of a magnet does not change, only its direction. To ensure this, the right hand side of the LLG equation must be perpendicular to the direction of $\hat{m}$. Actually there is an additional term due to the spin current

$$\alpha\hat{m}\times\frac{\vec{I}_s}{qN_s}$$

which is also perpendicular to $\hat{m}$ but we will ignore it since α is typically very small.

Starting from Eq.(14.31) we can project along the easy axis ($\hat{z}$) to obtain

$$\frac{dm_z}{dt} = \left(1-m_z^2\right)\left(\alpha\gamma\,\mu_0 H_K m_z + \frac{I_s}{qN_s}\right) \qquad (14.32)$$

We can argue exactly as we argued with Eq.(14.27) that the critical spin current needed to switch the magnet is given by

$$\left(\frac{I_s}{qN_s}\right)_{critical} = \alpha\gamma\,\mu_0 H_K \qquad (14.33)$$

This relation has a simple physical interpretation, noting from our earlier discussion leading to Eq.(14.28) that the right-hand side of Eq.(14.33) is ~ inverse of the time constant τ for switching. Eq.(14.33) tells us that the critical current needed is such that the total number of spins delivered by the spin current, $I_s\tau / q$, is approximately equal to the number of spins N_s comprising the magnet.

The critical current itself is proportional to the product of the number of spins and the inverse time constant ~ N_s / τ. Making use of Eq.(14.25) we have

$$\left(I_s\right)_{critical} = \frac{4q\alpha}{\hbar}\underbrace{\left(\frac{1}{2}\mu_0 H_K M_s \Omega\right)}_{Energy\ Barrier} \qquad (14.33)$$

The quantity appearing in parenthesis in Eq.(14.33) represents the energy barrier separating the two states of the magnet and in order for the magnet to exist stably in one of these, the barrier needs to be at least several tens of *kT*. Otherwise the magnet will keep flipping back and forth between the two states many times in time scales of interest.

For an energy barrier ~ *40 kT*, and α =0.01, Eq.(14.33) predicts a critical spin current of ~ 10 µA. The critical current is expected to be somewhat larger and experimentally values ~ 50-100 µA have been demonstrated.

14.4. Polarizers and Analyzers

Let me end this long Lecture with a few words pointing out a subtle aspect of spin that we have glossed over so far. This aspect can be appreciated by considering the non-local spin voltage measurement that we discussed earlier (see Fig.14.7) except that the probe magnet is neither parallel nor anti-parallel to the injecting probe, but instead makes some arbitrary angle θ with it (Fig.14.12)?

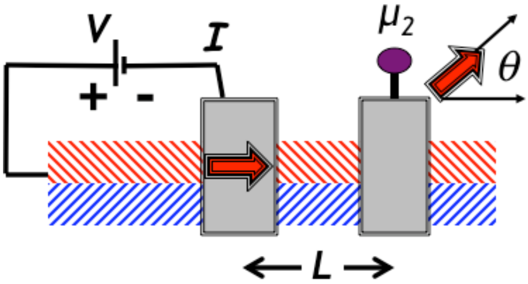
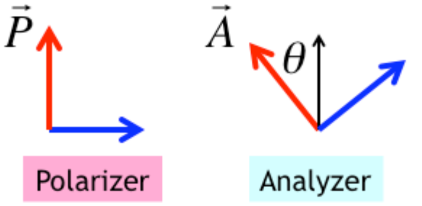


Fig.14.12. Same as Fig.14.7, except that the output probe is at an arbitrary angle to the injecting probe. What voltage would it measure?

The answer can be stated quite simply:

$$\mu_2 = \mu + \frac{\vec{P}_2 . \vec{\mu}_s}{2} \tag{14.34}$$

where the polarization $\vec{P}_2$ points in the direction of the output magnet, while the spin potential $\vec{\mu}_s$ points in the direction of the spin polarization in the channel, which we assume to be in the direction of the injecting magnet. Earlier (see Eq.(14.23)) we considered two special cases when the angle between the two magnets was either zero (parallel, P) or 180 degrees (antiparallel, AP).



$$\textit{Photon Flux} \sim (\vec{P} . \vec{A})^2 \sim \cos^2 \theta$$

How do we understand the general result in Eq.(14.34)? For those unfamiliar with electron spin, the simplest analogy is probably that of photon polarization. As we learn in freshman physics, a polarizer-analyzer combination lets through a flux proportional to $\cos^2\theta$ It is maximum when the two are parallel (θ=0 degrees), and a minimum when the two are perpendicular (θ=90 degrees).

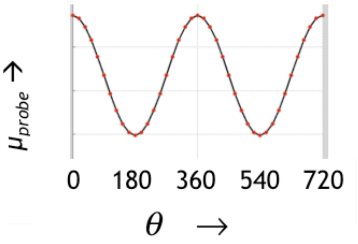
What about electrons? Suppose we have electrons that are all up, so that from Eq.(14.13a)

$$\mu_s \;=\; \mu_{up} \;=\; 2\mu \tag{14.35a}$$

then as we rotate the direction of magnetization of the probe θ, the measured voltage from Eq.(14.18) should change as

$$\frac{\mu_2}{\mu} \;=\; 1+P_2\cos\theta \tag{14.35b}$$

As with photons, the voltage is a maximum when the probe (analyzer) is parallel to the electron polarization (θ=0 degrees). But with electrons the ***minimum*** occurs, not when the two are perpendicular (θ=90 degrees) but when the two are ***antiparallel*** (θ=180 degrees).



Indeed if we assume a perfect voltage probe having $P_2 = 1$, Eq.(14.35b) reduces to

$$\frac{\mu_2}{\mu} \;=\; 1+\cos\theta \;=\; 2\cos^2\frac{\theta}{2}$$

showing that the analyzer lets through a fraction of electrons proportional to

$$\underbrace{\cos^2\frac{\theta}{2}}_{Electrons} \qquad \textit{rather than} \qquad \underbrace{\cos^2\theta}_{Photons}$$

This basic difference between electrons and photons is apparent from the spin valve experiment that we started this Lecture with (see Fig.14.1). The current is a minimum, not when the two magnets are perpendicular, but when they are anti-parallel. Orthogonal directions are not represented by say z and x that are 90 degrees apart. Rather they are represented by up and down that are 180 degrees apart. And that is why a proper description of electron spin requires ***spinors*** rather than vectors as we will discuss later in Lecture 22.

One point that causes some confusion is the following. It seems that if we had electrons in the channel whose spin direction we did not know, we could measure it using a magnet. As we turn the magnet the measured voltage should go through maxima and minima as shown in Fig.4.10, and the direction corresponding to a maximum tells us the direction of the electron spin.

But doesn't quantum mechanics teach us that the spin of an electron cannot be exactly measured? Yes, but that is true if we had just one electron. Here we are talking of an "army" of electrons identically prepared by an injecting contact and what our magnet measures is the average over many many such electrons. This is not in violation of any basic principle.

Lecture 15

Kubo formula

15.1. Kubo Formula for an Elastic Resistor
15.2. Onsager Relations

In our discussion we have stressed the non-equilibrium nature of the problem of current flow requiring contacts with different electrochemical potentials (see Fig.2.4). Just as heat flow is driven by a difference in temperatures, current flow is driven by a difference in electrochemical potentials. Our basic current expression (see Eqs.(3.3), (3.4))

$$I = q \int_{-\infty}^{+\infty} dE \, \frac{D(E)}{2t(E)} \left(f_1(E) - f_2(E)\right) \tag{15.1}$$

is applicable to arbitrary voltages but so far we have focused largely on the low bias approximation (see Eq.(3.1))

$$G = q^2 \int_{-\infty}^{+\infty} dE \left(-\frac{\partial f_0}{\partial E}\right) \frac{D(E)}{2t(E)} \tag{15.2}$$

Although we have obtained this result from the general non-equilibrium expression, it is interesting to note that the low bias conductance is really an *equilibrium property*. Indeed there is a fundamental theorem relating the low bias conductance for small voltages to the fluctuations in the current that occur at equilibrium when no voltage is applied. Let me explain.

Consider a conductor with no applied voltage (see Fig.15.1) so that both source and drain have the same electrochemical potential μ_0. There is of course no net current without an applied voltage, but even at equilibrium, every once in awhile, an electron crosses over from source to drain and on the average an equal number crosses over the other way from the drain to the source, so that

$$\langle I(t_0) \rangle_{eq} = 0$$

where the angular brackets $\langle \cdot\cdot \rangle$ denote either an "ensemble average" over many identical conductors or more straightforwardly a time average over the time t_0.

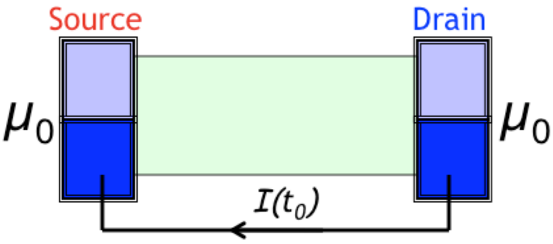


Fig.15.1.
At equilibrium both contacts have the same electrochemical potential μ_0. No net current flows, but there are equal currents I_0 from source to drain and back.

However, if we calculate the current correlation

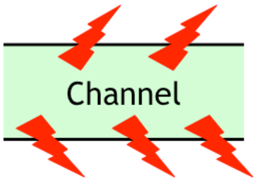
$$C_I = \int_{-\infty}^{+\infty} d\tau \ \langle I(t_0+\tau)\, I(t_0) \rangle_{eq} \tag{15.3}$$

we get a non-zero value even at equilibrium, and *the Kubo formula* relates this quantity to the low bias conductance :

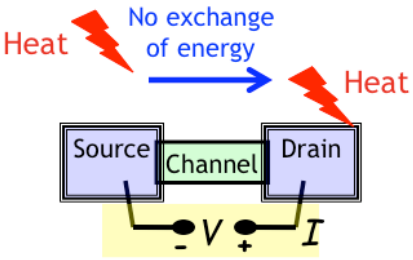
$$G = \frac{C_I}{2kT} = \frac{1}{2kT} \int_{-\infty}^{+\infty} d\tau \ \langle I(t_0+\tau)\, I(t_0) \rangle_{eq} \tag{15.4}$$

This is a very powerful result because it allows one to calculate the conductance by evaluating the current correlations using the methods of equilibrium statistical mechanics, which are in general more well-developed than the methods of non-equilibrium statistical mechanics. Indeed before the advent of mesoscopic physics in the late 1980's, the Kubo formula was the only approach used to model quantum transport. However, its use is limited to linear response. In these Lectures (Part three) we will stress the Non-Equilibrium Green's Function (NEGF) method for quantum transport, which allows us to address the non-equilibrium problem head on for quantum transport, just as the BTE discussed in Lecture 7 does for semiclassical transport.

In this Lecture, however, my purpose is primarily to connect our discussion to this very powerful and widely used approach. The Kubo formula in principle applies to large conductors with inelastic scattering, though in practice it may be difficult to evaluate the effect of complicated inelastic processes on the current correlation. The usual approach is to evaluate transport in long conductors with a high frequency alternating voltage, for which electrons can slosh back and forth without ever encountering the source or drain contacts. One could then obtain the zero frequency conductivity by letting the sample size L tend to infinity ***before*** letting the frequency tend to zero (see for example, Chapter 5 of Doniach and Sondheimer (1973)).



What we will do is something far simpler, namely look at the effect of contacts on the current correlations in an elastic resistor. We will show that applied to an elastic resistor the Kubo formula does lead to our old result (Eq.(15.2)) from Lecture 3. We will then discuss briefly how the Kubo formula leads to the Onsager relations mentioned in Lecture 13 (Eq.(13.19)).



15.1. Kubo Formula for an Elastic Resistor

15.1.1. One-Level Resistor

In the spirit of the bottom-up approach, consider first the one-level resistor from Chapter 3 connected to two contacts with the same electrochemical potential μ_0 and hence the same Fermi function $f_0(E)$ (see Fig.15.2).

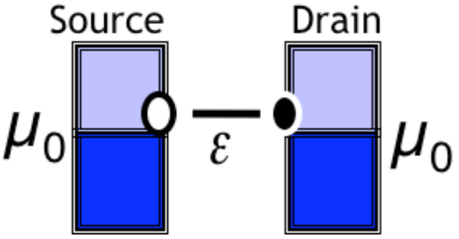
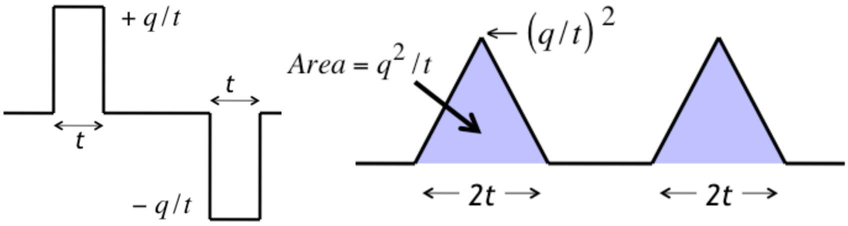


Fig.15.2.
At equilibrium with the same electrochemical potential in both contacts, there is no net current. But there are random pulses of current as electrons cross over in either direction.



There are random positive and negative pulses of current as electrons cross over from the source to the drain and from the drain to the source respectively. The average positive current is equal to the average negative current, which we call the equilibrium current I_0 and write it in terms of the transfer time t (see Eq.(3.2))

$$I_0 = \frac{q}{t} f_0(\varepsilon)\left(1 - f_0(\varepsilon)\right) \tag{15.5a}$$

where the factor $f_0(\varepsilon)\,(1-f_0(\varepsilon))$ is the probability that an electron will be present at the source ready to transfer to the drain but no electron will be present at the drain ready to transfer back. The correlation is obtained by treating the transfer of each electron from the source to the drain as an independent stochastic process.

The integrand in Eq.(15.3) then looks like a sequence of triangular pulses as shown each having an area of q^2/t, so that

$$C_I = \; 2\frac{q^2}{t}\,f_0(\varepsilon)\big(1-f_0(\varepsilon)\big) \tag{15.5b}$$

where the additional factor of 2 comes from the fact that I_0 only counts the positive pulses, while both positive and negative pulses contribute additively to C_I.

14.1.2 Elastic Resistor

We will now show that the Kubo formula (Eq.(15.4)) applied to an elastic resistor leads to the same conductance expression (Eq.(15.2)) that we obtained earlier. Generalizing our one-level results from Eqs.(15.5) to an elastic resistor with an arbitrary density of states, *D(E)* as before we have

$$I_0 \;=\; q\int_{-\infty}^{+\infty} dE\,\frac{D(E)}{2t(E)}\,f_0(E)\,(1-f_0(E)) \tag{15.6a}$$

$$C_I \;=\; 2q^2\int_{-\infty}^{+\infty} dE\,\frac{D(E)}{2t(E)}\,f_0(E)\,(1-f_0(E)) \tag{15.6b}$$

Note that $C_I \;=\; 2qI_0$. Making use of Eq.(15.4) we have for the conductance

$$G = \frac{C_I}{2kT} = q^2 \int_{-\infty}^{+\infty} dE\, \frac{f_0(E)(1-f_0(E))}{kT}\, \frac{D(E)}{2t(E)} \tag{15.7}$$

which is the same as our expression in Eq.(15.2), noting that

$$\left(-\frac{\partial f_0}{\partial E}\right) = \frac{f_0(E)\,(1-f_0(E))}{kT} \tag{15.8}$$

In summary, the Kubo formula (Eq.(15.4)) applied to an elastic resistor leads to the result (Eq.(15.2)) we obtained in Lecture 3 from elementary arguments. Interestingly, the identity in Eq.(15.8) is key to this equivalence, since our elementary arguments lead to a conductance proportional to

$$\frac{f_1 - f_2}{\mu_1 - \mu_2} \cong -\frac{\partial f_0}{\partial E}$$

while the current correlations in the Kubo formula lead to

$$\frac{f_0\,(1-f_0)}{kT}$$

Note how the current correlation requires us to invoke the exclusion principle for the $1\text{-}f_0$ factor, but the elementary argument does not. For phonons (Lecture 11) the elementary arguments lead to (see Eq.(11.8) for the Bose function, n)

$$\frac{n_1 - n_2}{\hbar\omega} \cong -\frac{\partial n}{\partial(\hbar\omega)} = \frac{n(1+n)}{kT} \tag{15.9}$$

and agreement with the corresponding Kubo formula would require a $1+n$ factor instead of the $1\text{-}f$ factor for electrons. We will talk a little more about Fermi and Bose functions in Lecture 16, but the point here is that the theory of noise is more intricate than the theory for the average current that we will focus on in these Lectures. However, I should mention that there is at present an extensive body of work on subtle correlation effects in elastic resistors some of which have been experimentally observed (see for example, Büttiker 2009).

15.2.Onsager Relations

A very important application of the Kubo formula is as a starting point for a very fundamental result like the Onsager relations mentioned in Lecture 13 (Eq.(13.17)).

$$G_{n,m}(+B) \quad = \quad G_{m,n}(-B) \qquad (15.10)$$

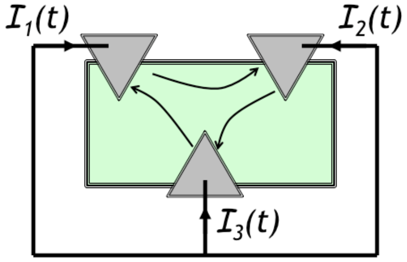
requiring the current at *n* due to a voltage at *m* to be equal to the current at m due to a voltage at *n* with any magnetic field reversed.

This is usually proved starting from the multiterminal version of the Kubo formula

$$G_{m,n} \quad = \quad \frac{1}{2kT} \int_{-\infty}^{+\infty} d\tau \quad \left\langle\, I_m(t_0+\tau)\, I_n(t_0)\right\rangle_{eq} \qquad (15.11)$$

involving the correlation between the currents at two different terminals.

Consider a three terminal structure with a magnetic field ($B > 0$) that makes electrons entering contact 1 bend towards 2, those entering 2 bend towards 3 and those entering 3 bend towards 1.

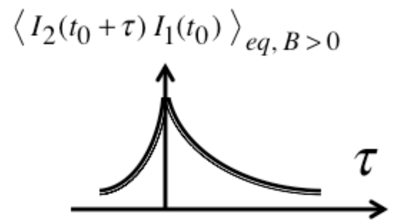


We would expect the correlation

$$\left\langle\, I_2(t_0+\tau)\, I_1(t_0)\,\right\rangle_{eq}$$

to look something like this sketch with the correlation extending further for positive τ .

This is because electrons go from 1 to 2, and so the current I_1 at time t_0 is strongly correlated to the current I_2 at a later time ($\tau > 0$), but not to the current at an earlier time.



If we reverse the magnetic field (*B < 0),* it is argued that the trajectories of electrons are reversed, so that

$$\begin{aligned}&\langle I_1(t_0+\tau)\, I_2(t_0) \rangle_{eq\,,\,B<0}\\ &\quad = \langle I_2(t_0+\tau)\, I_1(t_0) \rangle_{eq\,,\,B>0}\end{aligned} \tag{15.12}$$

This is the key argument. If we accept this, the Onsager relation (Eq.(15.10)) follows readily from the Kubo formula (Eq.(15.11)).

What we have discussed here is really the simplest of the Onsager relations for the generalized transport coefficients relating generalized forces to fluxes. For example, in Lecture 10 we discussed additional coefficients like G_S (see Eq.(10.3)) relating a temperature difference to the electrical current. There are generalized Onsager relations that require (at zero magnetic field) $G_P = T\, G_S$, G_P being the coefficient relating the heat current to the potential difference (see Eq.(10.12)).

This is of course not obvious and requires deep and profound arguments that have prompted some to call the Onsager relations the fourth law of thermodynamics (see for example, Yourgrau et al. 1966). Interestingly, however, in Lecture 11 we obtained transport coefficients (see Eqs.(11.5) and (11.6)) that satisfy this relation $G_P = T\, G_S$ straightforwardly without any profound or subtle arguments. We could cite this as one more example of the power and simplicity of the elastic resistor that comes from disentangling mechanics from thermodynamics.

Lecture 16

Second law

Back in Lecture 10, when discussing the heat current carried by electrons we drew a picture (Fig.10.8) showing the flow of electrons and heat in an elastic resistor consisting of a channel with two contacts (source and drain) with a voltage applied across it (Fig.16.1a). Fig.16.1b shows a slightly generalized version of the same picture that will be useful for the present discussion.

Fig.16.1b shows an elastic channel receiving N_1, N_2 electrons with contacts 1 and 2, held at potentials μ_1 and μ_2 respectively. Of course both N_1 and N_2 cannot be positive. If N_1 electrons enter the channel from one contact an equal number must leave from the other contact so that

$$N_1 \;+\; N_2 \;=\; 0 \qquad (16.1a)$$

For generality I have also shown an exchange of energy E_0 (but not electrons) with the surroundings at temperature T_0, possibly by the emission and absorption of phonons and/or photons. This exchange is absent in elastic resistors.

The principle of energy conservation requires that the total energy entering the channel is zero

$$E_1 \;+\; E_2 \;+\; E_0 \;=\; 0 \qquad (16.1b)$$

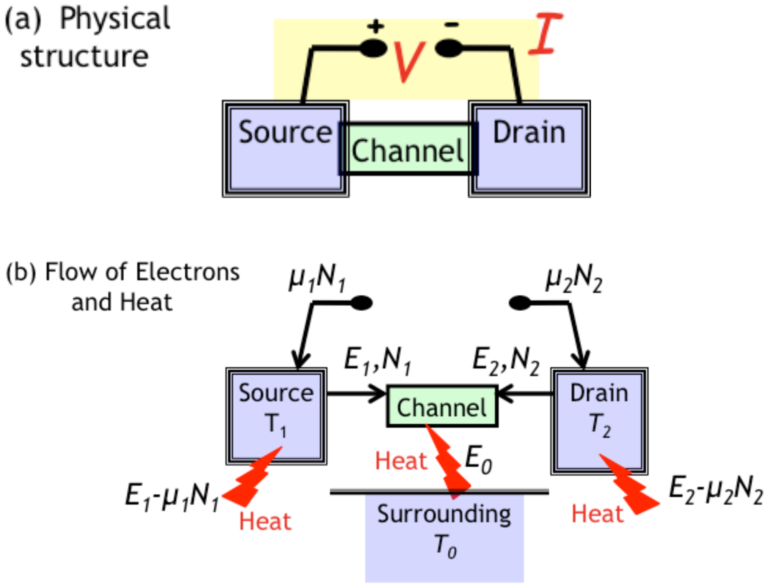


Fig.16.1. The flow of electrons and heat in the specific physical structure we have been discussing can be depicted in general terms as shown. For an elastic resistor, $E_0 = 0$.

This could be called an example of the ***first law*** of thermodynamics. However, there is yet another principle

$$\frac{E_1 - \mu_1 N_1}{T_1} + \frac{E_2 - \mu_2 N_2}{T_2} + \frac{E_0}{T_0} \leq 0 \qquad (16.2)$$

known as the ***second law*** of thermodynamics. Unlike the first law, the second law involves an inequality. While most people are comfortable with the first law or the principle of energy conservation, the second law still continues to excite debate and controversy.

And yet in some ways the second law embodies ideas that we know from experience. Suppose for example we assume all contacts to be at the same temperature ($T_2 = T_1 = T_0$). In this case Eq.(16.2) simply says that the total heat absorbed from the surroundings

$$(E_1 - \mu_1 N_1) \quad + \quad (E_2 - \mu_2 N_2) \quad + \quad E_0 \quad \leq \quad 0 \tag{16.3a}$$

Making use of Eq.(16.1b), this implies

$$\mu_1 N_1 \quad + \quad \mu_2 N_2 \quad \geq \quad 0 \tag{16.3b}$$

The total energy exchanged in the process $E_1 + E_2 + E_0$ has two parts: One that came from the thermal energy of the surroundings and the other that came from the battery. Eq. (16.3a) tells us that the former must be negative, and Eq.(16.3b) tells us that the latter must be positive. In other words, ***we can take energy from a battery and dissipate it as heat, but we cannot take heat from the surroundings and charge up our battery.***

This should come as no surprise to anybody. After all if we could use heat from our surroundings to charge a battery (perhaps even run a car!) then there would be no energy problem. But the point to note is that this is not prohibited by the first law since energy would still be conserved. It is the second law that makes a distinction between the energy stored in a battery and the thermal energy in our surroundings. The first is easily converted into the second, but not the other way around because thermal energy is distributed among many degrees of freedom. We can take energy from one degree of freedom and distribute it among many degrees of freedom, but we cannot take energy from many degrees of freedom and concentrate it all in one.

This intuitive feeling is quantified and generalized by the second law (Eq.(16.2)) based on solid experimental evidence. For example if we have multiple "contacts" at different temperatures then it is possible to take heat from the hotter contact, dump a part of it in the colder contact, use the difference to charge up a battery and still be compliance with the second law.

Are all the things we have discussed so far in compliance with the second law? The answer is yes. For the elastic resistor $E_0 = 0$, and we can write the second law from Eq.(16.1b) in the form

$$\frac{\varepsilon-\mu_1}{T_1}N_1 \quad + \quad \frac{\varepsilon-\mu_2}{T_2}N_2 \quad \leq \quad 0$$

where we have written $E_1 = \varepsilon N_1$ and $E_2 = \varepsilon N_2$, assuming that each electron entering and exiting the channel has an energy of ε. Making use of Eq.(16.1a) this means that

$$\left(\frac{\varepsilon-\mu_1}{T_1} \quad - \quad \frac{\varepsilon-\mu_2}{T_2}\right)N_1 \quad \leq \quad 0$$

Our description of the elastic resistor always meets this condition, since the flow of electrons is determined by $f_1 - f_2$, as we saw in Lecture 3. N_1 is positive indicating electron flow from source to drain if

$$f_1(\varepsilon) \quad > \quad f_2(\varepsilon)$$

that is, if

$$\frac{1}{1+e^{(\varepsilon-\mu_1)/kT}} \quad > \quad \frac{1}{1+e^{(\varepsilon-\mu_2)/kT}}$$

$$\frac{\varepsilon-\mu_1}{T_1} \quad < \quad \frac{\varepsilon-\mu_2}{T_2}$$

Similarly we can show that N_1 is negative if

$$\frac{\varepsilon-\mu_1}{T_1} \quad > \quad \frac{\varepsilon-\mu_2}{T_2}$$

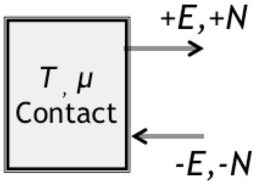
In either case we have

$$\left(\frac{\varepsilon-\mu_1}{T_1} \quad - \quad \frac{\varepsilon-\mu_2}{T_2}\right)N_1 \quad \leq \quad 0$$

thus ensuring that the second law is satisfied.

But what if we wish to go beyond the elastic resistor and include energy exchange within the channel. What would we need to ensure that we are complying with the second law?

16.1. Asymmetry of Absorption and Emission



The answer is that our model needs to ensure that for all processes involving the exchange of electrons with a contact held at a potential μ and temperature *T*, the probability of *absorbing E,N* be related to the probability of *emitting E,N* by the relation

$$\frac{P(+E,+N)}{P(-E,-N)} = e^{-(E-\mu N)/kT} \tag{16.4a}$$

If only energy is exchanged, but not electrons, then the relation is modified to

$$\frac{P(+E)}{P(-E)} = e^{-E/kT} \tag{16.4b}$$

To see how this relation (Eq.(16.4)) ensures compliance with the second law (Eq.(16.2)), consider the process depicted in Fig.16.1 involving energy and/or electron exchange with three different "contacts". Such a process should have a likelihood proportional to

$$P(E_1,N_1)\,P(E_2,N_2)\,P(E_0)$$

while the likelihood of the reverse process will be proportional to

$$P(-E_1,-N_1)\,P(-E_2,-N_2)\,P(-E_0)$$

In order for the former to dominate their ratio must exceed one:

$$\frac{P(+E_1,+N_1)\,P(+E_2,+N_2)\,P(+E_0)}{P(-E_1,-N_1)\,P(-E_2,-N_2)\,P(-E_0)} \geq 1$$

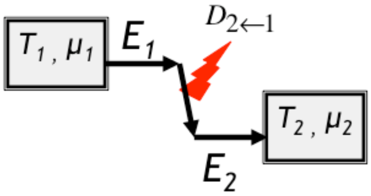
If all processes obey the relations stated in Eqs.(16.4a,b), we have

$$\exp\left(-\frac{E_1-\mu_1 N_1}{kT_1}\right)\exp\left(-\frac{E_2-\mu_2 N_2}{kT_2}\right)\exp\left(-\frac{E_0}{kT_0}\right) \geq 1$$

which leads to the second law stated in Eq.(16.2), noting that *exp (- x)* is greater than one, only if x is less than zero.

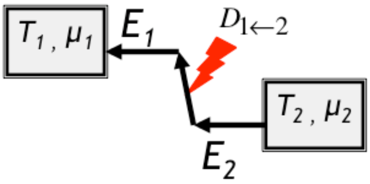
Note that the equality in Eq.(16.2) corresponds to the forward probability being only infinitesimally larger than the reverse probability, implying a very slow net forward rate. To make the “reaction” progress faster, the forward probability needs to exceed the reverse probability significantly, corresponding to the inequality in Eq.(16.2).

So how do we make sure our model meets the requirement in Eq.(16.4)? Consider for example a conductor with one inelastic scatterer in the middle separating a region having an energy level at E_1 from another having a level at E_2. Electrons flow from contact 1 to 2 by a process of emission whose probability is given by



$$D_{2\leftarrow 1}\, f_1(E_1)(1-f_2(E_2))$$

while the flow from 2 to 1 requires an absorption process with probability



$$D_{1\leftarrow 2}\, f_2(E_2)(1-f_1(E_1))$$

Since one process involves emission while the other involves absorption, the rates should obey the requirement imposed by Eq.(16.4b):

$$\frac{D_{2\leftarrow 1}}{D_{1\leftarrow 2}} = e^{(E_1-E_2)/kT_0} \qquad (16.5)$$

as we had stated earlier in Lecture 9 in a different context (Eq.(9.7)). T_0 is the temperature of the surroundings with which electrons exchange energy.

The current in such an inelastic resistor would be given by an expression of the form (suppressing the arguments E_1, E_2 for clarity)

$$I \sim \quad D_{2\leftarrow 1}\, f_1(1-f_2) - D_{1\leftarrow 2}\, f_2(1-f_1)$$

which reduces to the familiar form for elastic resistors

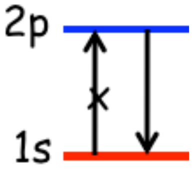
$$I \sim \quad (f_1 - f_2)$$

only if $$D_{2\leftarrow 1} \quad = \quad D_{1\leftarrow 2}$$

corresponding to elastic scattering $E_2 = E_1$. Ordinary resistors have both elastic and inelastic scatterers intertwined and there is no simple expression relating the current to f_1, f_2.

The bottom line is that any model that includes energy exchange in the channel should make sure that absorption and emission rates are related by Eq.(16.5) if the surroundings are in equilibrium with a fixed temperature. Any transport theory, semiclassical or quantum needs to make sure it complies with this requirement to avoid violating the second law.

16.2. Entropy



The asymmetry of emission and absorption expressed by Eqs.(16.4) is actually quite familiar to everyone, indeed so familiar that we may not recognize it. We all know that if we take a hydrogen atom and place its lone electron in an excited (say 2p) state, it will promptly emit light and descend to the 1s state. But an electron placed in the 1s state will stay there forever. We justify it by saying that the electron "naturally" goes to its lowest energy state.

But there is really nothing natural about this. Any mechanical interaction (quantum or classical) that takes an electron from 2p to 1s will also take it from 1s to 2p. The natural descent of an electron to its lowest energy state is driven by a force that is not mechanical in nature. It is "entropic" in origin, as we will try to explain.

Basically it comes from a ***property of the surroundings*** expressed by Eq.(16.4) which tells us that it much harder to absorb anything from a reservoir, compared to emitting something into it. At zero temperature, a system can only emit and never absorb, and so an electron in state 2p can emit its way to the lowest energy state 1s, but an electron in state 1s can go nowhere.

This behavior is of course quite well-established and does not surprise anyone. But it embodies the key point that makes transport and especially quantum transport such a difficult subject in general. Any theoretical model has to include entropic processes in addition to the familiar mechanical forces.

So where does the preferential tendency to lose energy rather than gain energy from any "reservoir" come from? Eq.(16.4) can be understood by noting that when the electron loses energy the contact gains in energy so that the ratio of the rate of losing energy to the rate of gaining energy is equal to the ratio of the density of states at $E_0 + \varepsilon$ to that at E_0 (Fig.16.2):

$$\frac{P(-\varepsilon)}{P(+\varepsilon)} = \frac{W(E_0+\varepsilon)}{W(E_0)}$$

Here *W(E)* represents the number of states available at an energy range *E* in the contact which is related to its entropy by the Boltzmann relation

$$S = k\,\ell n W \qquad (16.7)$$

so that

$$\frac{P(-\varepsilon)}{P(+\varepsilon)} = \exp\frac{S(E_0+\varepsilon)-S(E_0)}{k} \qquad (16.8)$$

Assuming that the energy exchanged ε is very small compared to that of the large contact, we can write

$$S(E_0+\varepsilon)-S(E_0) \approx \varepsilon\left(\frac{dS}{dE}\right)_{E=E_0} = \frac{\varepsilon}{T}$$

with the temperature defined by the relation

$$\frac{1}{T} = \left(\frac{dS}{dE}\right)_{E=E_0} \tag{16.9}$$

This is of course a very profound result saying that regardless of the detailed construction of any particular reservoir, as long as it is in equilibrium, *dS/dE* can be identified as its temperature.

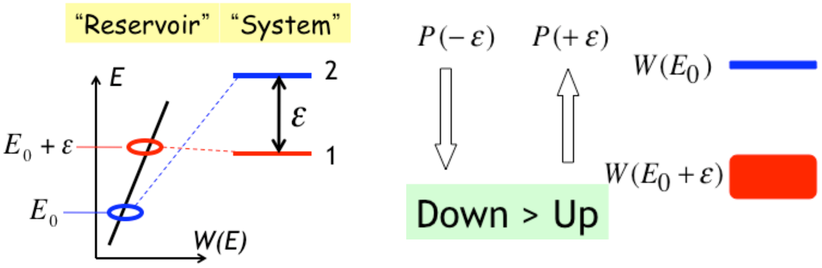


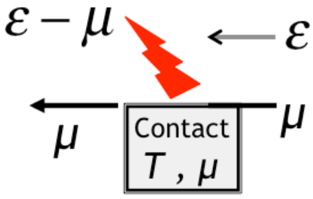
Fig.16.2. Electrons preferentially go down in energy because it means more energy for the "reservoir" with a higher density of states. It is as if the lower state has a far greater "weight" as indicated in the lower panel.

If we accept this, then Eq.(16.7) gives us the basic relation that governs the exchange of energy with any "reservoir" in equilibrium with a temperature T:

$$\frac{P(-\varepsilon)}{P(+\varepsilon)} = e^{\varepsilon/kT}$$

as we stated earlier (see Eq.(16.4b)).

If the emission of energy involves the emission of an electron which eventually leaves the contact with an energy μ, then ε should be replaced by $\varepsilon-\mu$, as indicated in Eq.(16.4a). The key idea is the same as what we introduced in Fig.10.8 when discussing thermoelectric effects, namely that when an electron is added to a reservoir with energy ε, an amount $\varepsilon - \mu$ is dissipated as heat, the remaining μ representing an increase in the energy of the contact due to the added electron. Indeed that is the definition of the electrochemical potential μ. Eventually the added electron leaves the contact as shown.

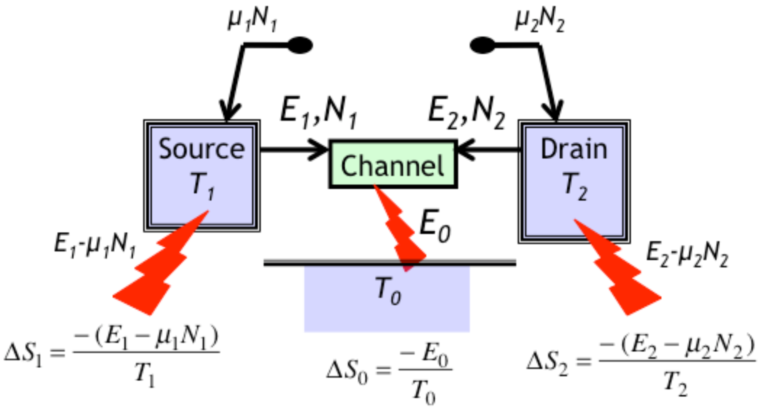


16.2.1. *Total Entropy Always Increases*

Now that we have defined the concept of entropy, we can use it to restate the second law from Eq.(16.2). If we look at Fig.16.1b we note that

$E_1 - \mu_1 N_1$ represents the energy exchange with a "reservoir" at T_1,
$E_2 - \mu_2 N_2$ represents the energy exchange with a "reservoir" at T_2,
E_0 represents the energy exchange with a "reservoir" at T_0.

Based on the definition of temperature in Eq.(16.9), we can write the corresponding changes in entropy $\Delta S_1, \Delta S_2, \Delta S_0$ as shown below



Note that these are exactly the same terms (except for the negative sign) appearing in Eq.(16.2), which can now be restated as

$$(\Delta S)_1 \quad + \quad (\Delta S)_2 \quad + \quad (\Delta S)_0 \quad \geq \quad 0 \qquad (16.10)$$

In other words, the second law requires the total change in entropy of all the reservoirs to be positive.

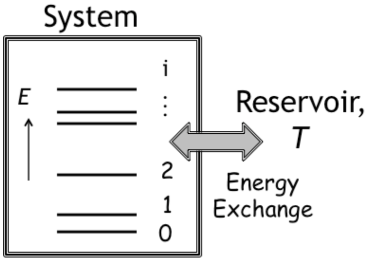
16.2.2. Free energy always decreases

At zero temperature, any system in coming to equilibrium with its surroundings, goes to its state having the lowest energy. This is because a reservoir at zero temperature will only allow the system to give up energy, but not to absorb any energy. Interestingly, at non –zero temperatures, one can define a quantity called the ***free energy***

$$F \quad = \quad E - TS \qquad (16.11)$$

such that at equilibrium a system goes to its state with minimum free energy. At *T=0,* the free energy, *F* is the same as the total energy, *E.*

To see this, consider a system that can exchange energy with a reservoir such that the total energy is conserved.



Using the subscript "R" for reservoir quantities we can write

$$\Delta E \quad + \quad \Delta E_R \quad = \quad 0 \qquad (16.12a)$$

$$\Delta S \quad + \quad \Delta S_R \quad \geq \quad 0 \qquad (16.12b)$$

which are basically the first and second laws of thermodynamics that we have been discussing. Noting that

$$\Delta S_R \quad = \quad \frac{\Delta E_R}{T}$$

we can combine Eqs.(16.12a,b) to write

$$\Delta F \equiv \quad \Delta E \quad - \quad T\Delta S \quad \leq \quad 0 \qquad (16.13)$$

which tells us that all energy exchange processes permitted by the first and second laws will cause the free energy to decrease, so that the final equilibrium state will be one with minimum free energy.

16.3. Law of Equilibrium

The preferential tendency to lose energy rather than gain energy from any surrounding "reservoir" as expressed in Eq.(16.4) leads to a universal law stating that any system in equilibrium having states i with energy E_i and with N_i particles will occupy these states with probabilities

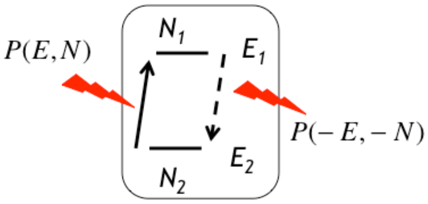
$$p_i = \frac{1}{Z} e^{-(E_i - \mu N_i)/kT} \qquad (16.14)$$

where Z is a constant chosen to ensure that all the probabilities add up to one.

To see this we note that all reservoirs in equilibrium have the property

$$\frac{P(+E,+N)}{P(-E,-N)} = e^{-(E-\mu N)/kT} \qquad \text{(same as 16.4a)}$$

$$E = E_1 - E_2\,, \quad N = N_1 - N_2$$



"Fock space"

Suppose we have a system with two states as shown exchanging energy and electrons with the surroundings. At equilibrium, we require upward transitions to balance downward transitions, so that

$$p_2\, P(E,N) \;=\; p_1\, P(-E,-N)$$

Making use of Eq.(16.4a), we have

$$\frac{p_1}{p_2} = \frac{P(+E,+N)}{P(-E,-N)} = e^{-((E_1-\mu N_1)-(E_2-\mu N_2))/kT}$$

It is straightforward to check that the probabilities given by Eq.(16.14) satisfy this requirement and hence represent an acceptable equilibrium solution.

How can we have a law of equilibrium so general that it can be applied to all systems irrespective of its details? Because as we noted earlier it comes from the ***property of the surroundings*** and not the system.

Eq.(16.14) represents the key principle or equilibrium statistical mechanics, Feynman (1965) called it the "summit". But it looks a little different from the two equilibrium distributions we introduced earlier, namely the Fermi function (Eq.(2.2)) and the Bose function (Eq.(11.8)).

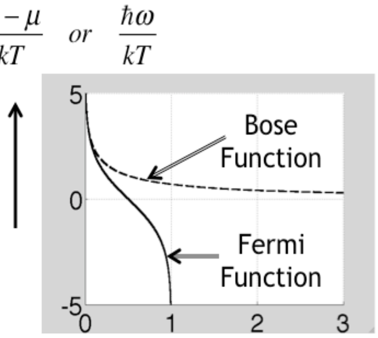


Fig.16.3. The Fermi function (Eq.(2.2)) and the Bose function (Eq.(11.8)).

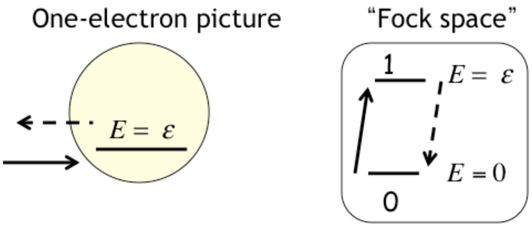
Fig.16.3 shows these two functions which look the same at high energies but deviate significantly at low energies. Electrons obey the exclusion principle and so the occupation *f(E)* is restricted to values between 0 and 1. The Bose function is not limited between 0 and 1 since there is no exclusion principle.

Interestingly, however, both the Bose function and the Fermi function are special cases of the general law of equilibrium in Eq.(16.14). To see this, however, we need to introduce the concept of ***Fock space*** since the energy levels appearing in Eq.(16.14) do not represent the one-electron states we have been using throughout these Lectures. They represent the so-called Fock space states, a new concept that needs some discussion.

16.4. Fock space states

Consider a simple system with just one energy level, ε. In the one electron picture we think of electrons going in and out of this level. In the Fock space picture we think of the two possible states of the system, one corresponding to an empty state with energy *E=0,* and one corresponding to a filled state with energy E = ε as shown.

When an electron comes in the system goes from the empty state *(0)* to the full state *(1),* while if an electron leaves, the system goes from *1* to *0.*



Applying the general law of equilibrium (Eq.(16.14)) to the Fock space states, we have

$$p_0 = 1/Z \quad and \quad p_1 = e^{-x}/Z$$

where

$$x \equiv (\varepsilon - \mu)/kT$$

Since the two probabilities p_0 and p_1 must add up to one, we have

$$Z = 1 + e^{-x}$$

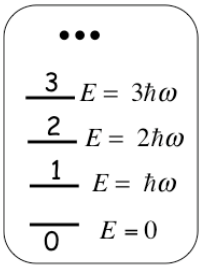
$$p_0 = \frac{1}{e^{-x}+1} = 1 - f_0(\varepsilon)$$

$$p_1 = \frac{e^{-x}}{e^{-x}+1} = \frac{1}{e^{x}+1} = f_0(\varepsilon)$$

The probability of the system being in the full state, p_1 thus equals the Fermi function while the probability of the system being in the empty state, p_0 equals one minus the Fermi function, as we would expect.

16.4.1. Bose function

The Bose function too follows from Eq.(16.14), but we need to apply it to a system where the number of particles go from zero to infinity. Fock space states for electrons on the other hand are restricted to just zero or one because of the exclusion principle.



“Fock space” for photons

Eq.(16.14) then gives us the probability of the system being in the N-photon state as

$$p_N = \frac{e^{-Nx}}{Z}\,, \quad where\ x \equiv \frac{\hbar\omega}{kT}$$

To ensure that all probabilities add up to one, we have

$$Z = \sum_{N=0}^{\infty} e^{-Nx} = \frac{1}{1-e^{-x}}$$

so that the average number of photons is given by

$$n = \sum_{N=0}^{\infty} N\,p_N = \frac{1}{Z}\sum_{N=0}^{\infty} N\,e^{-Nx}$$

Noting that $\displaystyle\sum_{N=0}^{\infty} N\,e^{-Nx} = -\frac{d}{dx}\sum_{N=0}^{\infty} e^{-Nx} = -\frac{d}{dx}Z$

we can show with a little algebra that $\displaystyle n = \frac{1}{e^{x}-1}$

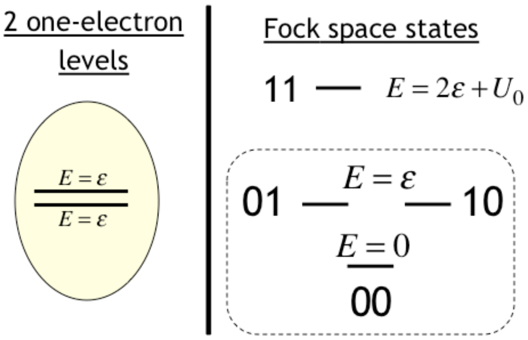
which is the Bose function stated earlier in Eq.(11.8).

The reason we have E-μ appearing in the Fermi function for electrons but not $\hbar\omega-\mu$ in the Bose function for photons or phonons is that the latter are not conserved. As we discussed in Section 16.2, when an electron enters the contact with energy E, it relaxes to an average energy of μ, and the energy dissipated is E-μ. But when a photon or a phonon with energy $\hbar\omega$ is emitted or absorbed, the energy dissipated is just that.

However, there are conserved particles (not photons or phonons) that also obey Bose statistics, and the corresponding Bose function has E-μ and not just E.

16.4.2. Interacting electrons

The general law of equilibrium (Eq.(16.14)) not only gives us the Fermi and Bose functions but in principle can also describe the equilibrium state of complicated interacting systems, if we are able to calculate the appropriate Fock space energies. Suppose we have an interacting system with two one-electron levels corresponding to which we have four Fock space states as shown labeled 00, 01, 10 and 11. The 11 state with both levels occupied has an extra interaction energy U_0 as indicated.



What is the average number of electrons if the system is in equilibrium with an electrochemical potential μ? Once again defining

$x \equiv \dfrac{\varepsilon-\mu}{kT}$, we have from Eq.(16.14)

$$p_{00} = \frac{1}{Z}$$

$$p_{01} = p_{10} = \frac{e^{-x}}{Z}$$

$$p_{11} = \frac{e^{-2x}}{Z}\, e^{-U_0/kT}$$

The average number of electrons is given by

$$n = \; 0.p_{00} + 1.p_{01} + 1.p_{10} + 2.p_{11}$$

$$= \; \frac{2(e^{-x} + e^{-2x}\; e^{-U_0/kT})}{Z}$$

We could work out the details for arbitrary interaction energy U_0, but it is instructive to look at two limits. Firstly, the non-interacting limit with $U_0 \rightarrow 0$ for which

$$Z = \; 1 + 2e^{-x} + e^{-2x} = \; (1 + e^{-x})^2$$

so that with a little algebra we have

$$n \; = \; \frac{2}{1 + e^{(\varepsilon-\mu)/kT}}\,, \quad U_0 \rightarrow 0$$

equal to the Fermi function times two as we might expect since there are two non-interacting states.

The other limit is that of strongly interacting electrons for which $Z = \; 1 + 2e^{-x}$ so that

$$n \; = \; \frac{1}{1 + \frac{1}{2}e^{(\varepsilon-\mu)/kT}}\,, \quad U_0 \rightarrow \infty$$

a result that does not seem to follow in any simple way from the Fermi function. With g one-electron states present, it takes a little more work to show that the number is

$$n \; = \; \frac{1}{1 + \frac{1}{g}e^{(\varepsilon-\mu)/kT}}\,, \quad U_0 \rightarrow \infty$$

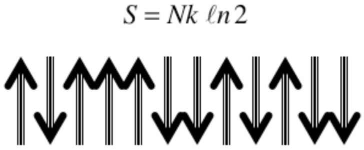
This result may be familiar to some readers in the context of counting electrons occupying localized states in a semiconductor.

Equilibrium statistical mechanics is a vast subject and we are of course barely scratching the surface. My purpose here is simply to give a reader a feel for the concept of Fock space states and how they relate to the one electron states we have generally been talking about.

This is important because the general law of equilibrium (Eq.(16.14)) and the closely related concept of entropy (Eq.(16.7)) are both formulated in terms of Fock space states. We have just seen how the law of equilibrium can be translated into one-electron terms for non-interacting systems. Next let us see how one does the same for entropy.

16.5.Alternative Expression for Entropy

Consider a system of independent localized spins, like magnetic impurities in the channel. At equilibrium, half the spins randomly point up and the other half point down. What is the associated entropy?



Eq.(16.7) defines the entropy S as $k\,\ell n W$, W being the total number of Fock space states accessible to the system. In the present problem we could argue that each spin has two possible states (up or down) so that a collection of N spins has a total of 2^N states:

$$W = 2^N \quad \rightarrow \quad S = k\ \ell n W = Nk\,\ell n 2 \tag{16.15}$$

This is correct, but there is an alternative expression that can be used whenever we have a system composed of a large number of identical independent systems, like the N spin collection we are considering:

$$S = \ -Nk \sum_i \tilde{p}_i\ \ell n \tilde{p}_i \tag{16.16}$$

where the $\tilde{p}_i$'s denote the probabilities of finding an individual system in its i^{th} state. An individual spin, for example has a probability of *1/2* for being in either an up or a down state, so that from Eq.(16.16) we obtain

$$S = \; -Nk\left(\frac{1}{2}\ell n\frac{1}{2}+\frac{1}{2}\ell n\frac{1}{2}\right) = \; Nk\,\ell n 2$$

exactly the same answer as before (Eq.(16.15)).

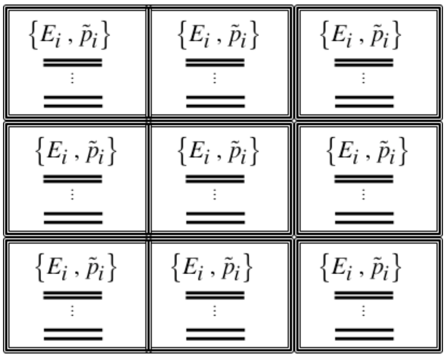
Eq.(16.16), however, is more versatile in the sense that we can use it easily even if the $\tilde{p}_i$'s happen to be say *1/4* and *3/4* rather than *1/2* for each. Besides it is remarkably similar to the expression for the Shannon entropy associated with the information content of a message composed of a string of *N* symbols each of which can take on different values *i* with probability $\tilde{p}_i$. In the next Lecture I will try to elaborate on this point further.

Let me end this Lecture simply by indicating how this new expression for entropy given in Eq.(16.16) is obtained from our old one that we used in Eq.(16.15). This is described in standard texts on statistical mechanics (see for example, Dill and Bromberg (2003)).

16.5.1. From Eq.(16.7) to Eq.(16.16)

Consider a very large number *N* of identical systems each with energy levels $\{E_i\}$ occupied according to probabilities $\{\tilde{p}_i\}$, such that the number of these syatems in state *i* is given by

$$N_i \;=\; N\,\tilde{p}_i$$



The total number of ways in which we can have a particular set of *{N_i}* should equal W, so that from standard combinatorial arguments we can write

$$W = \frac{N!}{N_1!N_2! \quad \cdots}$$

Taking the logarithm and using Stirling's approximation for large *n*

$$\ell n\, n! \cong n\, \ell n\, n - n$$

we have
$$\ell n W = \ell n N! - \ell n N_1! - \ell n N_2! - \quad \cdots$$
$$= N\,\ell n N - N\tilde{p}_1\,\ell n N\tilde{p}_1 - N\tilde{p}_2\,\ell n N\tilde{p}_2 - \cdots$$

Making use of the condition that all the probabilities$\{\, \tilde{p}_i \,\}$ add up to one, we have

$$\ell n W = -N\left(\, \tilde{p}_1\,\ell n \tilde{p}_1 + \tilde{p}_2\,\ell n \tilde{p}_2 + \quad \cdots \right) = -N \sum_i \tilde{p}_i\,\ell n \tilde{p}_i$$

This gives us *W* in terms of the probabilities, thus connecting the two expressions for entropy in Eq.(16.7) and Eq.(16.16).

16.5.1. Equilibrium Distribution from Minimizing Free Energy

One last observation before we move on. In general the system could be in some arbitrary state (not necessarily the equilibrium state) where each energy level E_i is occupied with some probability $\tilde{p}_i$. However, we have argued that for the equilibrium state, the probabilities $\tilde{p}_i$ are given by

$$\left[\tilde{p}_i\right]_{equilibrium} = \frac{1}{Z} e^{-E_i/kT} \equiv p_i \qquad \text{(see Eq.(16.14))}$$

where *Z* is a constant chosen to ensure that all the probabilities add up to one. We hasve also argued that the equilibrium state is characterized by a minimum in the free energy *F* = *E* – *TS*. Can we show that of all the possible choices for the probabilities $\{\tilde{p}_i\}$, the equilibrium distribution

$\{p_i\}$ is the one that will minimize the free energy ?

Noting that the energy of an individual system is given by

$$E \;=\; \sum_i E_i\,\tilde{p}_i$$

and using *S/N* from Eq.(16.16) we can express the free energy as

$$F \;=\; \sum_i \tilde{p}_i\;(E_i + kT\ell n\,\tilde{p}_i) \tag{16.17}$$

which can be minimized with respect to changes in $\{\tilde{p}_i\}$

$$\begin{aligned} dF \;=\; 0 \;&=\; \sum_i d\tilde{p}_i\;(E_i + kT\ell n\,\tilde{p}_i + kT) \\ &=\; \sum_i d\tilde{p}_i\;(E_i + kT\ell n\,\tilde{p}_i) \end{aligned}$$

noting that the sum of all probabilities is fixed, so that

$$\sum_i d\,\tilde{p}_i \;=\; 0$$

We can now argue that in order to ensure that *dF* is zero for arbitrary choices of $d\,\tilde{p}_i$ we must have

$$E_i \;+\; kT\,\ell n\,\tilde{p}_i \;=\; 0$$

which gives us the equilibrium probabilities in Eq.(16.14).

Even if the system is not in equilibrium we can use Eq.(16.17) to calculate the free energy *F* of an out-of equilibrium system if we know the probabilities $\tilde{p}_i$. But the answer should be a number larger than the equilibrium value.

In the next Lecture I will argue that in principle we can build a device that will harness the excess free energy

$$\Delta F = F - F_{eq}$$

of an out-of-equilibrium system to do useful work.

The excess free energy has two parts:

$$\underbrace{\Delta F}_{\substack{\text{excess free}\\ \text{energy}}} = \underbrace{\Delta E}_{\substack{\text{excess}\\ \text{energy}}} - T \underbrace{\Delta S}_{\text{information}}$$

The first part represents real energy, but the second represents information that is being traded to convert energy from the surrounding reservoirs into work. Let us now talk abut this "fuel value of information."

Chapter 17

Fuel value of information

A system in equilibrium contains no information, since the equilibrium state is independent of past history. Usually information is contained in systems that are stuck in some out of equilibrium state. We would like to argue that if we have such an out-of-equilibrium system, we can in principle construct a device that extracts an amount of energy less than or equal to

$$E_{available} \quad = \quad F - F_{eq} \tag{17.1}$$

where F is the free energy of the out-of-equilibrium system and F_{eq} is the free energy of the system once it is restored to its equilibrium state. Let me explain where this comes from.

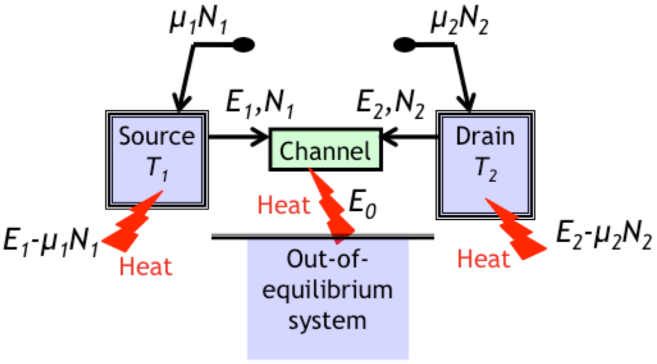


Fig.17.1. An out-of-equilibrium system can in principle used to construct a battery.

Consider the general scheme discussed in the last Lecture, but with both contacts at the same temperature T and with the electrons interacting with some metastable system. Since this system is stuck in an out-of-equilibrium state we cannot in general talk about its temperature.

For example a collection of independent spins in equilibrium would be randomly half up and half down at any temperature. So if we put them into an all-up state, as shown below, we cannot talk about the temperature of this system. But we could still use Eq.(16.16) to find its entropy, which would be zero.

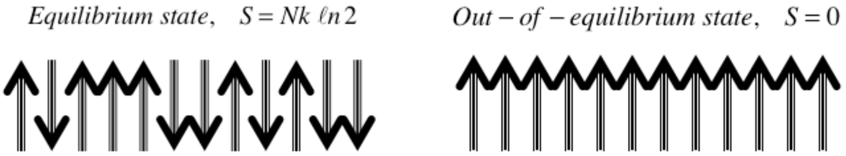


Fig.17.2. A collection of N independent spins in equilibrium would be randomly half up and half down, but could be put into an out-of-equilibrium state with all spins pointing up.

With this in mind we could rewrite the second law by replacing

$$\frac{E_0}{T_0} \quad with \quad -\Delta S$$

in Eq.(16.2) as discussed earlier (see Eq.(16.10)):

$$\frac{E_1 - \mu_1 N_1}{T_1} + \frac{E_2 - \mu_2 N_2}{T_2} - \Delta S \leq 0 \qquad (17.2a)$$

Energy conservation requires that

$$E_1 + E_2 = -E_0 \equiv \Delta E \qquad (17.2b)$$

where ΔE is the change in the energy of the metastable system.

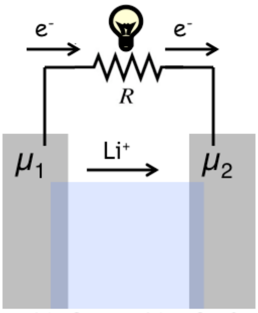
Combining Eqs.(17.2a,b), assuming $T_1 = T_2 = T$, and making use of $N_1 + N_2 = 0$, we have

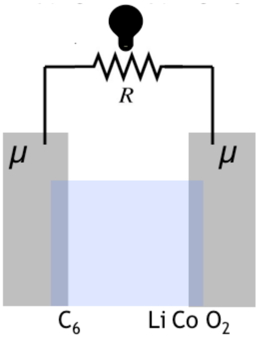
$$(\mu_1 - \mu_2)N_1 \quad \geq \quad \Delta E - T\Delta S = \Delta F \tag{17.3}$$

Ordinarily, ΔF can only be positive, since a system in equilibrium is at its minimum free energy and all it can do is to increase its F. In that case, Eq.(17.3) requires that N_1 have the same sign as μ_1 - μ_2, that is, electrons flow from higher to lower electrochemical potential, as in any resistor.

But a system in an out-of-equilibrium state can relax to equilibrium with a corresponding decrease in free energy, so that ΔF is negative, and N_1 could have a sign opposite to that of μ_1 - μ_2, without violating Eq.(17.3). Electrons could then flow from lower to higher electrochemical potential, as they do inside a battery. The key point is that a ***metastable non-equilibrium state can at least in principle be used to construct a battery***.

In a way this is not too different from the way real batteries work. Take the lithium ion battery for example. A charged battery is in a metastable state with excess Lithium ions intercalated in a carbon matrix at one electrode. As Lithium ions migrate out of the carbon electrode, electrons flow in the external circuit till the battery is discharged and the electrodes have reached the proper equilibrium state with the lowest free energy. The maximum energy that can be extracted is the change in the free energy.





Usually the change in the free energy F comes largely from the change in the real energy E (recall that $F = E - TS$). That does not sound too surprising. If a system starts out with an energy E that is greater than its equilibrium energy E_0, then as it relaxes, it seems plausible that a cleverly designed device could capture the extra energy $E\text{-}E_0$ and deliver it as useful work.

What makes it a little more subtle, is that the extracted energy could come from the change in entropy as well. For example the system of localized spins shown in Fig.17.2 in going from the all-up state to its equilibrium state suffers no change in the actual energy, assuming that the energy is the same whether a spin points up or down. In this case the entire decrease in free energy comes from the increase in entropy:

$$\begin{aligned} \Delta E &= 0 \\ \Delta S &= Nk\,\ell n 2 \\ \Delta F &= \Delta E - T\Delta S = -NkT\,\ell n 2 \end{aligned} \tag{17.5}$$

According to Eq.(17.3) we should be able to build a device that will deliver an amount of energy equal to $NkT\,\ell n 2$. In this Lecture I will describe a device based on the anti-parallel spin valve (Lecture 15) that does just that.

From a practical point of view, $NkT\,\ell n 2$, amounts to about 2.5 kiloJoules per mole, about two to three orders of magnitude lower than the available energy of real fuels like coal or oil which comes largely from ΔE.

But the striking conceptual point is that the energy we extract is not coming from the system of spins whose energy is unchanged. ***The energy comes from the surroundings.*** Ordinarily the second law stops us from taking energy from our surroundings to perform useful work. But the information contained in the non-equilibrium state in the form of "negative entropy" allows us to extract energy from the surroundings without violating the second law.

From this point of view we could use the relation *F=E-TS* to split up the right hand side of Eq.(17.1) into an actual energy and an info-energy that can be extracted from the surroundings by making use of the information available to us in the form of a deficit in entropy *S* relative to the equilibrium value S_{eq}:

$$E_{available} = \underbrace{(E - E_{eq})}_{Energy} + \underbrace{T\,(S_{eq} - S)}_{Info-Energy} \qquad (17.6)$$

For a set of independent localized spins in the all-up state, the available energy is composed entirely of info-energy: there is no change in the actual energy.

17.1.Information-Driven Battery

Let us see how we could design a device to extract the info-energy from a set of localized spins. Consider a perfect anti-parallel spin-valve device (Lecture 14) with a ferromagnetic source that only injects and extracts upspin electrons and a ferromagnetic drain that only injects and extracts downspin electrons from the channel (Fig.17.3).

These itinerant electrons interact with the localized spins through an exchange interaction of the form

$$u + D \quad \Leftrightarrow \quad U + d \qquad (17.7)$$

where *u, d* represent up and down channel electrons, while *U, D* represent up and down localized spins.

Ordinarily this “reaction” would be going equally in either direction. But by starting the localized spins off in a state with $U >> D$, we make the reaction go predominantly from right to left and the resulting excess itinerant electrons *u* are extracted by one contact while the deficiency in *d* electrons is compensated by electrons entering the other contact. After some time, there are equal numbers of localized *U* and *D* spins and the reaction goes in either direction and no further energy can be extracted.

But what is the maximum energy that can be extracted as the localized spins are restored from their all up state to the equilibrium state? The answer is $NkT\,\ell n2$ equal to the change in the free energy of the localized spins as we have argued earlier.

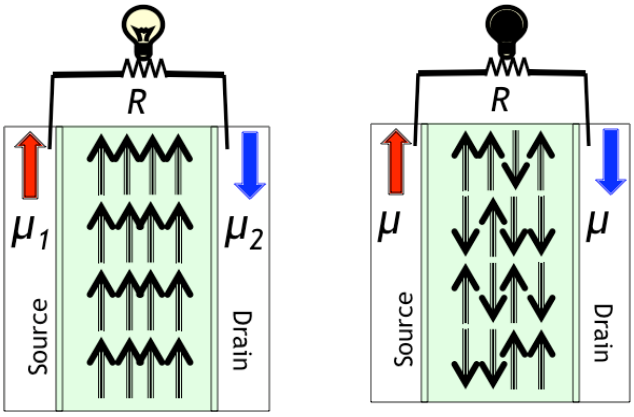


Fig.17.3.
An ***info-battery***: A perfect anti-parallel spin-valve device can be used to extract the excess free energy from a collection on N localized spins, all of which are initially up. Eventually the battery runs down when the spins have been randomized.

But let us see how we can get this result from a direct analysis of the device. Assuming that the interaction is weak we expect the upspin channel electrons (u) to be in equilibrium with contact 1 and the downspin channel electrons (d) to be in equilibrium with contact 2, so that

$$f_u(E) = \frac{1}{\exp\left(\dfrac{E-\mu_1}{kT}\right)+1}$$

and

$$f_d(E) = \frac{1}{\exp\left(\dfrac{E-\mu_2}{kT}\right)+1} \tag{17.8}$$

Assuming that the reaction

$$u + D \quad \Leftrightarrow \quad U + d \qquad \text{(same as Eq.(17.7))}$$

proceeds at a very slow pace so as to be nearly balanced, we can write

$$P_D\, f_u\,(1-f_d) \quad = \quad P_U\, f_d\,(1-f_u)$$

$$\frac{P_U}{P_D} \quad = \quad \frac{f_u}{1-f_u}\,\frac{1-f_d}{f_d} \quad = \quad e^{\Delta\mu/kT} \tag{17.9}$$

where
$$\Delta\mu \quad \equiv \quad \mu_1 - \mu_2$$

Here we assumed a particular potential $\mu_{1,2}$ and calculated the corresponding distribution of up and down localized spins. But we can reverse this argument and view the potential as arising from a particular distribution of spins.

$$\Delta\mu \quad = \quad kT\,\ell n \frac{P_U}{P_D} \tag{17.10}$$

Initially we have a larger potential difference corresponding to a preponderance of upspins (Fig.17.3, left), but eventually we end up with equal up and down spins (Fig.17.3, right) corresponding to $\mu_1 = \mu_2 = \mu$.

Looking at our basic reaction (Eq.(17.7)) we can see that everytime a D flips to an U, a *u* flips to a *d* which goes out through the drain. But when a *U* flips to a *D*, a *d* flips to a *u* which goes out through the source. So the net number of electrons transferred from the source to the drain equals half the change in the difference in the number of *U* and *D* spins:

$$n\,(Source \rightarrow Drain) = \quad -\,\Delta N_U$$

We can write the energy extracted as the potential difference times the number of electrons transferred

$$E = -\int_{Initial}^{Final} \Delta\mu \; dN_U$$

Making use of Eq.(17.10) we can write

$$E = -NkT \int_{Initial}^{Final} \left(\ell n P_U - \ell n P_D\right) dP_U$$

Noting that $dP_U + dP_D = 0$

and that $S = -Nk\,(P_U\,\ell n P_U + P_D\,\ell n P_D)$ (Eq.(16.16))

we can use a little algebra to rewrite the integrand as

$$\left(\ell n P_U - \ell n P_D\right) dP_U = d\left(P_U\,\ell n P_U + P_D\,\ell n P_D\right) = -dS / Nk$$

so that $E = T \int_{Initial}^{Final} dS = T\,\Delta S$

which is the basic result we are trying to establish, namely that the metastable state of all upspins can in principle be used to construct a battery that could deliver upto

$$T\,\Delta S = NkT\,\ell n 2$$

of energy to an external load.

17.2.Fuel Value Comes From Knowledge

A key point that might bother a perceptive reader is the following. We said that the state with all spins up has a higher free energy than that for a random collection of spins: $F > F_0$, and that this excess free energy can in principle be extracted with a suitable device.

But what is it that makes the random collection different from the ordered collection. As Feynman put it, we all realize that it is unusual to see a car with a license plate # 9999.

But it is just as remarkable to see a car with any specific predetermined number say 1439. Similarly if we really knew the spins to be in a very specific but seemingly random configuration like the one sketched here, its entropy would be zero, just like the all up configuration. The possibility of extracting energy comes not from the all up nature of the initial state, ***but from knowing exactly what state it is in.***

But how would we construct our conceptual battery to extract the energy from a random but known configuration? Consider a simple configuration that is not very random: The top half is up and the bottom half is down. Ordinarily this would not give us any open circuit voltage, the top half cancels the bottom half. But we could connect it as shown in Fig.17.4 reversing the contacts for the left and right halves and extract energy.

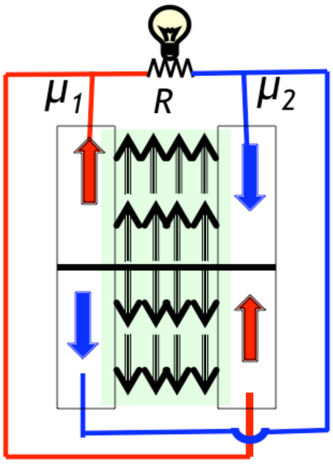


Fig.17.4. A suitably designed device can extract energy from any known configuration of spins.

Following the same principle we could construct a device to extract energy from a more random collection too. The key point is to know the exact configuration so that we can design the contacts accordingly.

Of course these devices would be much harder to build than the one we started with for the all-up configuration. But these devices are just intended to be conceptual constructs intended to illustrate a point. The point is that information consists of a system being in an out of equilibrium state and our knowing exactly which state it is in. This information can ***in principle*** be used to create a battery and traded for energy.

In the field of information theory, Shannon introduced the word entropy as a measure of the information content of a signal composed of a string of symbols *i* that appear with probability $\{p_i\}$

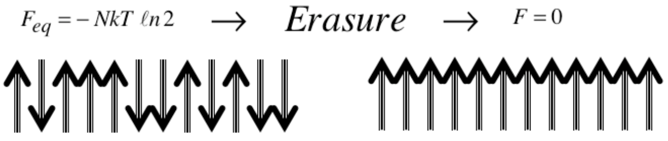
$$H = -\sum_{i} p_i \, \ell n \, p_i \tag{17.11}$$

This expression looks like the thermodynamic entropy (see Eq.(16.17)) except for the Boltzmann constant and there are many arguments to this day about the connection between the two. One could argue that if we had a system with states *i* with equilibrium probabilities $\{p_i\}$, then $k*H$ represents the entropy of an equilibrium system carrying no information. As soon as someone tells us which exact state it is in, the entropy becomes zero so that the entropy is lowered by $(Nk)*H$ increasing its free energy by $(NkT)*H$. In principle, at least we could construct a battery to extract this excess free energy $(NkT)*H$.

17.3. Landauer's Principle

The idea that a known metastable state can be used to construct a battery can be connected to Landauer's principle which talks about the minimum energy needed to erase one bit of information.

In our language, erasure consists of taking a system from an equilibrium state (F_{eq}) to a known standard state (F):



Is there a minimum energy needed to achieve this? We have just argued that once the spin is in the standard state we can construct a battery to extract $(F - F_{eq})$ from it. In a cyclic process we could spend E_{erase} to go from F_{eq} to F, and then construct a battery to extract $(F - F_{eq})$ from it, so that the total energy spent over the cycle equals

$$E_{erase} - \quad (F - F_{eq})$$

which must be greater than zero, or we would have a perpetual source of energy. Hence

$$E_{erase} \geq \quad F - F_{eq} \tag{17.12}$$

which in this case yields Landauer's principle:

$$E_{erase} \geq \quad NkT\, \ell n 2$$

It seems to us, that erasure need not necessarily mean putting the spins in an all-up state. More generally it involves putting them in a known state, analogous to writing a complicated musical piece on a magnetic disk. Also, the minimum energy of erasure need not necessarily be dissipated. It often ends up getting dissipated only because it is impractical to build an info-battery to get it back.

Fifty years ago Landauer asked deep questions that were ahead of his time. Today with the progress in nanoelectronics, the questions are becoming more and more relevant, and some of the answers at least seem fairly clear. Quantum mechanics, however, adds new features some of which are yet to be sorted out and are being actively debated at this time.

17.4. Maxwell's Demon

Our info-battery could be related to Maxwell's famous demon (see for example, Lex (2005)) who was conjectured to beat the second law by letting hot molecules (depicted black) go from left to right and cold molecules (depicted gray) go from right to left so that after some time the right hand side becomes hot and the left hand side becomes cold (Fig.17.5).

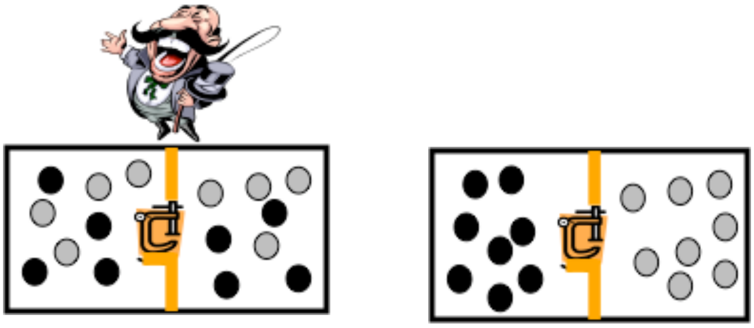


Fig.17.5. Maxwell's demon creates a temperature difference by letting hot molecules go preferentially to the right.

To see the connection with our "info-battery" in Fig.17.3 we could draw the following analogy:

$$\textit{Hot molecules} \quad \leftrightarrow \quad \textit{up} - \textit{spin electrons}$$

$$\textit{Cold molecules} \quad \leftrightarrow \quad \textit{down} - \textit{spin electrons}$$

$$\textit{Demon} \quad \leftrightarrow \quad \textit{Collection of Localized Spins each with two states}$$

$$\textit{Left}, \textit{Right of Box} \quad \leftrightarrow \quad \textit{Source}, \textit{Drain Contacts}$$

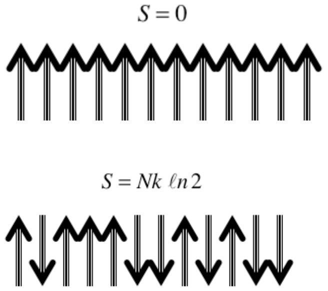
Our battery is run by a set of all up localized spins that flip electrons up and send them to the source, while replacing the down-spin from the drain. The demon sends hot molecules to the left and cold molecules to the right, which is not exactly the same process, but similar.

The key point, however, is that the demon is making use of information rather than energy to create a temperature difference just as our info-battery uses the low entropy sate of the localized spins to create a potential difference. Like our localized spins, the demon too must gradually transition into a high entropy state that will stop it from discriminating between hot and cold molecules. Or as Feynman (1963) put it in one of his Lectures,

> " .. if we build a finite-sized demon, the demon himself gets so warm, he cannot see very well after a while."

Like our info-battery (Fig.17.3), eventually the demon stops functioning when the entropy reaches its equilibrium value and all initial information has been lost.

We started in Lecture 1 by noting how transport processes combine two very different types of processes, one that is force-driven and another that is entropy-driven. In these last two Lectures, my objective has been to give readers a feeling for the concept of an ***"entropic force"*** that drives many everyday phenomena.



The fully polarized state with $S=0$ spontaneously goes to the unpolarized state with $S = Nk\,\ell n 2$, but to make it go the other way we need to connect a battery and do work on it.

This directed flow physically arises from the fact that the fully polarized state represents a single state while the unpolarized state represents numerous (2^N) possibilities. It is this sheer number that drives the impurities spontaneously from the low entropy to the high entropy state and not the other way. Many real life phenomena are driven by such entropic forces which are very different from ordinary forces that take a system from a single state to another single state.

What makes transport so complicated is the intertwining of these two types of phenomena in the channel region. We have seen how the elastic resistor allows us to separate the two, with entropic processes confined to the contacts, and the channel described by semi-classical mechanics. It is now time to look at the quantum version of the problem with the channel described by quantum mechanics.