WAVE PROPAGATION

Geert Brocks

March 13, 2020

Transport or scattering of waves is a prominent topic in many branches of physics. A typical experiment consists of a source that sends out a well-prepared wave onto a target whose properties we are interested in. The target transports/scatters the wave, and the transported/scattered waves are detected by probes set up by the experimenter. Such waves can be calculated by solving a wave equation, be it the classical wave equation for electromagnetic waves or for acoustic waves, or the Schrödinger equation for quantum mechanical waves. As any (partial) differential equation, a wave equation can be solved using numerical techniques based upon finite difference techniques.

Here we will use a technique that is somewhat closer to the physics of wave transport. Part I introduces some of the elementary folklore of scattering theory: the transfer matrix,[1] and its cousin, the scattering matrix. These matrices are at the hart of the numerical techniques we will use. These techniques and their pros and cons are also discussed in Part I. This part finishes with some more folklore of scattering theory. Such starred (*) sections are given as background. We will use some of the results presented there, and I give you the full derivation of those results, but the derivations are not necessary to do the computational project.[2]

The theoretical background is given in Part II. In this project we consider the transport of electron waves. Assuming elastic scattering (electrons bouncing around but not losing energy), the electric current and the conductance can be related to the transmission coefficient for electron waves. This basic relation is called the Landauer formula. Part II starts with a summary of current and conductance for electron waves, and (a derivation of) the Landauer formula. Initially the focus is on one-dimensional transport, which simplifies the mathematical formulation. Subsequently I will show you how to generalize the results to 3D systems consisting of layered materials, i.e., a stack of different layers that are homogeneous parallel to the layer direction. Generalization to materials that have 3D inhomogeneities is also possible in principle, but we will skip that here.[3]

Some experimental background is given in Part III. Of the many possible applications I have picked out two that are related to semiconductor devices: metal-insulator-metal diodes, and resonant-tunneling diodes. Starting from basic quantum mechanics, it is possible to calculate the full current-voltage (I/V) characteristics of these devices.

[1] Unfortunately, although standard, the term "transfer matrix" is not well chosen. Transfer matrix techniques exist in different branches of physics. Apart from the general idea, they not always have much in common. The meaning of the transfer matrix presented here is within a "scattering of waves" context.

[2] Sorry about that, a theorist's idiosyncrasy; I hate to present results without showing you where they come from, even if in the end only the results count.

[3] Nevertheless, this section is too long. Sorry about that, see it as background material. We need some of the results, and I give you the full derivations, but the latter are not required for the computational project.

Contents

Part I
Numerical background

I review some basic elements of scattering theory in one dimension, the transfer and scattering matrices in particular. These matrices are at the hart of the numerical algorithms used to calculate transmissions.

We are considering elastic scattering only, and we work at a fixed energy E. We require a solution to the time-independent Schrödinger equation

$$E\psi(x) + \frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} - V(x)\psi(x) = 0, \tag{1}$$

and assume that the potential is constant outside a finite region $[a;b]$, with $V(x) = V_L; x < a$ and $V(x) = V_R; x > b$. Inside the region $[a;b]$, $V(x)$ may have any behavior.[4] We are interested in energies $E > V_L, V_R$, such that states can propagate over the whole domain between $-\infty$ and ∞. We don't have any restrictions regarding E and $V(x)$ for $a < x < b$. $V(x)$ can be a potential well, a potential barrier, or any combination of these. Below I will use a potential barrier as an example, but I will show you how to generalize the results to any potential profile.

In principle, to solve Eq. 1 we can make use of the numerical techniques we have used to calculate bound states (see "excitons"). Constructing wave functions in the regions outside $[a;b]$ seems to be a waste of time and effort, however, as we know that in regions of constant potential the wave functions are plane waves. We know the wave numbers of those plane waves,[5] but we don't know the amplitudes of the reflected and the transmitted waves. The transfer and scattering matrix techniques are ways to calculate these amplitudes.

1 The transfer matrix

What a transfer matrix is, is best explained using a simple potential as an example: the rectangular barrier. Consider Fig. 1 and let the middle region run from $x = a$ to $x = b$. Since the potential is piecewise constant, we can write the solution as

$$\psi(x) = \begin{cases} Ae^{ikx} + Be^{-ikx}; & x < a \\ Ce^{\eta x} + De^{-\eta x}; & a < x < b \\ Fe^{ikx} + Ge^{-ikx}; & x > b \end{cases} \tag{2}$$

with

$$k = \sqrt{\frac{2mE}{\hbar}},\ \eta = \sqrt{\frac{2m(V_0 - E)}{\hbar}}. \tag{3}$$

For $0 < E < V_0$, both k and η are real numbers.

All of the equations presented below also hold for $0 < V_0 < E$, i.e., scattering over a barrier), as well as for $V_0 < 0 < E$, i.e., scattering over a potential well. In those cases $\eta = iq$ is a purely imaginary number.

[4] "Any" in the physical sense, not in the mathematical sense.

[5] Elementary quantum mechanics, see Sec. 5.1, Eqs. 79 and 81.

Here is the key point: provided we choose the constants A-G such, that the function $\psi(x)$ is continuous and differentiable everywhere, it is a solution to the Schrödinger equation for all x, including the boundaries $x = a$ and $x = b$. Continuity of $\psi(x)$ at $x = a$ gives the relation

$$Ae^{ika} + Be^{-ika} = Ce^{\eta a} + De^{-\eta a},$$

whereas continuity of $\frac{d}{dx}\psi(x)$ at $x = a$ gives

$$ikAe^{ika} - ikBe^{-ika} = \eta Ce^{\eta a} - \eta De^{-\eta a}.$$

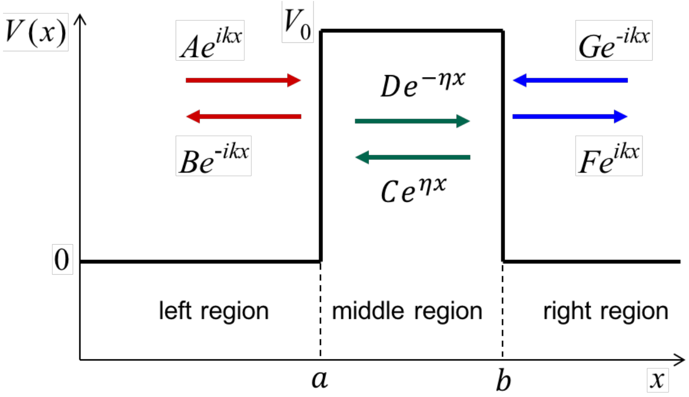

Figure 1: Scattering from a rectangular barrier. The transfer matrix $\mathbf{M}$ relates the coefficients of the waves at the right (blue) to the waves at the left (red) of the barrier.

These two equations can be combined into the matrix equation

$$\begin{aligned}\begin{pmatrix} A \\ B \end{pmatrix} &= \mathbf{M}^{(1)}(a)\begin{pmatrix} C \\ D \end{pmatrix} \\ \mathbf{M}^{(1)}(a) &= \begin{pmatrix} \left(\frac{ik+\eta}{2ik}\right) e^{-(ik-\eta)a} & \left(\frac{ik-\eta}{2ik}\right) e^{-(ik+\eta)a} \\ \left(\frac{ik-\eta}{2ik}\right) e^{(ik+\eta)a} & \left(\frac{ik+\eta}{2ik}\right) e^{(ik-\eta)a} \end{pmatrix}.\end{aligned} \tag{4}$$

For future reference we split this matrix as

$$\mathbf{M}^{(1)}(a) = \begin{pmatrix} e^{-ika} & 0 \\ 0 & e^{ika} \end{pmatrix} \frac{1}{2ik} \begin{pmatrix} ik+\eta & ik-\eta \\ ik-\eta & ik+\eta \end{pmatrix} \begin{pmatrix} e^{\eta a} & 0 \\ 0 & e^{-\eta a} \end{pmatrix} \tag{5}$$

In a similar way, continuity of $\psi(x)$ and $\frac{d}{dx}\psi(x)$ at $x = b$ gives the matrix equation

$$\begin{aligned}\begin{pmatrix} C \\ D \end{pmatrix} &= \mathbf{M}^{(2)}(b)\begin{pmatrix} F \\ G \end{pmatrix} \\ \mathbf{M}^{(2)}(b) &= \begin{pmatrix} \left(\frac{ik+\eta}{2\eta}\right) e^{(ik-\eta)b} & \left(\frac{\eta-ik}{2\eta}\right) e^{-(ik+\eta)b} \\ \left(\frac{\eta-ik}{2\eta}\right) e^{(ik+\eta)b} & \left(\frac{ik+\eta}{2\eta}\right) e^{-(ik-\eta)b} \end{pmatrix},\end{aligned} \tag{6}$$

or

$$\mathbf{M}^{(2)}(b) = \begin{pmatrix} e^{-\eta b} & 0 \\ 0 & e^{\eta b} \end{pmatrix} \frac{1}{2\eta} \begin{pmatrix} \eta+ik & \eta-ik \\ \eta-ik & \eta+ik \end{pmatrix} \begin{pmatrix} e^{ikb} & 0 \\ 0 & e^{-ikb} \end{pmatrix} \tag{7}$$

Combining Eqs. 4 and 6 gives an equation of the form

$$\begin{aligned}\begin{pmatrix} A \\ B \end{pmatrix} &= \mathbf{M}\begin{pmatrix} F \\ G \end{pmatrix} \\ \mathbf{M} &= \mathbf{M}^{(1)}\mathbf{M}^{(2)},\end{aligned} \tag{8}$$

where the 2×2 matrix $\mathbf{M}$ is the product of the two matrices of Eqs. 4 and 6.

In a scattering problem we are in direct control of our boundary conditions by choosing specific experimental conditions. We define an incoming wave Ae^{ikx}, with A a fixed value that gives the density of particles defined by our experimental source, see Sec. 5.1, Eq. 71. We set $G = 0$, which means that we have no incoming wave Ge^{-ikx} from the right (no wave is easily done experimentally, even for a theoretician). Comparing Eqs. 8 and 83, Sec. 5.1, we find for the transmission coefficient

$$T = \frac{v_R}{v_L}\frac{|F|^2}{|A|^2} = \frac{v_R}{v_L}\left|\frac{1}{M_{11}}\right|^2, \tag{9}$$

where $v_{L,R} = \hbar k_{L,R}/m$ is the velocity of the particles left and right of the barrier, see Sec. 5.1. Since the potential is symmetric, we actually have $v_R = v_L$ here, but we let the more general form of Eq. 9 stand for later. It is a simple exercise to work out the matrix element M_{11} and show that one obtains the usual textbook expression for the transmission coefficient[6]

$$T = \left[1 + \left(\frac{k^2+\eta^2}{2k\eta}\right)^2 \sinh^2\left[\eta(b-a)\right]\right]^{-1}. \tag{10}$$

Comparing Eqs. 8 and 78 gives the reflection coefficient as

$$R = \frac{|B|^2}{|A|^2} = \frac{|M_{21}|^2}{|M_{11}|^2}. \tag{11}$$

The matrix $\mathbf{M}$ of Eq. 8 is called the **transfer matrix**. It relates the waves in the left region to the waves in the right region. The square barrier can be viewed as a potential step up at $x = a$, followed by a potential step down at $x = b$. The transfer matrix is then expressed as a product of the transfer matrices of the individual steps, Eq. 8.

For completeness (not terribly important in my opinion, but some students complained when I omitted this case). The scheme outlined above breaks down in case $E = V_0$, i.e., if the energy of the incoming wave is exactly equal to the height of the barrier. We then have $\eta = 0$, see Eq. 3, which would give in Eq. 2, $\psi(x) = C + D$, for $a < x < b$. This is not correct. The Schrödinger equation for the square barrier in the region $a < x < b$ is

$$E\psi(x) + \frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} - V_0\psi(x) = 0, \tag{12}$$

which for $E = V_0$ becomes

$$\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} = 0, \tag{13}$$

whose general solution is

$$\psi(x) = C + Dx; \quad a < x < b. \tag{14}$$

[6] See Griffiths, problem 2.32.

We can use this in Eq. 2, while keeping the same waves for $x < a$ and for $x > b$. Following the same procedure as outlined above, we then find for the step up with $E = V_0$

$$\begin{aligned}\begin{pmatrix} A \\ B \end{pmatrix} &= \mathbf{M}^{(1)}(a) \begin{pmatrix} C \\ D \end{pmatrix} \\ \mathbf{M}^{(1)}(a) &= \begin{pmatrix} e^{-ika} & 0 \\ 0 & e^{ika} \end{pmatrix} \frac{1}{2ik} \begin{pmatrix} ik & ika+1 \\ ik & ika-1 \end{pmatrix}. \end{aligned} \tag{15}$$

Likewise, for the step down with $E = V_0$ we find

$$\begin{aligned}\begin{pmatrix} C \\ D \end{pmatrix} &= \mathbf{M}^{(2)}(b) \begin{pmatrix} F \\ G \end{pmatrix} \\ \mathbf{M}^{(2)}(b) &= \begin{pmatrix} 1-ikb & 1+ikb \\ ik & -ik \end{pmatrix} \begin{pmatrix} e^{ikb} & 0 \\ 0 & e^{-ikb} \end{pmatrix}, \end{aligned} \tag{16}$$

So, in case $E = V_0$ one has to use the matrices defined by Eqs. 15 and 16, instead of Eqs. 4 and 6.

1.1 The transfer matrix algorithm

The idea of finding a transfer matrix by multiplying transfer matrices of potential steps can be extended to a potential profile of a more general shape. All one has to do is to approximate the potential by a series of steps, as is illustrated in Fig. 2.

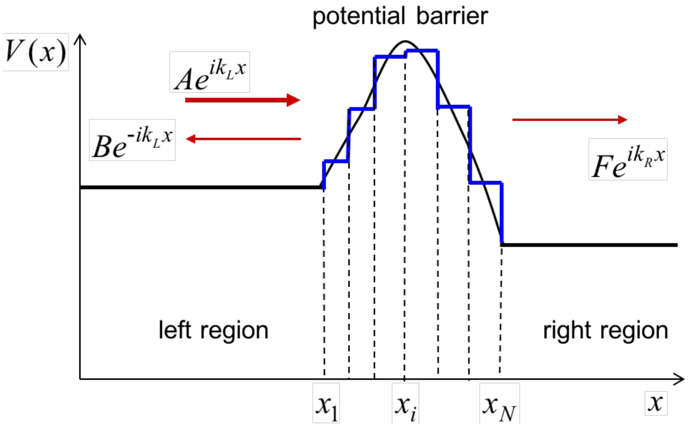

Figure 2: Approximating a barrier of any shape by a series of steps. The transfer matrix $\mathbf{M}$ is given by $\mathbf{M} = \mathbf{M}_1\mathbf{M}_1 \cdots \mathbf{M}_N$.

The transfer matrix $\mathbf{M}$ is given by the product of the transfer matrices of each step

$$\mathbf{M} = \mathbf{M}(x_1)\mathbf{M}(x_2)\cdots\mathbf{M}(x_N), \tag{17}$$

where each $\mathbf{M}(x_i)$ can be calculated by defining

$$\eta_i = \frac{\sqrt{2m\left[V(\xi_i) - E\right]}}{\hbar}, \quad \text{if} \quad E < V(\xi_i);\ x_i \leq \xi_i \leq x_{i+1}; \tag{18}$$

$$\eta_i = i\frac{\sqrt{2m\left[E - V(\xi_i)\right]}}{\hbar}, \quad \text{if} \quad E > V(\xi_i);\ x_i \leq \xi_i \leq x_{i+1}, \tag{19}$$

and using

$$\mathbf{M}(x_i) = \begin{pmatrix} e^{-\eta_{i-1}x_i} & 0 \\ 0 & e^{\eta_{i-1}x_i} \end{pmatrix} \frac{1}{2\eta_{i-1}} \begin{pmatrix} \eta_{i-1}+\eta_i & \eta_{i-1}-\eta_i \\ \eta_{i-1}-\eta_i & \eta_{i-1}+\eta_i \end{pmatrix} \begin{pmatrix} e^{\eta_i x_i} & 0 \\ 0 & e^{-\eta_i x_i} \end{pmatrix}, \tag{20}$$

or

$$\mathbf{M}(x_i) = \begin{pmatrix} \left(\frac{\eta_{i-1}+\eta_i}{2\eta_{i-1}}\right) e^{-(\eta_{i-1}-\eta_i)x_i} & \left(\frac{\eta_{i-1}-\eta}{2\eta_{i-1}}\right) e^{-(\eta_{i-1}+\eta_i)x_i} \\ \left(\frac{\eta_{i-1}-\eta_i}{2\eta_{i-1}}\right) e^{(\eta_{i-1}+\eta_i)x_i} & \left(\frac{\eta_{i-1}+\eta}{2\eta_{i-1}}\right) e^{(\eta_{i-1}-\eta_i)x_i} \end{pmatrix}. \tag{21}$$

Note that this expression is valid both for potential steps up and potential steps down, if one uses Eqs. 18 and 19, compare to Eqs. 5 and 7. The grid points x_i determine the positions of the potential steps. The points ξ_i, at which the potential is evaluated, Eqs. 18 and 19, determine the heights of the potential steps. One can choose $\xi_i = x_i$, but a different choice can improve the convergence, see the next section.

We can be a little bit cleverer and write out the product of Eq. 17 in terms of the matrices of Eq. 20 as

$$\mathbf{M} = \begin{pmatrix} e^{-\eta_0 x_1} & 0 \\ 0 & e^{\eta_0 x_1} \end{pmatrix} \mathbf{P}_1\mathbf{Q}_1\mathbf{P}_2\mathbf{Q}_2...\mathbf{Q}_{N-1}\mathbf{P}_N \begin{pmatrix} e^{\eta_N x_N} & 0 \\ 0 & e^{-\eta_N x_N} \end{pmatrix}, \tag{22}$$

with

$$\mathbf{P}_i = \frac{1}{2\eta_{i-1}} \begin{pmatrix} \eta_{i-1}+\eta_i & \eta_{i-1}-\eta_i \\ \eta_{i-1}-\eta_i & \eta_{i-1}+\eta_i \end{pmatrix}, \quad \mathbf{Q}_i = \begin{pmatrix} e^{\eta_i \Delta x_i} & 0 \\ 0 & e^{-\eta_i \Delta x_i} \end{pmatrix} \quad \text{and} \quad \Delta x_i = x_{i+1} - x_i. \tag{23}$$

The exponential factors now depend on the step width Δx_i and not on the absolute positions x_i. The grid points x_i don't have to be equidistant along the x-axis; using a non-uniform grid does not add any additional complexity. Also the height of the steps in $V(x)$ does not have to be uniform, so there is considerable freedom in choosing the optimal step profile.

Obviously this technique is most naturally suited for layered materials, i.e., systems that consist of layers of different materials stacked on top of one another, where the potential is constant within the layers, but has steps on the interfaces between the layers.

Eqs. 17-23 constitute the **transfer matrix algorithm**. Once the transfer matrix is calculated, physical properties can be derived from its matrix elements, such as the transmission and reflection coefficients, see Eqs. 9 and 11. Transfer matrices are used frequently in numerical calculations solving wave scattering problems in various branches of physics, such as quantum mechanics, optics, and acoustics.

For completeness. (not terribly important in my opinion, but some students complained when I omitted this case). As discussed above, the scheme outlined above breaks down for cases where $E = V(\xi_i)$, as then $\eta_i = 0$ in Eqs. 18 and 19. Going from a region where $\eta_{i-1} \neq 0$ to a region where $\eta_i = 0$, then according to the analysis that has lead to Eq. 15, the matrix of Eq. 20 has to be replaced by

$$\mathbf{M}(x_i) = \begin{pmatrix} e^{-\eta_{i-1}x_i} & 0 \\ 0 & e^{\eta_{i-1}x_i} \end{pmatrix} \frac{1}{2\eta_{i-1}} \begin{pmatrix} \eta_{i-1} & \eta_{i-1}x_i + 1 \\ \eta_{i-1} & \eta_{i-1}x_i - 1 \end{pmatrix}.$$

Going from a region where $\eta_{i-1} = 0$ to a region where $\eta_i \neq 0$, then according to the analysis that has lead to Eq. 16, the matrix of Eq. 20 has to be replaced by

$$\mathbf{M}(x_i) = \begin{pmatrix} 1 - \eta_i x_i & 1 + \eta_i x_i \\ \eta_i & -\eta_i \end{pmatrix} \begin{pmatrix} e^{\eta_i x_i} & 0 \\ 0 & e^{-\eta_i x_i} \end{pmatrix}.$$

These substitutions destroy my beautifully symmetric expression of Eq. 22, which is then not valid anymore. A pragmatic approach would actually be to avoid a situation where $E = V(\xi_i)$ by a careful choice of grid points ξ_i or energies E. That way it would always be possible to use Eqs. 20-23.

1.2 Convergence

I will not try a formal analysis of the convergence. So, just as a statement: for a well-behaved potential $V(x)$, the algorithm of Eqs. 17-21 converges to a unique transfer matrix $\mathbf{M}_V$, if we increase the density of sampling points x_i. This is irrespective of the choice of the points ξ_i at which the potential is evaluated, see Eqs. 18 and 19, provided $x_i \leq \xi_i \leq x_{i+1}$.

Nevertheless, given a specific choice of sampling points x_i, the positions of the points ξ_i influences how well a calculated transfer matrix $\mathbf{M}$ approximates the converged $\mathbf{M}_V$. One obvious choice is $\xi_i = x_i$ (and use $V(x_i)$ in Eqs. 18 and 19). This amounts to approximating the true potential with a sequence of potential steps as shown in Fig. 3(a). Another possible choice is $\xi_i = x_{i+1}$ (using $V(x_{i+1})$ in Eqs. 18 and 19), which leads to the potential step profile shown in Fig. 3(b). Although both choices converge to the same transfer matrix $\mathbf{M}_V$ if we increase the density of sampling points x_i, one can appreciate that neither the potential step profile of Fig. 3(a), nor that of Fig. 3(b), are the best possible approximations to the true potential, given a fixed choice of sampling points x_i.

One has the option of choosing intermediate points $x_i < \xi_i < x_{i+1}$ to evaluate the potential, $V(\xi_i)$, in Eqs. 18 and 19. A popular choice is the **midpoint rule**

$$\xi_i = \frac{x_i + x_{i+1}}{2}. \tag{24}$$

This leads to the potential step profile shown in Fig. 3(c). Even without a formal analysis, one might expect that this leads to a better approximation of the potential.

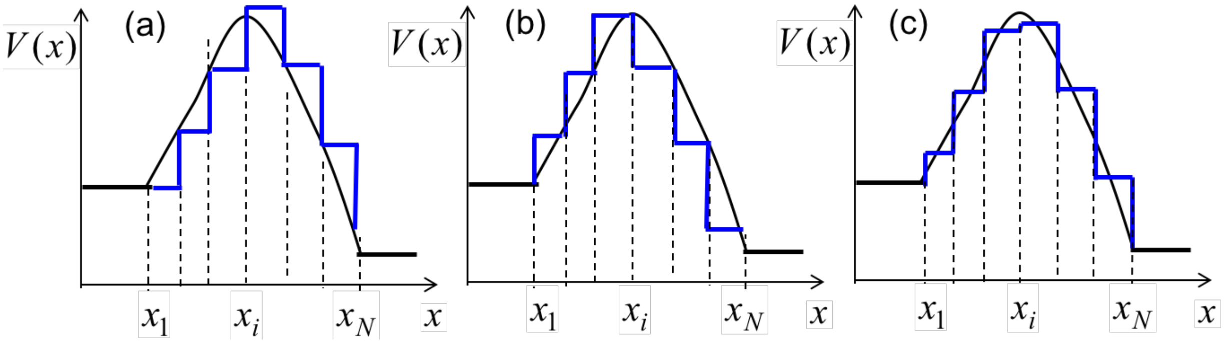

Figure 3: Given a choice of sampling points x_i, one has some freedom to choose the points ξ_i at which the potential is evaluated. (a) With $\xi_i = x_i$ one obtains the sequence of potential steps given in blue; (b) likewise for $\xi_i = x_{i+1}$, and (c) for $\xi_i = (x_i + x_{i+1})/2$. The true potential $V(x)$ is represented by the black curve.

1.3 Numerical stability

The transfer matrix algorithm often works satisfactory, but there can be numerical instabilities in this approach. The simplest one can be identified from the expression of the transfer matrix of Eq. 20 or 21. If the η_i's are imaginary numbers, i.e., for traveling waves, see Eq. 19, we have no problems. But if the η_i's are real, Eq. 18, as for the evanescent waves that describe tunneling through a potential barrier, the matrix contains the real factors $\exp[\eta x]$ and $\exp[-\eta x]$, which can become very large, respectively very small, if the barrier is high (large η), or far away (large x). With $\exp[\eta x]$ you run the risk on *overflow*, i.e. numbers that become too large to be represented on the computer, and with $\exp[-\eta x]$ there is the risk on *underflow*, i.e., numbers that are too small for the computer. In particular overflow is catastrophic, and the algorithm breaks down. In practice, this is most often not a major problem as the range of real numbers numbers that can be represented on the computer is pretty large. Moreover, to avoid this problem, we can use Eqs. 22 and 23 to calculate the transfer matrix, where the exponential factors $\exp[\eta_i \Delta x_i]$ in $\mathbf{Q}_i$ depend on the step width Δx_i and not on the absolute positions x_i. As typically $\Delta x_i \ll |x_i|$, this sharply diminishes the risk on overflow.

There is, however, another numerical instability in the transfer matrix algorithm, which is more subtle and is not easily avoided. We can work it out for a single barrier as an example. A single barrier consists of a step up, followed by a step down. In the notation of Eq. 22, we can write the transfer matrix as as

$$\mathbf{M} = \begin{pmatrix} e^{-\eta_0 x_1} & 0 \\ 0 & e^{\eta_0 x_1} \end{pmatrix} \mathbf{P}_1 \mathbf{Q}_1 \mathbf{P}_2 \begin{pmatrix} e^{\eta_2 x_2} & 0 \\ 0 & e^{-\eta_2 x_2} \end{pmatrix}. \tag{25}$$

The numbers η_0 and η_2 are imaginary, Eq. 19, otherwise we would have no traveling waves at all. This means that the first and the last matrices in Eq. 25 contain numbers with absolute value one and zero, which don't give numerical problems of the sort discussed above. Let us work out the remaining matrix product

$$\mathbf{P}_1 \mathbf{Q}_1 \mathbf{P}_2 = \frac{1}{4\eta_0\eta_1} \begin{pmatrix} a_0 a_1 e^{\eta_1 \Delta x_1} + b_0 b_1 e^{-\eta_1 \Delta x_1} & a_0 b_1 e^{\eta_1 \Delta x_1} + b_0 a_1 e^{-\eta_1 \Delta x_1} \\ b_0 a_1 e^{\eta_1 \Delta x_1} + a_0 b_1 e^{-\eta_1 \Delta x_1} & b_0 b_1 e^{\eta_1 \Delta x_1} + a_0 a_1 e^{-\eta_1 \Delta x_1} \end{pmatrix}, \tag{26}$$
$$\text{with} \quad a_0 = \eta_0 + \eta_1; \; a_1 = \eta_1 + \eta_2; \; b_0 = \eta_0 - \eta_1; \; b_1 = \eta_1 - \eta_2.$$

In case of a barrier, the number η_1 is real, Eq. 18, and if the barrier is high (η_1 large) or wide (Δx_1 large), the exponential factors $\exp(\eta_1 \Delta x_1)$ and $\exp(-\eta_1 \Delta x_1)$ differ by many orders of magnitude.

On most computer systems one uses a fixed number of bits to represent a real number, which means that summing two numbers that differ by a large factor becomes impossible. As an example, $1.0 + 1.4 \times 10^{-2} = 1.014$. If one has only two digits available, the result cannot be stored. Cut the result to two digits, 1.0, means one has lost the 10^{-2} information completely. Even when summing two numbers differing by a significant factor one loses accuracy. As an example, $1.0 + 1.4 \times 10^{-1} = 1.14$. If one has only two digits available, the "4" will be dropped and lost. On a typical computer one uses ~ 16 digits to store a number, so the example is somewhat unrealistic. Nevertheless, you get the point; if two numbers differ by a large factor, you lose the information stored in the smallest number, when you add the two numbers.

The question is of course: does it matter? The answer to that depends on what exactly you want to calculate. Suppose you want to calculate the transmission according to Eq. 9

$$T = \frac{v_R}{v_L} \frac{1}{|M_{11}|^2}. \tag{27}$$

One needs the element M_{11} of the transmission matrix, Eq. 25. Suppose that $\exp(\eta_1 \Delta x_1) \gg \exp(-\eta_1 \Delta x_1)$, and in summing the two exponentials as in Eq. 26 on the computer, one loses all information concerning the $\exp(-\eta_1 \Delta x_1)$ terms. The absolute value of the required matrix element then becomes

$$|M_{11}| = \frac{1}{4\,|\eta|_0\,\eta_1} \left(|\eta_0|^2 + \eta_1^2\right)^{\frac{1}{2}} \left(|\eta_2|^2 + \eta_1^2\right)^{\frac{1}{2}} e^{\eta_1 \Delta x_1}, \tag{28}$$

where I have used that η_0 and η_2 are pure imaginary numbers. Using this expression in Eq. 27 then gives a transmission coefficient. One can compare it to the exact expression of Eq. 10 for a symmetric barrier, where $\eta_0 = \eta_2 = ik$, $\eta_1 = \eta$, and $\Delta x_1 = b - a$. If in Eq. 10 one uses the limit $\eta(b-a) \gg 1$ (i.e., a high and/or wide barrier), then the transmission becomes

$$T \approx \left(\frac{4kq}{k^2+q^2}\right)^2 e^{-2\eta(b-a)}. \tag{29}$$

This is *exactly* the expression one gets if one uses Eqs. 27 and 28 to calculate the transmission coefficient! In other words, losing all $\exp(-\eta_1 \Delta x_1)$ terms has *no consequences* for calculating the transmission.

This conclusion is not universally true, however. Suppose we want to calculate the transmission from the other side of the barrier

$$T' = \frac{v_L}{v_R} \frac{|B|^2}{|G|^2}, \tag{30}$$

see Fig. 1, and Sec. 5.1. In Secs. 2 and 3 it is shown that T' can be expressed as

$$T' = \frac{v_L}{v_R} |S_{12}|^2 \quad \text{with} \quad S_{12} = M_{22} - \frac{M_{21} M_{12}}{M_{11}}, \tag{31}$$

with M_{ij} the elements of the transmission matrix of Eq. 25. Using the expressions of Eq. 25 and 26, and assuming, as in the previous paragraph, that all $\exp(-\eta_1 \Delta x_1)$ terms are lost, one obtains

$$S_{12} = \alpha e^{\eta_1 \Delta x_1}, \tag{32}$$

where you can figure out the coefficient α yourself. *The expression of Eq. 32 is* **blatant nonsense***!* A transmission through a barrier cannot *increase* with the thickness Δx_1, that is unphysical. Instead it should *decrease* with the thickness Δx_1. In fact, if you do the calculation properly and keep all $\exp(-\eta_1 \Delta x_1)$ terms, you will find

$$S_{12} = \beta e^{-\eta_1 \Delta x_1}, \tag{33}$$

where again you figure out the coefficient β yourself. Eq. 33 has the expected physical form, which means that in the calculation of Eq. 31 all $\exp(\eta_1 \Delta x_1)$ cancel out, and only the $\exp(-\eta_1 \Delta x_1)$ terms remain. Unfortunately, in a computer calculation it is those latter terms that can get lost, which means that you are **losing essential information**.

The central problem lies in Eq. 26, which shows that all the elements of the transfer matrix contain a sum over $\exp[\eta_i \Delta x_i]$ and $\exp[-\eta_i \Delta x_i]$ terms. Physically this makes sense, as one needs both the $\exp[\eta x]$ and $\exp[-\eta x]$ functions to construct a proper wave function in the potential barrier, see Eq. 2. The transfer matrix algorithm propagates all solutions in one direction through the barrier, in this case from right to left, irrespective of whether these solutions represent exponentially decaying of growing evanescent waves. Numerically, the growing waves can easily overwhelm the decaying ones, which can be disastrous if the latter carry the information one needs. It would be nice if one had a method to numerically separate the contributions from the growing and the decaying waves. This can be done by using the scattering matrix, as is discussed below.

2 The scattering or S matrix

We discuss another elementary concept in scattering theory, called the scattering matrix. Whereas the transfer matrix relates the modes on the right to the modes on the left, the **scattering matrix** relates outgoing modes to incoming modes. For the example shown in Fig. 1, the modes that are coming into the barrier are Ae^{ikx} (from the left) and Ge^{-ikx} (from the right) and the modes that are going out from the barrier are Be^{-ikx} (to the left) and Fe^{ikx} (to the right). The scattering matrix is defined by the relation

$$\begin{pmatrix} B \\ F \end{pmatrix} = \mathbf{S} \begin{pmatrix} A \\ G \end{pmatrix}. \tag{34}$$

Comparison to the transfer matrix, Eq. 8, gives for the matrix elements of the scattering matrix

$$\begin{aligned} S_{11} &= \frac{M_{21}}{M_{11}}; \;\; S_{12} = M_{22} - \frac{M_{21}M_{12}}{M_{11}}; \\ S_{21} &= \frac{1}{M_{11}}; \;\; S_{22} = -\frac{M_{12}}{M_{11}}. \end{aligned} \tag{35}$$

For the step-up potential, Eq. 4, this gives

$$\mathbf{S}^{(1)} = \begin{pmatrix} \left(\frac{ik-\eta}{ik+\eta}\right) e^{2ika} & \left(\frac{2\eta}{ik+\eta}\right) e^{(ik-\eta)a} \\ \left(\frac{2ik}{ik+\eta}\right) e^{(ik-\eta)a} & -\left(\frac{ik-\eta}{ik+\eta}\right) e^{-2\eta a} \end{pmatrix}. \tag{36}$$

The scattering matrix for the step-down potential, Eq. 6, can be obtained by interchanging ik and η in this expression, and replacing a by b.

2.1 The scattering matrix algorithm

One can use the notation of Eqs. 18-21 to define a scattering matrix for a potential step (up or down) at position x_i

$$\mathbf{S}(x_i) = \begin{pmatrix} \left(\frac{\eta_{i-1}-\eta_i}{\eta_{i-1}+\eta_i}\right) e^{2\eta_{i-1}x_i} & \left(\frac{2\eta_i}{\eta_{i-1}+\eta_i}\right) e^{(\eta_{i-1}-\eta_i)x_i} \\ \left(\frac{2\eta_{i-1}}{\eta_{i-1}+\eta_i}\right) e^{(\eta_{i-1}-\eta_i)x_i} & -\left(\frac{\eta_{i-1}-\eta_i}{\eta_{i-1}+\eta_i}\right) e^{-2\eta_i x_i} \end{pmatrix}. \tag{37}$$

In this notation, the waves that are coming into the step are then $Ae^{\eta_{i-1}x}$ (from the left) and $Ge^{-\eta_i x}$ (from the right) and the waves that are going out from the step are $Be^{-\eta_{i-1}x}$ (to the left) and $Fe^{\eta_i x}$ (to the right). If η_i, η_{i-1} are real and > 0, then all incoming and outgoing waves are growing functions.[7] As all waves have the same character now, we may expect to

[7] $\exp[\eta x]$ going to the right (increasing x) is a growing wave, and $\exp[-\eta x]$ going to the left (decreasing x) is a growing wave.

have solved the problem we had with the transfer matrix, i.e., mixing growing and decaying waves.

The multiplication rule for scattering matrices

$$\mathbf{S} = \mathbf{S}^{(1)} \circ \mathbf{S}^{(2)} \tag{38}$$

can be derived from the inverse of Eq. 35

$$\begin{aligned} M_{11} &= \frac{1}{S_{21}}; \;\; M_{12} = -\frac{S_{22}}{S_{21}}; \\ M_{21} &= \frac{S_{11}}{S_{21}}; \;\; M_{22} = S_{12} - \frac{S_{11}S_{22}}{S_{21}}, \end{aligned} \tag{39}$$

and from the multiplication of the transfer matrices,

$$\begin{aligned} S_{11} = S_{11}^{(1)} + S_{12}^{(1)} S_{11}^{(2)} \left[1 - S_{22}^{(1)} S_{11}^{(2)}\right]^{-1} S_{21}^{(1)}; \;\; S_{12} = S_{12}^{(1)} \left[1 - S_{11}^{(2)} S_{22}^{(1)}\right]^{-1} S_{12}^{(2)}; \\ S_{21} = S_{21}^{(2)} \left[1 - S_{22}^{(1)} S_{11}^{(2)}\right]^{-1} S_{21}^{(1)}; \;\; S_{22} = S_{22}^{(2)} + S_{21}^{(2)} \left[1 - S_{22}^{(1)} S_{11}^{(2)}\right]^{-1} S_{22}^{(1)} S_{12}^{(2)}. \end{aligned} \tag{40}$$

This multiplication rule is a lot more complicated than the simple matrix product we used for transfer matrices. The main advantage is that it leads to a much more stable algorithm.[8]

The scattering matrix algorithm consists of

$$\mathbf{S} = \mathbf{S}(x_1) \circ \mathbf{S}(x_2) \circ ... \circ \mathbf{S}(x_N), \tag{41}$$

with the matrices defined by Eq. 37, the factors η_i by Eq. 18, and the product rule (from right to left) given by Eq. 40. A comparison of Eq. 35 with Eqs. 9 and 11 shows that the reflection and transmission coefficients are given by

$$R = |S_{11}|^2; \;\; T = \frac{v_R}{v_L} |S_{21}|^2. \tag{42}$$

2.2 Numerical stability

Let us see whether indeed we have solved the stability problem for large barriers discussed in Sec. 1.3. Ignoring the prefactors and focusing on the exponential factors we write the matrix of Eq. 36 as

$$\mathbf{S}(x_i) \sim \begin{pmatrix} e^{2\eta_{i-1}x_i} & e^{(\eta_{i-1}-\eta_i)x_i} \\ e^{(\eta_{i-1}-\eta_i)x_i} & -e^{-2\eta_i x_i} \end{pmatrix}. \tag{43}$$

This gives for the product

$$\mathbf{S}(x_1) \circ \mathbf{S}(x_2) \sim \begin{pmatrix} e^{2\eta_0 x_1} \frac{1+2e^{2\eta_1(x_2-x_1)}}{1+e^{2\eta_1(x_2-x_1)}} & e^{\eta_0 x_1 - \eta_2 x_2} \frac{e^{\eta_1(x_2-x_1)}}{1+e^{2\eta_1(x_2-x_1)}} \\ e^{\eta_0 x_1 - \eta_2 x_2} \frac{e^{\eta_1(x_2-x_1)}}{1+e^{2\eta_1(x_2-x_1)}} & -e^{-2\eta_2 x_2} \frac{1+2e^{2\eta_1(x_2-x_1)}}{1+e^{2\eta_1(x_2-x_1)}} \end{pmatrix}. \tag{44}$$

Assuming the bad case that $e^{2\eta_1(x_2-x_1)} \gg 1$ (a wide barrier), and ignoring prefactors again, we can write

$$\mathbf{S}(x_1) \circ \mathbf{S}(x_2) \sim \begin{pmatrix} e^{2\eta_0 x_1} & e^{(\eta_0-\eta_2)x_1} e^{-(\eta_1+\eta_2)(x_2-x_1)} \\ e^{(\eta_0-\eta_2)x_1} e^{-(\eta_1+\eta_2)(x_2-x_1)} & -e^{-2\eta_2 x_2} \end{pmatrix}. \tag{45}$$

[8] As the famous Dutch philosopher Johan Cruijff said: “elluk naadeil hep sun foordeil”.

The off-diagonal matrix elements S_{21} and S_{12} lead to the transmission coefficients for waves incoming from the left and the right, respectively, see Eq. 42 and Sec. 3. Both contain a factor $\exp\left[-\eta_1(x_2-x_1)\right] \ll 1$, as for a barrier η_1 is real, and for a high/wide barrier $\eta_1(x_2-x_1) \gg 1$. These exponentials, decaying with the width (x_2-x_1) of the barrier, are exactly what one expects to see for a transmission, so we don't have the numerical problems discussed in Sec. 1.3.

Products with further scattering matrices provide multiplications with additional factors of this type. Numerically this is fine, as in multiplications one does not loose additional accuracy. Using the example I used before, $1.0 \times 1.4 \times 10^{-2} = 1.4 \times 10^{-2}$. Two digits suffice to store the result.[9] Technically, the scattering matrix solves the problem, because it propagates the incoming waves from both sides into the barrier. So it treats the transmission of waves incoming from the right and from the left on the same footing. This is unlike the transmission matrix, which always propagates from left to right.

3 Physical meaning of the S matrix*

It is custom to write the scattering matrix as

$$\mathbf{S} = \begin{pmatrix} r & t' \\ t & r' \end{pmatrix}. \tag{46}$$

From Eqs. 9, 11 and 35 it follows that the transmission and reflection coefficients are given by

$$T = \frac{v_R}{v_L}\left|t\right|^2; \quad R = \left|r\right|^2, \tag{47}$$

hence the name transmission and reflection amplitudes for t and r, respectively. All the matrix elements of the S-matrix have a physical meaning, which can be deduced from the definition of the matrix, see Eq. 34. The physical meaning is explained in Fig. 4.

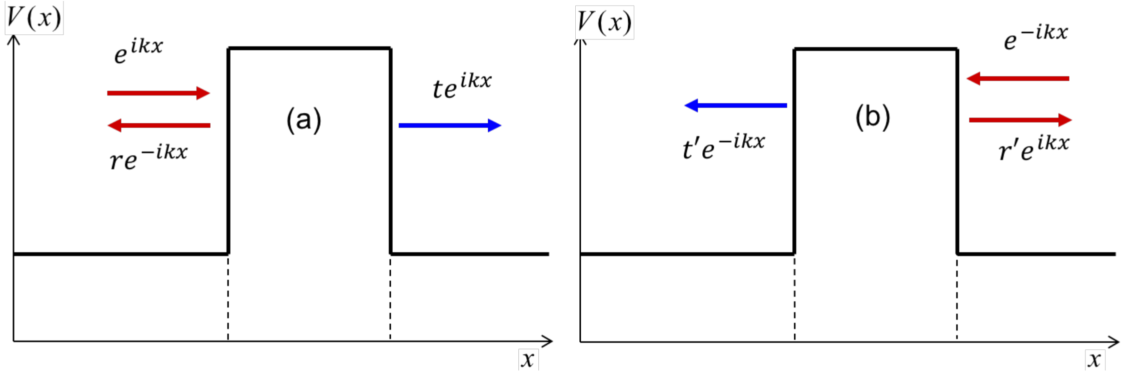

Figure 4: (a) The matrix elements r and t of the S-matrix, Eq. 46, are the reflection amplitude respectively the transmission amplitude, when the incoming wave is from the left. (b) The matrix elements $r'$and t' of the S-matrix, Eq. 46, are the reflection amplitude respectively the transmission amplitude, when the incoming wave is from the right.

[9] $1.1 \times 1.4 \times 10^{-2} = 1.54 \times 10^{-2}$. Using two digits to store the result, one looses the "4". But this is the "ordinary" error one makes when representing a number on the computer by a fixed number of bits. In $1.1 + 1.4 \times 10^{-2} = 1.114 \rightarrow 1.1$ one looses all the information regarding the 1.4×10^{-2}.

Pictures like these also help to understand the physical meaning of the multiplication rule of Eq. 40. Writing the latter in the notation of reflection and transmission amplitudes, we have

$$\begin{aligned} r = r_1 + t_1' r_2 \left[1 - r_1' r_2\right]^{-1} t_1; \;\; t' = t_1' \left[1 - r_1' r_2\right]^{-1} t_2'; \\ t = t_2 \left[1 - r_1' r_2\right]^{-1} t_1; \;\; r' = r_2' + t_2 \left[1 - r_1' r_2\right]^{-1} r_1' t_2', \end{aligned} \tag{48}$$

where r, t, r', t' are the matrix elements of $\mathbf{S}$, r_1, t_1, r_1', t_1' the matrix elements of $\mathbf{S}^{(1)}$, and r_2, t_2, r_2', t_2' the matrix elements of $\mathbf{S}^{(2)}$, see Eq. 40. Focus on the expression of t, and write it as a Taylor expansion

$$t = t_2 \left[1 - r_1' r_2\right]^{-1} t_1 = t_2 \left[1 + r_1' r_2 + \left(r_1' r_2\right)^2 + ...\right] t_1 = t_2 t_1 + t_2 r_1' r_2 t_1 + t_2 r_1' r_2 r_1' r_2 t_1 + ... \tag{49}$$

The terms in this expansion have a clear physical meaning, which is explained for the case of a double barrier potential in Fig. 5.

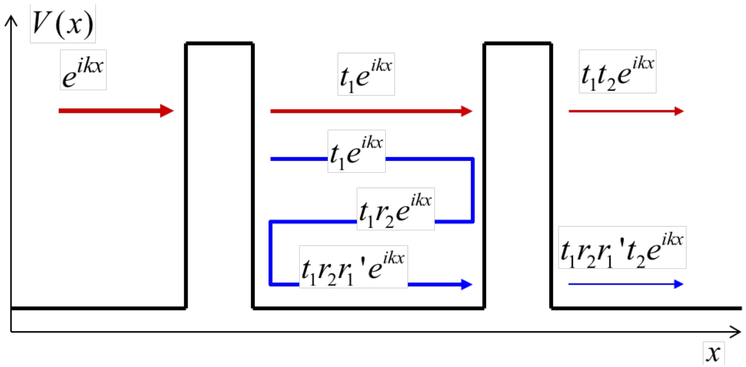

Figure 5: Transmission of double barrier potential. $t_1 t_2 \exp\left[ikx\right]$ is the wave directly transmitted through the two barriers (red arrows). Crossing the first barrier gives a transmitted wave $t_1 \exp\left[ikx\right]$, which acts as incoming wave to the second barrier. Multiplying with t_2 then gives the wave transmitted by the second barrier. $t_1 r_2 r_1' t_2 \exp\left[ikx\right]$ is the wave (blue arrows) transmitted by the first barrier (t_1), reflected by the second barrier (r_2), back-reflected by the first barrier (r_1'), before finally transmitted by the second barrier (t_2).

The term $t_1 t_2$ represents that part of the wave that is directly transmitted by the two barriers. The term $t_2 r_1' r_2 t_1$ represents the wave that is reflected and back-reflected once between the two barriers, before it is transmitted; the term $t_2 r_1' r_2 r_1' r_2 t_1$ represents the wave that is reflected and back-reflected twice, etcetera. The total transmission amplitude t is then the sum over all these transmissions, each representing a certain number of multiple reflections. Note that the reflection and transmission amplitudes are complex numbers. They contain phase factors, see Eqs. 36 and 37, and these phases contain the position of the barriers. The sum over all transmissions of multiple reflections encompasses the interference between those terms due to the phases build up along each reflection path.

Depending on whether this interference is constructive or destructive, the total transmission $T = |t|^2$ is large or small. With total constructive interference, the transmission can even be total, $T = 1$. Total transmission is called a **resonance**.[10] As the phase difference

[10] A standard, but not a very insightful term. Many phenomena in physics are called resonances if they are unexpectedly large.

depends on the wave number k, see Fig. 5, and the wave number depends on the energy E, one may expect that resonances occur only at very specific energies. At other energies the interference will be (partly) destructive, and the transmission is much lower.

A small note for experts*

For people who will/have read other texts, the scattering matrix is ordinarily defined as

$$\begin{aligned} S'_{11} &= S_{11};\;\; S'_{12} = \sqrt{\frac{v_L}{v_R}} S_{12}; \\ S'_{21} &= \sqrt{\frac{v_R}{v_L}} S_{21};\;\; S'_{22} = S_{22}. \end{aligned} \tag{50}$$

The reason for including the $\sqrt{v}$ factors is because one wants the scattering matrix $\mathbf{S}'$ to reflect a basic conservation law, namely the conservation of current. In a stationary problem the current going out must be the same as the current coming in, since otherwise one would get an accumulation or depletion of particles, as discussed in Part II.

$$\begin{aligned} J_{out} &= J_{in} \Rightarrow \\ v_L |B|^2 + v_R |F|^2 &= v_L |A|^2 + v_R |G|^2 \Rightarrow \\ \left\| \begin{pmatrix} \sqrt{v_L} B \\ \sqrt{v_R} F \end{pmatrix} \right\|^2 &= \left\| \begin{pmatrix} \sqrt{v_L} A \\ \sqrt{v_R} G \end{pmatrix} \right\|^2 . \end{aligned} \tag{51}$$

From Eqs. 34 and 50 it is easily shown that this holds only if

$$\mathbf{S}'^{\dagger}\mathbf{S}' = \mathbf{S}'\mathbf{S}'^{\dagger} = \mathbf{I}. \tag{52}$$

In other words, **conservation of current** means that the **scattering matrix is unitary**. Of course, if $v_L = v_R$, as in case of a symmetric barrier, $\mathbf{S}' = \mathbf{S}$.

4 The WKB approximation*

In the days before computers (oh yes, there has been such an age in ancient prehistory), or the use of computers was restricted to the happy few (computational physicists, for instance, in more recent prehistory), there was a need for computational techniques that could executed by hand. Obviously such techniques involve making further approximations. One of the techniques that was popular in handling scattering problems is called the WKB approximation.[11] WKB is a semiclassical approximation, which not only is of use if a computer is not at hand (unlikely today), but also if you need analytical expressions (still handy today). As such it is used today even by people who were not born in the dark ages. I am not going to explain WKB in detail, but only that part that is useful to us now.

Suppose we have a system where the internal reflection and back-reflection coefficients are small, $|r_i| \ll 1$, $|r'_i| \ll 1$. This typically happens when calculating the transmission of a thick barrier at energies sufficiently far below the top of the barrier. The physical

[11] WKB stands for "Wenzel, Kramers, Brillouin", a German-Dutch-French collaboration dating from 1926. In the anglo-saxon literature you also find JWKB, where the extra J stands for "Jeffreys". The English, what can you say; after brexit they will call it the J approximation.

situation is sketched in Fig. 6. In such a situation all multiple-reflected waves have a very small amplitude compared to the directly transmitted waves, and we may neglect their contribution. Eq. 49 then becomes

$$t \approx t_1 t_2 \quad \text{or} \quad T \approx T_1 T_2, \tag{53}$$

with $T_i = |t_i|^2$. This is essentially the WKB approximation in the scattering context. It is called a semiclassical approximation, because part of it is quantum, as tunneling through a barrier is a quantum phenomenon, but there is no interference between (reflected) waves. As interference is a key characteristic of waves, one could say the wave character is lost, as in classical mechanics.

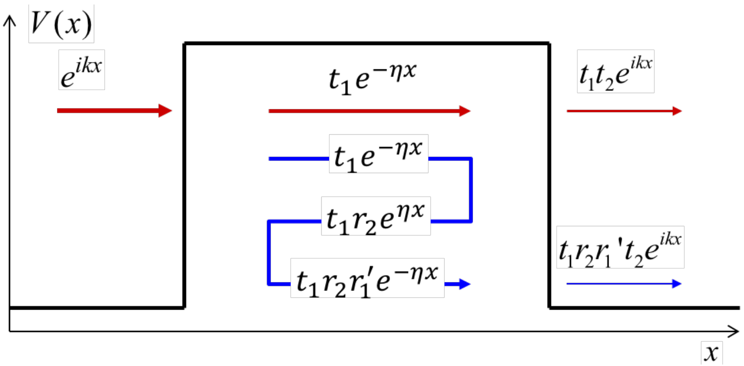

Figure 6: Transmission of a thick barrier at an energy where the waves inside the barrier are evanescent. As in Fig. 5, $t_1 t_2 \exp[ikx]$ is the wave directly transmitted through the two barriers (red arrows), whereas $t_1 r_2 r_1' t_2 \exp[ikx]$ is the wave transmitted after one reflection and back reflection between the potential steps (blue arrows). As the wave is decaying all along its path ($\exp[-\eta x]$ decays for x going from left to right, and $\exp[\eta x]$ decays for x going from right to left), the (back-)reflection coefficients are small.

As $t_1 = 1/M_{11}^{(1)}$ and $t_2 = 1/M_{11}^{(2)}$, see Eqs. 35 and 46, one has in this approximation

$$M_{11} \approx M_{11}^{(1)} M_{11}^{(2)}. \tag{54}$$

The transmission of a square barrier becomes

$$T_{\text{WKB}} \approx \frac{1}{\left|M_{11}^{(1)} M_{11}^{(2)}\right|^2} = \left(\frac{4k\eta}{k^2+\eta^2}\right)^2 e^{-2\eta(b-a)}, \tag{55}$$

using Eqs. 9, 4 and 6.[12] This expression shows the exponential decay one expects for tunneling through a thick barrier.

The procedure can be extended to a barrier of any shape, see Fig. 2. Using the expressions of Eqs. 22 and 23, and keeping only the matrix elements $(\mathbf{P}_i)_{11}$ and $(\mathbf{Q}_i)_{11}$, one obtains[13]

$$M_{11} \approx e^{i\phi} \prod_{i=1}^{N} \frac{\eta_{i-1}+\eta_i}{2\eta_{i-1}} \exp[\eta_i \Delta x_i]. \tag{56}$$

[12] The same result can be obtained from Eq. 10 by taking the limit $\eta(b-a) \gg 1$.

[13] $i\phi = -\eta_0 x_1 + 2\eta_N x_N - \eta_N x_{N+1}$ is an irrelevant phase factor. Note that, because we have traveling incoming waves and transmitted waves, both η_0 and η_N must be imaginary numbers, see Eq. 19.

One can take the continuum limit of this expression $\lim_{\Delta x_i \to 0}$. If the potential is a continuous function, then $\lim_{\Delta x_i \to 0} \left[(\eta_{i-1} + \eta_i) \, / \, (2\eta_{i-1}) \right] = 1$, see Eq. 18, and

$$\lim_{\Delta x_i \to 0} M_{11} \approx e^{i\phi} \lim_{\Delta x_i \to 0} \prod_{i=1}^{N} \exp\left[\eta_i \Delta x_i\right] = e^{i\phi} \lim_{\Delta x_i \to 0} \exp\left[\sum_{i=1}^{N} \eta_i \Delta x_i\right],$$

which obviously can be rewritten as

$$M_{11} \approx e^{i\phi} \exp\left(\int_{x_1}^{x_N} \eta(x) dx\right) \quad \text{with} \quad \eta(x) = \frac{\sqrt{2m\left[V(x) - E\right]}}{\hbar}. \tag{57}$$

The transmission can then be written as

$$T_{\text{WKB}} = \frac{v_R}{v_L} \frac{1}{|M_{11}|^2} \approx \frac{v_R}{v_L} \exp\left(-2 \int_{x_{\text{in}}}^{x_{\text{out}}} \eta(x) dx\right) \quad \text{with} \quad V(x_{\text{in}}) = V(x_{\text{out}}) = E, \tag{58}$$

the inner and outer classical turning points. The step from Eqs. 57 to 58 goes as follows. As $T \propto 1/|M_{11}|^2$, if $V(x) < E$, then $\eta(x)$ is purely imaginary, so $\exp\left[\eta(x)dx\right]$ is a pure phase factor, and $|\exp\left[\eta(x)dx\right]| = 1$. This means that we only get a non-trivial contribution to T from points where $V(x) > E$, since then $\eta(x)$ is a real number. The points where $V(x) > E$ are in the interval $x_{\text{in}} < x < x_{\text{out}}$.[14] Eq. 58 is called the **WKB approximation** for tunneling.

It does not hold for potentials that are discontinuous, such as the square barrier of Fig. 6, but that is easily mended in view of Eq. 56. Suppose there is a discontinuity at $x = a$, then one has to multiply the expression by

$$\left|\frac{2\eta_{\uparrow}}{\eta_{\uparrow} + \eta_{\downarrow}}\right|^2 \quad \text{where} \quad \eta_{\uparrow} = \lim_{x \uparrow a} \eta(x) \text{ and } \eta_{\downarrow} = \lim_{x \downarrow a} \eta(x). \tag{59}$$

One has to do this for each discontinuity. For a square barrier there are two discontinuities, one at the start and one at the end of the barrier. Using Eq. 59 as a recipe for these two, and multiplying with Eq. 58 then gives the expression of Eq. 55.

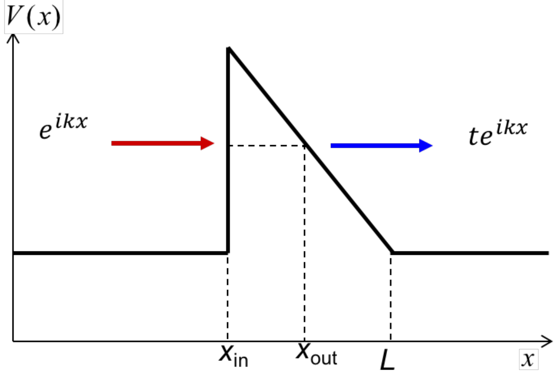

Figure 7: Transmission of a triangular barrier. x_{in} and x_{out} mark the classical turning points of a particle with energy $E = \hbar^2 k^2/(2m)$.

[14] In classical physics the kinetic energy $K \geq 0$. As $K + V = E$, this means that $V \leq E$ in classical physics. The points x where $V(x) = E$ are called the *turning points*, as a classical particle reverts its motion there. A quantum particle can penetrate the "forbidden" region, where $V > E$ (the kinetic energy $K < 0$ there), which is called *tunneling*.

The WKB approximation is particularly handy if one needs a quick calculation, as Eq. 58 essentially only involves the calculation of one integral. Consider the case shown in Fig. 7. This potential has a triangular shape

$$V(x) = \begin{cases} 0 & x < x_{\text{in}} \\ \frac{V_0}{L-x_{\text{in}}}(L-x) & x_{\text{in}} < x < L \ . \\ 0 & x > L \end{cases} \tag{60}$$

Of course one can calculate the transmission through this barrier numerically "exact", approximating it by a sequence of potential steps, as in Fig. 2. The WKB approximation gives a quick approximative answer

$$T \approx \left| \frac{2ik}{ik+\eta} \right|^2 \exp\left(-\frac{2\sqrt{2m}}{\hbar} \int_{x_{\text{in}}}^{x_{\text{out}}} \sqrt{V(x)-E}\, dx \right).$$

The integral can be done easily, giving

$$T_{\text{FN}} = \frac{4E}{V_0} \exp\left(-\frac{4\sqrt{2m}}{3\hbar} \frac{\Phi_b^{\frac{3}{2}}}{e\mathcal{E}} \right) \quad \text{with} \quad \Phi_b = V_0 - E, \quad \text{and} \quad e\mathcal{E} = \frac{V_0}{L-x_{\text{in}}}. \tag{61}$$

Here Φ_b is the effective barrier height, and $\mathcal{E}$ is the electric field across the barrier. Eq. 61 is called the **Fowler-Nordheim equation**. It gives a characteristic dependence of the transmission (and hence of the tunnel current, see Part II) as function of the electric field strength $\mathcal{E}$ as $T \propto \exp[-\alpha/\mathcal{E}]$. It is used to describe the field emission of electron guns, for instance, or field-induced injection currents in electronic devices. Considering all the approximations we have made, it is of course far from exact, but experimentalists like it a lot because of its simplicity.[15]

Part II

Theoretical background

5 Quantum currents

The **probability current** $J(x,t)$

$$J(x,t) = \frac{i\hbar}{2m}\left(\Psi \frac{\partial \Psi^*}{\partial x} - \Psi^* \frac{\partial \Psi}{\partial x} \right) \tag{62}$$

is defined such that

$$\frac{dP_{ab}(t)}{dt} = J(a,t) - J(b,t) \tag{63}$$

describes the change in the probability $P_{ab}(t)$ of finding a particle in the region $a < x < b$ at time t, where

$$P_{ab}(t) = \int_a^b |\Psi(x,t)|^2\, dx, \tag{64}$$

[15] Experimentalists tend to prefer simplicity over correctness.

provided the wave function $\Psi(x,t)$ describing the particle is normalized. Proving this is trivial:

$$\begin{aligned}
\frac{dP_{ab}(t)}{dt} &= \int_a^b \left[\Psi^*(x,t)\frac{\partial\Psi(x,t)}{\partial t} + \frac{\partial\Psi^*(x,t)}{\partial t}\Psi(x,t)\right] dx \\
&= \frac{i\hbar}{2m}\int_a^b \left[\Psi^*(x,t)\frac{\partial^2\Psi(x,t)}{\partial x^2} - \frac{\partial^2\Psi^*(x,t)}{\partial x^2}\Psi(x,t)\right] dx \\
&= \frac{i\hbar}{2m}\left[\Psi^*(x,t)\frac{\partial\Psi(x,t)}{\partial x} - \frac{\partial\Psi^*(x,t)}{\partial x}\Psi(x,t)\right]_a^b .
\end{aligned}$$

Going from the first to the second line one uses the Schrödinger equation $i\hbar\frac{\partial\Psi(x,t)}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V(x)\Psi(x,t)$ and its complex conjugate (the potential terms cancel), and from the second to the third line one uses partial integration. Setting $b = a + dx$ with dx infinitesimal, allows one to write Eq. 63 as

$$\frac{\partial\rho(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}, \tag{65}$$

with $\rho(x,t) = |\Psi(x,t)|^2$ the probability density. You might recognize Eq. 65 as a continuity equation, which describes the relation between a density and a current.

Probability currents may seem rather abstract, but they are easily related to something more familiar. Suppose the particle has a charge q, then the expected charge found in the region $a < x < b$ at time t is $Q_{ab}(t) = qP_{ab}(t)$.[16] Defining the **electrical current** as $I(x,t) = qJ(x,t)$, Eq. 63 can be rewritten as

$$\frac{dQ_{ab}(t)}{dt} = I(a,t) - I(b,t). \tag{66}$$

The rate of change of charge is given by the difference between the current flowing in from one side minus the current flowing out from the other side.

A nice thing is that, even if the *wave function cannot be normalized*, like the wave function of a free particle, the *probability current* according to Eq. 62 is still a *well-defined* quantity. Free particles often enter in scattering problems, where we are interested in quantities like reflection and transmission coefficients. Since the latter can be directly defined in terms of probability currents, we can get away with using non-normalizable wave functions.[17]

5.1 Stationary states

Suppose now that $\Psi(x,t)$ describes a **stationary state**, i.e.

$$\Psi(x,t) = \psi(x)e^{-\frac{i}{\hbar}Et}. \tag{67}$$

[16] This is an expectation value in the quantum mechanical sense. One starts the wave at time 0 and at time t one measures whether the particle is in the region $a < x < b$. By repeating this "experiment" over and over, one can calculate the probability $P_{ab}(t)$. $Q_{ab}(t)$ is then the average charge found in this region from these repeated "experiments". If you have problems imagining this, then think of a particle emitter that sends out a pulse of many (independent) particles. The averaging is then done automatically and the average charge is what you measure.

[17] For quantum purists: one *can* work with normalizable wave packets. The math then becomes ugly, and, in the proper limit, the physical results will be the same as when using plane waves.

Then one finds from Eq. 64 $dP_{ab}(t)/dt = 0$ and from Eqs. 63 and 62

$$J(x,t) = J \ = \ \text{constant}. \tag{68}$$

The probability current is constant, i.e. independent of position and time.

For example, consider a free particle with the wave function

$$\psi(x) = A\, e^{ikx}. \tag{69}$$

From Eq. 64 we calculate

$$P_{ab} = |A|^2\,(b-a). \tag{70}$$

Since P_{ab} is the probability of finding the particle in the interval between $x = a$ and $x = b$, i.e. an interval of length $b - a$, we can interpret $|A|^2$ as the probability density per unit length. It is also called the particle density[18]

$$\rho = |A|^2\,. \tag{71}$$

The probability current is easily calculated from its definition, Eq. 62,

$$J = \frac{\hbar k}{m}\,|A|^2 = \frac{\hbar k}{m}\rho. \tag{72}$$

According to de Broglie's relation $p = \hbar k$ is the momentum of the particle and

$$v = \frac{\hbar k}{m} = \frac{p}{m} \tag{73}$$

is then the velocity of the particle. The electrical current is given by

$$I = qJ = qv\rho, \tag{74}$$

which is the usual definition of an electrical current, namely charge×velocity×density. For the wave function of Eq. 69 both velocity and density are constant, so the wave function describes a uniform current. Suppose $q > 0$; then if $k > 0$, the current flows to the right, if $k < 0$, the current flows to the left. From now on we assume that $k > 0$.

Now let's go to the more complicated wave function

$$\psi(x) = Ae^{ikx} + Be^{-ikx}, \tag{75}$$

with A, B constants. The associated probability current is

$$J = \frac{\hbar k}{m}\,|A|^2 - \frac{\hbar k}{m}\,|B|^2\,, \tag{76}$$

which is interpreted as a right going current minus a left going current. In a scattering problem one would interpret the first term on the right-hand side of Eq. 75 as the incident wave and the second term as the reflected wave. Eq. 76 is then interpreted as the difference between incident and reflected currents

$$J = J_{in} - J_R. \tag{77}$$

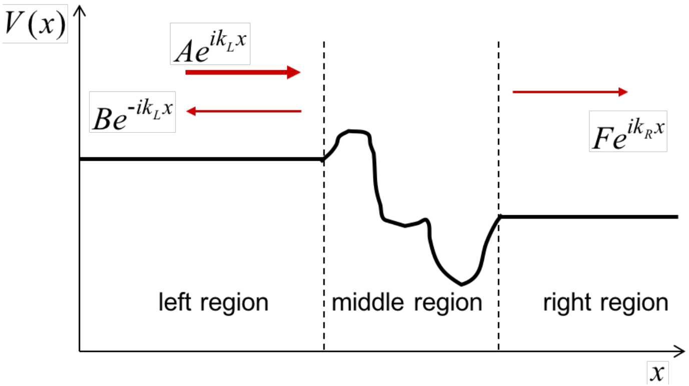

Figure 8: Cartoon representing a general one-dimensional scattering problem. In the left region the potential is constant $V(x) = V_L$, in the middle region the potential $V(x)$ can be anything, and in the right region the potential is constant $V(x) = V_R$. The middle region is called the **scattering region**. The left and right regions are called the **left and right leads**. In the left lead we have an incoming wave $Ae^{ik_L x}$ and a reflected wave $Be^{-ik_L x}$ and in the right lead we have a transmitted wave $Fe^{ik_R x}$.

The reflection coefficient R is defined as the ratio between reflected and incident currents

$$R = \frac{J_R}{J_{in}} = \frac{|B|^2}{|A|^2}. \tag{78}$$

Consider the scattering problem shown in Fig. 8. In the left region we assume that the potential is constant $V(x) = V_L$, in the middle region the potential can have any shape $V(x)$, and in the right region the potential is again constant $V(x) = V_R$. The solution in the left region is given by Eq. 75 with k replaced by k_L

$$k_L = \frac{\sqrt{2m(E - V_L)}}{\hbar}. \tag{79}$$

The solution in the right region is given by the transmitted wave

$$\psi(x) = Fe^{ik_R x}\,;\; x \text{ in right region}, \tag{80}$$

with

$$k_R = \frac{\sqrt{2m(E - V_R)}}{\hbar}. \tag{81}$$

One can calculate the transmitted current as

$$J_T = \frac{\hbar k_R}{m} |F|^2. \tag{82}$$

The transmission coefficient T is defined as the ratio between transmitted and incident currents

$$T = \frac{J_T}{J_{in}} = \frac{v_R}{v_L} \frac{|F|^2}{|A|^2}, \tag{83}$$

[18] Using a beam of N particles in which each particle is independent of the others and is described by the wave function of Eq. 69, the particle density is $N|A|^2$, which is the number of particles to be found per unit length.

using Eq. 73.

From the fact that the current has to be independent of position *everywhere*, see Eq. 68, it follows from Eqs. 77 and 82 that

$$\begin{aligned} J_{in} - J_R &= J_T \Leftrightarrow \\ J_{in} &= J_R + J_T. \end{aligned} \tag{84}$$

This relation expresses the conservation of current, or: "current in = current out" (reflected plus transmitted). No matter how weird the potential in the middle region is, the current going into it has to be equal to the total current coming out of it. No particles magically appear or disappear in the middle region. From the definitions of Eqs. 78 and 83 it is shown that Eq. 84 is equivalent to

$$1 = R + T, \tag{85}$$

i.e., the reflection and transmission coefficients add up to 1. Since these coefficients denote the probabilities that a particle is reflected or transmitted, this simply states that particles are either reflected or transmitted.

6 Quantum conductance

The device shown in Fig. 9 is called a tunnel junction. The left and right regions consist of metals and the middle region consists of an insulator material, e.g., a metal-oxide.[19] Such devices can be made in a very controlled way with the middle region having a thickness of a few nm. One is interested in electrical currents, i.e. the transport of electrons through such junctions, or, more generally, in the current-voltage characteristics of such a device.[20] On this small, nanometer length scale electrons have to be considered as waves and quantum tunneling is important. **Nano-electronics** is the general name of the field where one designs and studies special devices that make use of this kind of electron wave behavior.

We start with the simplest possible one-dimensional model of a tunnel junction. We approximate the potential for electrons in a material by a constant.[21] The constant potential depends on the atoms a material is composed of, so it is different for every material. The potential in the tunnel junction of Fig. 9 along the transport direction can then be represented by a square barrier, as shown in Fig. 10.

6.1 The Landauer formula in linear response

According to Eq. 83 the transmitted electrical current is given by

$$I_T = I_{in} T, \tag{86}$$

[19] Scanning tunneling microscopy (STM) uses a tunnel junction between the probe tip and a surface, where the middle region is simply vacuum.

[20] The device is called MIM, which stands for metal-insulator-metal. Using magnetic metals the device can be applied as a magnetic field sensor, or in MRAMs (magnetic random access memories).

[21] This approximation is often used for simple metals such as alkali's, aluminium, silver and gold. The constant potential approximation is also called the **jellium approximation**. It does not hold for complicated metals such as the transition metals or for insulators. In all honesty, it doesn't even hold very well for simple metals if one is interested in quantitative results. But, never let reality interfere with a good story.

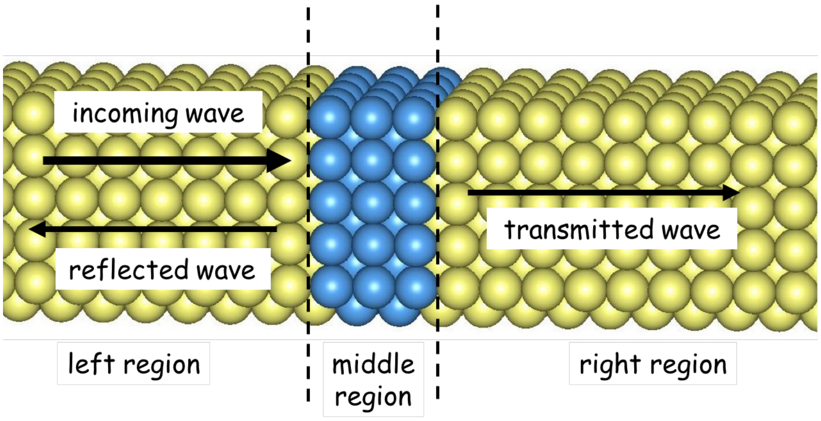

Figure 9: Schematic representation of a tunnel junction. The yellow balls represent atoms of a metal, the blue balls represent atoms of an insulator. The left and right regions stretch macroscopically far into the left and right, respectively. The electron waves in the metal are reflected or transmitted by the insulator in the middle region.

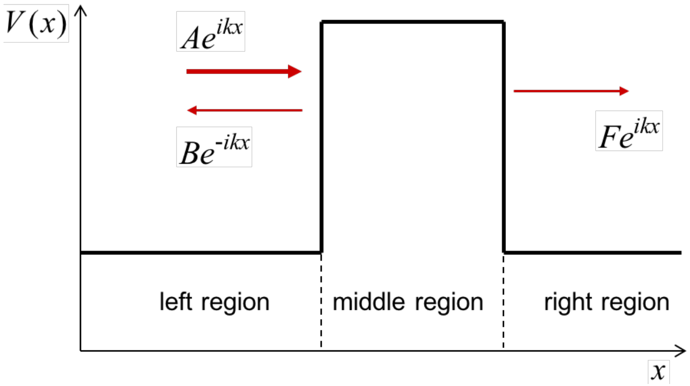

Figure 10: Simple approximation of the potential along the transport direction of a tunnel junction, see Fig. 9. In the metal (left and right regions) the potential is constant, $V(x) = V_1$. In the insulator the potential is also constant, $V(x) = V_0$, where $V_0 > V_1$. The incoming, reflected and transmitted waves are given by Ae^{ikx}, Be^{-ikx} and Fe^{ikx}.

using the definition $I = qJ$ (the charge q of an electron is $-e$). The incoming current I_{in} is given by Eq. 74 as

$$I_{in} = -ev\rho. \tag{87}$$

How large are the velocity v and density ρ of the incoming electrons? To answer that question we must ask a more basic one: how is the incoming current created in a device? In an experimental setup this is done by applying a voltage difference U between the left and right regions. The left and right regions are metals, which can be connected to the two ends of a battery, for instance. This results in a potential drop $\Delta V = -eU$ between left and right regions, as shown in Fig. 11.[22] We suppose that the temperature is low and the metals are

[22] By "potential" (ΔV) I mean "potential energy", dimension "eV" (or "J", if you insist on SI). By "voltage" (U) I mean "potential" in the electrostatic sense, dimension "V". Unfortunately there is a confusing difference in sign between the two, because some idiot in the past has chosen the charge of an electron as negative, and we idiots follow him.

non-magnetic, so we have spin degeneracy. Then $v\rho$ is given by the simple expression

$$v\rho = \frac{\Delta V}{\pi\hbar} = -\frac{eU}{\pi\hbar}. \tag{88}$$

This expression is derived in the next section.

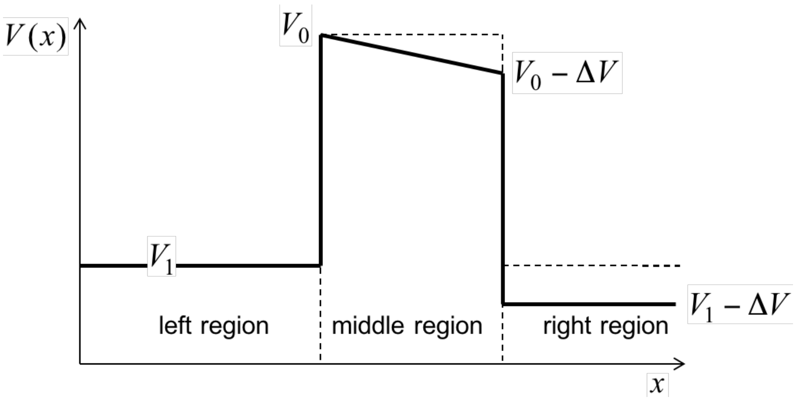

Figure 11: The potential profile when a bias voltage U is applied between the left and right leads. This changes the potential of the right region by $\Delta V = -eU$ to $V_1 - \Delta V$ with respect to the left region; see Fig. 10. The potential drop is indicated schematically, using the sign convention that $\Delta V > 0$ means that the left region has a lower potential (see footnote on the previous page). If the bias voltage is small, i.e. $\Delta V \ll V_0 - V_1$, then we can still use the transmission coefficient T calculated for the unbiased square barrier (given by the dashed line).

If the potential drop ΔV in Fig. 11 is small compared to the barrier height $V_0 - V_1$, we can use the unbiased square barrier potential from Fig. 10 to calculated the transmission coefficient T. This is the so-called **linear response regime**. Eqs. 86-88 then give for the transmitted electrical current, also called the **tunneling current**

$$I_T = \frac{e^2}{\pi\hbar} U\, T, \tag{89}$$

which is a remarkably simple expression. If we define the **conductance** $\mathcal{G}$ as current divided by voltage, we get

$$\mathcal{G} = \frac{I_T}{U} = \frac{e^2}{\pi\hbar} T. \tag{90}$$

Since T is just a dimensionless number between 0 and 1, $e^2/\pi\hbar$ has the dimension of conductance. It is the fundamental **quantum of conductance**; its value is $e^2/\pi\hbar \approx 7.75 \times 10^{-5}$ Ω^{-1}.[23]

Eq. 90 is called the **Landauer formula**; it plays a central role in nano-electronics.[24] As it stands here, it is valid for a one-dimensional, spin-degenerate system at low voltage and at not too high a temperature, but it can be generalized.

[23] If you are more used to working with resistances, the resistance $\mathcal{R}$ is the inverse of the conductance, i.e. $\mathcal{R} = 1/\mathcal{G}$, so the quantum of resistance is $\pi\hbar/e^2 \approx 12.9\text{k}\Omega$.

[24] R. Landauer, *Philosophical Magazine* **21**, 863 (1970). Also called the Landauer-Büttiker formalism. The first idea to relate the conductance of a small system to its scattering properties comes from Landauer. Büttiker made a generalization to multi-terminal devices with more than two electrodes.

6.2 Simple derivation of the Landauer formula*

I will give a very simple derivation of Eq. 88. We have to do a little bit of solid state physics, but I use simple introductory quantum mechanics language only. Spin degeneracy means that each energy level can be filled with two electrons. The non spin-degenerate case is relevant for magnetic materials. I let you work out that case yourself.

The Pauli exclusion principle and the Fermi energy

The left and right regions of a tunnel junction consist of metal wires, see Fig. 9. These wires are supposed to be very, very long compared to the size of the middle region. In a simple-minded model the potential of a long metal wire looks like Fig. 12. The potential is approximately constant inside the wire and it has steps at the beginning and end of the wire to keep the electrons in. The energy levels of this square well potential are,[25]

$$E_n = \frac{\hbar^2 k_n^2}{2m} \quad \text{with} \quad k_n = n\frac{\pi}{L}. \tag{91}$$

The spacing between the energy levels, $E_n - E_{n-1}$, scales as $1/L^2$ with the length L of the wire. If L is large, the spacing becomes very small, so from a distance the energy level spectrum almost looks like a continuum, as illustrated by Fig. 12.

The wave functions are given by

$$\psi_n(x) = \sqrt{\frac{2}{L}} \sin k_n x = \frac{1}{i\sqrt{2L}} \left(e^{ik_n x} - e^{-ik_n x} \right). \tag{92}$$

These are not exactly what we need, because they correspond to standing waves, whereas we need traveling waves to describe currents, see Eq. 69. For the incoming current we only need the $\exp(ik_n x)$ part. Setting $A = 1/(i\sqrt{2L})$, the corresponding electron density according to Eq. 71 is

$$\rho = |A|^2 = \frac{1}{2L}. \tag{93}$$

The wire is full of electrons since each of the atoms in the wire brings at least one electron with it. Filling the energy levels according to the Pauli principle, and having N electrons in total, the highest occupied level is $E_{\frac{N}{2}}$. The highest occupied level in the ground state is called the **Fermi energy** or E_F.[26]

Incoming and tunnel currents

Go back to the tunnel junction and fill its potential profile, Fig. 10, with electrons from the left and right wires. This is shown in Fig. 13. The Fermi energy E_F in the left and right regions is the same. The exclusion principle then tells us that there can be no flow of current. Any electron on the left side that would try to go to the right side or vice versa finds an energy level that is already occupied by an electron, which excludes any other electron

[25] This is actually the solution for an infinitely deep square well, whereas you might think that we need the solution for a square well with a finite depth. However, if the well is very wide and not too shallow, the infinite well is a good approximation.

[26] In an ordinary metal the Fermi energy is of order 10 eV with respect to the lowest energy level of the valence electrons.

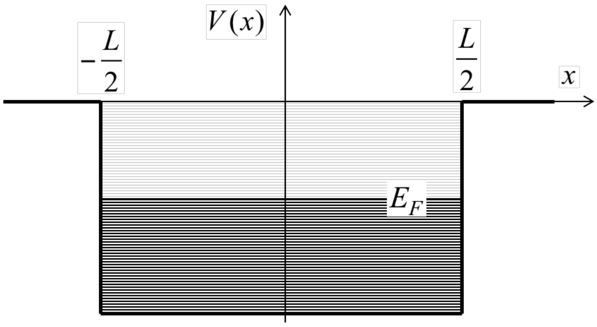

Figure 12: Schematic drawing of the potential and the energy levels of a long wire. The points $-L/2$ and $L/2$ mark the beginning and the end of the wire. The spacing between the energy levels is so small that the energy spectrum almost looks like a continuum. E_F marks the Fermi energy, i.e. the highest level that is occupied in the ground state by an electron.

from going there. This confirms what we know from everyday life; in an unbiased system no current flows.[27]

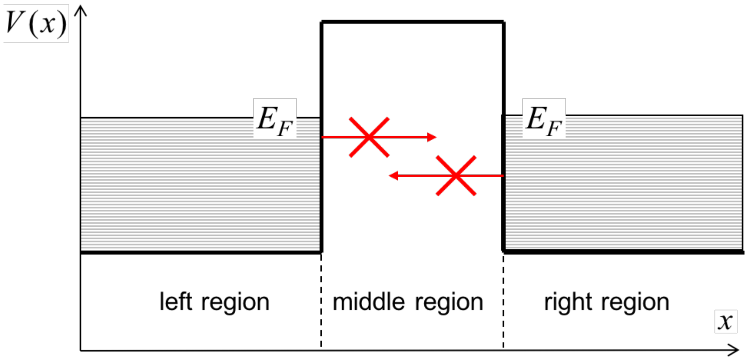

Figure 13: Tunnel junction where left and right regions are filled with electrons. The Fermi energies E_F on the left and right side are identical. The exclusion principle forbids electrons to trespass from left to right or vice versa.

Now apply a bias voltage between left and right regions as in Fig. 11. The result is shown in Fig. 14. The bias voltage lowers the right region in energy. Suddenly the electrons on the left side that occupy energy levels E_n with $E_F - \Delta V < E_n < E_F$ find empty levels with that energy on the right side. They can tunnel through the barrier to occupy these levels. Provided the potential drop ΔV is small, we can approximate the transmission coefficients of all these electrons by T at an energy $E = E_F$.[28] The incoming current of Eq. 87 has contributions from <u>*all*</u> electrons with energies between $E_F - \Delta V$ and E_F.

$$I_{in} = -2e\rho \sum_{E_F - \Delta V < E_n < E_F} v_n. \tag{94}$$

[27] One can also reverse the argument. If the Fermi energies on the left and right side would be different, then a current would flow. However this current would be short-lived. By sending electrons from left to right one occupies a level on the right side, and de-occupies a level on the left side. This would go on until the highest occupied levels on the left and right are the same, .e., the Fermi level is the same everywhere, which marks equilibrium.

[28] Again, this is valid only in the linear response regime.

The factor of 2 is there because there are two electrons in each level (one spin up, and one spin down).

This sum in Eq. 94 is rather awkward, but using a trick we can turning it into an integral

$$\begin{aligned}\sum v_n &= \frac{L}{\pi}\sum v_n \frac{\pi}{L} = \frac{L}{\pi}\sum v_n \Delta k \\ &\approx \frac{L}{\pi}\int v dk, \end{aligned} \tag{95}$$

where

$$\Delta k = k_n - k_{n-1} = \frac{\pi}{L}, \tag{96}$$

see Eq. 91. Turning the sum into an integral is allowed because L is very large, so Δk is tiny. The lower and upper bound of the integral in Eq. 95 should correspond to the energies $E_F - \Delta V$ and E_F, whereas the integral is over dk, which is again awkward. We can however turn it into an integral over dE, using the following trick

$$v = \frac{\hbar k}{m} = \frac{1}{\hbar}\frac{d\left(\frac{\hbar^2 k^2}{2m}\right)}{dk} = \frac{1}{\hbar}\frac{dE}{dk}. \tag{97}$$

In Eq. 95 it gives

$$\begin{aligned}\int_{E_F-\Delta V}^{E_F} v dk &= \int_{E_F-\Delta V}^{E_F} \frac{1}{\hbar}\frac{dE}{dk}dk = \int_{E_F-\Delta V}^{E_F} \frac{1}{\hbar}dE \\ &= \frac{1}{\hbar}\left(E_F - (E_F - \Delta V)\right) = \frac{1}{\hbar}\Delta V. \end{aligned} \tag{98}$$

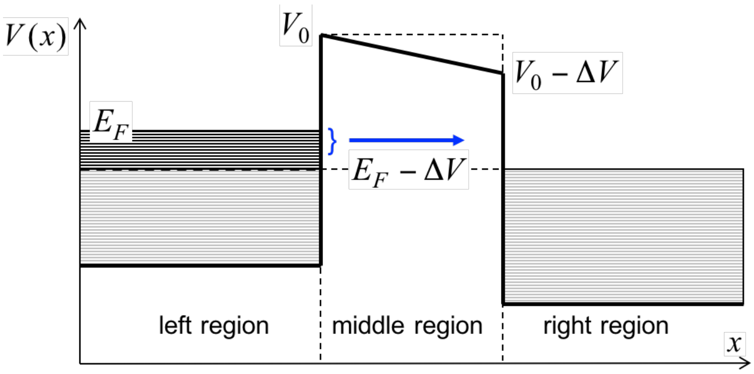

Figure 14: Tunnel junction with an applied bias potential $\Delta V > 0$. All the levels occupied by electrons in the left region with an energy $E_F - \Delta V < E_n < E_F$ correspond to empty energy levels in the right region. The electrons in these levels can tunnel from though the barrier from left to right.

Collecting Eqs. 98, 95 and Eq. 93 in Eq. 94, we find for the incoming current

$$I_{in} = -e\frac{\Delta V}{\pi\hbar}. \tag{99}$$

This is the required expression for the incoming current, see Eqs. 87 and 88. The tunnel current is then given by

$$I_T = I_{in}T = -\frac{e\Delta V}{\pi\hbar}T, \tag{100}$$

where the transmission coefficient T needs to be calculated for the energy $E = E_F$. Note that with $\Delta V = -eU$, where U is the potential difference (in Volts), this corresponds to Eq. 89. The Landauer formula, Eq. 90, is then derived straightforwardly.

6.3 Finite bias

The Landauer formula is typically used if the bias ΔV is small. Two conditions need to be fulfilled: (i) the transmission coefficient is approximately constant for energies $E_F - \Delta V < E < E_F$, and (ii) the potential barrier is not appreciably distorted by the applied bias (the slanted line between V_0 and $V_0 - \Delta V$ in Fig. 14 must be close to the dashed line). In the most general case, the transmission coefficient T depends on the energy E, as well as on the shape of the barrier, represented by ΔV. If the conditions (i) and (ii) are met, however, we can approximate $T(E, \Delta V) \approx T(E_F, 0)$, which is the transmission coefficient used in the linear response regime discussed in Sec. 6.1.

In experiment it is quite possible to achieve situations where conditions (i) and (ii) do not hold. Following the reasoning of the previous section, the generalization of Eq. 100 for the case depicted in Fig. 14 ($\Delta V > 0$) is given by

$$\boxed{I_T = -\frac{e}{\pi\hbar}\int_{E_F-\Delta V}^{E_F} T(E, \Delta V)dE,} \tag{101}$$

with T the energy dependent transmission, calculated for the distorted barrier. Note that, if $T(E, \Delta V) \approx T(E_F, 0)$, we get Eq. 100 again. Eq. 101 holds as long as $\Delta V < E_F$, see Fig. Fig. 14; for $\Delta V > E_F$, the lower bound of the integral has to be replaced by 0, as obviously there are no states below the bottom of the energy band.

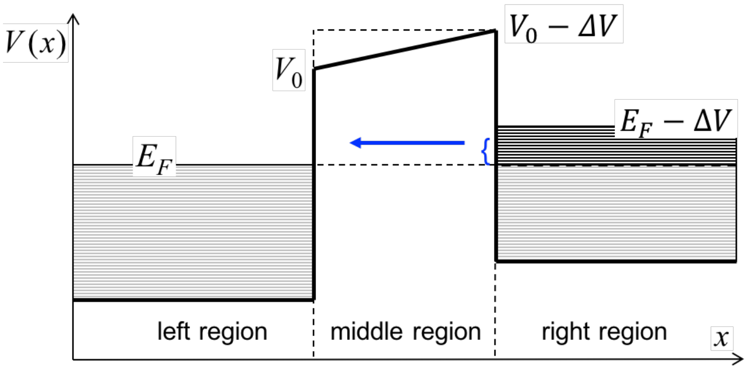

Figure 15: Tunnel junction with an applied bias potential $\Delta V < 0$. All the levels occupied by electrons in the right region with an energy $E_F < E_n < E_F - \Delta V$ correspond to empty energy levels in the left region. The electrons in these levels can tunnel from though the barrier from right to left.

One of the nice things about Eq. 101 is that it is also valid for negative applied bias potentials, $\Delta V < 0$. The situation is sketched in Fig. 15. Electrons with energies $E_F < E < E_F - \Delta V$ now tunnel from right to left (instead of from left to right, as in Fig. 14). This change in current direction is covered by the change of sign of the integral of Eq. 101. Furthermore, at fixed energy $T_{\text{left}\rightarrow\text{right}}(E, \Delta V) = T_{\text{right}\rightarrow\text{left}}(E, \Delta V)$, i.e., the transmission coefficient for a wave incident from the left is the same as that for a wave incident from the right.[29] It means that when calculating the current according to Eq. 101, one can use either $T_{\text{left}\rightarrow\text{right}}$ or $T_{\text{right}\rightarrow\text{left}}$.

Also in case $\Delta V < 0$ one has to be a bit careful with the bounds of the integral of Eq. 101. If $-\Delta V > E_F$, then the upper bound of the integral has to be replaced by $-\Delta V$, as there are no states below the bottom of the band, see Fig. 15.

6.4 Finite temperature*

One can write Eq. 101 in a slightly different form

$$I_T = -\frac{e}{\pi\hbar}\int_{-\infty}^{\infty} T(E,\Delta V)\Theta\left(E - E_F + \Delta V\right)\left[1 - \Theta\left(E - E_F\right)\right] dE, \tag{102}$$

where Θ is the (Heaviside) step function

$$\Theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}. \tag{103}$$

The step function is related to the zero temperature limit of the Fermi-Dirac distribution

$$f(E, E_F) = \frac{1}{\exp\left[\left(E - E_F\right)/k_B T\right] + 1},$$

because

$$\lim_{T\rightarrow 0} f(E, E_F) = 1 - \Theta\left(E - E_F\right).$$

The Fermi-Dirac distribution $0 \leq f(E, E_F) \leq 1$ gives the occupation number of each energy level E. At $T = 0$, all levels up to $E = E_F$ are fully occupied, and all levels with higher energy are completely empty. At $T > 0$, the energy levels close to E_F become partially occupied.

It strongly suggests that we can generalize Eq. 102 to a finite temperature as

$$I_T = -\frac{e}{\pi\hbar}\int_{-\infty}^{\infty} T(E,\Delta V)\left[1 - f\left(E, E_F - \Delta V\right)\right] f\left(E, E_F\right) dE. \tag{104}$$

Looking at Fig. 14, the interpretation of this expression is straight-forward. Only electrons that tunnel from left to right contribute to the current. These electrons start in states on the left described by an occupation number $f\left(E, E_F\right)$. However they can only tunnel to states on the right, if these are empty, as the Pauli principle forbids a double occupancy of states. The emptiness of a state is measured by $1 - f\left(E, E_F\right)$, but because the right side is shifted in energy by ΔV, this factor becomes $1 - f\left(E, E_F - \Delta V\right)$. Finally, off course, the probability of tunneling is given by the transmission coefficient T.

[29] If you like, this is a consequence of a fundamental symmetry that the Schrödinger equation obeys: time-reversal symmetry, also called motion reversal symmetry. The classical wave equation for optics or acoustics also obeys this symmetry.

7 Landauer for layered 3D materials

We start with a generalization of the square barrier potential of Fig. 10. A two-dimensional example is shown in Fig. 16. The potential $V(x, y)$ is separable; the barrier is in the x-direction and the potential is independent of y. The straightforward extension to three dimensions is a potential that is independent of y and z. Such a potential landscape is a simple model for a thin layer of an insulator sandwiched between two metals, in other words, a tunnel junction.

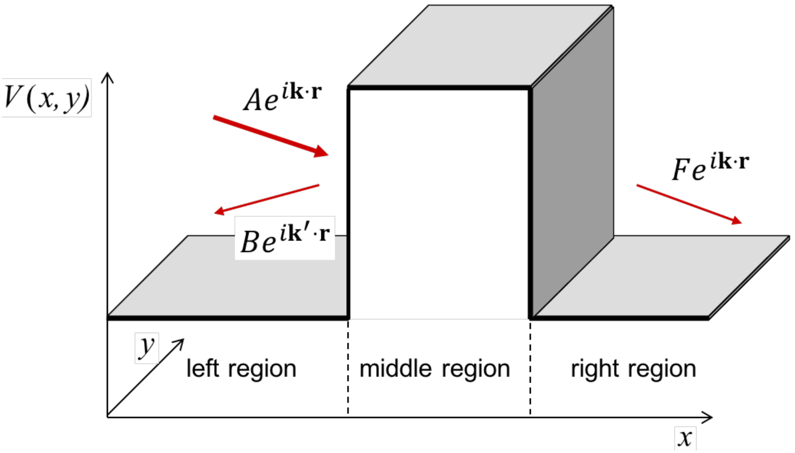
$V(x,y)$
$Ae^{i\mathbf{k}\cdot\mathbf{r}}$
$Be^{i\mathbf{k}'\cdot\mathbf{r}}$
$Fe^{i\mathbf{k}\cdot\mathbf{r}}$
y
left region
middle region
right region
x

Figure 16: A square barrier in two dimensions. In the left and right regions the potential is constant, $V(x, y) = V_1$. In the middle region the potential is also constant, $V(x, y) = V_0$, where $V_0 > V_1$. The incoming, reflected and transmitted waves are given by $Ae^{i\mathbf{k}\cdot\mathbf{r}}$, $Be^{i\mathbf{k}'\cdot\mathbf{r}}$ and $Fe^{i\mathbf{k}\cdot\mathbf{r}}$.

The Schrödinger equation

$$E\psi(\mathbf{r}) + \frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) - V(\mathbf{r})\psi(\mathbf{r}) = 0 \tag{105}$$

is separable in Cartesian coordinates

$$\left[E + \left\{\frac{\hbar^2}{2m}\frac{d^2}{dx^2} - V(x)\right\} + \frac{\hbar^2}{2m}\frac{d^2}{dy^2} + \frac{\hbar^2}{2m}\frac{d^2}{dz^2}\right]\psi(x, y, z) = 0,$$

and its solutions can be written as

$$\psi(x, y, z) = \phi(x)e^{ik_y y}e^{ik_z z} = \phi(x)e^{i\mathbf{k}_\parallel\cdot\mathbf{r}};\ \mathbf{k}_\parallel = \begin{pmatrix} 0 \\ k_y \\ k_z \end{pmatrix}, \tag{106}$$

where $e^{i\mathbf{k}_\parallel\cdot\mathbf{r}}$ describes a free particle in the direction parallel to the barrier. The function $\phi(x)$ is the solution of the one-dimensional scattering problem

$$\left[E_x + \left\{\frac{\hbar^2}{2m}\frac{d^2}{dx^2} - V(x)\right\}\right]\phi(x) = 0, \tag{107}$$

with the energy

$$E_x = E - \frac{\hbar^2 k_\parallel^2}{2m}. \tag{108}$$

Note that $\mathbf{k}_{\|}$ is a good quantum number,[30] which together with the energy E fixes the wave function. We say that $\mathbf{k}_{\|}$ defines a **mode** of the system.

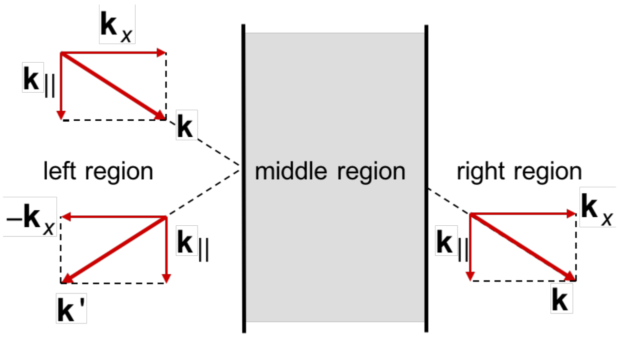

Figure 17: Scattering from a planar square barrier as viewed in the $(k_x, \mathbf{k}_{\|})$ plane.

A view of the scattering geometry in the $(k_x, \mathbf{k}_{\|})$ plane is given in Fig. 17. In the left and right regions the wave function is

$$\psi_{\mathbf{k}_{\|}}(\mathbf{r}) = \begin{cases} A\left[e^{i\mathbf{k}\cdot\mathbf{r}} + r_{\mathbf{k}_{\|}}(E)e^{i\mathbf{k}'\cdot\mathbf{r}}\right] \ ; \ \mathbf{r} \text{ in left region} \\ A\left[t_{\mathbf{k}_{\|}}(E)e^{i\mathbf{k}\cdot\mathbf{r}}\right] \ ; \ \mathbf{r} \text{ in right region} \end{cases} \tag{109}$$

with $\mathbf{k} = (k_x, \mathbf{k}_{\|})$, $\mathbf{k}' = (-k_x, \mathbf{k}_{\|})$, and the transmission and reflection amplitudes are $t_{\mathbf{k}_{\|}}(E)$ and $r_{\mathbf{k}_{\|}}(E)$, respectively. In three dimensions the probability current is a vector

$$\mathbf{J}(\mathbf{r},t) = \frac{i\hbar}{2m}\left[\Psi(\mathbf{r},t)\nabla\Psi^*(\mathbf{r},t) - \Psi^*(\mathbf{r},t)\nabla\Psi(\mathbf{r},t)\right] \tag{110}$$

As in the one-dimensional case, for a stationary problem, $\Psi(\mathbf{r},t) = \psi(\mathbf{r})e^{-\frac{i}{\hbar}Et}$, the current is constant. In three dimensions it is actually appropriate to use the phrase “current density” for $\mathbf{J}$, as in classical electrodynamics.

Here is the key point: for the experiments we are interested in, **only** J_x **matters**, i.e. the x-component of the current. For devices like that shown in Fig. 16 (and all other devices that we will consider), one attaches macroscopically large electrodes to the left and right regions and applies a bias voltage between those electrodes. Only the current along the x-direction is then measured.[31,32]

7.1 Derivation of the Landauer formula in 3D*

So we are interested in

$$J_x = \frac{i\hbar}{2m}\left[\psi(\mathbf{r})\frac{\partial\psi^*(\mathbf{r})}{\partial x} - \psi^*(\mathbf{r})\frac{\partial\psi(\mathbf{r})}{\partial x}\right]. \tag{111}$$

[30] a set of two quantum numbers, actually.

[31] For those of you who are experts in scattering phenomena, note that this is different from angle resolved three-dimensional scattering experiments in free space. There one is usually interested in all three components of the current. (Moreover, if the scatterer is localized in space, one applies a transformation to spherical coordinates.)

[32] More complicated measurements are possible. In “multiprobe” experiments more than two electrodes are attached to the device. This is often done in combination with applying an external magnetic field, as in measuring the Hall effect. The Landauer formalism can be extended to include multiprobe measurements. This extension is due to Büttiker.

From Eq. 109 one can easily show that the transmitted current carried by one mode $\mathbf{k}_{\|}$ is given by

$$J_{T,\mathbf{k}_{\|}} = \rho \left| t_{\mathbf{k}_{\|}}(E) \right|^2 v_x \tag{112}$$

with the density $\rho = |A|^2$ and the velocity in the x-direction $v_x = \frac{\hbar k_x}{m}$. Deriving an expression for the conductance follows the same steps as in Sec. 6. Applying a small bias $\Delta V = -eU$ between left and right regions, the transmitted current is carried by all modes that have an energy $E_{\mathbf{k}}$ in the range from $E_F - \Delta V$ to E_F, see Fig. 14.

$$J_T = 2\rho \sum_{E_F - \Delta V < E_{\mathbf{k}} < E_F} \left| t_{\mathbf{k}_{\|}}(E) \right|^2 v_x, \tag{113}$$

where the factor 2 accounts for the spin degeneracy. We can use the same tricks as in Sec. 6. Using states that are normalized in a 3D, we have $\rho = \frac{1}{2L^3}$; compare to Eq. 93.[33] We write $\sum_{E_F-\Delta V<E_{\mathbf{k}}<E_F} = \sum_{\mathbf{k}_{\|}} \sum_{k_x}$, where one has to sum only over those states that have their energy in the indicated interval. Convert $\sum_{k_x}$ into a one-dimensional integral, use $v_x = \frac{1}{\hbar}\frac{dE_x}{dk_x}$, and assume that $t_{\mathbf{k}_{\|}}(E)$ is independent of the energy in the small energy range ΔV. The algebra is

$$\begin{aligned} J_T &= \frac{1}{L^3} \sum_{\mathbf{k}_{\|}} \sum_{k_x} \left| t_{\mathbf{k}_{\|}}(E) \right|^2 v_x = \frac{1}{L^3} \sum_{\mathbf{k}_{\|}} \frac{L}{\pi} \int \left| t_{\mathbf{k}_{\|}}(E) \right|^2 v_x dk_x \\ &= \frac{1}{L^2}\frac{1}{\pi} \sum_{\mathbf{k}_{\|}} \int \left| t_{\mathbf{k}_{\|}}(E) \right|^2 \frac{1}{\hbar}\frac{dE_x}{dk_x} dk_x = \frac{1}{L^2}\frac{1}{\pi\hbar} \sum_{\mathbf{k}_{\|}} \int_{E_F - \Delta V - \frac{\hbar^2 k_{\|}^2}{2m}}^{E_F - \frac{\hbar^2 k_{\|}^2}{2m}} \left| t_{\mathbf{k}_{\|}}(E_x + \frac{\hbar^2 k_{\|}^2}{2m}) \right|^2 dE_x \\ &= \frac{1}{L^2}\frac{\Delta V}{\pi\hbar} \sum_{\mathbf{k}_{\|}} \left| t_{\mathbf{k}_{\|}}(E_F) \right|^2 . \end{aligned} \tag{114}$$

The transmitted current is $I_T = \int \mathbf{J}_T \cdot d\mathbf{S} = J_T L^2$ and the conductance $\mathcal{G} = I_T/U$ is then given by

$$\mathcal{G} = \frac{e^2}{\pi\hbar} \sum_{\mathbf{k}_{\|}} \left| t_{\mathbf{k}_{\|}}(E_F) \right|^2 = \frac{e^2}{\pi\hbar} T(E_F). \tag{115}$$

The expression is the Landauer formula, see Eq. 90, but now for layered systems in 3D. The total transmission T is expressed as a sum over the transmissions of the individual modes $\mathbf{k}_{\|}$. One has to sum over the $\mathbf{k}_{\|}$ that contribute to the transmission at the Fermi energy E_F.

7.2 The Fermi surface*

The Fermi surface is defined by the relation $E_{\mathbf{k}} = E_F$, the Fermi energy being a property of the material. It can be visualized as a surface in reciprocal space, i.e., in three-dimensional $\mathbf{k}$-space. For free electrons or electrons in a constant potential

$$E_{\mathbf{k}} = \frac{\hbar^2}{2m}\left(k_x^2 + k_y^2 + k_z^2\right) = \frac{\hbar^2}{2m}\left(k_x^2 + k_{\|}^2\right) = E_F.$$

[33] The function $\frac{1}{L^2} e^{i\mathbf{k}_{\|}\cdot\mathbf{r}}$ is normalized in a 2D box of size L, as you can easily check yourself. The function $\phi(x)$ is treated as in Eq. 93 and gets the normalization factor $\frac{1}{2L}$. The product gives the normalization factor of $\psi(x,y,z)$, see Eq. 106.

the Fermi surface is the surface of a sphere, as shown in Fig. 18(a). The Fermi surface can help to visualize which modes contribute to the transmission. The latter can be enumerated by projecting the part of Fermi surface with $k_x > 0$ (a hemisphere in this case) onto the $\mathbf{k}_\| = (k_y, k_z)$ plane, as shown in Fig. 18(b). All the modes $\mathbf{k}_\|$ within this projection exist at the Fermi energy E_F and they contribute to the transmission in Eq. 115. The scattering geometry in real space can be deduced from Fig. 17. $\mathbf{k}_\| = (0,0)$[34] corresponds to a wave with normal incidence to the barrier. The larger $k_\| = \left|\mathbf{k}_\|\right|$, the more glancing the incidence of the corresponding wave. At the edge of the circle in Fig. 18(b) one has $\frac{\hbar^2}{2m}k_\|^2 = E_F$, so $k_x = 0$, and the corresponding wave propagates parallel to the barrier.

For a simple square barrier one can easily calculate the transmission. If the Fermi energy is less than the barrier height, the transmission of a single mode is given by

$$\left|t_{\mathbf{k}_\|}(E_F)\right|^2 = T_{1D}(E_x) = T_{1D}(E_F - E_\|) \quad \text{with} \quad E_\| = \frac{\hbar^2 k_\|^2}{2m}, \tag{116}$$

and T_{1D} given by Eqs. 10 and 3. The result is visualized in Fig. 18(c). Since T_{1D} is a monotonically increasing function of the energy, the maximal transmission is for $\mathbf{k}_\| = (0,0)$, i.e. for normal incidence. The transmission decreases monotonically with increasing $k_\|$, i.e. with the angle of incidence, until it is zero for parallel incidence. Such a simple Fermi surface and a simple transmission are typical for free electrons.

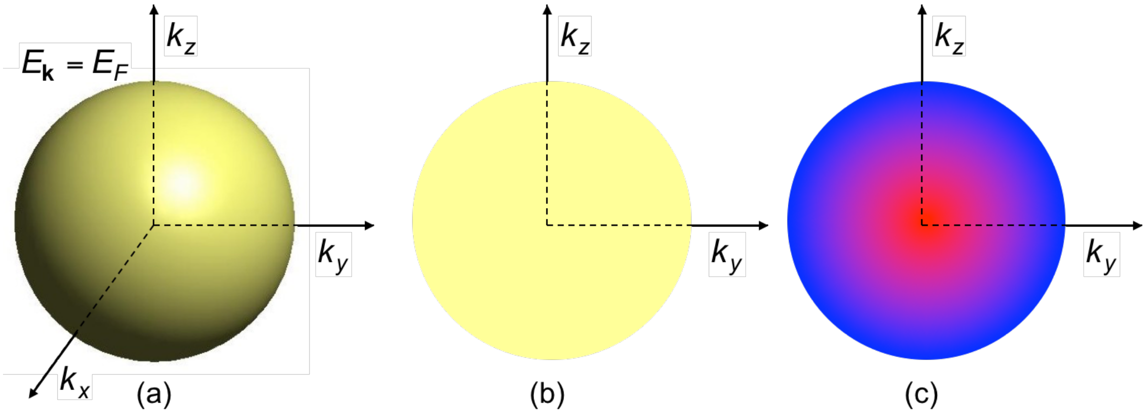

Figure 18: (a) The Fermi surface, as defined by $E_\mathbf{k} = E_F$, of electrons in a constant potential is a sphere. (b) The projection of the surface in the $\mathbf{k}_\| = (k_y, k_z)$ plane. The shaded area denotes all the $\mathbf{k}_\|$ modes that contribute to the transmission at E_F. (c) The transmission $|t_{\mathbf{k}_\|}|^2$ as function of $\mathbf{k}_\|$. Red indicates a high transmission and blue a low transmission. The highest transmission is for $\mathbf{k}_\| = (0,0)$, i.e., normal incidence. It decreases to 0 towards the edge of the circle.

7.3 Finite bias

For $\Delta V > 0$, as depicted in Fig. 14, Eq. 102 can be generalized for layered 3D materials to

$$\mathcal{J}_T = -\frac{e}{\pi\hbar}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T_{1D}(E_x, \Delta V)\Theta\left(E_x + E_\| - E_F + \Delta V\right)\left[1 - \Theta\left(E_x + E_\| - E_F\right)\right] g_{2D}(E_\|) dE_x dE_\|, \tag{117}$$

[34] called the Γ-point in solid state physics folklore.

where $g_{2D}(E_{\parallel})$is the density of states of electrons in $\mathbf{k}_{\parallel}$ modes. Note the dimension of $\mathcal{J}_T$ is A/m^2, i.e., it is a current density. The density of states of free electrons in two dimensions is dead simple[35]

$$g_{2D}(E) = \frac{m}{\pi\hbar^2}\Theta(E). \tag{118}$$

This means we can do the integral over $E_{\parallel}$ by hand and obtain

$$\boxed{\mathcal{J}_T = -\frac{em}{\pi^2\hbar^3}\left[\Delta V \int_0^{E_F-\Delta V} T_{1D}(E_x,\Delta V)dE_x + \int_{E_F-\Delta V}^{E_F} (E_F - E_x)\, T_{1D}(E_x,\Delta V)dE_x\right], \quad (119)}$$

making use of the fact that $E_x \geq 0$. The expression of Eq. 119 is called the **Tsu-Esaki formula**.[36] Note that E_x runs over the whole interval from 0 to E_F, which is in contrast the the 1D case, Eq. 101. Obviously, if $\Delta V > E_F$, then the first integral of Eq. 119 disappears, and the lower bound of the second integral has to be replaced by 0, as $E_x \geq 0$, see the discussion in Sec. 6.3.

To be able to approximate $T_{1D}(E,\Delta V) \approx T_{1D}(E_F, 0)$, one needs to obey two criteria (i) the transmission coefficient is approximately constant for energies $0 < E < E_F$, and (ii) the potential barrier is not appreciably distorted by the applied bias ΔV. The first criterium (i) is particularly hard to meet. In practice it only occurs if $E_F \ll V_0$, see Figs. 13 and 14, which is hardly ever the case if metals are used in the left and right regions of the tunnel junction. Using (doped) semiconductors makes life different, however. The Fermi levels in semiconductors can be finely controlled by the concentration of dopant atoms, making it possible to achieve $E_F \ll V_0$.

Unfortunately Eq. 119 only holds for $\Delta V > 0$, see Fig. 14. For $\Delta V < 0$ the expression needs to be modified slightly to

$$\boxed{\mathcal{J}_T = -\frac{em}{\pi^2\hbar^3}\left[\Delta V \int_{-\Delta V}^{E_F} T_{1D}(E_x,\Delta V)dE_x - \int_{E_F}^{E_F-\Delta V} (E_F - \Delta V - E_x)\, T_{1D}(E_x,\Delta V)dE_x\right], \quad (120)}$$

see also Fig. 15. In this case, if $-\Delta V > E_F$, the first integral disappears, and the lower bound of the second integral has to be replaced by $-\Delta V$.

Part III
Experimental background

8 Metal-Insulator-Metal (MIM) diodes

The Metal-Insulator-Metal (MIM) diode is a device with strongly nonlinear I/V characteristics that show some similarity to those of a semiconductor diode. MIMs are capable of

[35] See your introduction to solid state physics course.

[36] Leo Esaki won a physics Nobel prize in 1973 for inventing semiconductor devices based upon electron tunneling.

very fast operation with switching frequencies in the THz range. As thin-film diodes (TFDs) they are also in use in flat panel displays (such as active matrix LCDs).

A MIM diode consists of two different metals, separated by an insulating layer. This layer is sufficiently thin, such that electrons can tunnel through it from one metal to the other. In general, the work functions of two different metals are different.[37] This means that the Fermi energies of the two metals are different, see Fig. 19(a). If one brings the two metals in contact via a permeable insulator layer, then electrons will flow from one metal to the other. This stops after a common Fermi level is established, i.e., when the system is in equilibrium.

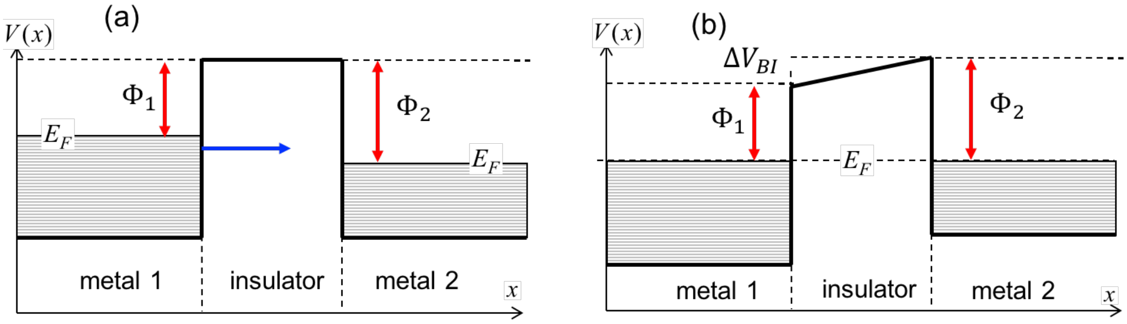

Figure 19: Schematic energy diagram of a metal-insulator-metal (MIM) device. (a) If the Fermi energies E_F on the left and right metal are different, then electrons will flow from left to right, until (b) a common Fermi level is established. The difference between the work functions of the two metals $\Phi_2 - \Phi_1 = \Delta V_{BI}$ is found as build-in potential difference across the insulator.

If electrons are added to a metal, its potential (energy) rises as it becomes negatively charged, see Fig. 19(b). Likewise, if electrons are extracted, the metal becomes positively charged, and its potential (energy) drops. Few electrons are actually required to change a potential substantially; in fact, negligibly few, compared to the number of electrons available in a bulk material. The added electrons or removed electrons (the holes), and the negative and positive charges resulting from it, reside at the surfaces of the conductors, or in this case, at the interfaces of the metals with the insulator. This means that Φ_1 and Φ_2 do not change going from Fig. 19(a) to (b). As in a parallel plate capacitor, the surface charges in the two metals then give a potential step ΔV across the insulator that compensates for the difference in work functions,

$$\Delta V_{BI} = \Phi_2 - \Phi_1. \tag{121}$$

This is called *the build-in potential*.

To operate the device, one applies a bias ΔV between the two metals, where $\Delta V > 0$ is called *forward bias*, and $\Delta V < 0$ is called *reverse bias*, see Fig. 20. The I/V characteristics as a function of bias are a topic in this project.

[37] The work function Φ of a metal is the minimum energy it takes to extract an electron from the metal, so $\Phi = V_{\text{vac}} - E_F$, where V_{vac} is the potential in vacuum, sufficiently far from the metal. This seemingly simple expression is somewhat misleading, as hidden in it is also a contribution from the potential step one always has at the interface between two materials, in this case the interface between the metal and vacuum. Unfortunately, the size of this potential step very much depends on the details of the charge distribution at the interface. Even perfectly clean and flat surfaces of the same material have different work functions, if they have different crystallographic orientations. A change in orientation gives a change in surface termination, which changes the charge distribution at the interface with vacuum.

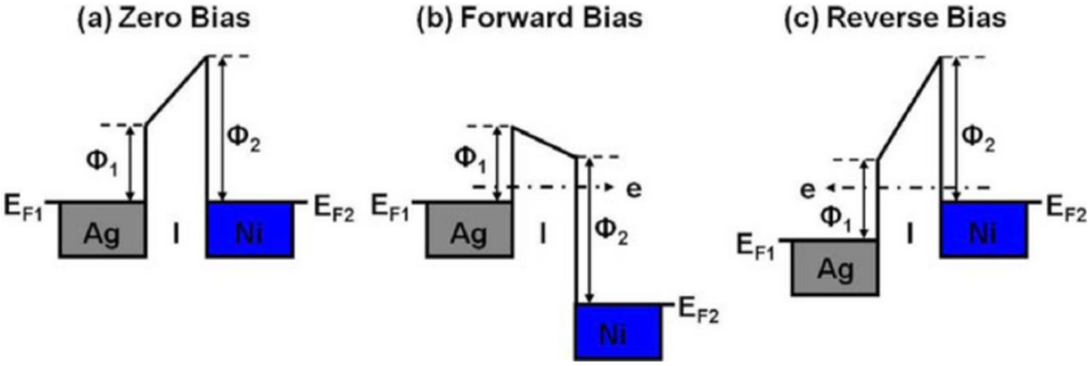

Figure 20: Schematic energy diagram of a metal-insulator-metal (MIM) diode based on two different metals, Ni and Ag in (a) zero bias, $E_{F1} = E_{F2}$, (b) forward bias, $\Delta V = E_{F1} - E_{F2} > 0$, and (c) reverse bias, $\Delta V = E_{F1} - E_{F2} < 0$. [*C. Zhuang et al. ECS Solid State Lett. 4, 39-42 (2015).*]

9 Resonant-tunneling diodes (RTDs)

With modern deposition techniques it is possible to grow multilayers of different materials in a very controlled way. Fig. 21 shows the schematic energy diagram of a ABA′BA layered structure with A and B two different materials. The A′ region in the middle acts as a potential well. As you know, a potential well supports bound states, which have discrete energy levels E_r. Because an electron can tunnel through the walls of material B that surrounds the well, it won't stay in a bound state forever. We say that the electron in such a state in A′ has a finite lifetime τ_r. Hence the states are called quasi-bound states. Because of the Heisenberg uncertainty principle, their energy levels are not infinitely sharp, as for truly bound states, but they have a spread in energy $\Delta E_r \approx \hbar/\tau_r$.

The wave functions of these quasi-bound states are essentially standing (electron) waves. So, in order that these can be formed, the electron waves must remain coherent in the A′ region. This means that scattering within the A′ region must be small, which implies using very clean materials (no impurity scattering) at a sufficiently low temperature (no phonon scattering). It also helps if the wave length λ_r of the electron waves is comparable to the thickness of the A′ region. If that thickness is of the order of several nm, this points towards using doped semiconductors instead of metals.[38] Indeed such multilayer structures are now almost routinely grown from semiconductor materials.

[38] The characteristic wave length of an electron in a metal is the Fermi wave length $\lambda_F = 2\pi/k_F$, which is typically of the order of the interatomic spacing, i.e., a few Å. In a semiconductor we can make the Fermi energy $E_F = \hbar^2 k_F^2/2m$ essentially as small as we wish through controlled doping. This means that we can get λ_F as large as we wish. A practical limit is set by the fact that we also want to have a minimum number of charge carriers (otherwise the current becomes too small for detection), which sets a minimum E_F. In practice, attaining a λ_F of several tens of nm is no problem.

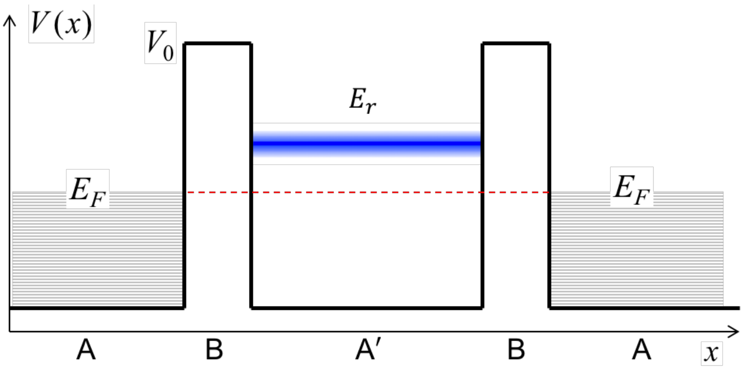

Figure 21: Schematic energy diagram of a double-barrier resonant-tunneling diode at zero bias. The region between the two barriers supports quasi-bound states at energies E_r, called resonance levels.

For instance, in the family of III-V semiconductors there are several materials that can be grown on top of one another without introducing any stacking faults in the crystal lattice (called epitaxial growth). Such stacking faults are examples of defects in the crystal lattice. Defects commonly act as traps for charge carriers, and thus need to be avoided. The regions A in Fig. 21 can for instance be heavily n-doped GaAs. Heavy n-doping gives a Fermi level inside the conduction band (which is called a degenerate semiconductor), and the thick black line in Fig. 21 then represents the bottom of the conduction band. The B material can for instance be the (undoped, i.e., intrinsic) alloy Ga$_x$Al$_{1-x}$As, which has the same crystal structure as GaAs, but has its conduction band at a higher energy. The middle region A′ then consists of undoped GaAs. Similar structures can be made with other members of the III-V family, but also with members of the IV family, Si, Ge, and Si$_x$Ge$_{1-x}$.

The structure of Fig. 21 is actually a device, called a **resonant-tunneling diode (RTD)**. It works as follows. Obviously, no current flows at zero bias as in Fig. 21. Now imagine we apply a bias between the two outer regions A. Then electrons will tunnel from left to right, as in Fig. 22. If the bias is small, then the resonance level E_r is not at the right energy to take up any electrons, as the electrons that tunnel from left to right have an energy $E_F < E < E_F - \Delta V$. In other words, the whole region BA′B acts as one thick tunnel barrier, comparable to Fig. 14, and the tunneling current is quite low.

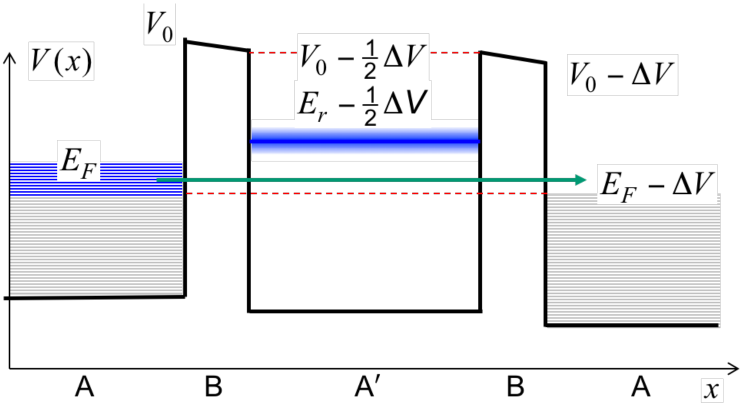

Figure 22: Schematic energy diagram of a double-barrier resonant-tunneling diode at a small bias ΔV. In this case, it is assumed that the potential drops occur in the barrier regions. Electrons can tunnel from left to right via the the middle BA′B region.

If we increase the bias ΔV, the resonance level can come in a position where

$$E_F < E_r - \frac{1}{2}\Delta V < E_F - \Delta V, \tag{122}$$

see Fig. 23. This is called the **resonant condition**. Tunneling from left to right can now take place via the resonance level in the A′ region. In effect, only the two B regions now act as tunnel barriers, and, as these regions are quite thin, the tunnel current is now much larger.

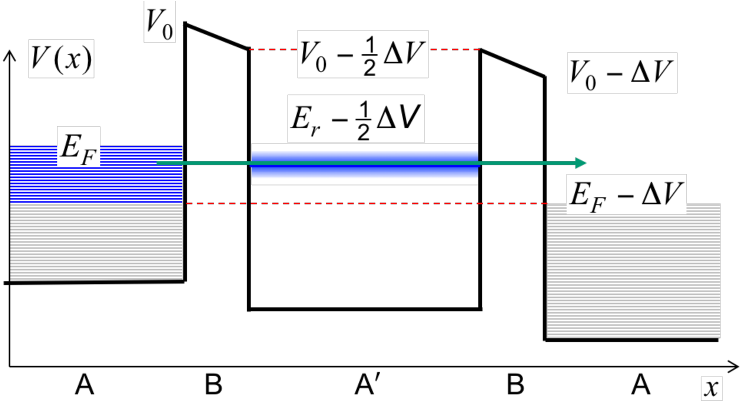

Figure 23: Schematic energy diagram of a double-barrier resonant-tunneling diode at a larger bias ΔV. In this case, it is assumed that the potential drops occur in the barrier regions. Electrons can now tunnel from left to right via the resonance level in the middle A′ region.

This resonant condition can disappear again upon increasing the bias even further, and the tunnel current becomes smaller again. In other words, the current is not a monotonically increasing function of the bias. It can decrease upon increasing the bias. This effect is called **negative differential resistance (NDR)**. How and when this happens is a topic within this project.