Lessons from Nanoelectronics
B. Quantum Transport

Supriyo Datta
Purdue University

World Scientific (2012)
Manuscript copy

Second Edition
To be published in 2017
List of Available Video Lectures
Quantum Transport

This book is based on a set of two online courses originally offered in 2012 on nanoHUB-U and more recently in 2015 on edX. These courses are now available in self-paced format at nanoHUB-U (https://nanohub.org/u) along with many other unique online courses.

Additional information about this book along with questions and answers is posted at the book website.

In preparing the second edition we decided to split the book into parts A and B following the two online courses available on nanoHUB-U entitled *Fundamentals of Nanoelectronics*

- **Part A: Basic Concepts**
- **Part B: Quantum Transport**

Also of possible interest in this context: *NEGF: A Different Perspective*

Following is a detailed list of *video lectures* available at the course website corresponding to different sections of this volume (Part B: Quantum Transport).

Unless mentioned otherwise, the section numbers refer to


To access MATLAB codes

QTAT stands for

- *Quantum Transport: Atom to Transistor*, Ch.5-11, Cambridge (2005)

To access MATLAB codes
<table>
<thead>
<tr>
<th>Lecture in Course website</th>
<th>Topic</th>
<th>Discussed in this book</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit 1: L1.1</td>
<td>Introduction</td>
<td>Chapter 18</td>
</tr>
<tr>
<td>Unit 1: L1.2</td>
<td>Wave Equation</td>
<td>Section 18.1</td>
</tr>
<tr>
<td>Unit 1: L1.3</td>
<td>Differential to Matrix Equation</td>
<td>Section 18.3</td>
</tr>
<tr>
<td>Unit 1: L1.4</td>
<td>Dispersion Relation</td>
<td>Section 18.4.1</td>
</tr>
<tr>
<td>Unit 1: L1.5</td>
<td>Counting States</td>
<td>QTAT 5.1</td>
</tr>
<tr>
<td>Unit 1: L1.6</td>
<td>Beyond 1D</td>
<td>Section 18.4.2</td>
</tr>
<tr>
<td>Unit 1: L1.7</td>
<td>Matrix Equation With Basis</td>
<td>QTAT 5.1</td>
</tr>
<tr>
<td>Unit 1: L1.8</td>
<td>Graphene</td>
<td>Section 18.4.4</td>
</tr>
<tr>
<td>Unit 1: L1.9</td>
<td>Reciprocal Lattice/Valleys</td>
<td>QTAT 5.2</td>
</tr>
<tr>
<td>Unit 1: L1.10</td>
<td>Summary</td>
<td>Chapter 18</td>
</tr>
<tr>
<td>Unit 2: L2.1</td>
<td>Introduction</td>
<td>Chapters 19,20</td>
</tr>
<tr>
<td>Unit 2: L2.2</td>
<td>Semiclassical Model</td>
<td>Section 19.1.1</td>
</tr>
<tr>
<td>Unit 2: L2.3</td>
<td>Quantum Model</td>
<td>Section 19.1.2, 19.1.3</td>
</tr>
<tr>
<td>Unit 2: L2.4</td>
<td>NEGF Equations</td>
<td>Section 19.2.1-19.2.3</td>
</tr>
<tr>
<td>Unit 2: L2.5</td>
<td>Current Operator</td>
<td>Section 19.2.4, 19.3</td>
</tr>
<tr>
<td>Unit 2: L2.6</td>
<td>Scattering Theory</td>
<td>Section 20.1</td>
</tr>
<tr>
<td>Unit 2: L2.7</td>
<td>Transmission</td>
<td>Section 20.1.1</td>
</tr>
<tr>
<td>Unit 2: L2.8</td>
<td>Resonant Tunneling</td>
<td>Section 20.2</td>
</tr>
<tr>
<td>Unit 2: L2.9</td>
<td>Dephasing</td>
<td>Sections 19.4, 20.3</td>
</tr>
<tr>
<td>Unit 2: L2.10</td>
<td>Summary</td>
<td>Chapters 19, 20</td>
</tr>
<tr>
<td>Unit 3: L3.1</td>
<td>Introduction</td>
<td>Chapter 21</td>
</tr>
<tr>
<td>Unit 3: L3.2</td>
<td>Quantum Point Contact</td>
<td>21.1</td>
</tr>
<tr>
<td>Unit 3: L3.3</td>
<td>Self-Energy</td>
<td>21.2</td>
</tr>
<tr>
<td>Unit 3: L3.4</td>
<td>Surface Green's Function</td>
<td>Section 21.2.2</td>
</tr>
<tr>
<td>Unit 3: L3.5</td>
<td>Graphene</td>
<td>Section 21.2.3</td>
</tr>
<tr>
<td>Unit 3: L3.6</td>
<td>Magnetic Field</td>
<td>Section 21.3</td>
</tr>
<tr>
<td>Unit 3: L3.7</td>
<td>Golden Rule</td>
<td>QTAT 10.1</td>
</tr>
<tr>
<td>Unit 3: L3.8</td>
<td>Inelastic Scattering</td>
<td>QTAT 10.3</td>
</tr>
<tr>
<td>Unit 3: L3.9</td>
<td>Can NEGF Include Everything</td>
<td>Chapter 23</td>
</tr>
<tr>
<td>Unit 3: L3.10</td>
<td>Summary</td>
<td>Chapter 21</td>
</tr>
<tr>
<td>Unit 4: L4.1</td>
<td>Introduction</td>
<td>Chapter 22</td>
</tr>
<tr>
<td>Unit 4: L4.2</td>
<td>Magnetic Contacts</td>
<td>Section 22.1</td>
</tr>
<tr>
<td>Unit 4: L4.3</td>
<td>Rotating Contacts</td>
<td>Section 22.2</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Unit 4: L4.4</td>
<td>Vectors and Spinors</td>
<td>Section 22.4</td>
</tr>
<tr>
<td>Unit 4: L4.5</td>
<td>Spin-orbit Coupling</td>
<td>Section 22.3.2</td>
</tr>
<tr>
<td>Unit 4: L4.6</td>
<td>Spin Hamiltonian</td>
<td>Section 22.3</td>
</tr>
<tr>
<td>Unit 4: L4.7</td>
<td>Spin Density/Current</td>
<td>Section 22.6.1</td>
</tr>
<tr>
<td>Unit 4: L4.8</td>
<td>Spin Voltage</td>
<td>Section 22.6.2</td>
</tr>
<tr>
<td>Unit 4: L4.9</td>
<td>Spin Circuits</td>
<td>Section 22.6.3</td>
</tr>
<tr>
<td>Unit 4: L4.10</td>
<td>Summary</td>
<td>Chapter 22</td>
</tr>
</tbody>
</table>
III. Contact-ing Schrödinger

18. The Model 252
19. NEGF Method 277
20. Can Two Offer Less Resistance than One? 303
21. Quantum of Conductance 319
22. Rotating an Electron 337
23. Does NEGF Include “Everything?” 370
24. The Quantum and the Classical 389
Lecture 18

The Model

18.1. Schrödinger Equation
18.2. Electron-Electron Interactions
18.3. Differential to Matrix Equation
18.4. Choosing Matrix Parameters

Over a century ago Boltzmann taught us how to combine Newtonian mechanics with entropy-driven processes

\[
\text{Classical Dynamics} + \text{\textcolor{red}{\textbullet}} = \text{BTE}
\]

and the resulting Boltzmann transport equation (BTE) is widely accepted as the cornerstone of semiclassical transport theory. Most of the results we have discussed so far can be (and generally are) obtained from the Boltzmann equation, but the concept of an elastic resistor makes them more transparent by spatially separating force-driven processes in the channel from the entropy-driven processes in the contacts.

In this part of these lecture notes I would like to discuss the quantum version of this problem, using the non-equilibrium Green’s function (NEGF) method to combine quantum mechanics described by the Schrödinger equation with "contacts"

\[
\text{Quantum Dynamics} + \text{\textcolor{red}{\textbullet}} = \text{NEGF}
\]

much as Boltzmann taught us how to combine classical dynamics with "contacts".
The NEGF method originated from the classic works in the 1960’s that used the methods of many-body perturbation theory to describe the distributed entropy-driven processes along the channel. Like most of the work on transport theory (semiclassical or quantum) prior to the 1990’s, it was a “contact-less” approach focused on the interactions occurring throughout the channel, in keeping with the general view that the physics of resistance lay essentially in these distributed entropy generating processes.

As with semiclassical transport, our bottom-up perspective starts at the other end with the elastic resistor with entropy-driven processes confined to the contacts. This makes the theory less about interactions and more about "connecting contacts to the Schrödinger equation", or more simply, about contact-ing Schrödinger.
Lessons from Nanoelectronics

But let me put off talking about the NEGF model till the next Lecture, and use subsequent lectures to illustrate its application to interesting problems in quantum transport. As indicated in Fig.18.1b the NEGF method requires two types of inputs: the Hamiltonian, \( [H] \) describing the dynamics of an elastic channel, and the self-energy \( [\Sigma] \) describing the connection to the contacts, using the word “contacts” in a broad figurative sense to denote all kinds of entropy-driven processes. Some of these contacts are physical like the ones labeled “1” and “2” in Fig.18.1b, while some are conceptual like the one labeled “0” representing entropy changing processes distributed throughout the channel.

In this Lecture let me just try to provide a super-brief but self-contained introduction to how one writes down the Hamiltonian \([H]\). The \([\Sigma]\) can be obtained by imposing the appropriate boundary conditions and will be described in later Lectures when we look at specific examples applying the NEGF method.

We will try to describe the procedure for writing down \([H]\) so that it is accessible even to those who have not had the benefit of a traditional multi-semester introduction to quantum mechanics. Moreover, our emphasis here is on something that may be helpful even for those who have this formal background. Let me explain.

Most people think of the Schrödinger equation as a differential equation which is the form we see in most textbooks. However, practical calculations are usually based on a discretized version that represents the differential equation as a matrix equation involving the Hamiltonian matrix \([H]\) of size NxN, N being the number of “basis functions” used to represent the structure.

This matrix \([H]\) can be obtained from first principles, but a widely used approach is to represent it in terms of a few parameters which are chosen to match key experiments. Such semi-empirical approaches are often used because of their convenience and because they can often explain a wide range of experiments beyond the key ones that are used as input, suggesting that they capture a lot of essential physics.
In order to follow the rest of the Lectures it is important for the readers to get a feeling for how one writes down this matrix $[H]$ given an accepted energy-momentum $E(p)$ relation (Lecture 5) for the material that is believed to describe the dynamics of conduction electrons with energies around the electrochemical potential.

But I should stress that the NEGF framework we will talk about in subsequent lectures goes far beyond any specific model that we may choose to use for $[H]$. The same equations could be (and have been) used to describe say conduction through molecular conductors using first principles Hamiltonians.

### 18.1. Schrödinger Equation

We started these Lectures by noting that the key input needed to understand current flow is the density of states, $D(E)$, telling us the number of states available for an electron to access on its way from the source to the drain. Theoretical models for $D(E)$ all start from the Schrödinger equation which tells us the available energy levels. However, we managed to obtain expressions for $D(E)$ in Chapter 5 without any serious brush with quantum mechanics by (1) starting from a given energy-momentum relation $E(p)$, (2) relating the momentum to the wavelength through the de Broglie relation ($p = \hbar / \text{wavelength}$) and then (3) requiring an integer number of half wavelengths to fit into the conductor, the same way acoustic waves fit on a guitar string.

This heuristic principle is mathematically implemented by writing a wave equation which is obtained from a desired energy-momentum relation by making the replacements.
where the latter stands for

\[ p_x \rightarrow -i\hbar \frac{\partial}{\partial x}, \quad p_y \rightarrow -i\hbar \frac{\partial}{\partial y}, \quad p_z \rightarrow -i\hbar \frac{\partial}{\partial z}, \]

Using this principle, the classical energy-momentum relation

\[ E_{\text{classical}}(\tilde{p}) = \frac{p_x^2 + p_y^2 + p_z^2}{2m} \]

leads to the wave equation

\[ i\hbar \frac{\partial \tilde{\psi}}{\partial t} = -\frac{\hbar^2}{2m} \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \tilde{\psi}(x,y,z,t) \]

whose solutions can be written in the form of exponentials of the form

\[ \tilde{\psi}(x,y,z,t) = \psi_0 e^{+ikx} e^{+iky} e^{+ikz} e^{-iEt/h} \]

where the energy \( E \) is related to the wavevector \( \tilde{k} \) by the dispersion relation

\[ E(\tilde{k}) = \frac{\hbar^2(k_x^2 + k_y^2 + k_z^2)}{2m} \]

Eq.(18.4) looks just like the classical energy-momentum relation (Eq.(18.2a)) of the corresponding particle with

\[ \tilde{p} = \hbar \tilde{k} \]

which relates the particulate property \( \tilde{p} \) with the wavelike property \( \tilde{k} \). This can be seen to be equivalent to the de Broglie relation (\( p = h/\text{wavelength} \)) noting that the wavenumber \( k \) is related to the wavelength through

\[ k = \frac{\pi}{\lambda} \]
The principle embodied in Eq. (18.1) ensures that the resulting wave equation has a group velocity that is the same as the velocity of the corresponding particle

$$\frac{1}{\hbar} \mathbf{V}_k E_{\text{Wave}} = \mathbf{V}_p E_{\text{Particle}}$$

18.1.1. Spatially Varying Potential

The wave equation Eq.(18.2b) obtained from the energy-momentum relation describes free electrons. If there is a force described by a potential energy $U(\bar{r})$ so that the classical energy is given by

$$E_{\text{classical}}(\bar{r}, \bar{p}) = p_x^2 + p_y^2 + p_z^2 \frac{2m}{2m} + U(x,y,z) \quad (18.6a)$$

then the corresponding wave equation has an extra term due to $U(\bar{r})$

$$i\hbar \frac{\partial}{\partial t} \psi = -\frac{\hbar^2}{2m} \nabla^2 \psi + U(\bar{r}) \bar{\psi} \quad (18.6b)$$

where $\bar{r} \equiv (x,y,z)$ and the Laplacian operator is defined as

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

Solutions to Eq.(18.6b) can be written in the form

$$\bar{\psi}(\bar{r},t) = \psi(\bar{r}) e^{-iE \hbar}$$

where $\psi(\bar{r})$ obeys the time-independent Schrödinger equation

$$E \psi(\bar{r}) = H_{\text{op}} \psi(\bar{r}) \quad (18.7a)$$
where $H_{\text{op}}$ is a differential operator obtained from the classical energy function in Eq.(18.6a), using the replacement mentioned earlier (Eq.(18.1)):

$$H_{\text{op}} = -\frac{\hbar^2}{2m} \nabla^2 + U(\vec{r}) \quad (18.7b)$$

Quantum mechanics started in the early twentieth century with an effort to “understand” the energy levels of the hydrogen atom deduced from the experimentally observed spectrum of the light emitted from an incandescent source. For a hydrogen atom Schrödinger used the potential energy

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

where the atomic number $Z = 1$, due to a point nucleus with charge $+q$, and solved Eq.(18.7) analytically for the allowed energy values $E_n$ (called the eigenvalues of the operator $H_{\text{op}}$) given by

$$E_n = -\frac{Z^2 q^2}{n^2} \frac{1}{8\pi \varepsilon_0 a_0}$$

with

$$a_0 = \frac{4\pi \varepsilon_0 \hbar^2}{m q^2}$$

and the corresponding solutions

$$\psi_{nlm}(\vec{r}) = R_{nl}(\vec{r}) Y_l^m(\theta, \phi)$$

obeying the equation

$$E_n \psi_{nlm} = \left(-\frac{\hbar^2}{2m} \nabla^2 - \frac{Z q^2}{4\pi \varepsilon_0 r} \right) \psi_{nlm}(\vec{r})$$
Energy levels in atoms are catalogued with three indices $n, \ell, m$.

The energy eigenvalues in Eq. (18.8) were in extremely good agreement with the known experimental results, leading to general acceptance of the Schrödinger equation as the wave equation describing electrons, just as acoustic waves, for example, on a guitar string are described by

$$\omega^2 u(z) = -\frac{\partial^2}{\partial z^2} u$$

A key point of similarity to note is that when a guitar string is clamped between two points, it is able to vibrate only at discrete frequencies determined by the length $L$. Similarly electron waves when “clamped” have discrete energies and most quantum mechanics texts start by discussing the corresponding “particle in a box” problem.

Shorter the length $L$, higher the pitch of a guitar and hence the spacing between the harmonics. Similarly smaller the box, greater the spacing between the allowed energies of an electron. Indeed one could view the hydrogen atom as an extremely small 3D box for the electrons giving rise to the discrete energy levels shown in Fig. 18.2. This is of course just a qualitative picture. Quantitatively, we have to solve the time-independent Schrödinger equation (Eq. (18.7)).
There is also a key dissimilarity between classical waves and electron waves. For acoustic waves we all know what the quantity $u(z)$ stands for: it is the displacement of the string at the point $z$, something that can be readily measured. By contrast, the equivalent quantity for electrons, $\psi(\mathbf{r})$ (called its wavefunction), is a complex quantity that cannot be measured directly and it took years for scientists to agree on its proper interpretation. The present understanding is that the real quantity $\psi \psi^*$ describes the probability of finding an electron in a unit volume around $\mathbf{r}$. This quantity, when summed over many electrons, can be interpreted as the average electron density.

18.2. Electron-electron interactions and the scf method

After the initial success of the Schrödinger equation in “explaining” the experimentally observed energy levels of the hydrogen atom, scientists applied it to increasingly more complicated atoms and by 1960 had achieved good agreement with experimentally measured results for all atoms in the periodic table (Herman and Skillman (1963)). It should be noted, however, that these calculations are far more complicated primarily because of the need to include the electron-electron (e-e) interactions in evaluating the potential energy (hydrogen has only one electron and hence no e-e interactions).

For example, Eq.(18.9) gives the lowest energy for a hydrogen atom as $E_1 = -13.6 \text{ eV}$ in excellent agreement with experiment. It takes a photon with at least that energy to knock the electron out of the atom ($E > 0$), that is to cause photoemission. Looking at Eq.(18.8) one might think that in Helium with $Z=2$, it would take a photon with energy $\sim 4*13.6 \text{ eV} = 54.5 \text{ eV}$ to knock an electron out. However, it takes photons with far less energy $\sim 30 \text{ eV}$ and the reason is that the electron is repelled by the other electron in Helium. However, if we were to try to knock the second electron out of Helium, it would indeed take photons with energy $\sim 54 \text{ eV}$, which is known as the second ionization potential. But usually what we want is the first ionization potential or a related quantity called the electron affinity. Let me explain.
Current flow involves adding an electron from the source to the channel and removing it into the drain. However, these two events could occur in either order.

The electron could first be added and then removed so that the channel evolves as follows

A. \( N \rightarrow N+1 \rightarrow N \) electrons (Affinity levels)

But if the electron is first removed and then added, the channel would evolve as

B. \( N \rightarrow N-1 \rightarrow N \) electrons (Ionization levels)

In the first case, the added electron would feel the repulsive potential due to \( N \) electrons. Later when removing it, it would still feel the potential due to \( N \) electrons since no electron feels a potential due to itself. So the electron energy levels relevant to this process should be calculated from the Schrödinger equation using a repulsive potential due to \( N \) electrons. These are known as the affinity levels.

In the second case, the removed electron would feel the repulsive potential due to the other \( N-1 \) electrons. Later when adding an electron, it would also feel the potential due to \( N-1 \) electrons. So the electron energy levels relevant to this process should be calculated from the Schrödinger equation using a repulsive potential due to \( N-1 \) electrons. These are known as the ionization levels.

The difference between the two sets of levels is basically the difference in potential energy due to one electron, called the single electron charging energy \( U_0 \). For something as small as a Helium atom it is \( \sim 25 \)
eV, so large that it is hard to miss. For large conductors it is often so small that it can be ignored, and it does not matter too much whether we use the potential due to \( N \) electrons or due to \( N-1 \) electrons. For small conductors, under certain conditions the difference can be important giving rise to single-electron charging effects, which we will ignore for the moment and take up again later in Lecture 24.

Virtually all the progress that has been made in understanding "condensed matter," has been based on the self-consistent field (scf) method where we think of each electron as behaving quasi-independently feeling an average self-consistent potential \( U(\vec{r}) \) due to all the other electrons in addition to the nuclear potential. This potential depends on the electron density \( n(\vec{r}) \) which in turn is determined by the wavefunctions of the filled states. Given the electron density how one determines \( U(\vec{r}) \) is the subject of much discussion and research. The “zero order” approach is to calculate \( U(\vec{r}) \) from \( n(\vec{r}) \) based on the laws of electrostatics, but it is well-established that this so-called Hartree approximation will overestimate the repulsive potential and there are various approaches for estimating this reduction. The density functional theory (DFT) has been spectacularly successful in describing this correction for equilibrium problems and in its simplest form amounts to a reduction by an amount proportional to the cube root of the electron density

\[
U(\vec{r}) = U_{\text{Hartree}} - \frac{q^2}{4\pi\epsilon} \left( n(\vec{r}) \right)^{1/3}
\]  

(18.9)

Many are now using similar corrections for non-equilibrium problems like current flow as well, though we believe there are important issues that remain to be resolved.

We should also note that there is a vast literature (both experiment and theory) on a regime of transport that cannot be easily described within an scf model. It is not just a matter of correctly evaluating the self-consistent potential. The very picture of quasi-independent electrons
The Model 263

moving in a self-consistent field needs revisiting, as we will see in Lecture 23.

18.3. Differential to Matrix Equation

All numerical calculations typically proceed by turning the differential equation in Eq.(18.11) into a matrix equation of the form

\[ E[S] \{ \psi \} = [H] \{ \psi \} \tag{18.10} \]

or equivalently

\[ E \sum_m S_{nm} \psi_m = \sum_m H_{nm} \psi_m \]

by expanding the wavefunction in terms of a set of known functions \( u_m(\mathbf{r}) \) called the basis functions:

\[ \psi(\mathbf{r}) = \sum_m \psi_m u_m(\mathbf{r}) \tag{18.11a} \]

The elements of the two matrices \([S]\) and \([H]\) are given respectively by

\[ S_{nm} = \int d\mathbf{r} u_n^*(\mathbf{r}) u_m(\mathbf{r}) \]

\[ H_{nm} = \int d\mathbf{r} u_n^*(\mathbf{r}) H_{op} u_m(\mathbf{r}) \tag{18.11b} \]

These expressions are of course by no means obvious, but we will not go into it further since we will not really be making any use of them. Let me explain why.

18.3.1. Semi-empirical tight-binding (TB) models

There are a wide variety of techniques in use which differ in the specific basis functions they use to convert the differential equation into a matrix equation. But once the matrices \([S]\) and \([H]\) have been evaluated, the eigenvalues \( E \) of Eq.(18.10) (which are the allowed energy levels) are determined using powerful matrix techniques that are widely available.
In modeling nanoscale structures, it is common to use basis functions that are spatially localized rather than extended functions like sines or cosines. For example, if we were to model a hydrogen molecule, with two positive nuclei as shown (see Fig.18.3), we could use two basis functions, one localized around the left nucleus and one around the right nucleus. One could then work through the algebra to obtain $[H]$ and $[S]$ matrices of the form

$$H = \begin{bmatrix} \varepsilon & t \\ t & \varepsilon \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} 1 & s \\ s & 1 \end{bmatrix} \quad (18.12)$$

where $\varepsilon$, $t$ and $s$ are three numbers.

The two eigenvalues from Eq.(18.10) can be written down analytically as

$$E_1 = \frac{\varepsilon - t}{1 - s} \quad \text{and} \quad E_2 = \frac{\varepsilon + t}{1 + s}$$

Fig.18.3. To model a Hydrogen molecule with two positive nuclei, one could use two basis functions, one localized around the left nucleus and one around the right nucleus.

What we just described above would be called a first-principles approach. Alternatively one could adopt a semi-empirical approach treating $\varepsilon$, $t$ and $s$ as three numbers to be adjusted to give the best fit to our “favorite” experiments. For example, if the energy levels $E_{1,2}$ are known from experiments, then we could try to choose numbers that match these. Indeed, it is common to assume that the $[S]$ matrix is just an identity matrix ($s=0$), so that there are only two parameters $\varepsilon$ and $t$ which are then adjusted to match $E_{1,2}$. Basis functions with $s = 0$ are said to be “orthogonal.”
18.3.2. Size of matrix, \( N = n \times b \)

What is the size of the \([H]\) matrix? Answer: \((N \times N)\), \( N \) being the total number of basis functions. How many basis functions? Answer: Depends on the approach one chooses. In the tight-binding (TB) approach, which we will use, the basis functions are the atomic wavefunctions for individual atoms, so that \( N = n \times b \), \( n \) being the number of atoms and \( b \), the number of basis functions per atom. What is \( b \)? Let us look at specific examples.

Suppose we want to model current flow through graphene consisting of carbon atoms arranged in a two dimensional hexagonal sheet (see Fig. 18.4). Carbon \((Z=6)\) has six electrons which are accommodated in the 1s, 2s and 2p levels as shown. The electrons in the highest levels that is the 2s and 2p levels are the so called valence electrons that move around and carry current. So in the simplest theories, it is common to use the 2s and 2p levels on each atom as the basis functions, with \( b=4 \).

The same is true of say silicon \((Z=14)\), the most common semiconductor for electronic devices. Its fourteen electrons are accommodated as shown with the valence electrons in the 3s, 3p levels. Once again in the simplest models \( b=4 \), though some models include five 3d levels and/or the two 4s levels as part of the basis functions too.

One of the nice things about graphene is that the 2s,\( p_x,p_y \) orbitals are in the simplest approximation completely decoupled from the 2p\(_z\) orbitals, and for understanding current flow, one can get a reasonable description with just one 2p\(_z\) orbital for every carbon atom, so that \( b=1 \).
Lessons from Nanoelectronics


In these simplest models, the matrix $[H]$ is of size $(nxn)$, $n$ being the total number of carbon atoms. Its diagonal elements have some value $\epsilon$, while the matrix element $H_{nm}$ equals some value $t$ if $n$ and $m$ happen to be nearest neighbors. If they are not nearest neighbors then one expects the value to be smaller since the functions $\phi_m$ and $\phi_n$ appearing in Eq.(18.14) do not overlap as much. In nearest neighbor tight-binding models it is common to set all such matrix elements to zero, so that we are finally left with just two parameters $\epsilon$ and $t$ which are then adjusted to match known results.

18.4. Choosing Matrix Parameters

One common way to select the parameters is to fit the known energy dispersion relation $E(k)$, also called the energy-momentum relation $E(p)$ (Note that $p = \hbar k$) as discussed in Lecture 5. These relations have been arrived at through years of work combining careful experimental measurements with sophisticated first-principles calculations. If we can get our semi-empirical model to fit the accepted dispersion relation for a material, we have in effect matched the whole set of experiments that contributed to it.

18.4.1. One-Dimensional Conductor

Suppose we have a one-dimensional conductor that we would like to model with a nearest neighbor orthogonal tight-binding model with two
parameters $\epsilon$ and $t$ representing the diagonal elements and the nearest neighbor coupling (Fig.18.5).

![Diagram of a one-dimensional array of atoms spaced by “a” modeled with a nearest neighbor orthogonal tight-binding model with two parameters $\epsilon$ and $t$ representing the diagonal elements and the nearest neighbor coupling.](image)

How would we choose $\epsilon$ and $t$ so that we approximate a parabolic dispersion relation

$$E(k) = E_c + \frac{\hbar^2 k^2}{2m} ?$$

(18.13)

The answer is that our model represents a set of algebraic equations (see Eq.(18.10)) which for the orthogonal model reduces to

$$E \psi_n = \sum_m H_{nm} \psi_m \quad \Rightarrow \quad E = \sum_m H_{nm} \frac{\psi_m}{\psi_n}$$

If we assume a solution of the form

$$\psi_n = \exp(ikna)$$

we obtain the $E(k)$ relation corresponding to Eq.(18.10):

$$E(k) = \sum_n H_{nn} \exp(ik(n-n)a)$$

(18.14)

Can we always assume a solution of this form? No. In general Eq.(18.14) will give us different results for $E(k)$ depending on what value we choose for $n$ when doing the summation and what we get for some particular choice of $n$ is not very helpful. But if the structure is “translationally
invariant” such that we get the same answer for all $n$ then we get a unique $E(k)$ relation and Eq.(18.13) indeed represents an acceptable solution to our set of equations.

For our particular nearest neighbor model Eq.(18.14) yields straightforwardly

$$E(k) = \varepsilon + t \exp(+ika) + t \exp(-ika)$$

$$= \varepsilon + 2t \cos ka$$

(18.15)

How would we make this match the desired parabolic relation in Eq.(18.13)? Clearly one could not match them for all values of $k$, only for a limited range. For example, if we want them to match over a range of $k$-values around $k=0$, we can expand the cosine in a Taylor series around $ka = 0$ to write

$$\cos ka = 1 - \frac{(ka)^2}{2}$$

so that the best match is obtained by choosing

$$t = -\frac{\hbar^2}{2ma^2}$$

$$\varepsilon = E_C - 2t$$

(18.16)

Fig.18.6. A spatially varying potential $U(x)$ along the channel is included by adding the local value of $U$ to the diagonal element $\varepsilon$. 
Finally I should mention that when modeling a device there could be a spatially varying potential $U(x)$ along the channel which is included by adding the local value of $U$ to the diagonal element as indicated in Fig.18.6. We now no longer have the “translational invariance” needed for a solution of the form $\exp(ikx)$ and the concept of a dispersion relation $E(k)$ is not valid. But a Hamiltonian of the form just described (Fig.18.6) can be used for numerical calculations and appear to be fairly accurate at least for potentials $U(x)$ that do not vary too rapidly on an atomic scale.

### 18.4.2. Two-Dimensional Conductor

A two-dimensional array of atoms (Fig.18.7) can be modeled similarly with a nearest neighbor orthogonal TB model, with the model parameters $\varepsilon$ and $t$ chosen to yield a dispersion relation approximating a standard parabolic effective mass relation:

$$E(k_x,k_y) = E_C + \frac{\hbar^2(k_x^2 + k_y^2)}{2m}$$

(18.17)

In this case we can assume a solution of the form

$$\psi_n = \exp(i\vec{k} \cdot \vec{r}_n)$$

where $\vec{k} = k_x\hat{x} + k_y\hat{y}$ and $\vec{r}_n$ denotes the location of atom $n$. Substituting into Eq.(18.10) we obtain the dispersion relation.
which for our nearest neighbor model yields

\begin{align}
E(\vec{k}) &= \sum_{m} H_{mn} \exp\left(i\vec{k} \cdot (\vec{r}_m - \vec{r}_n)\right) \\
&= \varepsilon + t \exp(+ik_x a) + t \exp(-ik_x a) \\
&\quad + t \exp(+ik_y a) + t \exp(-ik_y a) \\
&= \varepsilon + 2t \cos(k_x a) + 2t \cos(k_y a)
\end{align}

Following the same arguments as in the 1-D case, we can make this match the parabolic relation in Eq.(18.17) by choosing

\begin{align}
t &= -\hbar^2 / 2ma^2 \\
\varepsilon &= E_c - 4t
\end{align}

18.4.3. TB parameters in B-field

It is shown in Appendix C that if we replace \( \vec{p} \) with \( \vec{p} + q\vec{A} \) in Eq.(18.6a)

\[ E_{\text{classical}}(\vec{r},\vec{p}) = \frac{(\vec{p} + q\vec{A}) \cdot (\vec{p} + q\vec{A})}{2m} + U(\vec{r}) \]

yields the correct classical laws of motion of a particle of charge \(-q\) in a vector potential \( \vec{A} \). The corresponding wave equation is obtained using the replacement in Eq.(18.1): \( \vec{p} \rightarrow -i\hbar \vec{\nabla} \).

To find the appropriate TB parameters for the Hamiltonian in a B-field we consider the homogeneous material with constant \( E_c \) and a constant vector potential. Consider first the 1D problem with

\[ E(p_x) = E_c + \frac{(p_x + qA_x)(p_x + qA_x)}{2m} \]

so that the corresponding wave equation has a dispersion relation
The Model

\[ E(k_x) = E_c + \frac{(\hbar k_x + qA_x)(\hbar k_x + qA_x)}{2m} \]

which can be approximated by a cosine function

\[ E(k_x) = \varepsilon + 2t \cos\left(\frac{k_x a + qA_x a}{\hbar}\right) \]

with \( \varepsilon \) and \( t \) chosen according to Eq.(18.16). This means that we can model it with the 1-D lattice shown here, which differs from our original model in Fig.18.5 by the extra phase \( qA_x a / \hbar \).

Similar arguments lead to a similar phase in the \( y \)-direction as well. This is included in the 2D tight-binding model by modifying the nearest neighbor coupling elements to include an appropriate phase in the nearest neighbor coupling elements as shown in Fig.18.8 with

\[ \varphi_x = \frac{qA_x a}{\hbar}, \quad \varphi_y = \frac{qA_y a}{\hbar} \]

Fig.18.8. The effect of a magnetic field in the \( z \)-direction is included in a tight-binding model by introducing a phase in the nearest neighbor coupling elements as discussed in the text.
To include a $B$-field we have to let the vector potential vary spatially from one lattice point to the next such that

$$\vec{B} = \vec{\nabla} \times \vec{A}$$

For example, a $B$-field in the $z$-direction described in general by a vector potential $A_z(y)$ and/or $A_x(x)$ such that

$$B_z = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$$

For a given $B$-field the potential $A$ is not unique, and it is usually convenient to choose a potential that does not vary along the direction of current flow.

18.4.4. Lattice with a “Basis”

We have seen how for any given TB model we can evaluate the $E(\vec{k})$ relation from Eq.(18.18) and then fit it to a desired function. However, Eq.(18.18) will not work if we have a “lattice with a basis”. For example if we apply it to the graphene lattice shown in Fig.18.9, we will get different answers depending on whether we choose “n” to be the left carbon atom or the right carbon atom. The reason is that in a lattice like this these two carbon atoms are not in identical environments: One sees two bonds to the left and one bond to the right, while the other sees one bond to the left and two bonds to the right. We call this a lattice with a basis in the sense that two carbon atoms comprise a unit cell: if we view a pair of carbon atoms (marked A and B) as a single entity then the lattice looks translationally invariant with each entity in an identical environment.

We can then write the set of equations in Eq.(18.10) in the form

$$E \{\psi\}_n = \sum_m [H]_{nm} \{\psi\}_m$$

(18.20)
where \( \{\psi_n\} \) is a (2x1) column vector whose components represent the two atoms comprising unit cell number \( n \). Similarly \([H_{nm}]\) is a (2x2) matrix representing the coupling between the two components of unit cell \( n \) and unit cell \( m \) (see Fig.18.9).

\[
\begin{bmatrix}
0 & t \\
0 & 0
\end{bmatrix}
\]

Fig.18.9: If we view two carbon atoms as a single entity then the lattice in Fig.18.4 looks translationally invariant with each entity in an identical environment. Viewing the two atoms in each unit cell as a single entity we can write the set of equations in the form shown in Eq.(18.20) with \([H]_{nm}\) given by (2x2) matrices as shown.

Now if we write the solution in the form

\[
\{\psi\}_n = \{\psi\}_0 \exp(ik \cdot \vec{r}_n)
\]

we obtain from Eq.(18.20)

\[
E \{\psi\}_0 = [h(\vec{k})]\{\psi\}_0
\]

\[
[h(\vec{k})] = \sum_m [H]_{nm} \exp(i\vec{k} \cdot (\vec{r}_m - \vec{r}_n))
\]
Note that \([h(\vec{k})]\) obtained from Eq.(18.23) is a (2x2) matrix, which can be shown after some algebra to be
\[
[h(\vec{k})] = \begin{bmatrix} \epsilon & h_0^* \\ h_0 & \epsilon \end{bmatrix} \tag{18.24}
\]
where
\[h_0 \equiv t(1 + 2 \cos(k_y a) \exp(i k_x b)) \tag{18.25}\]
Eq.(18.22) yields two eigenvalues for the energy \(E\) for each value of \(\vec{k}\):
\[E(\vec{k}) = \epsilon \pm |h_0(\vec{k})| \tag{18.26}\]
Eq.(18.26) and (18.25) give a widely used dispersion relation for graphene. Once again to obtain a simple polynomial relation we need a Taylor series expansion around the \(k\)-value of interest. In this case the \(k\)-values of interest are those that make
\[h_0(\vec{k}) = 0, \]
so that
\[E(\vec{k}) = \epsilon \]
This is because the equilibrium electrochemical potential is located at \(\epsilon\) for a neutral sample for which exactly half of all the energy levels given by Eq.(18.26) are occupied.

It is straightforward to see that this requires
\[h_0(\vec{k}) = 0 \quad \rightarrow \quad k_x a = 0, k_y b = \pm 2\pi / 3 \quad (18.27)\]
Alternatively one could numerically make a grayscale plot of the magnitude of \(h_0(\vec{k})\) as shown below and look for the dark spots where it is a minimum. Each of these spots is called a valley and one can do a Taylor expansion around the minimum to obtain an approximate dispersion relation valid for that valley. Note that two of the dark spots correspond to the points in Eq.(18.27), but there are other spots too and it requires some discussion to be convinced that these additional valleys do
not need to be considered separately (see for example, Chapter 5, Datta (2005)).

A Taylor expansion around the points in Eq.(18.27) yields

\[ h_0(k) = \pm i a (k_x \mp i \beta_y), \]

where

\[ \beta_y \equiv k_y \mp 2\pi / 3b \quad (18.28) \]

Using this approximate relation we obtain a simple dispersion relation:

\[ E = \epsilon \pm at \sqrt{k_x^2 + k_y^2} \quad (18.29) \]

which corresponds to the energy-momentum relation

\[ E = v_0 p \]

that we stated in Lecture 5, if we set \( \epsilon = 0 \). The two valleys correspond to the two values of \( k_y b \) in Eq.(18.27).

In summary, although the differential form of the Schrödinger equation (Eq.(18.2)) is the well-known one that appears in most textbooks as well as on T-shirts, practical calculations are usually based on a discretized version that represents the Hamiltonian operator, \( H_{ap} \) (Eq.(18.8)) as a
Lessons from Nanoelectronics

matrix of size N x N, N being the number of basis functions used to represent the structure.

Given a set of basis functions, the matrix $[H]$ can be obtained from first principles, but a widely used approach is to use the principles of bandstructure to represent the matrix in terms of a few parameters which are chosen to match key experiments. Such semi-empirical approaches are often used because of their convenience and can explain a wide range of experiments beyond the key ones used that are used as input, suggesting that they capture a lot of essential physics.

Our approach in these Lectures will be to

1. take accepted energy-momentum $E(p)$ relations that are believed to describe the dynamics of conduction electrons with energies around the electrochemical potential $\mu_0$,

2. extract appropriate parameters to use in tight-binding model by discretizing it.

Knowing the $[H]$, we can obtain the $[\Sigma_{1,2}]$ describing the connection to the physical contacts and possible approaches will be described when discussing specific examples in Lectures 20 through 22. A key difference between the $[H]$ and $[\Sigma]$ matrices is that the former is Hermitian with real eigenvalues, while the latter is non-Hermitian with complex eigenvalues, whose significance we will discuss in the next Lecture.

As we mentioned at the outset, there are many approaches to writing $[H]$ and $[\Sigma_{1,2}]$ of which we have only described the simplest versions. But regardless of how we chose to write these matrices, we can use the NEGF-based approach to be described in the next Lecture.
Lecture 19

NEGF Method

19.1. One-level Resistor
19.2. Multi-level Resistors
19.3. Conductance Functions for Coherent Transport
19.4. Elastic Dephasing

In the last Lecture I tried to provide a super-brief but hopefully self-contained introduction to the Hamiltonian matrix [H] whose eigenvalues tell us the allowed energy levels in the channel. However, [H] describes an isolated channel and we cannot talk about the steady-state resistance of an isolated channel without bringing in the contacts and the battery connected across it. In this Lecture I will describe the NEGF-based transport model that can be used to model current flow, given [H] and the [Σ]’s (Fig.19.1).

![Diagram](image)

Fig.19.1. The NEGF-based quantum transport model described here allows us to model current flow given the Hamiltonian matrix [H] describing the channel, the self-energy matrices [Σ] describing the connection of the channel to the contacts, and [Σ₀] describing interactions within the channel.
As I mentioned in Lectures 1 and 18, the NEGF method originated from the seminal works of Martin and Schwinger (1959), Kadanoff and Baym (1962), Keldysh (1965) and others who used the methods of many-body perturbation theory (MBPT) to describe the distributed entropy-generating processes along the channel which were believed to constitute the essence of resistance. Since MBPT is an advanced topic requiring many semesters to master, the NEGF method is generally regarded as an esoteric tool for specialists.

In our bottom-up approach we will start with elastic resistors for which energy exchange is confined to the contacts, and the problem of resistance can be treated within a one-electron picture by connecting contacts to the Schrödinger equation. Indeed our approach will be to start from the usual time-independent Schrödinger equation $E \{\psi\} = \{H\} \{\psi\}$ and add two terms to it representing the outflow and inflow from the contact

$$E \{\psi\} = \{H\} \{\psi\} + \left[ \sum \right] \{\psi\} + \{s\}_{\text{OUTFLOW}}$$

$$+ \{s\}_{\text{INFLOW}}$$

These two terms arise from imposing open boundary conditions on the Schrödinger equation with an incident wave from the contact as shown in Chapters 8,9 of Datta (2005). Some readers may notice the similarity of the additional terms here with those appearing in the Langevin equation used to describe Brownian motion, but we will not go into it.

Using this modified Schrödinger equation, the wavefunction can be written as

$$\{\psi\} = \left[ EI - H - \Sigma \right]^{-1} \{s\}$$

We will then argue that since the inflow from multiple sources $\{s\}$ are incoherent, one cannot superpose the resulting $\{\psi\}$’s and it is more convenient to work in terms of quantities like (superscript ‘+’ denotes conjugate transpose)
NEGF Method

\[ [G^n] \sim \{\psi\} \{\psi\}^+ \]
\[ [\Sigma^{in}] \sim \{s\} \{s\}^+ \]

which can be superposed. Defining

\[ G^R = [E{I} - H - \Sigma]^{-1} \]

(19.1)

and

\[ G^A = [G^R]^+ \]

we can write

\[ \{\psi\} = [G^R] \{s\} \]

so that

\[ \frac{\{\psi\} \{\psi\}^+}{G^n} = \frac{[G^R]}{\Sigma^{in}} \frac{\{s\} \{s\}^+ [G^A]}{\Sigma^{in}} \]

giving us the second NEGF equation

\[ G^n = G^R \Sigma^{in} G^A \]

(19.2)

Though we have changed the notation somewhat, writing \( \Sigma \) for \( \Sigma^R \), and (see Chapter 8, Datta 1995)

\[ G^n \quad \text{for} \quad -iG^< \quad \text{and} \quad \Sigma^{in} \quad \text{for} \quad -i\Sigma^< \]

Eqs.(19.1, 19.2) are essentially the same as Eqs.(75-77) in Keldysh (1965), which is one of the seminal founding papers on the NEGF method that obtained these equations using MBPT. Although for simplicity we have only discussed the time-independent version here, a similar derivation could be used for the time-dependent version too (See Appendix, Datta (2005)).

How could we obtain these results using elementary arguments, without invoking MBPT? Because we are dealing with an elastic resistor where all entropy-generating processes are confined to the contacts and can be handled in a relatively elementary manner. But should we call this NEGF?
It seems to us that NEGF has two aspects, namely
A. Eqs.(19.1), (19.2) and
B. calculating $[\Sigma], [\Sigma^{int}]$ that appear in Eqs.(19.1), (19.2).

For historical reasons, these two aspects, A and B, are often intertwined in the literature, but they need not be. Indeed these two aspects are completely distinct in the Boltzmann formalism (Lecture 7). The Boltzmann transport equation (BTE)

$$\frac{\partial f}{\partial t} + \mathbf{v}\cdot\nabla f + \mathbf{F}\cdot\nabla p f = S_{op} f \quad \text{(same as Eq.(7.5))}$$

is used to describe semiclassical transport in many different contexts, but the evaluation of the scattering operator $S_{op}$ has evolved considerably since the days of Boltzmann and varies widely depending on the problem at hand.

Similarly it seems to me that the essence of NEGF is contained in Eqs.(19.1), (19.2) while the actual evaluation of the $[\Sigma]$’s may well evolve as we look at more and more different types of problems. The original MBPT–based approach may or may not be the best, and may need to be modified even for problems involving electron-electron interactions.

Above all we believe that by decoupling Eqs.(19.1) and (19.2) from the MBPT method originally used to derive them, we can make the NEGF method more transparent and accessible so that it can become a part of the standard training of physics and engineering students who need to apply it effectively to a wide variety of basic and applied problems that require connecting contacts to the Schrödinger equation.

I should also note briefly the relation between the NEGF method applied to elastic resistors with the scattering theory of transport or the transmission formalism widely used in mesoscopic physics. Firstly, the scattering theory works directly with the Schrödinger equation with open
boundary conditions that effectively add the inflow and outflow terms we mentioned:

$$E \{ \psi \} = [H] \{ \psi \} + \left[ \Sigma \right] \{ \psi \} + \begin{cases} \{ s \} \text{OUTFLOW} \\ \{ s \} \text{INFLOW} \end{cases}$$

However, as we noted earlier it is then important to add individual sources incoherently, something that the NEGF equation (Eq.(19.2)) takes care of automatically.

The second key difference is the handling of dephasing processes in the channel, something that has no classical equivalent. In quantum transport randomization of the phase of the wavefunction even without any momentum relaxation can have a major impact on the measured conductance. The scattering theory of transport usually neglects such dephasing processes and is restricted to \textit{phase-coherent elastic resistors}.

Incoherence is commonly introduced in this approach using an insightful observation due to Büttiker that dephasing processes essentially remove electrons from the channel and re-inject them just like the voltage probes discussed in Section 12.2 and so one can include them phenomenologically by introducing conceptual contacts in the channel.

This method is widely used in mesoscopic physics, but it seems to introduce both phase and momentum relaxation and I am not aware of a convenient way to introduce pure phase relaxation if we wanted to. In the NEGF method it is straightforward to choose \left[ \Sigma_0 \right] so as to include phase relaxation with or without momentum relaxation as we will see in the next Lecture. In addition, the NEGF method provides a rigorous framework for handling all kinds of interactions in the channel, both elastic and inelastic, using MBPT. Indeed that is what the original work from the 1960’s was about.

Let me finish up this long introduction by briefly mentioning the two other key equations in NEGF besides Eqs.(19.1) and (19.2). As we will see, the quantity \( \{ G \} \) appearing in Eq.(19.2) represents a matrix version of the electron density (times \( 2\pi \)) from which other quantities of interest
can be calculated. Another quantity of interest is the matrix version of the density of states (again times $2\pi$) called the spectral function $A$ given by

$$A = G^R \Gamma G^A = G^A \Gamma G^R$$

$$= i[G^R - G^A]$$

(19.3a)

where $G^R, G^A$ are defined in Eq.(19.1) and the $[\Gamma]$'s represent the anti-Hermitian parts of the corresponding $[\Sigma]$'s

$$\Gamma = i[\Sigma - \Sigma^+]$$

(19.3b)

which describe how easily the electrons in the channel communicate with the contacts.

There is a component of $[\Sigma], [\Gamma], [\Sigma^{in}]$ for each contact (physical or otherwise) and the quantities appearing in Eqs.(19.1-19.3) are the total obtained summing all components. The current at a specific contact $m$, however, involves only those components associated with contact $m$:

$$\bar{I}_m = \frac{q}{\hbar} \text{Trace}[\Sigma_m^{in} A - \Gamma_m G^n]$$

(19.4)

Note that $\bar{I}_m(E)$ represents the current per unit energy and has to be integrated over all energy to obtain the total current. In the following four Lectures we will look at a few examples designed to illustrate how Eqs.(19.1)-(19.4) are applied to obtain concrete results.

But for the rest of this Lecture let me try to justify these equations. We start with a one-level version for which all matrices are just numbers (Section 19.1), then look at the full multi-level version (Section 19.2), obtain an expression for the conductance function $G(E)$ for coherent transport (Section 19.3) and finally look at the different choices for the dephasing self-energy $[\Sigma_0]$ (Section 19.4).
19.1. One-level resistor

To get a feeling for the NEGF method, it is instructive to look at a particularly simple conductor having just one level and described by a $1 \times 1$ matrix $\{H\}$ that is essentially a number: $[H] = \varepsilon$.

Starting directly from the Schrödinger equation we will see how we can introduce contacts into this problem. This will help set the stage for Section 19.3 when we consider arbitrary channels described by $(N \times N)$ matrices instead of the simple one-level channel described by $(1 \times 1)$ matrices.”

19.1.1. Semiclassical treatment

It is useful to first go through a semiclassical treatment as an intuitive guide to the quantum treatment. Physically we have a level connected to two contacts, with two different occupancy factors

$f_1(\varepsilon)$ and $f_2(\varepsilon)$

Let us assume the occupation factor to be one for the source and zero for the drain, so that it is only the source that is continually trying to fill up the level while the drain is trying to empty it. We will calculate the resulting current and then multiply it by

$f_1(\varepsilon) - f_2(\varepsilon)$

to account for the fact that there is injection from both sides and the net current is the difference.
With \( f_1 = 1 \) in the source and \( f_2 = 0 \) in the drain, the average number \( N \) of electrons \( (N < 1) \) should obey an equation of the form

\[
\frac{dN}{dt} = -(\nu_1 + \nu_2)N + S_1 + S_2 \tag{19.5}
\]

where \( \nu_1 \) and \( \nu_2 \) represent the rates (per second) at which an electron escapes into the source and drain respectively, while \( S_1 \) is the rate at which electrons try to enter from the source. The steady state occupation is obtained by setting

\[
\frac{dN}{dt} = 0 \rightarrow N = \frac{S_1 + S_2}{\nu_1 + \nu_2} \tag{19.6}
\]

We can fix \( S_1 \), by noting that if the drain were to be disconnected, \( N \) should equal the Fermi function \( f_1(\varepsilon) \) in contact 1, which we will assume one for this discussion. This means

\[
\frac{S_1}{\nu_1} = f_1(\varepsilon) \quad \text{and} \quad \frac{S_2}{\nu_2} = f_2(\varepsilon) \tag{19.7}
\]

The current can be evaluated by writing Eq.(19.5) in the form

\[
\frac{dN}{dt} = (S_1 - \nu_1 N) + (S_2 - \nu_2 N) \tag{19.8}
\]

and noting that the first term on the right is the current from the source while the second is the current into the drain. Under steady state conditions, they are equal and either could be used to evaluate the current that flows in the circuit:

\[
I = q(S_1 - \nu_1 N) = q(\nu_2 N - S_2) \tag{19.9}
\]

From Eqs.(19.6), (19.7) and (19.9), we have

\[
N = \frac{\nu_1 f_1(\varepsilon) + \nu_2 f_2(\varepsilon)}{\nu_1 + \nu_2} \tag{19.10a}
\]
\[ I = q \frac{V_{1}V_{2}}{V_{1} + V_{2}} \left( f_{1}(\varepsilon) - f_{2}(\varepsilon) \right) \]  \hspace{1cm} (19.10b)

and

**19.1.2. Quantum treatment**

Let us now work out the same problem using a quantum formalism based on the Schrödinger equation. In the last Chapter we introduced the matrix version of the time-independent Schrödinger equation

\[ E \{ \psi \} = [H] \{ \psi \} \]

which can be obtained from the more general time-dependent equation

\[
i \hbar \frac{\partial}{\partial t} \{ \psi(t) \} = [H] \{ \psi(t) \} \]

(19.11a)

by assuming

\[
\{ \psi(t) \} = \{ \psi \} e^{-iEt/\hbar} \]

(19.11b)

For problems involving steady-state current flow, the time-independent version is usually adequate, but sometimes it is useful to go back to the time-dependent version because it helps us interpret certain quantities like the self-energy functions as we will see shortly.

In the quantum formalism the squared magnitude of the electronic wavefunction \( \tilde{\psi}(t) \) tells us the probability of finding an electron occupying the level and hence can be identified with the average number of electrons \( N (\ll 1) \). For a single isolated level with \( [H] = \varepsilon \), the time evolution of the wavefunction is described by

\[
i \hbar \frac{d}{dt} \tilde{\psi} = \varepsilon \tilde{\psi} \]

which with a little algebra leads to

\[
\frac{d}{dt} (\tilde{\psi} \tilde{\psi}^*) = 0
\]
showing that for an isolated level, the number of electrons $\psi \psi^*$ does not change with time.

Our interest, however, is not in isolated systems, but in channels connected to two contacts. Unfortunately the standard quantum mechanics literature does not provide much guidance in the matter, but we can do something relatively simple using the rate equation in Eq.(19.4) as a guide.

We introduce contacts into the Schrödinger equation by modifying it to read

\[ i\hbar \frac{d}{dt} \bar{\psi} = \left( \varepsilon - i \frac{\gamma_1 + \gamma_2}{2} \right) \bar{\psi} \] (19.12a)

so that the resulting equation for

\[ \frac{d}{dt} \bar{\psi} \psi^* = -\left( \frac{\gamma_1 + \gamma_2}{\hbar} \right) \bar{\psi} \psi^* \] (19.12b)

looks just like Eq.(19.5) except for the source term $S_i$ which we will discuss shortly.

We can make Eq.(19.12b) match Eq.(19.5) if we choose

\[ \gamma_1 = \hbar \nu_1 \] (19.13a)

\[ \gamma_2 = \hbar \nu_2 \] (19.13b)

We can now go back to the time-independent version of Eq.(19.12a):

\[ E\psi = \left( \varepsilon - i \frac{\gamma_1 + \gamma_2}{2} \right) \psi \] (19.14)

obtained by assuming a single energy solution:

\[ \bar{\psi}(t) = \psi(E) e^{-iEt/\hbar} \]
Eq. (19.14) has an obvious solution $\psi = 0$, telling us that at steady-state there are no electrons occupying the level which makes sense since we have not included the source term $S_1$. All electrons can do is to escape into the contacts, and so in the long run the level just empties to zero.

**Introducing a source term** into Eq. (19.11) and defining $\gamma \equiv \gamma_1 + \gamma_2$, we have

$$E\psi = \left(\varepsilon - i\frac{\gamma}{2}\right)\psi + s_1 \quad (19.15)$$

Unlike the semiclassical case (Eq. (19.5)) we are introducing only one source rather than two. The reason is subtle and we will address it later at the end of this Section. From Eq. (19.15), we can relate the wavefunction to the source

$$\psi = \frac{s_1}{E - \varepsilon + i(\gamma/2)} \quad (19.16)$$

Note that the wavefunction is a maximum when the electron energy $E$ equals the energy $\varepsilon$ of the level, as we might expect. But the important point about the quantum treatment is that the wavefunction is not significantly diminished as long as $E$ differs from $\varepsilon$ by an amount less than $\gamma$. This is an example of “broadening” or energy uncertainty that a semiclassical picture misses.

To obtain the strength of the source we require that the total number of electrons on integrating over all energies should equal our rate equation result from Eq. (19.5). That is,
Lessons from Nanoelectronics

\[ \int_{-\infty}^{+\infty} dE \psi \psi^* = \frac{v_1}{v_1 + v_2} = \frac{\gamma_1}{\gamma_1 + \gamma_2} \quad (19.17) \]

where we have made use of Eq.(19.13). We now use Eqs.(19.16), (19.17) to evaluate the right hand side in terms of the source

\[ \int_{-\infty}^{+\infty} dE \psi \psi^* = \int_{-\infty}^{+\infty} dE \frac{s_1 s_1^*}{(E - \varepsilon)^2 + \left(\frac{\gamma}{2}\right)^2} = \frac{2\pi s_1 s_1^*}{\gamma} \quad (19.18) \]

where we have made use of a standard integral

\[ \int_{-\infty}^{+\infty} dE \frac{\gamma}{(E - \varepsilon)^2 + \left(\frac{\gamma}{2}\right)^2} = 2\pi \quad (19.19) \]

From Eqs.(19.17) and (19.18) we obtain, noting that

\[ 2\pi s_1 s_1^* = \gamma_1 \quad (19.20) \]

The strength of the source is thus proportional to the escape rate which seems reasonable: if the contact is well coupled to the channel and electrons can escape easily, they should also be able to come in easily.

Just as in the semiclassical case (Eq.(19.9)) we obtain the current by looking at the rate of change of \( N \) from Eq.(19.12b)

\[ \frac{d}{dt} \psi \psi^* = \text{Inflow from 1} - \frac{\gamma_1}{\hbar} \psi \psi^* - \frac{\gamma_2}{\hbar} \psi \psi^* \]

where we have added a term “Inflow from 1” as a reminder that Eq.(19.12a) does not include a source term. Both left and right hand sides of this equation are zero for the steady-state solutions we are considering. But just like the semiclassical case, we can identify the current as either the first two terms or the last term on the right:
\[
\frac{I}{q} = (\text{Inflow from 1}) \quad -\frac{\gamma_1}{\hbar} \psi \psi^* = \frac{\gamma_2}{\hbar} \psi \psi^*
\]

Using the second form and integrating over energy we can write

\[
I = q \int_{-\infty}^{+\infty} dE \frac{\gamma_2}{\hbar} \psi \psi^*
\]  \hspace{1cm} (19.21)

so that making use of Eqs.(19.16) and (19.20), we have

\[
I = \frac{q \gamma_1 \gamma_2}{\hbar} \int_{-\infty}^{+\infty} dE \frac{1}{(E - \varepsilon)^2 + (\gamma/2)^2}
\]  \hspace{1cm} (19.22)

which can be compared to the semiclassical result from Eq.(19.10) with \( f_1=1, f_2=0 \) (note: \( \gamma = \gamma_1 + \gamma_2 \))

\[
I = \frac{q \gamma_1 \gamma_2}{\hbar} \frac{\gamma_1 + \gamma_2}{\hbar}
\]

19.1.3. Quantum broadening

Note that Eq.(19.22) involves an integration over energy, as if the quantum treatment has turned the single sharp level into a continuous distribution of energies described by a density of states \( D(E) \):

\[
D = \frac{\gamma / 2\pi}{(E - \varepsilon)^2 + (\gamma / 2)^2}
\]  \hspace{1cm} (19.23)

Quantum mechanically the process of coupling inevitably spreads a single discrete level into a state that is distributed in energy, but integrated over all energy still equals one (see Eq.(19.15)). One could call it a consequence of the uncertainty relation

\[
\gamma t \geq \hbar
\]
relating the length of time $t$ the electron spends in a level to the uncertainty "$\gamma$" in its energy. The stronger the coupling, shorter the time and larger the broadening.

Is there any experimental evidence for this energy broadening (Eq.(19.23)) predicted by quantum theory? A hydrogen molecule has an energy level diagram like the one-level resistor we are discussing and experimentalists have measured the conductance of a hydrogen molecule with good contacts and it supports the quantum result (Smit et al. 2002). Let me elaborate a little.

Comparing Eq.(19.22) with Eq.(3.3) for elastic resistors we can write the conductance function for a one-level device including quantum broadening as

$$G(E) = \frac{q^2}{h} \frac{\gamma_1 \gamma_2}{(E - \varepsilon)^2 + \left(\frac{\gamma}{2}\right)^2}$$

If we assume (1) equal coupling to both contacts:

$$\gamma_1 = \gamma_2 = \frac{\gamma}{2}$$

and (2) a temperature low enough that the measured conductance equals $G(E = \mu_0)$, $\mu_0$ being the equilibrium electrochemical potential, we have
So the quantum theory of the one-level resistor says that the measured conductance should show a maximum value equal to the quantum of conductance \( q^2 / h \) when \( \mu_0 \) is located sufficiently close to \( \varepsilon \). The experimentally measured conductance is equal to \( 2q^2 / h \), the extra factor of 2 being due to spin degeneracy, since levels come in pairs and what we have is really a two-level rather than a one-level resistor.

19.1.4. Do Multiple Sources Interfere?

In our quantum treatment we considered a problem with electrons injected only from the source \((f_1 = 1)\) with the drain empty \((f_2 = 0)\) (Eq.(19.15)), unlike the semiclassical case where we started with both sources \(S_1\) and \(S_2\) (Eq.(19.5)).

This is not just a matter of convenience. If instead of Eq.(19.15) we start from

\[
E\psi = \left( \varepsilon - i \frac{\gamma}{2} \right) \psi + s_1 + s_2
\]

we obtain

\[
\psi = \frac{s_1 + s_2}{E - \varepsilon + i \frac{\gamma}{2}}
\]

so that

\[
\psi \psi^* = \frac{1}{(E - \varepsilon)^2 + \left( \frac{\gamma}{2} \right)^2} \left( s_1 s_1^* + s_2 s_2^* + s_1 s_2^* + s_2 s_1^* \right)
\]

which has two extra interference terms that are never observed experimentally because the electrons injected from separate contacts have uncorrelated phases that change randomly in time and average to zero.
The first two terms on the other hand add up since they are positive numbers. It is like adding up the light from two light bulbs: we add their powers not their electric fields. Laser sources on the other hand can be coherent so that we actually add electric fields and the interference terms can be seen experimentally. Electron sources from superconducting contacts too can be coherent leading to Josephson currents that depend on interference. But that is a different matter.

Our point here is simply that normal contacts like the ones we are discussing are incoherent and it is necessary to take that into account in our models. The moral of the story is that we cannot just insert multiple sources into the Schrödinger equation. We should insert one source at a time, calculate bilinear quantities (things that depend on the product of wavefunctions) like electron density and current and add up the contributions from different sources. Next we will describe the non-equilibrium Green function (NEGF) method that allows us to implement this procedure in a systematic way and also to include incoherent processes.

19.2. Quantum transport through multiple levels

We have seen how we can treat quantum transport through a one-level resistor with a time-independent Schrödinger equation modified to include the connection to contacts and a source term:

\[ E \psi = \left( \epsilon - i \frac{\gamma}{2} \right) \psi + s \]

How do we extend this method to a more general channel described by an \( N \times N \) Hamiltonian matrix \( [H] \) whose eigenvalues give the \( N \) energy levels?

For an \( N \)-level channel, the wavefunction \( \{ \psi \} \) and source term \( \{ s_1 \} \) are \( N \times 1 \) column vectors and the modified Schrödinger equation looks like

\[ E \{ \psi \} = [H + \Sigma_1 + \Sigma_2] \{ \psi \} + \{ s_1 \} \]  \hspace{1cm} (19.24)
where $\Sigma_1$ and $\Sigma_2$ are $N \times N$ non-Hermitian matrices whose anti-Hermitian components

$$
\Gamma_1 = i(\Sigma_1 - \Sigma_1^\dagger)
$$
$$
\Gamma_2 = i(\Sigma_2 - \Sigma_2^\dagger)
$$

play the roles of $\gamma_{1,2}$ in our one-level problem.

In Lecture 18 we discussed how for different structures we can write down the channel Hamiltonian $[H]$ and in the next few Lectures I will present examples to show how the $[\Sigma]$ are obtained.

For the moment, let us focus on how the basic NEGF equations summarized earlier (Eqs.(19.1)-(19.4)) follow from our contact-ed Schrödinger equation, Eq.(19.24).

**19.2.1. Obtaining Eqs.(19.1)**

From Eq.(19.24) it is straightforward to write

$$
\{\psi\} = [G^R]\{s_1\}
$$

where $G^R$ is given by Eq.(19.1) with

$$
\Sigma = \Sigma_1 + \Sigma_2
$$

(19.25)
19.2.2. Obtaining Eqs.(19.2)

The matrix electron density, \( G^n \), defined as

\[
G^n \rightarrow 2\pi \{\psi\}{\psi}^+ = 2\pi [G^R] \{s_1\}{s_1}^+ [G^A]
\]

where the superscript "+" stands for conjugate transpose, and \( G^A \) stands for the conjugate transpose of \( G^R \).

For the one-level problem \( 2\pi s_1 s_1^* = \gamma_1 \) (see Eq.(19.20)): the corresponding matrix relation is

\[
2\pi \{s_1\}{s_1}^+ = [\Gamma_1]
\]

so that

\[
G^n = [G^R][\Gamma_1][G^A]
\]

This is for a single source term. For multiple sources, the electron density matrices, unlike the wavefunctions, can all be added up with the appropriate Fermi function weighting to give Eq.(19.2),

\[
G^n = [G^R][\Sigma^{in}][G^A]
\] (same as 19.2)

with \( \Sigma^{in} \) representing an incoherent sum of all the independent sources:

\[
[\Sigma^{in}] = [\Gamma_1] f_1(E) + [\Gamma_2] f_2(E) \quad (19.26)
\]

19.2.3. Obtaining Eq.(19.3)

Eq.(19.2) gives us the electron density matrix \( G^n \), in terms of the Fermi functions \( f_1 \) and \( f_2 \) in the two contacts. But if both \( f_1 \) and \( f_2 \) are equal to one then all states are occupied, so that the matrix electron density becomes equal to the matrix density of states, called the spectral function matrix \( [A] \). Setting \( f_1 = 1 \) and \( f_2 = 1 \), in Eq.(19.2) we have

\[
[A] = [G^R][\Gamma][G^A] \quad (19.27)
\]
since \( \Gamma = \Gamma_1 + \Gamma_2 \). This gives us part of Eq.(19.3). The rest of Eq.(19.3) can be obtained from Eq.(19.1) using straightforward algebra as follows:

\[
[G^R]^{-1} = EI - H - \Sigma \quad (19.28a)
\]

Taking conjugate transpose of both sides

\[
\left([G^R]^{-1}\right)^* = \left([G^R]^+\right)^{-1} = EI - H - \Sigma^+ \quad (19.28b)
\]

Subtracting Eq.(19.28b) from (19.28a) (note that \( G^A \) stands for \( [G^R]^+ \)) and making use of Eq.(19.3b)

\[
[G^R]^{-1} - [G^A]^{-1} = i [\Gamma] \quad (19.28c)
\]

Multiplying with \( [G^R] \) from the left and \( [G^A] \) from the right we have

\[
i \left([G^R]-[G^A]\right) = G^R \Gamma G^A
\]

thus giving us another piece of Eq.(19.3). The final piece is obtained by multiplying Eq.(19.28c) with \( [G^R] \) from the left and \( [G^A] \) from the right.

19.2.4. Obtaining Eq.(19.4): The Current Equation

Like the semiclassical treatment and the one-level quantum treatment, the current expression is obtained by considering the time variation of the number of electrons \( N \). Starting from

\[
i \hbar \frac{d}{dt} \{\psi\} = [H + \Sigma] \{\psi\} + \{s\}
\]

and its conjugate transpose (noting that \( H \) is a Hermitian matrix)

\[
- i \hbar \frac{d}{dt} \{\psi\}^+ = \{\psi\}^+ [H + \Sigma^+] + \{s\}^+
\]
we can write

\[ \text{i} \hbar \frac{d}{dt} \{ \psi \} \{ \psi \}^+ = \left( \text{i} \hbar \frac{d}{dt} \{ \psi \} \right) \{ \psi \}^+ + \{ \psi \left( \text{i} \hbar \frac{d}{dt} \{ \psi \}^+ \right) \}
\]

\[ = \left( \{H + \Sigma\} \{ \psi \} + \{s\} \{ \psi \}^+ \right) - \{ \psi \} \left( \{ \psi \}^\dagger [H + \Sigma^\dagger] + \{s\}^\dagger \right)
\]

\[ = [(H + \Sigma) \psi \psi^\dagger - \psi \psi^\dagger (H + \Sigma^\dagger)] + [ss^\dagger G^A - G^R ss^\dagger]
\]

where we have made use of the relations

\[ \{ \psi \} = [G^R] \{ s \} \quad \text{and} \quad \{ \psi \}^+ = \{ s \}^+ [G^A]
\]

Since the trace of \([ \psi \psi^\dagger ]\) represents the number of electrons, we could define its time derivative as a matrix current operator whose trace gives us the current. Noting further that

\[ 2\pi \{ \psi \} \{ \psi \}^+ = [G^n] \quad \text{and} \quad 2\pi \{ s \} \{ s \}^+ = [\Gamma]
\]

we can write

\[ I^{op} = \frac{[HG^n - G^n H] + [\Sigma G^n - G^n \Sigma^\dagger] + [\Sigma^n G^A - G^R \Sigma^A]}{2\pi \hbar} \] (19.29)

This current operator provides a good starting point for discussing more subtle quantities like spin currents, but for the moment we will focus on ordinary currents requiring just the trace of this operator which tells us the time rate of change of the number of electrons \( N \) in the channel

\[ \frac{dN}{dt} = \frac{-i}{\hbar} \text{Trace} \left( [\Sigma G^n - G^n \Sigma^\dagger] + [\Sigma^n G^A - G^R \Sigma^A] \right)
\]

noting that \( \text{Trace} [AB] = \text{Trace} [BA] \). Making use of Eq.(19.3b)

\[ \frac{dN}{dt} = \frac{1}{\hbar} \text{Trace} \left( \Sigma^n A - \Gamma G^n \right)\]
Now comes a tricky argument. Both the left and the right hand sides of Eq.(19.29) are zero, since we are discussing steady state transport with no time variation. The reason we are spending all this time discussing something that is zero is that the terms on the left can be separated into two parts, one associated with contact 1 and one with contact 2. They tell us the currents at contacts 1 and 2 respectively and the fact that they add up to zero is simply a reassuring statement of Kirchhoff’s law for steady-state currents in circuits.

With this in mind we can write for the current at contact $m$ ($m=1,2$)

$$\tilde{I}_m = \frac{1}{h} \text{Trace} \left[ \Sigma_m^A - \Gamma_m G^n \right]$$

as stated earlier in Eq.(19.4). This leads us to the picture shown in Fig.19.5 where we have also shown the semiclassical result for comparison.

---

Fig.19.5: Filling and emptying a channel: Classical and quantum treatment.
19.3. Conductance Functions for Coherent Transport

Finally we note that using Eqs. (19.2)-(19.3) we can write the current from Eq. (19.4) a little differently

$$ I(E) = \frac{q}{h} \text{Trace}[\Gamma_1 G^R \Gamma_2 G^A](f_1(E) - f_2(E)) $$

which is very useful for it suggests a quantum expression for the conductance function $G(E)$ that we introduced in Lecture 3 for all elastic resistors:

$$ G(E) = \frac{q^2}{h} \text{Trace} \left[ \Gamma_1 G^R \Gamma_2 G^A \right] \quad (19.30) $$

More generally with multiterminal conductors we could introduce a self-energy function for each contact and show that

$$ \tilde{I}_m = \frac{q}{h} \sum_r \tilde{T}_{mn} (f_m(E) - f_n(E)) \quad (19.31) $$

with

$$ \tilde{T}_{mn} = \text{Trace} \left[ \Gamma_m G^R \Gamma_n G^A \right] \quad (19.32) $$

For low bias we can use our usual Taylor series expansion from Eq. (2.8) to translate the Fermi functions into electrochemical potentials so that Eq. (19.31) looks just like the Buttiker equation (Eq. (12.3)) with the conductance function given

$$ G_{m,n}(E) = \frac{q^2}{h} \text{Trace} \left[ \Gamma_m G^R \Gamma_n G^A \right] \quad (19.33) $$

which is energy-averaged in the usual way for elastic resistors (see Eq. (3.1)).

$$ G_{m,n} = \int_{-\infty}^{+\infty} dE \left( -\frac{\partial f_0}{\partial E} \right) G_{m,n}(E) $$
19.4. Elastic Dephasing

So far we have focused on the physical contacts described by $[\Sigma_{1,2}]$ and the model as it stands describes coherent quantum transport where electrons travel coherently from source to drain in some static structure described by the Hamiltonian $[H]$ without any interactions along the channel described by $[\Sigma_0]$ (Fig. 19.1). In order to include $[\Sigma_0]$, however, no change is needed as far as Eqs. (19.1) through (19.4) is concerned. It is just that an additional term appears in the definition of $\Sigma$, $\Sigma^{in}$:

$$\Sigma = \Sigma_1 + \Sigma_2 + \Sigma_0$$

$$\Gamma = \Gamma_1 + \Gamma_2 + \Gamma_0$$

$$[\Sigma^{in}] = [\Gamma_1]f_1(E) + [\Gamma_2]f_2(E) + [\Sigma_0^{in}]$$ (19.34)

What does $[\Sigma_0]$ represent physically? From the point of view of the electron a solid does not look like a static medium described by $[H]$, but like a rather turbulent medium with a random potential $U_R$ that fluctuates on a picosecond time scale. Even at fairly low temperatures when phonons have been frozen out, an individual electron continues to see a fluctuating potential due to all the other electrons, whose average is modeled by the scf potential we discussed in Section 18.2. These fluctuations do not cause any overall loss of momentum from the system of electrons, since any loss from one electron is picked up by another. However, they do cause fluctuations in the phase leading to fluctuations in the current. What typical current measurements tell us is an average flow over nanoseconds if not microseconds or milliseconds. This averaging effect needs to be modeled if we wish to relate to experiments.
As we mentioned earlier, the NEGF method was originally developed in the 1960's to deal with the problem of including inelastic processes into a quantum description of large conductors. For the moment, however, we will look at simple elastic dephasing processes leaving more general interactions for Lecture 23.

For such processes the self-energy functions are given by

\[
\Sigma_0 = D \times [G^R] \quad \text{(19.35a)}
\]

\[
\Sigma_{0i} = D \times [G^j] \quad \text{(19.35b)}
\]

where \( \times \) denotes element by element multiplication. Making use of the relations in Eqs.(19.3), it is straightforward to show from Eq.(19.35a) that

\[
\Gamma_0 = D \times [A] \quad \text{(19.35c)}
\]

The elements of the matrix \([D]\) represent the correlation between the random potential at location “i” and at location “j”:

\[
D_{ij} = \left\langle U_{Ri} U_{Rj} \right\rangle \quad \text{(19.36)}
\]

Two cases are of particular interest. The first is where the random potential is well-correlated throughout the channel having essentially the same value at all points “i” so that the every element of the matrix \([D]\) has essentially the same value \(D_0\).
Model A: \( D_{ij} = D_0 \) \hspace{1cm} (19.37)

The other case is where the random potential has zero correlation from one spatial point \( i \) to another \( j \), so that

Model B: \( D_{ij} = D_0, \quad i = j \quad \text{and} \quad = 0, \quad i \neq j \) \hspace{1cm} (19.38)

Real processes are usually somewhere between the two extremes represented by models A and B.

To see where Eqs.(19.35) come from we go back to our contact-ed Schrödinger equation

\[
E \{ \psi \} = [H + \Sigma_1 + \Sigma_2] \{ \psi \} + \{ s_1 \}
\]

and noting that a random potential \( U_R \) should lead to an additional term that could be viewed as an additional source term

\[
E \{ \psi \} = [H + \Sigma_1 + \Sigma_2] \{ \psi \} + U_R \{ \psi \} + \{ s_1 \}
\]

with a corresponding inscattering term given by

\[
\Sigma_0^{in} = 2\pi U_R U_R^* \{ \psi \} \{ \psi \}^+ = D_0 G^n
\]

corresponding to Model A (Eq.(19.37)) and a little more careful argument leads to the more general result in Eq.(19.36). That gives us Eq.(19.35b). How about Eq.(19.35a) and (19.35c) ?

The simplest way to justify Eq.(19.35c) is to note that together with Eq.(19.35b) (which we just obtained) it ensures that the current at terminal \( 0 \) from Eq.(19.4) equals zero:

\[
I_0 = \frac{q}{\hbar} \text{Trace}[\Sigma_0^{in} A - \Gamma_0 G^n] = 0
\]
This is a required condition since terminal $\theta$ is not a physical contact where electrons can actually exit or enter from.

Indeed a very popular method due to Büttiker introduces incoherent processes by including a fictitious probe (often called a Büttiker probe) whose electrochemical potential is adjusted to ensure that it draws zero current. In NEGF language this amounts to assuming

$$\Sigma^i_0 = \Gamma_0 f_P$$

with the number $f_P$ is adjusted for zero current. This would be equivalent to the approach described here if the probe coupling $\Gamma_0$ were chosen proportional to the spectral function $fA$ as required by Eq.(19.35c).

Note that our prescription in Eq.(19.35) requires a “self-consistent evaluation” since $\Sigma$, $\Sigma^i_0$ depend on $G^R$ and $G^A$ which in turn depend on $\Sigma$, $\Sigma^i_0$ respectively (see Eqs.(19.1), (19.2)).

Also, Model A (Eq.(19.37)) requires us to calculate the full Green’s function which can be numerically challenging for large devices described by large matrices. Model B makes the computation numerically much more tractable because one only needs to calculate the diagonal elements of the Green’s functions which can be done much faster using powerful algorithms.

In these Lectures, however, we focus on conceptual issues using “toy” problems for which numerical issues are not the “show stoppers.” The important conceptual distinction between Models A and B is that the former destroys phase but not momentum, while the latter destroys momentum as well [Golizadeh-Mojarad et al. 2007].

The dephasing process can be viewed as extraction of the electron from a state described by $[G^o]$ and reinjecting it in a state described by $D \times G^o$. Model A is equivalent to multiplying $[G^o]$ by a constant so that the electron is reinjected in exactly the same state that it was extracted in, causing no loss of momentum, while Model B throws away the off-
diagonal elements and upon reinjection the electron is as likely to go on way or another. Hopefully this will get clearer in the next Lecture when we look at a concrete example.

Another question that the reader might raise is whether instead of including elastic dephasing through a self-energy function \( \Sigma_0 \) we could include a potential \( U_R \) in the Hamiltonian itself and then average over a number of random realizations of \( U_R \). The answer is that the two methods are not exactly equivalent though in some problems they could yield similar results. This too should be a little clearer in the next lecture when we look at a concrete example.

For completeness, let me note that in the most general case \( D_{ijkl} \) is a fourth order tensor and the version we are using (Eq.(19.35)) represents a special case for which \( D_{ijkl} \) is non-zero only if \( i=k, j=l \) (see Appendix E).
Lessons from Nanoelectronics

Lecture 20

Can Two Offer Less Resistance than One?

20.1. Modeling 1D Conductors
20.2. Quantum Resistors in Series
20.3. Potential Drop Across Scatterer(s)

In the next three Lectures we will go through a few examples of increasing complexity which are interesting in their own right but have been chosen primarily as “do it yourself” problems that the reader can use to get familiar with the quantum transport model outlined in the last Lecture. The MATLAB codes are all included in Appendix F.

In this Lecture we will use 1D quantum transport models to study an interesting question regarding multiple scatterers or obstacles along a conductor. Are we justified in neglecting all interference effects among them and assuming that electrons diffuse like classical particles as we do in the semiclassical picture?

This was the question Anderson raised in his 1958 paper entitled "Absence of Diffusion in Certain Random Lattices" pointing out that diffusion could be slowed significantly and even suppressed completely due to quantum interference between scatterers. “Anderson localization” is a vast topic and we are only using some related issues here to show how the NEGF model provides a convenient conceptual framework for studying interesting physics.

For any problem we need to discuss how we write down the Hamiltonian $[H]$ and the contact self-energy matrices $[\Sigma]$. Once we have these, the computational process is standard. The rest is about understanding and enjoying the physics.
20.1. Modeling 1D Conductors

For the one-dimensional examples discussed in this Lecture, we use the 1-D Hamiltonian from Fig.18.6, shown here in Fig.20.1. As we discussed earlier for a uniform wire the dispersion relation is given by

\[ E(k) = \varepsilon + 2t \cos ka \] (20.1a)

which can approximate a parabolic dispersion

\[ E = E_c + \frac{\hbar^2 k^2}{2m} \] (20.1b)

by choosing

\[ E_c = \varepsilon + 2t \] (20.2a)

and

\[ -t \equiv t_0 = \frac{\hbar^2}{2ma^2} \] (20.2b)

It is straightforward to write down the \([H]\) matrix with \(\varepsilon\) on the diagonal and “t” on the upper and lower diagonals. What needs discussion are the self-energy matrices. The basic idea is to replace an infinite conductor described by the Hamiltonian \([H]\) with a finite conductor described by \([H + \Sigma_1 + \Sigma_2]\) assuming open boundary conditions at the ends, which means that electron waves escaping from the surface do not give rise to any reflected waves, as a good contact should ensure.

Fig.20.1. For the one-dimensional examples discussed in this Lecture, we use the 1-D Hamiltonian from Fig.18.6.
For a one-dimensional lattice the idea is easy to see. We start from the original equation for the extended system

\[ E\psi_n = t\psi_{n-1} + \epsilon\psi_n + t\psi_{n+1} \]

and then assume that the contact has no incoming wave, just an outgoing wave, so that we can write

\[ \psi_{n+1} = \psi_n e^{ika} \]

which gives

\[ E\psi_n = t\psi_{n-1} + (\epsilon + te^{ika})\psi_n \]

In other words the effect of the contact is simply to add \( t\exp(+ika) \) to \( H_\text{nn} \) which amounts to adding the self-energy

\[
\Sigma_1 = \begin{bmatrix}
t e^{ika} & 0 & 0 & \ldots \\
0 & 0 & 0 & \\
0 & 0 & 0 & \ldots \\
\ldots & \ldots & \ldots & \\
\end{bmatrix}
\]

(20.1a)

\[
\Sigma_2 = \begin{bmatrix}
\epsilon & t & 0 & 0 & \ldots \\
0 & 0 & 0 & \\
0 & 0 & 0 & \ldots \\
0 & 0 & 0 & \\
0 & 0 & 0 & t e^{ika} \\
\end{bmatrix}
\]

(20.1b)

Note the only non-zero element is the \((1,1)\) element. Similarly at the other contact we obtain

In short, the self-energy function for each contact has a single non-zero element corresponding to the point that is connected to that contact.
20.1.1. 1D ballistic conductor

A good test case for any theory of coherent quantum transport is the conductance function for a length of uniform ballistic conductor: If we are doing things right, the conductance function $G(E)$ should equal the quantum of conductance $q^2/h$ times an integer equal to the number of modes $M(E)$ which is one for 1D conductors (neglecting spin). This means that the transmission (see Eq.(19.30))

$$\bar{T}(E) = \text{Trace} \left[ \Gamma_1 G^R \Gamma_2 G^A \right]$$

(20.3)

should equal one over the energy range

$$0 < E - E_c < 4t_0$$

covered by the dispersion relation

$$E = \varepsilon + 2t \cos ka = E_c + 2t_0(1 - \cos ka)$$

(20.4)

but zero outside this range (see Fig.20.2 below with $U=0$). This is a relatively simple but good example to try to implement numerically when getting started. Obtaining a constant conductance across the entire band is usually a good indicator that the correct self-energy functions are being used and things have been properly set up.

Fig.20.2. Transmission through a single point scatterer in a 1D wire.
20.1.2. 1D conductor with one scatterer

Another good example is that of a conductor with just one scatterer whose effect is included in the Hamiltonian \([H]\) by changing the diagonal element corresponding to that point to \(\epsilon + U\):

\[
H = \begin{pmatrix}
\epsilon & t & 0 \\
\vdots & \epsilon + U & t \\
0 & t & \epsilon \\
\vdots & \epsilon & \ddots
\end{pmatrix}
\]

Fig.20.2 shows the numerical results for \(U=0\) (ballistic conductor) and for \(U = 2t_0\). Actually there is a simple analytical expression for the transmission through a single point scatterer

\[
\bar{T}(E) = \frac{(2t\sin ka)^2}{U^2 + (2t\sin ka)^2} = \frac{(h\nu/a)^2}{U^2 + (h\nu/a)^2}
\]

that we can use to check our numerical results. This expression is obtained by treating the single point where the scatterer is located as the channel, so that all matrices in the NEGF method are \((1x1)\) matrices, that is, just numbers:
It is easy to see that the Green’s function is given by

\[ G^R(E) = \frac{1}{E - (\varepsilon + U) - 2t e^{i\kappa a}} = \frac{1}{-U - i2t \sin \kappa a} \]

making use of Eq.(20.2). Hence

\[ \Gamma_1 G^R \Gamma_2 G^A = \frac{(2t \sin \kappa a)^2}{U^2 + (2t \sin \kappa a)^2} \]

giving us the stated result in Eq.(20.3). The second form is obtained by noting from Eq.(20.2) that

\[ h\nu = \frac{dE}{dk} = -2at \sin \kappa a \]  

(20.6)

Once you are comfortable with the results in Fig.20.2 and are able to reproduce it, you should be ready to include various potentials into the Hamiltonian and reproduce the rest of the examples in this Lecture.

**20.2. Quantum Resistors in Series**

In Lecture 12 we argued that the resistance of a conductor with one scatterer with a transmission probability \( T \) can be divided into a scatterer resistance and an interface resistance (see Eqs.(12.1), (12.2))

\[ R_i = \frac{h}{q^2 M} \left( \frac{1 - T}{T_{scatter}} + \frac{1}{T_{interface}} \right) \]

What is the resistance if we have two scatterers each with transmission \( T \)?
Lessons from Nanoelectronics

Fig. 20.3.
Channel with (a) one scatterer, and (b) two scatterers.

We would expect the scatterer contribution to double:

\[ R_2 = \frac{h}{q^2 M} \left\{ 2 \frac{1 - T}{T_{\text{scatterer}}} + \frac{1}{T_{\text{interface}}} \right\} \]

\[ = \frac{h}{q^2 M} \frac{2 - T}{T} \]

We can relate the two resistances by the relation:

\[ R_2 = R_1 (2 - T) \]

If \( T \) is close to one we have the ballistic limit with \( R_2 = R_1 \): two sections in series have the same resistance as one of them, since all the resistance comes from the interfaces.

If \( T \ll 1 \), we have the Ohmic limit with \( R_2 = 2R_1 \): two sections have twice the resistance as one of them, since all the resistance comes from the channel.
But can \( R_2 \) ever be less than \( R_1 \)? Not as long as electrons behave like classical particles. There is no way an extra roadblock on a classical highway can increase the traffic flow. But on a quantum highway this is quite possible due to wave interference.

We could use our 1D model to study problems of this type. Fig. 20.4 shows the transmission functions \( \bar{T}(E) \) calculated numerically for a conductor with one scatterer and a conductor with two scatterers.

![Diagram showing transmission functions \( \bar{T}(E) \) for a conductor with one scatterer and a conductor with two scatterers.]

If the electrochemical potential happens to lie at an energy like the one marked "B", \( R_2 \) will be even larger than the Ohmic result \( R_1 \). But if the electrochemical potential lies at an energy like the one marked "A", \( R_2 \) is less than \( R_1 \).

At such energies, the presence of the second scatterer creates a reflection that cancels the reflection from the first one, because they are spaced a quarter wavelength apart. Such quarter wave sections are widely used to create anti-reflection coatings on optical lenses and are well-known in the world of waves, though they are unnatural in the world of particles.
Actually there is a class of devices called resonant tunneling diodes that deliberately engineer two strategically spaced barriers and make use of the resulting sharp peaks in conductance to achieve interesting current-voltage characteristics like the one sketched here where over a range of voltages, the slope $dI/dV$ is negative (“negative differential resistance, NDR”). We could use our elastic resistor model for the current from Eq.(3.3) and along with the conductance function from NEGF

$$G(E) = \frac{q^2}{h} \mathcal{T}(E) = \frac{q^2}{h} \text{Trace} \left[ \Gamma_1 G R \Gamma_2 G_A \right]$$

to model devices like this, but it is important to include the effect of the applied electric field on the $[H]$ as mentioned earlier (see Fig.18.6). In these Lectures we will focus more on low bias response for which this aspect can be ignored.

Consider for example a resistor with scatterers distributed randomly throughout the channel. If we were to use the quantum formalism to calculate the conductance function for a single-moded wire with random scatterers we would find that once the classical transmission $M \lambda / L$ drops below one, the quantum conductance is extremely low except for occasional peaks at specific energies (Fig.20.5). The result marked semiclassical is obtained by calculating $T$ for a single scatterer and then increasing the scatterer contribution by a factor of six:

$$R_6 = \frac{h}{q^2 M} \left( 6 \frac{1 - T}{T_{\text{scattering}}} + 1 \frac{1}{T_{\text{interface}}} \right) = \frac{h}{q^2 M} \frac{6 - 5T}{T}$$
Comparing the classical and quantum results suggests that such conductors would generally show very high resistances well in excess of Ohm's law, with occasional wild fluctuations. In a multi-moded wire too quantum calculations show the same behavior once the classical transmission $M \lambda / L$ drops below one. Such conductors are often referred to as being in the regime of strong localization. Interestingly, even when $M \lambda / L$ is well in excess of one, the quantum conductance is a little (~ approximately one) less than the classical value and this is often referred to as the regime of weak localization.

However, localization effects like these are usually seen experimentally only at low temperatures. At room temperature there is seldom any evidence of deviation from Ohm's law. Consider for instance a copper wire with a cross-section of $10nm \times 10nm$ which should contain approximately 1000 atoms and hence should have $M \sim 1000$ (see discussion at end of Chapter 4). Assuming a mean free path of $40 nm$ this suggests that a copper wire any longer than $M \lambda \sim 40 \mu m$ should exhibit strange non-Ohmic behavior, for which there is no experimental evidence. Why?
The answer is that localization effects arise from quantum interference and will be observed only if the entire conductor is phase-coherent. A copper wire 40 µm long is not phase coherent, certainly not at room temperature. Conceptually we can think of the real conductor as a series of individual coherent conductors, each of length equal to the phase coherence length $L_P$ and whether we see localization effects will depend not on $M\lambda/L$, but on $M\lambda/L_P$.

The bottom line is that to describe real world experiments especially at room temperature it is often important to include a certain degree of dephasing processes as described at the end of the last Lecture. Unless we include an appropriate degree of dephasing our quantum models will show interference effects leading to resonant tunneling or strong localization which under certain conditions may represent real world experiments, but not always. Just because we are using quantum mechanics, the answer is not automatically more "correct."

This can be appreciated by looking at the potential variation along the channel using NEGF and comparing the results to our semiclassical discussion from Lecture 12.

20.3. Potential Drop across Scatterer(s)

In Lecture 12 we discussed the spatial variation of the occupation factor which translates to a variation of the electrochemical potential for low bias. A conductor with one scatterer in it (Fig.20.6), can be viewed (see Fig.12.5) as a normalized interface resistance of one in series with a normalized scatterer resistance of $(1-T)/T$, which can be written as

$$(Normalised) \quad R_{scatterer} = \left(\frac{Ua}{\hbar}\right)^2 \quad (20.7)$$

using Eq.(20.5). The semiclassical potential profile in Fig.20.6 is then obtained by noting that since the current is the same everywhere, each section shows a potential drop proportional to its resistance.
The quantum profile is obtained using an NEGF model to calculate the effective occupation factor throughout the channel by looking at the ratio of the diagonal elements of $G^n$ and $A$ which are the quantum versions of the electron density and density of states respectively:

$$f(j) = \frac{G^n(j,j)}{A(j,j)}$$  \hspace{1cm} (20.8)

For low bias, this quantity translates linearly into a local electrochemical potential as noted in Lecture 2 (see Eq.(2.8)). If we choose $f = 0$ at one contact, $f=1$ at another contact corresponding to $qV$, then the $f(j)$ obtained from Eq.(20.8) is simply translated into an electrochemical potential $\mu$ at that point:

$$\mu(j) = qV f(j)$$  \hspace{1cm} (20.9)

The occupation $f(j)$ shows oscillations due to quantum interference making it hard to see the potential drop across the scatterer (see solid black line marked NEGF).

Fig.20.6. Potential drop across a scatterer calculated from the quantum formalism. (a) Physical structure, (b) Coherent NEGF calculation at $E = t_0$. 

Can Two Offer Less Resistance than One?
Experimentalists have measured profiles such as these using scanning probe microscopy (SPM) and typically at room temperature the quantum oscillations are not seen, because of the dephasing processes that are inevitably present at room temperature. This is another example of the need to include dephasing in order to model real world experiments especially at room temperature.

Indeed if we include pure phase relaxation processes (Eq.(19.37)) in the NEGF model we obtain a clean profile looking a lot like what we would expect from a semiclassical picture (see Fig.20.7a).

Fig.20.7. Potential drop for the structure in Fig.20.6 calculated from the NEGF method at $E=t_0$ with dephasing, (a) Phase-relaxation only, Eq.(19.37), (b) Phase and momentum relaxation, Eq.(19.38).

Interestingly, if we use a momentum relaxing model for $\Sigma_0$ (Eq.(19.38)), the potential drops linearly across the structure (see Fig.20.7b), exactly what we would expect for a distributed classical resistor. The resistance per lattice site for this distributed resistor due to $D_0$ can be obtained by replacing $U^2$ with $D_0$ in Eq.(20.7):

\[
\text{(Normalized)} \quad R = \left( \frac{a}{\hbar v} \right)^2 D_0 \frac{L}{\ell_{\text{lat}}} \]

where $\ell_{\text{lat}}$ is the number of lattice sites.
Another interesting example is that of the two quantum resistors in series that we started with. We noted then that at energies corresponding to points A and B in Fig.20.2 we have constructive and destructive interference respectively. This shows up clearly in the potential profile for coherent transport with $D_0 = 0$ (see Fig.20.8). At $E = 0.6t_0$ corresponding to destructive interference, the profile looks like what we might expect for a very large resistor showing a large drop in potential around it, along with some sharp spikes superposed on it. At $E = 0.81t_0$ corresponding to constructive interference, the profile looks like what we expect for a ballistic conductor with all the drop occurring at the two contacts and none across the scatterers.

![Potential drop across two scatterers in series calculated from the NEGF method without dephasing at two energies, $E = 0.81t_0$ and $E = 0.6t_0$ corresponding to points marked “A” and “B” respectively in Fig.20.4.](image)

Clearly at $E = 0.81t_0$ the answer to the title question of this Lecture is yes, two scatterers can offer less resistance than one. And this strange result is made possible by quantum interference. And once we introduce sufficient phase relaxation into the model using a non-zero $D_0$, the profile at both energies look much the same like any semiclassical resistor (Fig.20.9).
Before we move on, let me note that although it is straightforward to include dephasing into toy calculations like this, for large devices described by large matrices, it can be numerically challenging. This is because with coherent NEGF ($D_0 = 0$) or with the momentum relaxing model (Eq.(19.38)), it is often adequate to calculate just the diagonal elements of the Green’s functions using efficient algorithms. But for pure phase relaxation (Eq.(19.37), it is necessary to calculate the full Green’s function increasing both computational and memory burdens significantly.

![Phase Relaxation Only: $D_0 = 0.09 t_0^2$](image)

*Fig. 20.9. Potential drop across two scatterers in series calculated from the NEGF method with pure phase relaxation at two energies, $E = 0.81 t_0$ and $E = 0.6 t_0$ corresponding to points marked “A” and “B” respectively in Fig.20.4.*

So a natural question to ask is whether instead of including dephasing through $\Sigma_0$ we could include the potential $U_R$ in the Hamiltonian itself and then average our quantity of interest over a number of random realizations of $U_R$. Would these results be equivalent?

For short conductors like the one shown in Fig.20.4, this seems to be true, but for long conductors like the one in Fig.20.5 this may not be true. With a conductor in the regime of strong localization (Fig.20.5) it is hard to see how averaging the coherent quantum result over many configurations can lead to the semiclassical result.
NEGF with dephasing does not just average over many configurations, it also averages over different sections of the same configuration and that is why it is able to capture the semiclassical result which often describes real world experiments at room temperature quite well.

But could NEGF capture the localization effects observed at low temperatures through a proper choice of $\Sigma_0$? We believe so, but it would involve going beyond the simple dephasing models (technically known as the self-consistent Born approximation) for evaluating $\Sigma_0$ described in Section 19.4.

One last comment before we move on. Note that our discussion of the potential profile has focused entirely on the electrochemical potential based on the transport equation, rather than the electrostatic potential obtained by self-consistently solving the Poisson equation with appropriate corrections if needed.

The reason is that the electrostatic potential is approximately given by the electrochemical potential convolved with a screening function which is very sharp in metals, but more diffuse in semiconductors. It thus does not add much to our understanding of potential profiles, though a detailed consideration may be important in interpreting specific experiments.

These issues were discussed at length in McLennan et al. (1991) referred in the last lecture and also in Chapter 2 of Datta (1995).
Lecture 21

Quantum of Conductance

21.1. 2D Conductor as 1D Conductors in Parallel
21.2. Contact self-energy for 2D Conductors
21.3. Quantum Hall Effect

As I mentioned our primary objective in Lectures 20-22 is to help the reader get familiar with the NEGF model through “do it yourself” examples of increasing complexity. The last Lecture used 1D examples. In this Lecture we look at 2D examples which illustrate one of the key results of mesoscopic physics, namely the observation of conductances that are an integer multiple of the conductance quantum $q^2/h$.

21.1. 2D Conductor as 1D Conductors in Parallel

Among the seminal experiments from the 1980’s that gave birth to mesoscopic physics was the observation that the conductance of a ballistic 2D conductor went down in integer multiples of $2q^2/h$ as the width of the narrow region was decreased.

To understand this class of experiments we need a 2D model (Fig.21.1). As with 1D, two inputs are required: the Hamiltonian $[H]$ and the contact self-energy matrices $[\Sigma]$. Once we have these, the rest is standard.

For $[H]$, we use the 2-D Hamiltonian from Fig.18.7 for conductors described by parabolic $E(k)$ relations. As we discussed earlier for a uniform wire the dispersion relation is given by

$$E(k_x,k_y) = \varepsilon + 2t \cos k_x a + 2t \cos k_y a \quad (21.1a)$$

which can approximate a parabolic dispersion
Quantum of Conductance

\[ E = E_c + \frac{\hbar^2 k^2}{2m} \]  
(21.1b)

by choosing \[ E_c = \epsilon + 4t \]  
(21.2a)

and \[ -t \equiv t_0 = \frac{\hbar^2}{2ma^2} \]  
(21.2b)

Fig. 21.1.
(a) Schematic of structure for measuring the conductance of a short constriction created in a two dimensional conductor. (b) 2D model used for NEGF-based calculation and discussion. (c) Numerically computed transmission shows steps as a function of energy.
Once again what needs discussion are the *self-energy matrices*, \([\Sigma]\), but before we get into it let us look at the transmission function

\[
\tilde{T}(E) = \text{Trace} \left[ \Gamma_1 G^R \Gamma_2 G^A \right] \quad \text{(same as Eq.(20.3))}
\]

obtained directly from the numerical model (Fig.21.1), which shows steps at specific energies. How can we understand this?

The elementary explanation from Section 5.5 is that for a ballistic conductor the transmission function is just the number of modes \(M(E)\) which equals the number of half de Broglie wavelengths that fits into the width \(W\) of the conductor (\(\text{Int}(x)\) denotes the highest integer less than \(x\))

\[
M = \text{Int} \left( \frac{2W}{h/p} \right) = \text{Int} \left( \frac{2W}{h} \sqrt{\frac{2m(E-E_c)}{p^2}} \right)
\]

where we have used the parabolic relation \(E - E_c = \frac{p^2}{2m}\). To compare with our numerical results we should use the cosine dispersion relation.

Experimentally what is measured at low temperatures is \(M(E=\mu_0)\) and the steps are observed as the width is changed as first reported in van Wees et al. (1988) and Wharam et al. (1988). To compare with experimental plots, one could take a fixed energy \(E=t_0\) and plot the transmission as a function of the number of points along the width to get something like this.
Why does our numerical model show these steps? One way to see this is to note that our 2D model can be visualized as a linear 1D chain as shown in the adjoining figure where the individual elements $\alpha$ of the chain represent a column. For example if there are three sites to each column, we would have

$$\alpha = \begin{bmatrix} \varepsilon & t & 0 \\ t & \varepsilon & t \\ 0 & t & \varepsilon \end{bmatrix} \quad (21.3a)$$

while the coupling $\beta$ from one column to the next is diagonal:

$$\beta = \begin{bmatrix} t & 0 & 0 \\ 0 & t & 0 \\ 0 & 0 & t \end{bmatrix} \quad (21.3b)$$

Note that the matrix $\alpha$ describing each column has off-diagonal elements $t$, but we can eliminate these by performing a basis transformation to diagonalize it:

$$[\tilde{\alpha}] = [V]^+ [\alpha] [V] \quad (21.3c)$$

where $[V]$ is a matrix whose columns represent the eigenvectors of $\alpha$.

Since the matrix $\beta$ is essentially an identity matrix it is unaffected by the basis transformation, so that in this transformed basis we can visualize the 2D conductors as a set of independent 1D conductors, each of which has a different diagonal element $\varepsilon_1, \varepsilon_2, \varepsilon_3$. 
equal to the eigenvalues of $\alpha$. Each of these 1D conductors has a transmission of one in the energy range ($t_0 = |f|$)

$$
\epsilon_n - 2t_0 < E < \epsilon_n + 2t_0
$$
as sketched below. Adding all the individual transmissions we obtain the transmission showing up-steps in the lower part and down-steps in the upper part.

Usually when modeling n-type conductors we use the lower part of the band as shown in Fig.21.1, and so we see only the up-steps occurring at

$$
\epsilon_n - 2t_0
$$

Now the $\epsilon_n$’s are the eigenvalues of $\alpha$ (see Eq.(21.3a)) which are given by

$$
\epsilon_n = \varepsilon - 2t_0 \cos k_n a, \quad \text{with} \quad k_n a = \frac{n\pi}{N+1} \quad (21.4)
$$

where $N$ is the number of points along the width which determines the size of $\alpha$. This result is not obvious, but can be shown analytically or checked easily using MATLAB.
Using Eq.(21.2) and (21.4) we can write the location of the steps as

$$
\varepsilon_n - 2t_0 = E_c + 2t_0 \left( 1 - \cos \frac{n\pi}{N+1} \right)
$$

which matches the numerical result obtained with $N=25$ very well as shown.

21.1.1. Modes or Subbands

The approach we just described of viewing a 2D (or 3D) conductor as a set of 1D conductors in parallel is a very powerful one. Each of these 1D conductors is called a mode (or subband) and has a dispersion relation

$$
E_n(k_x) = \varepsilon_n - 2t_0 \cos k_x a
$$

as shown below. These are often called the subband dispersion relations obtained from the general dispersion relation in Eq.(21.1a) by requiring $k_y$ to take on quantized values given by

$$
k_y a = \frac{n\pi}{N+1}
$$

where each integer $n$ gives rise to one subband as shown. If we draw a horizontal line at any energy $E$, then the number of dispersion relations it crosses is equal to twice the number of modes $M(E)$ at that energy, since
each mode gives rise to two crossings, one for a state with positive velocity, and one for a state with negative velocity.

\[
\frac{E_n(k_x) - E_c}{t_0}
\]

21.2. Contact Self-Energy for 2D Conductors

Let us now address the question we put off, namely how do we write the self-energy matrices for the contacts. Ideally the contact regions allow electrons to exit without any reflection and with this in mind, a simple way to evaluate \([\Sigma]\) is to assume the contacts to be just uniform extensions of the channel region and that is what we will do here.

21.2.1. Method of basis transformation

The viewpoint we just discussed in Section 21.1 allows us to picture a 2D conductor as a set of decoupled 1D conductors, by converting from the usual lattice basis to an abstract mode basis through a basis transformation:

\[
\begin{bmatrix}
\tilde{X} \\
\end{bmatrix}_{\text{Mode Basis}} = [V]^+ \begin{bmatrix}
X \\
\end{bmatrix}_{\text{Lattice Basis}} [V] \quad (21.5a)
\]
being any matrix in the regular lattice basis. A unitary transformation like this can be reversed by transforming back:

\[
\begin{bmatrix}
X
\end{bmatrix}_{\text{Lattice Basis}} = \begin{bmatrix}
V
\end{bmatrix}_{\text{Mode Basis}} \begin{bmatrix}
\tilde{X}
\end{bmatrix}_{\text{Lattice Basis}} \begin{bmatrix}
V
\end{bmatrix}^+_{\text{Mode Basis}}
\] (21.5b)

In our present problem we can easily write down the self-energy in the mode basis for each independent 1D wire and then connect them together.

\[E_n = \varepsilon_n - 2t_0 \cos k_n a\]

so that overall we could write
and then transform it back to the lattice basis as indicated in Eq.(21.5b):

\[
[\Sigma_1] = [V] [\tilde{\Sigma}_1] [V]^+
\]

21.2.2. General Method

The method of basis transformation is based on a physical picture that is very powerful and appealing. However, I believe it cannot always be used at least not as straightforwardly, since in general it may not be possible to diagonalize both $\alpha$ and $\beta$ simultaneously.

For the square lattice $\beta = t [I]$ (Eq.(21.3b)) making it “immune” to basis transformations. But in general this may not be so. The graphene lattice from Fig.18.9 pictured on the right is a good example. How do we write $[\Sigma]$ in such cases?

Any conductor with a uniform cross-section can be visualized as a linear 1-D chain of “atoms” each having an on-site matrix Hamiltonian $[\alpha]$ coupled to the next “atom” by a matrix $[\beta]$. Each of these matrices is of size (nxn), $n$ being the number of basis functions describing each unit.
The *self-energy matrix* is zero except for the last (nxn) block at the surface

\[ \Sigma_2(E) = \begin{bmatrix}
  \alpha & \beta & \cdots & \beta \\
  \beta & \alpha & \cdots & \beta \\
  \vdots & \vdots & \ddots & \vdots \\
  \beta & \beta & \cdots & \beta \\
\end{bmatrix} \]

The non-zero block is given by

\[ \beta g_2 \beta^+ \]  

where \( g_2 \) is called the surface Green’s function for contact 2, and is obtained by iteratively solving the equation:

\[ [g_2]^{-1} = (E + i0^+)I - \alpha - \beta^+ g_2 \beta \]  

for \( g_2 \), where \( 0^+ \) represents a positive infinitesimal. Eq.(21.6) is of course not meant to be obvious, but we have relegated the derivation to Appendix E. We will not go into the significance of the infinitesimal \( i0^+ \) (see for example Datta (1995), Chapter 3 or Datta (2005), Chapter 8).

For the moment let me just note that for a 1-D conductor with

\[ \alpha = \epsilon \quad \text{and} \quad \beta = t, \]

Eq.(21.2) reduces to an ordinary quadratic equation:

\[ g_2(E + i0^+ - \epsilon - t^2 g_2) = 1 \]

whose solution gives two possible solutions \( i e^{\pm i k a} \) for the self-energy, and the one we want is that with the *negative imaginary part*, for which the corresponding broadening \( \Gamma \) is positive. More generally, we have a matrix quadratic equation (Eq.(21.6b)) and the infinitesimal \( i0^+ \) ensures that a numerical iterative solution converges on the solution for which \( \Gamma \) has all positive eigenvalues.
21.2.3. Graphene: Ballistic Conductance

As an example we have shown in Fig.21.3 the transmission \( \bar{T}(E) \) calculated numerically for two common orientations of graphene, the so-called zigzag and armchair configurations with dimensions chosen so as to have roughly equal widths. Since these are ballistic conductors, the transmission is equal to the number of modes \( M(E) \) and can be approximately described by the number of wavelengths that fit into the widths. The actual energy dependence is different from that obtained for the square lattice (see Eq.(21.3)) because of the linear \( E(k) \) relation:

\[
E = \hbar \nu_0 k = \nu_0 p
\]

This applies equally to any orientation of graphene. Both the orientations shown have the same overall slope, but the details are quite different. For example, at \( E = 0 \), the armchair is non-conducting with \( M=0 \) while the zigzag is conducting with non-zero \( M \).

For large dimensions the steps are close together in energy (compared to \( kT \)) and both appear to be semi-metallic. But for small dimensions the steps are much larger than \( kT \). The zigzag now shows zero transmission \( \bar{T}(E)=0 \) at \( E=0 \) (“semiconducting”) while the armchair shows non-zero conductance (“metallic”). These are clear observable differences that show up in experiments on samples of small width at low temperatures.

Another interesting observable difference is that between a flat graphene sheet and a cylindrical carbon nanotube (CNT). Mathematically, they are both described by the same Hamiltonian \( [H] \) but with different boundary conditions. Graphene like most normal conductors requires “hardwall boundary conditions” (HBC) where the lattice ends abruptly at the edges. CNT’s on the other hand are among the few real conductors that require “periodic boundary conditions” (PBC) with no edges.
Fig. 21.3. $\bar{T}(E)$ calculated from NEGF-based model for a ballistic (a) graphene sheet with armchair and zigzag edges as shown with roughly equal widths ($24*2b \approx 14*3a$), (b) carbon nanotube (CNT) obtained by rolling up the graphene sheet along the width.
The results for CNT are relatively easy to understand analytically, while those for graphene require a more extensive discussion. (see for example Brey and Fertig, 2006). As we mentioned in Lecture 5, PBC is mathematically simpler and that is why it is used so extensively for large conductors where it is known experimentally that the exact boundary conditions are not very relevant. But this of course is not true of small conductors and the difference is evident in Fig.21.2 for small conductors only a few nanometers in width. We will not go into this further. Our objective here is simply to show how easily our quantum transport formalism captures all the known physics.

The power of the numerical method lies in being able to calculate $M(E)$ automatically even before one has “understood” the results. However, one should use numerical calculations not as a substitute for understanding, but as an aid to understanding.

21.3. Quantum Hall Effect

The Hall effect (Lecture 13) provides another good example for a two-dimensional application of the quantum transport model. The basic structure involves a long conductor with side probes designed to measure the transverse Hall voltage developed in the presence of a magnetic field.

We use the same 2D Hamiltonian from Fig.21.1 but now including a magnetic field as explained in Section 18.4.3. As discussed in Lecture 13, the Hall resistance is given by the ratio of the Hall voltage to the current. In a theoretical model we could calculate the Hall voltage in one of two ways. We could attach a voltage probe to each side and use Büttiker’s multiterminal method to find the potentials they float to.

Alternatively we could do what we explained in Section 20.3, namely calculate the fractional occupation of the states at any point $j$ by looking at the ratio of the diagonal element of the electron density $G^<$ and the density of states $A$ and use the low bias Taylor expansion (Eq.(2.8)) to translate the occupation factor profile into a potential profile.
Fig. 21.4 shows the calculated Hall resistance (normalized to the quantum of resistance $h/q^2$) as a function of the magnetic field. The striking result is of course the occurrence of plateaus at high fields known as the quantum Hall effect (von Klitzing et al. 1980). But first let us note the low field regime where the calculated result matches the Hall resistance expected from semiclassical theory

$$R_H = B/qn$$  \hspace{1cm} \text{(same as Eq. (13.3))}$$

The dashed line in Fig. 21.3 is obtained from Eq. (13.4) assuming
and noting that the numerical calculation is carried out at $E = t$, corresponding to $ka = \pi / 3$, with $a = 2.5 \text{ nm}$.

The semiclassical theory naturally misses the high field results which arise from the formation of Landau levels due to quantum effects. These are evident in the numerical plots of the local density of states at high $B$-field (20 T) shown in Fig.21.5.

Fig.21.5. Grayscale plot of local density of states, obtained from the diagonal elements of $[A(E)]$ calculated at $B=20T$ from the NEGF method. Also shown on the right are the diagonal elements of $[G_n(E)]$ calculated assuming $f_1 = 1$, $f_2 = 0$.

Usually the density of states varies relatively gently with position, but in the quantum Hall regime, there is a non-trivial modification of the local density of states which can be plotted from the NEGF method by looking at the diagonal elements of the spectral function $A(j,j;E)$. Fig.21.5 is a grayscale plot of $A(j,j;E)$ with energy $E$ on the horizontal axis and the position $j$ along the width on the vertical axis. The white streaks indicate a high density of states corresponding to the energy of Landau levels,
which increase in energy along the edge forming what are called edge states.

As we mentioned in Lecture 13, the edge states can be pictured semiclassically in terms of “skipping orbits” that effectively isolate oppositely moving electrons from each other giving rise to a “divided highway” that provides an incredible level of ballisticity. This is evident if we plot the electron density from the diagonal elements of $[G]$ under non-equilibrium conditions assuming $f_1 = 1, f_2 = 0$. Only the edge states on one side of the sample are occupied. If we reverse the current flow assuming $f_1 = 0, f_2 = 1$, we would find the edge states on the other side of the conductor occupied.

The energies of the Landau levels are given by

$$E_n = \left( n + \frac{1}{2} \right) \hbar \omega_c$$  \hspace{1cm} (21.8)

where $n$ is an integer, $\omega_c$ being the cyclotron frequency (see Eq.(13.7)). We expect the streaks to be spaced by

$$\hbar \omega_c = \frac{\hbar qB}{m} = \frac{2qBa^2}{\hbar} - t_0$$

where we have made use of Eq.(13.7) and (21.2b). Noting that $B=20T$, $a=2.5$ nm, we expect a spacing of $\sim 0.37t_0$ between the streaks in approximate agreement with Fig.21.5.

Eq.(21.8) is a quantum result that comes out of the Schrödinger equation including the vector potential which is part of our numerical model. One can understand it heuristically by noting that semiclassically electrons describe circular orbits in a magnetic field, completing one orbit in a time (see Eq.(13.7))
Lessons from Nanoelectronics

\[ t_c = \frac{2\pi}{\omega_c} = \frac{2\pi p}{qB} \]

so that the circumference of one orbit of radius \( r_c \) is given by

\[ 2\pi r_c = v t_c = \frac{2\pi p}{qB} \]

If we now impose the quantum requirement that the circumference equal an integer number of de Broglie wavelengths \( h/p \), we have

\[ \frac{2\pi p}{qB} = \text{integer} \cdot \frac{h}{p} \]

Semiclassically an electron can have any energy \( E = p^2/2m \). But the need to fit an integer number of wavelengths leads to the condition that

\[ p^2 = \text{integer} \cdot hqB \]

suggested that the allowed energies should be given by

\[ E = \text{integer} \cdot \frac{hqB}{2m} = \text{integer} \cdot \frac{h\omega_c}{2} \]

which is not exactly the correct answer (Eq.(21.8)), but close enough for a heuristic argument.

The resulting current equals

\[ \frac{q^2}{h} V \left( \text{Number of Edge States} \right) \]

while the Hall voltage simply equals the applied voltage since one edge of the sample is in equilibrium with the source and other with the drain.
This leads to a quantized Hall resistance given by

\[ \frac{h}{q^2} \frac{1}{\text{Number of Edge States}} \]

giving rise to the plateaus of 1/4, 1/3, 1/2, 1 seen in Fig.21.4, as the magnetic field raises the Landau levels, changing the number of edge states at an energy \( E=t_0 \) from 4 to 3 to 2 to 1.

I should mention that the theoretical model does not include the two spins and so gives a resistance that is twice as large as the experimentally observed values which look more like

\[ \frac{h}{2q^2} \frac{1}{\text{Number of Edge States}} \]

because edge states usually come in pairs, except at high \( B \)-fields.

Also, we have not talked at all about the fractional quantum Hall effect observed in pure samples at larger \( B \)-fields with Hall resistances that look like

\[ \frac{h}{q^2} \frac{1}{a \text{ fraction}} \]

This is a vast and rich area of research on its own beyond the scope of the simple NEGF model discussed here. As it stands it captures only the integer Hall effect though innovative extensions could take it beyond this regime.
Lessons from Nanoelectronics

Lecture 22

Rotating an electron

22.1. One-level Spin Valve
22.2. Rotating Magnetic Contacts
22.3. Spin Hamiltonians
22.4. Vectors and Spinors
22.5. Spin Precession
22.6. From NEGF to Diffusion

Back in Lecture 14, we discussed how magnets can be used to create spin potentials inside a non-magnetic channel that extend even outside the path of the current and can be measured using another magnetic voltage probe with a polarization $\vec{P}$ (Fig.22.1). This led naturally to the concept of a charge potential $\mu$ and a spin potential $\mu_s$ such that

$$\mu_{\text{probe}} = \mu + \frac{P \mu_s}{2}$$

At the end of the Lecture we stated the more general result

$$\mu_{\text{probe}} = \mu + \frac{\vec{P} \cdot \vec{\mu}_s}{2}$$  \hspace{1cm} (22.1)

that can be used even when the magnet polarization $\vec{P}$ and the spin potential $\vec{\mu}_s$ in the channel are not collinear.

I am not sure if anyone has actually done the experiment shown in Fig.22.1, namely inject spins with a fixed magnet and measure the voltage with a magnet whose direction is rotated. What has been done, however, is to keep both magnets fixed and rotate the electron spin inside
the channel using an external magnetic field and more recently, an electric field.

All of these effects are interesting in their own right, but here I will use them primarily to illustrate how spin effects are included in the NEGF method described in Lecture 19. We will end by showing how the NEGF method leads to the result stated in Eq.(22.1). As we will see the quantum method requires a (2x2) matrix potential which can also be described in terms of a charge potential $\mu$ supplemented by a vector spin potential $\vec{\mu}_s$ as indicated in Eq.(22.1).

As we mentioned in Lecture 14, electron spin is a lot like photon polarization, except for one key difference, namely that orthogonal directions are not represented by say $z$ and $x$ that are 90 degrees apart. Rather they are represented by up and down that are 180 degrees apart, which is why the maxima and minima of the oscillations in Fig.22.1 are separated by 180 degrees.

And that is why we need something other than vectors to represent electron spin, namely spinors. A vector $\hat{n}$ is described by three real components, namely the components along $x$, $y$ and $z$, but spinors are described by two complex components, which are its components along up and down:
Nevertheless we visualize the spinor as an object pointing in some direction just like a vector. How do we reconcile the visual picture with the 2-component complex representation?

A spinor pointing along a direction described by a unit vector

\[ \hat{n} = \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix} \]  

(22.2)

has components given by

\[ \begin{align*} 
\cos \frac{\theta}{2} e^{-i\phi/2} & \equiv c \\
\sin \frac{\theta}{2} e^{+i\phi/2} & \equiv s
\end{align*} \]  

(22.3)

This is of course not obvious and later in the Lecture I will try to explain why Eqs.(22.2) and (22.3) represent isomorphic (more correctly “homomorphic”) ways to represent an abstract rotatable object pointing in some direction. For the moment let us accept Eq.(22.3) for the components of a spinor and work out some of its consequences.

Although these subtleties of visualization and interpretation take some time to get used to, formally it is quite straightforward to incorporate spin into the quantum transport formalism from Lecture 19. The basic equations from Eq.(19.1) through (19.4) remain the same, but all the matrices like

\[ [H], \ [\Sigma], \ [G^n], \ [A] \]
become twice as big (Fig. 22.2).

Ordinarily these matrices are of size (N x N), if N is the number of grid points (or more formally the number of basis functions) used to describe the channel. Inclusion of spin basically doubles the number of basis functions: every grid point turns into two points, an up and a down (Fig. 22.2).

How would we write down \( [H] \) including spin? We can visualize the TB parameters (See Fig.18.7) exactly as before except that each on-site element \([\alpha]\) and the coupling elements \([\beta]\) are each (2x2) matrices (Fig. 22.3).

In the simplest case, we can imagine a “spin-innocent” channel that treats both spin components identically. Such a channel can be modeled by choosing the TB parameters as follows:

\[
\alpha = \epsilon [I] \\
\beta_x = t [I] , \; \beta_y = t [I] \quad (22.4)
\]
where $[I]$ is the (2x2) identity matrix. We effectively have two identical decoupled Hamiltonians that includes no new physics.

Similarly we can write the self-energy $[\Sigma]$ for ordinary contacts that treat both spin components identically simply by taking our usual values and multiplying by $[I]$. This would again be in the category of a trivial extension that introduces no new physics. The results should be the same as what we would get if we worked with one spin only and multiplied by two at the end.

All spin-related phenomena like the ones we discussed in Lecture 14 arise either from non-trivial contacts described by $[\Sigma]$ with spin-related properties or from channels described by $[H]$ with spin-related properties or both.

Let us now try to get a feeling for spin transport problems by applying the NEGF method to a series of examples, starting with a simple one-level version of the spin valve we started Lecture 14 with. From a computational point of view the only question is how to write down $[H]$, $[\Sigma]$. Once we have these, the rest is standard. One can then proceed to understand and enjoy the physics.

### 22.1. One-level Spin Valve

As we discussed in Lecture 14, a spin valve (Fig.22.3) shows different conductances $G_P$ and $G_{AP}$ depending on whether the magnetic contacts have parallel (P) or anti-parallel (AP) magnetizations. Using a simple model we showed in Lecture 14 that the magnetoresistance (MR) can be expressed as

$$MR \equiv \frac{G_P}{G_{AP}} - 1 = \frac{P^2}{1 - P^2}$$

where the polarization $P$ was defined in terms of the interface resistances. In that context we noted that the standard expression for the
MR for magnetic tunnel junctions (MTJ’s) has an extra factor of two (Eq.(14.3b))

\[ MR \equiv \frac{G_P}{G_{AP}} - 1 = \frac{2P^2}{1-P^2} \quad (22.5) \]

which could be understood if we postulated that the overall resistance was proportional to the product of the interface resistances and not their sum.

![Diagram](image1.png)

Fig.22.3. The spin-valve has different conductances \( G_P \) and \( G_{AP} \) depending on whether the magnetic contacts have parallel (P) or anti-parallel (AP) magnetization.

We could obtain this result (Eq.(22.5)) including the factor of two directly from our NEGF model if we apply it to a one-level resistor and assume that the equilibrium electrochemical potential \( \mu_0 \) is located many \( kT \)'s below the energy \( \varepsilon \) of the level as sketched.

![Diagram](image2.png)

Fig.22.4 summarizes the 2x2 matrices \([H], [\Sigma]\) for this device. Also shown for comparison are the corresponding 1x1 “matrices” (that is, just numbers) for the same device without spin. Note that the channel is
assumed to treat both spins identically so that \([H]\) is essentially an identity matrix, but the \([\Sigma]\)'s have different values for the up and downspin components.

Using these matrices it is straightforward to obtain the Green’s function

\[
\begin{bmatrix}
G^R \\
\end{bmatrix} = \begin{bmatrix}
E - \epsilon + \frac{i}{2}(\gamma_{1u} + \gamma_{2u}) & 0 \\
0 & E - \epsilon + \frac{i}{2}(\gamma_{1d} + \gamma_{2d})
\end{bmatrix}^{-1}
\]

and hence the transmission

\[
\bar{T} = \text{Trace} \Gamma_1 G^R \Gamma_2 G^A \\
= \frac{\gamma_{1u} \gamma_{2u}}{(E - \epsilon)^2 + \left(\frac{\gamma_{1u} + \gamma_{2u}}{2}\right)^2} + \frac{\gamma_{1d} \gamma_{2d}}{(E - \epsilon)^2 + \left(\frac{\gamma_{1d} + \gamma_{2d}}{2}\right)^2}
\]

Fig. 22.4. One-level spin-valve: Modifying the \([H], [\Sigma]\) for a spin-less one-level device to represent a one-level spin valve.
For the *parallel (P)* configuration we can assume both contacts to be identical so that we can write \((\alpha > \beta)\)

\[\gamma_{1u} = \gamma_{2u} = \alpha \]  
(22.6a)

\[\gamma_{1d} = \gamma_{2d} = \beta \]  
(22.6b)

while for the *anti-parallel (AP)* configuration the second contact has the roles of \(\alpha\) and \(\beta\) reversed with respect to the former:

\[\gamma_{1u} = \gamma_{2d} = \alpha \]  
(22.7a)

\[\gamma_{1d} = \gamma_{2u} = \beta \]  
(22.7b)

Inserting Eqs.(22.6a,b) into (22.5) we have the P- transmission

\[\overline{T}_P = \frac{\alpha^2}{(E - \epsilon)^2 + \alpha^2} + \frac{\beta^2}{(E - \epsilon)^2 + \beta^2} \]

while using Eqs.(22.7a,b) in Eq.(22.5) gives the AP- transmission

\[\overline{T}_{AP} = \frac{2\alpha\beta}{(E - \epsilon)^2 + \left(\frac{\alpha + \beta}{2}\right)^2} \]

The measured conductance depends on the average transmission over a range of energies of a few \(kT\) around \(\mu_0\). Assuming that

\[\epsilon - \mu_0 >> kT, \alpha, \beta \]

we can write

\[G_P \sim \overline{T}_P(E = \mu_0) = \frac{\alpha^2}{(\mu_0 - \epsilon)^2 + \alpha^2} + \frac{\beta^2}{(\mu_0 - \epsilon)^2 + \beta^2} = \frac{\alpha^2 + \beta^2}{(\mu_0 - \epsilon)^2} \]

and

\[G_{AP} \sim \overline{T}_{AP}(E = \mu_0) = \frac{2\alpha\beta}{(\mu_0 - \epsilon)^2} \]
This gives us

\[ MR \equiv \frac{G_p}{G_{Ap}} - 1 = \frac{\alpha^2 + \beta^2}{2\alpha\beta} - 1 = \frac{2P^2}{1 - P^2} \]

as stated earlier in Eq.(22.5) with the polarization defined as

\[ P = \frac{\alpha - \beta}{\alpha + \beta} \quad (22.8) \]

Actually we could also obtain the result (Eq.(14.3a)) obtained from the resistor model in Lecture 14, if we assume that \( \mu_0 \) is located right around the level \( \epsilon \), with \( kT \gg \alpha, \beta \). But we leave that as an exercise. After all this is just a toy problem intended to get us started.

### 22.2. Rotating Magnetic Contacts

We argued in the last Section that for an anti-parallel spin valve, the second contact simply has the roles of \( \alpha \) and \( \beta \) reversed relative to the first, so that we can write

\[ \Gamma_1 = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \quad \Gamma_2 = \begin{bmatrix} \beta & 0 \\ 0 & \alpha \end{bmatrix} \]

But how would we write the corresponding matrix for a contact if it were pointing along some arbitrary direction defined by a unit vector \( \hat{n} \)? The answer is

\[ \Gamma = \frac{\alpha + \beta}{2} [I] + \frac{\alpha - \beta}{2} \begin{bmatrix} n_z & n_x - in_y \\ n_x + in_y & -n_z \end{bmatrix} \quad (22.9) \]

where \( n_x, n_y, n_z \) are the components of the unit vector \( \hat{n} \) along \( x, y \) and \( z \) respectively. This result is of course not obvious and we will try to justify it shortly. But it is reassuring to note that the results for both the parallel and the anti-parallel contact come out as special cases of this general result (Eq.(22.9)).
Rotating an electron

If \( n_z = +1, n_x = n_y = 0 \): \( \Gamma = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \)

If \( n_z = -1, n_x = n_y = 0 \): \( \Gamma = \begin{bmatrix} \beta & 0 \\ 0 & \alpha \end{bmatrix} \)

One way to understand where Eq.(22.9) comes from is to note that the appropriate matrix describing a magnet pointing along \( \hat{n} \) would be

\[
\tilde{\Gamma} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}
\]

(22.10)

if we were to take \( \hat{n} \) and \( -\hat{n} \) as our reference directions instead of \( \hat{z} \) and \( -\hat{z} \) as we normally do. How could we then transform the \( \tilde{\Gamma} \) from Eq.(22.10) into the usual \( \pm \hat{z} \) basis?

Answer: Transform from the \( \pm \hat{n} \) to the \( \pm \hat{z} \) basis

\[
\begin{bmatrix} \hat{n} & -\hat{n} \\ \hat{z} & -\hat{z} \end{bmatrix} \begin{bmatrix} \hline c & -s^* \\ s & c^* \hline \end{bmatrix} \begin{bmatrix} \hat{n} \\ -\hat{n} \end{bmatrix} = \begin{bmatrix} \hat{z} & -\hat{z} \end{bmatrix} \begin{bmatrix} \alpha c^* & \alpha s^* \\ -s^* & c^* \end{bmatrix}
\]

(22.11)

using the unitary transformation matrix \( [V] \) whose columns represent the components of a spinor pointing along \( \pm \hat{n} \). The first column follows from the result we stated earlier in Eq.(22.3), while the second can be obtained from Eq.(22.3) if we set \( \theta \rightarrow \pi - \theta, \phi \rightarrow \pi + \phi \)

and remove a common phase factor from the two components.

Multiplying out the three matrices in Eq.(22.11) we have

\[
\Gamma = \begin{bmatrix} c & -s^* \\ s & c^* \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c^* & s^* \\ -s & c \end{bmatrix} = \begin{bmatrix} c & -s^* \\ s & c^* \end{bmatrix} \begin{bmatrix} \alpha c^* & \alpha s^* \\ -s^* & c^* \end{bmatrix} = \begin{bmatrix} \alpha c^* & \alpha s^* \\ -s^* & c^* \end{bmatrix} \begin{bmatrix} \beta s & \beta c \\ -\beta s & \beta c \end{bmatrix} \]
Lessons from Nanoelectronics

\[ \begin{bmatrix} \alpha cc^* + \beta ss^* & (\alpha - \beta) cs^* \\ (\alpha - \beta) sc^* & \alpha ss^* + \beta cc^* \end{bmatrix} \]

Making use of the definitions of \( c, s \) from Eq.(22.3) and some common trigonometric identities like

\[ 2\cos^2 \theta = 1 + \cos \theta, \quad 2\sin^2 \theta = 1 - \cos \theta, \]

\[ \text{and} \quad 2\sin \frac{\theta}{2} \cos \frac{\theta}{2} = \sin \theta \]

(22.12)

we can rewrite this as

\[ \Gamma = \frac{1}{2} \begin{bmatrix} (\alpha + \beta) + (\alpha - \beta) \cos \theta & (\alpha - \beta) \sin \theta e^{-i\phi} \\ (\alpha - \beta) \sin \theta e^{i\phi} & (\alpha + \beta) - (\alpha - \beta) \cos \theta \end{bmatrix} \]

which leads to the result stated earlier in Eq.(22.9) if we make use of Eq.(22.2) for the \( x, y, z \) components of a unit vector.

Finally let me note that if we define the polarization as a vector whose magnitude is given by Eq.(22.8) and direction is given by \( \hat{n} \):

\[ \vec{P} \equiv P \hat{n} = \frac{\alpha - \beta}{\alpha + \beta} \hat{n} \]

(22.13)

then we could rewrite Eq.(22.9) as

\[ \Gamma = \frac{\alpha + \beta}{2} \left[ I \right] + \begin{bmatrix} P_z \quad P_x - iP_y \\ P_x + iP_y \quad -P_z \end{bmatrix} \]

(22.14)

which can be rearranged as shown

\[ \frac{[\Gamma]}{(\alpha + \beta)/2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \]
Any 2x2 matrix can be expressed in terms of the four matrices appearing here consist of the identity matrix \([I]\) along with the three Pauli spin matrices

\[
\sigma_x = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \quad \text{and} \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}
\] (22.15)

which are widely used in the spin-related literature.

Making use of the Pauli spin matrices, we could write Eq.(22.14) compactly in the form

\[
\Gamma = \frac{\alpha + \beta}{2} \left( [I] + [\sigma_x]P_x + [\sigma_y]P_y + [\sigma_z]P_z \right)
\]

\[
= \frac{\alpha + \beta}{2} \left( [I] + [\sigma].\vec{P} \right)
\] (22.16)

This result applies to the self-energy matrices as well. For example, if

\[
\hat{\Sigma} = -\frac{i}{2} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}
\]

in the \(\pm \hat{n}\) basis, then in the \(\pm \hat{z}\) basis it is given by

\[
\Sigma = -i \frac{\alpha + \beta}{4} [I] - i \frac{\alpha - \beta}{4} [\sigma].\hat{n}
\]

\[
= -i \frac{\alpha + \beta}{4} \left( [I] + [\sigma].\vec{P} \right)
\]

22.3. Spin Hamiltonians

Now that we have seen how to describe contacts with spin-dependent properties, let us talk briefly about channels with spin-dependent properties.
22.3.1. Channel with Zeeman Splitting

The commonest example is the Zeeman splitting that causes the energies of the up-spin state to go up by $\mu_{el}B$ and that of the down spin states to go down by $\mu_{el}B$, $\mu_{el}$ being the effective magnetic moment of the electron discussed in Section 14.3.

If the magnetic field points along $+\hat{n}$, then in the $\pm \hat{n}$ basis the corresponding Hamiltonian should look like

$\mu_{el} \begin{bmatrix} +B & 0 \\ 0 & -B \end{bmatrix}$

Following our discussion in the last Section we can write it in the $\pm \hat{z}$ basis as

$H_B = \mu_{el} \hat{\sigma} \cdot B$

The overall Hamiltonian is obtained by adding this to the spin-independent part multiplied by $[I]$. For parabolic dispersion this gives

$H = \frac{\hbar^2}{2m} \left( k_x^2 + k_y^2 \right) [I] + \mu_{el} \hat{\sigma} \cdot B$ \hspace{1cm} (22.17)

while for a 2D square lattice we have (see Eq.(18.19))

$H = \left( \varepsilon + 2t \cos k_x a + 2t \cos k_y a \right) [I] + \mu_{el} \hat{\sigma} \cdot B$ \hspace{1cm} (22.18)

The corresponding parameters for the 2D lattice in Fig.22.3 (also shown here for convenience) are given simply by

$\alpha = \varepsilon [I] + \mu_{el} \hat{\sigma} \cdot \vec{B}$

$\beta_x = t [I] , \beta_y = t [I]$ \hspace{1cm} (22.19)
Only the on-site parameter $\alpha$ is changed relative to the spin independent channel (Eq.(22.4)).

### 22.3.2. Channel with Rashba Interaction

A more complicated example is that of the Rashba spin-orbit coupling described by a Hamiltonian of the form

$$H_R = \eta \hat{z} \cdot (\hat{\mathbf{\sigma}} \times \hat{k}) = \eta (\sigma_x k_y - \sigma_y k_x) \quad (22.20)$$

whose effect has been observed in 2D surface conduction channels assumed to lie in the $x$-$y$ plane. This is believed to be a relativistic effect whereby the extremely high atomic scale electric fields (that exist even at equilibrium) are perceived as an effective magnetic field by the electron and the resulting “Zeeman splitting” is described by $H_R$.

We will not go into the underlying physics of this effect any further here and simply address the question of how to include it in our 2D lattice model. With this in mind we approximate the linear terms with sine functions

$$H_R = \frac{\eta}{a} (\sigma_x \sin k_y a - \sigma_y \sin k_x a) \quad (22.21)$$

which are written in terms of exponentials:

$$H_R = \frac{\eta}{2ia} \sigma_x (e^{i k_y a} - e^{-i k_y a}) - \frac{\eta}{2ia} \sigma_y (e^{i k_x a} - e^{-i k_x a})$$

Clearly $H_R$ can be described by a Hamiltonian with

$$\beta_x = \frac{i \eta}{2a} \sigma_y, \quad \beta_x^+ = -\frac{i \eta}{2a} \sigma_y$$

$$\beta_y = -\frac{i \eta}{2a} \sigma_x, \quad \beta_y^+ = \frac{i \eta}{2a} \sigma_x$$
in order to ensure that if we write down the dispersion relation for the lattice we will indeed get back the original result in Eq.(22.17). Adding this to the usual spin-independent part from Eq.(22.4) along with any real magnetic field $B$ we have the overall parameters:

$$\alpha = \varepsilon I + \mu_{el} \bar{\sigma} \cdot \bar{B}$$

$$\beta_x = t I + \frac{i \eta}{2a} \sigma_y, \quad \beta_x^+ = t I - \frac{i \eta}{2a} \sigma_y$$

$$\beta_y = t I - \frac{i \eta}{2a} \sigma_x, \quad \beta_y^+ = t I + \frac{i \eta}{2a} \sigma_x \quad (22.22)$$

### 22.4. Vectors and Spinors

One of the important subtleties that takes some time to get used to is that we represent spin with two complex components, but we visualize it as a rotatable object pointing in some direction, which we have learnt to represent with a vector having three real components. To see the connection between the spinor and the vector, it is instructive to consider the precession of a spin in a magnetic field from both points of view.

Consider the one-level device with $\varepsilon = 0$, and with a magnetic field in the $z$-direction so that the Schrödinger equation can be written as

$$\frac{d}{dt} \begin{pmatrix} \psi_u \\ \psi_d \end{pmatrix} = \frac{\mu_{el} B_z}{i \hbar} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \psi_u \\ \psi_d \end{pmatrix} \quad (22.23)$$

These are two separate differential equations whose solution is easily written down:

$$\psi_u(t) = \psi_u(0) e^{-i \omega t/2}$$

$$\psi_d(t) = \psi_d(0) e^{+i \omega t/2}$$
where
\[ \omega = \frac{2 \mu_B B_z}{\hbar} \quad (22.24) \]

So if the electron starts out at some angle \((\theta, \phi)\) with a wavefunction
\[
\begin{align*}
\psi_u(0) &= \cos \frac{\theta}{2} e^{-i\phi/2} \\
\psi_d(0) &= \sin \frac{\theta}{2} e^{+i\phi/2}
\end{align*}
\]
at \(t = 0\), then at a later time it will have a wavefunction given by
\[
\begin{align*}
\psi_u(t) &= \cos \frac{\theta}{2} e^{-i\phi/2} e^{-i \omega t/2} \\
\psi_d(t) &= \sin \frac{\theta}{2} e^{+i\phi/2} e^{+i \omega t/2}
\end{align*}
\]
which means that the spin will be rotating around the \(z\)-axis such that the angle \(\theta\) remains fixed while the angle \(\phi\) increases linearly with time:
\[ \phi(t) = \phi(0) + \omega t \quad (22.25) \]

Making use of Eq.(22.2) for the \(x, y\) and \(z\) components of the vector \(\hat{n}\) we can write
\[
\begin{align*}
n_x &= \sin \theta \cos \phi(t) \\
n_y &= \sin \theta \sin \phi(t)
\end{align*}
\]
and
\[ n_z = \cos \theta \quad (22.26) \]

From Eqs.(22.25) and (22.26) we can show that
\[
\begin{align*}
\frac{dn_x}{dt} &= -\omega n_y \\
\frac{dn_y}{dt} &= +\omega n_x
\end{align*}
\]
which can be written in matrix form
For comparison we have rewritten the Schrödinger equation we started with (see Eq.(22.23)) in terms of the rotation frequency $\omega$:

$$\frac{d}{dt} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \omega \begin{bmatrix} 0 & -1 & 0 \\ +1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix}$$  \hspace{1cm} (22.27a)

If we wanted to describe the rotation of an electron due to a $B$-field pointing in the $x$-direction, it is easy to see how we would modify Eq.(22.27a): Simply interchange the coordinates, $x \rightarrow y, y \rightarrow z, z \rightarrow x$.

$$\frac{d}{dt} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & +1 & 0 \end{bmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix}$$  \hspace{1cm} (22.27b)

and we obtain $R_x$ in place of $R_z$. But it is not as clear how to modify Eq.(22.27b). The correct answer is to replace $\sigma_z$ with $\sigma_x$ (Eq.(22.15))

$$\frac{d}{dt} \begin{bmatrix} \psi_u \\ \psi_d \end{bmatrix} = \frac{\omega}{2i} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \psi_u \\ \psi_d \end{bmatrix}$$

but the reason is not as obvious.

Eqs.(22.27a) and (22.27b) both describe the same physics, namely the rotation of a spin about the $z$-axis due to an applied $B$-field in the $z$-direction, one in terms of three real components and the other in terms of two complex components.
Rotating an electron

But what do matrices like $R$ in Eq.(22.27a) have in common with matrices like $\sigma$ in Eq.(22.27b) that makes them “isomorphic” allowing them to describe the same physics? Answer: They obey the same “commutation relations”. Let me explain.

It is easy to check that the matrices

$$R_x = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & +1 & 0 \end{bmatrix}, \quad R_y = \begin{bmatrix} 0 & 0 & +1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$$

$$R_z = \begin{bmatrix} 0 & -1 & 0 \\ +1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

obey the relations

$$R_x R_y - R_y R_x = R_z$$

$$R_y R_z - R_z R_y = R_x$$  \hspace{1cm} (22.28a)

$$R_z R_x - R_x R_z = R_y$$

The Pauli spin matrices obey a similar relationship with $R$ replaced by $\sigma/2i$:

$$\sigma_x \sigma_y - \sigma_y \sigma_x = 2i \sigma_z$$

$$\sigma_y \sigma_z - \sigma_z \sigma_y = 2i \sigma_x$$  \hspace{1cm} (22.28b)

$$\sigma_z \sigma_x - \sigma_x \sigma_z = 2i \sigma_y$$

The standard textbook introduction to spin starts from these commutation relations and argues that they are a property of the “rotation group”. In order to find a mathematical representation with two components for a rotatable object, one must first write down three (2x2) matrices obeying
these commutation properties which would allow us to rotate the spinor around each of the three axes respectively.

What are the components of a spinor that points along \( z \)? Since rotating it around the \( z \)-axis should leave it unchanged, it should be an eigenvector of \( \sigma_z \) that is,

\[
\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]

which indeed represent an upspin and a downspin along \( z \). Similarly if we want the components of a spinor pointing along \( x \), then we should look at the eigenvectors of \( \sigma_x \), that is,

\[
\begin{pmatrix} +1 \\ +1 \end{pmatrix}/\sqrt{2} \quad \text{or} \quad \begin{pmatrix} +1 \\ -1 \end{pmatrix}/\sqrt{2}
\]

which represent up and down spin along \( +x \). If we consider a spinor pointing along an arbitrary direction described by a unit vector \( \hat{n} \) (see Eq.(22.2)) and wish to know what its components are, we should look for the eigenvectors of

\[
\vec{\sigma} \cdot \hat{n} = \sigma_x \sin \theta \cos \phi + \sigma_y \sin \theta \sin \phi + \sigma_z \cos \theta
\]

\[
= \begin{bmatrix} \cos \theta & \sin \theta e^{-i\phi} \\ \sin \theta e^{+i\phi} & -\cos \theta \end{bmatrix}
\]

(22.29a)

which can be written as (\( c \) and \( s \) defined in Eq.(22.3))

\[
\begin{pmatrix} c \\ s \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -s^* \\ c^* \end{pmatrix}
\]

(22.29b)

In short, the rigorous approach to finding the spinor representation is to first determine a set of three matrices with the correct commutation relations and then look at their eigenvectors. Instead in this Lecture I
adopted a reverse approach stating the spinor components at the outset and then obtaining the matrices through basis transformations.

22.5. Spin Precession

We have already discussed how to write \([H], [\Sigma]\) including non-trivial spin-dependent effects and we could set up numerical models to calculate the electron density \(G\), or the density of states \(A\), or the current using the standard NEGF equations from Lecture 19. Consider for example, the non-local spin potential measurement we started this Lecture with (see Fig.22.5).

Fig.22.5. Spin potential measurement can be modeled with a 1D channel Hamiltonian having four contacts, two of which are magnetic described by \(\Sigma_1, \Sigma_2\).

Fig.22.6 shows the result obtained from the numerical model which supports the basic result stated in Eq.(22.1). The measured voltage oscillates as a function of the angle of magnetization of the voltage probe. It has a constant part independent of the angle and an oscillatory component proportional to the polarization \(P_v\) of the voltage probe which
can be understood in terms of Eq.(22.1) stated at the beginning of this Lecture.

As I mentioned earlier, I am not sure if the experiment shown in Fig.22.5 has been done, but what has been done is to keep both magnets fixed and rotate the electron spin inside the channel.

![Diagram](image)

**Fig.22.6.** Voltage probe signal as the magnetization of the probe is rotated calculated from NEGF model.

How do we rotate the spin? One method that has been widely used is an external magnetic field $B$ which causes the spin direction to precess around the magnetic field as we discussed in Section 22.4 with an angular frequency given by

$$\omega = \frac{2\mu_0 B_z}{\hbar} \quad \text{(same as Eq.(22.24))}$$
This means that the spin voltage at the point where the probe is connected will rotate by an angle $\omega \tau$ where $\tau$ is the time it takes for the electron to travel from the point of injection to the voltage probe. Writing $\tau = L/v$, we have from Eq.(22.1) using Eq.(22.24)

$$\mu_{probe} = \mu + P_2 \mu_s \cos \frac{2 \mu_el}{h} B_z$$  \hspace{1cm} (22.31)

One would expect to see an oscillatory signal as a function of the magnetic field. However, one is usually not dealing with ballistic transport, and there is a large spread in the time $\tau$ spent by an electron between injection and detection so that the average value of this signal over all $\tau$ is essentially zero. What is typically observed is not an oscillatory signal as a function of the B-field but a reduction in the signal from $P\mu_s$ down to zero, which is referred to as the Hanle signal. However, Hanle signals showing several oscillations have also been observed, but this requires that the spread in $\tau$ be much less than its mean value (see for example, Huang et al. 2007).

Another possible approach to rotating electron spins is to use the Rashba effect in materials with strong spin-orbit coupling. In many semiconductors, it is now well established that a surface electric field along $z$ (Fig.22.6) leads to an effective magnetic field that depends on the electron momentum. This can be seen by comparing the Hamiltonians for the B-field (Eq.(22.16))

$$H_B = \mu_{el} \vec{\sigma} \cdot \vec{B}$$

with that for the Rashba interaction (22.20) which can be rewritten as

$$H_R = \eta \hat{\vec{z}} \cdot (\vec{\sigma} \times \vec{k}) \rightarrow \eta \vec{\sigma} \cdot (\hat{\vec{z}} \times \vec{k})$$

suggesting that the effective B-field due to the Rashba interaction is given by

$$\mu_{el} \vec{B}_{eff} = \eta \hat{\vec{z}} \times \vec{k}$$  \hspace{1cm} (22.32)

so that from Eq.(22.31) we expect an oscillatory signal of the form
\[ \mu_{\text{probe}} = \mu + P_2 \mu_s \cos \frac{2\eta kL}{\hbar v} \]  \hspace{1cm} (22.33)

with a period \( \Delta \eta \) defined by

\[ \frac{2kL}{\hbar v} \Delta \eta = 2\pi \rightarrow \Delta \eta = \frac{2\pi a t_0}{kL} \sin k a \]

Fig. 22.7. In materials with a large Rashba coefficient, a gate voltage should lead to an oscillatory output, if the source and drain magnets point along \( x \), but not if they point along \( z \).

This is in approximate agreement with the numerical result obtained from the NEGF method (Fig. 22.7) using an energy \( E \) corresponding to
rotating an electron

\[ ka = \frac{\pi}{3} \]

and a distance of about \( L=40a \) between the injector and the detector.

In the structure shown in Fig.22.7 the electrons traveling along \(+x\) should feel an effective \( B \)-field along \( y \). Since the injected spins have a spin voltage \( \vec{\mu} \), pointing along the source and drain magnets \((x)\) it should be rotated. Note that the oscillation should not be observed if the source and drain magnets point along \( y \) rather than along \( x \). There is some recent experimental evidence for this effect (Koo et al 2009, Wunderlich et al. 2010).

22.5.1. Spin-Hall Effect

Spin-diffusion in non-magnetic materials like copper have been studied extensively and the basics are now fairly clear as we saw in Lecture 14. What is relatively less understood is the nature of spin diffusion in materials with high spin-orbit coupling.

For example, an interesting experiment that has attracted a lot of attention in the last decade is the spin-Hall effect. Back in Lecture 13 we discussed the Hall effect where in the presence of a magnetic field, electrons from the source curve “upwards” while those from the drain curve downwards creating a measurable potential difference in the \( y \)-direction (Fig.22.8).

Fig.22.8. The spin-Hall effect can be understood in terms of an effective force that makes \(+z\)-spins and \(-z\)-spins veer in opposite directions, unlike the ordinary Hall effect that makes both spins veer in the same direction.
Since electrons feel an effective magnetic field (Eq.(22.32)) in conductors with spin-orbit coupling, it seems natural to ask whether we could see a Hall effect without an external $B$-field simply due to this effective field.

The answer indeed is yes, but the phenomenon is subtle with $+z$-spins and $-z$-spins veering in opposite directions, unlike the ordinary Hall effect that makes both spins veer in the same direction (Eq.(13.1)). As a result, there is no buildup of charge and hence no ordinary Hall voltage that can be measured with ordinary contacts. But there is a spin voltage that can be measured with magnetic contacts. The first experiments detected this spin buildup using optical probes in large conductors (see for example Sih et al. 2006) but more recently magnetic contacts have been used to measure spin-Hall effects of this type in small conductors (see for example, Wunderlich et al. 2010).

Our NEGF-based transport model seems to capture this effect as shown in Fig.22.9 for a 2D conductor with spin-orbit coupling like the one shown in Fig.22.7 with $\eta = 1e - 11eV - m$, assuming an energy $E = 0.05\, t_0$. We use Eq.(22.31) to extract the $z$-component of the spin density from the $G^n$ as discussed. The numerical results show a build-up of opposite spins on opposite sides of the conductor which reverses on reversing the current, in accordance with what we discussed above.

Fig.22.9. NEGF-Based transport model for spin-Hall effect. Numerical results show a pile-up of $+z$-spins and $-z$-spins on opposite sides of the sample which reverses on reversing the current.
There is much activity currently in this general area of spin transport in materials with strong spin-orbit coupling like the “topological insulators” (see for example Xiu et al. 2011 and references therein) which exhibit a more striking version of the spin Hall effect.

The main point I want to make is that many striking non-intuitive results come automatically from the NEGF model, allowing one to explore new physics that we may not yet have understood clearly. However, I should mention that we have not discussed the calculation and interpretation of spin currents which are needed in order to couple the transport problem with the LLG equation describing the magnet dynamics (see Fig.14.10). We will not go into this and simply suggest a few references for interested readers.

22.6. From NEGF to Diffusion

Let me end this Lecture by talking a little about the connection between the diffusion equation approach from Lecture 14 with the NEGF-based approach of this Lecture. We have seen that the NEGF-based numerical examples in Fig.22.6, 7 are well described by Eq.(22.1) which follows heuristically from the discussions in Lecture 14. But the reason for this agreement is not obvious, especially since the formal method is based on two-component complex spinors, while $\mu_s$ is an old-fashioned three-component real vector.

Earlier we talked about the connection between vector $\hat{n}$ along which a spin points and the wavefunction $\psi$ representing it. To relate Eq.(22.1) to the NEGF method we need to consider quantities like $G^n \sim \psi \psi^+$ rather than the wavefunction $\psi$, since the NEGF is formulated in terms of $G^n$. Besides it is $G^n$ and not $\psi$ that is observable and can be related to experiment.
22.6.1. Matrix Electron Density

We have often referred to \( [G^n] \) as the matrix electron density whose diagonal elements tell us the number of electrons at a point. With spin included, \( [G^n] \) at a point is a (2x2) matrix and the elements of this matrix tell us the number of electrons \( N \) or the net number of spins \( \vec{S} \).

To see this consider an electron pointing in some direction \( \hat{n} \) represented by a spinor wavefunction of the form (see Eq.(22.3))

\[
\psi = \begin{pmatrix}
\cos\frac{\theta}{2} e^{-i\phi/2} \\
\sin\frac{\theta}{2} e^{i\phi/2}
\end{pmatrix} \equiv \begin{pmatrix}
c \\
s
\end{pmatrix}
\]

so that the corresponding (2x2) electron density \( G^n \) is given by

\[
\psi^\dagger \psi = \begin{pmatrix}
c & s \\
\bar{s} & \bar{c}
\end{pmatrix} \begin{pmatrix}
c \bar{c}^* & \bar{s} s^* \\
sc^* & ss^*
\end{pmatrix}
\]

Making use of Eq.(22.2) and Eq.(22.12) we have

\[
\psi^+\psi^+ = \begin{pmatrix}
1+n_z & n_x - in_y \\
n_x + in_y & 1-n_z
\end{pmatrix} = \frac{1}{2} \left[I + \vec{\sigma} \cdot \hat{n}\right]
\]

For a collection of \( N \) electrons we can add up all their individual contributions to \( \psi^+\psi^+ \) to obtain the net \( [G^n] \) given by

\[
\frac{G^n}{2\pi} = \frac{1}{2} \begin{pmatrix}
N+S_z & S_x - iS_y \\
S_x + iS_y & N-S_z
\end{pmatrix} = \frac{1}{2} \left[N[I] + \vec{\sigma} \cdot \vec{S}\right] \tag{22.30}
\]

Given a \( [G^n] \) we can extract these quantities from the relations
Rotating an electron

\[ N = \frac{1}{2\pi} \text{Trace}[G^n], \quad \hat{S} = \frac{1}{2\pi} \text{Trace}[\hat{\sigma} G^n] \tag{22.31} \]

which follow from Eq.(22.30) if we make use of the fact that all three matrices (Eq.(22.15)) have zero trace, along with the following properties of the Pauli spin matrices that are easily verified.

\[ \sigma_x^2 = \sigma_y^2 = \sigma_z^2 = I \tag{22.32a} \]

\[ \sigma_x \sigma_y = -\sigma_y \sigma_x = i \sigma_z \tag{22.32b} \]

\[ \sigma_y \sigma_z = -\sigma_z \sigma_y = i \sigma_x \tag{22.32c} \]

\[ \sigma_z \sigma_x = -\sigma_x \sigma_z = i \sigma_y \tag{22.32d} \]

along with the that is the sum of their diagonal elements is zero.

In summary, all the information contained in the 2x2 Hermitian matrix \([G_n]\) can be expressed in terms of four real quantities consisting of a scalar \(N\) and the three components of a vector \(\hat{S}\) which can be extracted using Eq.(22.31).

22.6.2. Measuring the Spin Potential

That brings us to the question of relating the result stated at the beginning of this Lecture

\[ \mu_{\text{probe}} = \mu + \frac{\hat{P} \cdot \hat{\mu}_s}{2} \quad (\text{Same as Eq.(22.1))} \]

to the NEGF method. The scalar version of this result was obtained in Lecture 14 (see Eq.(14.23)) using the semiclassical model for a probe shown and setting the probe current to zero. Now we can obtain the general vector version by starting from the NEGF model for a probe (Fig.22.10) with the current given by (see Eq.(19.4))
$I \sim \text{Trace} \left[ \Gamma \left[ f_{probe} [A] - [G^n] \right] \right]$

so that for zero probe current we must have

$$f_{probe} = \frac{\text{Trace} \left[ \Gamma \right] [G^n]}{\text{Trace} \left[ \Gamma \right] [A]}$$

Fig. 22.10. Model for a probe connected to a channel.

Making use of Eq. (22.30) for $[G^n]$, assuming that the density of states (D) is spin-independent

$$\frac{A}{2\pi} = \frac{D}{2} [I]$$

writing the probe coupling in the form (see Eq. (22.16))

$$\Gamma = \gamma \left[ I + \vec{\sigma} \cdot \vec{P} \right]$$

and noting that the Pauli matrices all have zero trace, we obtain

$$f_{probe} = \text{Trace} \left[ I + \vec{\sigma} \cdot \vec{P} \right] \left[ \frac{N}{D} I + \vec{\sigma} \cdot \frac{\vec{S}}{D} \right]$$

(22.33)
Once again there is an identity that can be used to simplify this expression: For any two vectors $\vec{P}$ and $\vec{B}$, it is straightforward (but takes some algebra) to verify that

$$[\vec{\sigma} \cdot \vec{P}] \, [\vec{\sigma} \cdot \vec{B}] = (\vec{P} \cdot \vec{B}) [I] + i \vec{\sigma} \cdot [\vec{P} \times \vec{B}] ,$$

so that

$$[I + \vec{\sigma} \cdot \vec{P}] \, \left[ b \, I + \vec{\sigma} \cdot \vec{B} \right] = (b + \vec{P} \cdot \vec{B}) [I] + \vec{\sigma} \cdot [\vec{P} + \vec{B} + i \vec{P} \times \vec{B}] \quad (22.34)$$

Making use of this identity and noting once again that the Pauli matrices have zero trace, we can write from Eq.(12.33)

$$f_{\text{probe}} = \frac{N}{D} + \vec{P} \cdot \vec{S} \quad = \quad f + \vec{P} \cdot \vec{f}_{\text{s}} \quad (22.35)$$

where the charge occupation $f$ and the spin occupation $f_{\text{s}}$ are each defined such that their maximum possible values are one. We can translate this relation in terms of occupation into the relation in terms of potentials stated earlier (see Eq.(22.1)) using the linear relation between the two for small bias (see Eq.(2.8)).

### 22.6.3. Four-Component Diffusion

The key point we want to stress is that the quantum formalism naturally leads to a $2 \times 2$ complex matrix $[G]$ at each point, but it is straightforward to translate it into four physically transparent components, like $N$, $\vec{S}$ or $\mu$, $\vec{\mu}$.

Back in Lecture 14, we saw how many spin transport phenomena can be understood in terms of the Valet-Fert equation along with a model for the spin-dependent interface resistances. However, this approach is limited to problems that involve spins in one direction (the $z$-direction) only.

Now we have the full NEGF model that can be used for spins pointing in any direction, but it is computationally more demanding and seems conceptually unrelated to the old approach. By translating the information in $[G]$ into $N$, $\vec{S}$ we can bridge this gap. For example, the
spin-dependent interface conductances of Lecture 14 could be replaced by a 4x4 conductance matrix that relates the four components of the potential to the four components of the current:

\[
\begin{bmatrix}
I \\
I_{sx} \\
I_{sy} \\
I_{sz}
\end{bmatrix} = \begin{bmatrix}
4 \times 4
\end{bmatrix}
\begin{bmatrix}
\Delta \mu \\
\Delta \mu_{sx} \\
\Delta \mu_{sy} \\
\Delta \mu_{sz}
\end{bmatrix}
\]

Similarly the two component Valet-Fert equation for \( \mu, \mu_s \) can be extended to a four component diffusion equation for

\[
\mu, \quad \mu_{sx}, \mu_{sy}, \mu_{sz}
\]

The question one could ask is whether these four-component formulations are equivalent to the NEGF method discussed here. The answer we believe is that they capture a subset of the effects contained in the NEGF and there are many problems where this subset may be adequate. Let me explain.

At the beginning of this Lecture I mentioned that including spin increases the size of the matrices by a factor of two since every point along \( z \) effectively becomes two points, an up and a down. So if there are three points in our channel, the matrix \( [G''] \) will be of size (6x6).
Rotating an electron elements between two spatial elements which as we saw in Lecture 20
gives rise to quantum interference effects. But we may not be missing
much, since as we discussed, dephasing processes often destroy these
interference effects anyway.

Spin information is usually more robust. While phase relaxation times
are often sub-picosecond, spin relaxation times are much longer, in
nanoseconds. And so it is important to retain the information in the (2x2)
diagonal blocks, even if we are throwing away the rest.

Formally we could do that starting from the NEGF method by defining a
suitable D-matrix of the type discussed in Lecture 19 relating the
incattering to the electron density (\( \times \) denotes element by element
multiplication)

\[
[\Sigma^{in}] = D \times [G^n] \quad \text{(Same as Eq.(19.35b))}
\]

The dephasing process can be viewed as extraction of the electron from a
state described by \( [G^n] \) and reinjecting it in a state described by \( D \times G^n \).

We introduced two models A and B with \( D \) defined by Eqs.(19.37) and
(19.38) respectively. Model A was equivalent to multiplying \( [G^n] \) by a
constant so that the electron was reinjected in exactly the same state that
it was extracted in, causing no loss of momentum, while Model B threw
away the off-diagonal elements causing loss of momentum as we saw in
the numerical example in Fig.20.7. We could define a Model C having a
D-matrix that retains spin information while destroying momentum:

\[
\frac{[D]}{D_0} = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 \\
\end{pmatrix}
\quad \text{(22.36)}
\]
One could view this as Model B-like with respect to the lattice, but Model A-like with respect to spin. We could rewrite the NEGF equation

\[ G^n = G^R \Sigma^{in} G^A \]

as

\[ [G^n]_{i,j} = \sum_j [G^R]_{i,j} [\Sigma^{in}]_{j,j} [G^A]_{j,j} \]

\[ = D_0 \sum_j [G^R]_{i,j} [G^n]_{j,j} [G^A]_{j,j} \quad (22.37) \]

where the indices \( i, j \) refer to lattice points and we have made use of the fact that in our Model C, \( \Sigma^{in} \) is diagonal as far as the lattice is concerned.

We have seen earlier that at any point on the lattice the 2x2 matrix \( [G^n] \) can be expressed in terms of four components, namely \( N \) and \( \rho \) so that with a little algebra we could rewrite Eq.(22.37) in the form

\[
\begin{bmatrix}
N \\
S_x \\
S_y \\
S_z
\end{bmatrix}_i 
= \sum_j \begin{bmatrix}
4 \times 4 \\
"Hopping" \\
Matrix
\end{bmatrix}_{i,j} 
\begin{bmatrix}
N \\
S_x \\
S_y \\
S_z
\end{bmatrix}_j 
\quad (22.38)
\]

where the (4x4) matrix could be viewed as describing the probability of the \((N, \rho)\) at a point “j” hopping to a point “i” in one time step. Indeed the (1x1) version of Eq.(22.38) resembles the standard description of Brownian motion on a lattice that leads to the drift-diffusion equation.

Spin diffusion equations based on alternative approaches like the Kubo formalism have been discussed in the past (see for example, Burkov et al. 2004). The main point I want to convey is that NEGF-based approaches can also be used to justify and benchmark spin diffusion models which could well capture the essential physics and provide insights that a purely numerical calculation misses. In the last lecture I will briefly explain how one might be able to adapt this approach to more complicated spin-like objects as well. But that is a topic for future research.
Back in Lecture 18 we used this picture to introduce our quantum transport model representing an elastic channel described by a Hamiltonian $[H]$ and self-energies $[\Sigma]$ describing the exchange of electrons with the physical contacts and energy with the surroundings which can be viewed as additional conceptual “contacts”.

Given these inputs, the basic NEGF equations (see Eqs.(19.1)-(19.4)) tell us how to analyze any given structure. Since then we have been looking at various examples illustrating how one writes down $[H]$ and $[\Sigma]$ and uses the NEGF equations to extract concrete results and investigate the physics. One major simplification we have adopted is in our treatment of the interactions in the channel represented by $\Sigma_0$ which we have either ignored (coherent transport) or treated as an elastic dephasing process described by Eqs.(19.39).
This choice of self-energy functions leads to no exchange of energy with the surroundings, but it has an effect on transport due to the exchange of momentum and “phase”. Basically we have been talking about elastic resistors like the ones we started these Lectures with, except that we are now including quantum mechanical effects. One could say that in the last few Lectures we have applied the general Non-Equilibrium Green’s Function (NEGF) method to an elastic resistor, just as in Part one we applied the general Boltzmann Transport Equation (BTE) to an elastic resistor.

So how do we go beyond elastic resistors? For semiclassical transport, it is clear in principle how to include different types of interaction into the BTE for realistic devices and much progress has been made in this direction. Similarly for quantum transport, the NEGF tells us how to evaluate the self-energy $\Sigma_0$ for any given microscopic interaction. In these lectures we have talked only about elastic dephasing which is a small subset of the interactions considered in the classic work on NEGF (see for example, Danielewicz 1984 or Mahan 1987).

In practice, however, it remains numerically challenging to go beyond elastic resistors and approximate methods continue to be used widely. Readers interested in the details of device analysis at high bias may find an old article (Datta (2000)) useful. This article has a number of concrete results obtained using MATLAB codes that I had offered to share with anyone who asked me for it. Over the years many have requested these codes from me which makes me think they may be somewhat useful and we plan to have these available on our website for these notes.

I should mention that many devices are rather forgiving when it comes to modeling the physics of inelastic scattering correctly. Devices with energy levels that are equally connected to both contacts (Fig.9.5b.) do not really test the deep physics of inelastic transport and cannot distinguish between a good theory and a bad one. A good test for inelastic scattering models is the device shown in Fig.9.5a for which the entire terminal current is driven by inelastic processes. Only a
fundamentally sound theory will predict results that comply with the requirements of the second law.

But practical issues apart, can the NEGF method model "everything", at least in principle?

The formal NEGF method developed in the 1960’s was based on many-body perturbation theory (MBPT) which provided clear prescriptions for evaluating the self-energy functions

\[ \Sigma, \Sigma^{in} \]

for a given microscopic interaction up to any order in perturbation theory. It may seem that using MBPT we can in principle include everything. However, I believe this is not quite true since it is basically a perturbation theory which in a broad sense amounts to evaluating a quantity like \((1-x)^{l}\) by summing a series like \(1+x+x^2+x^3+\ldots\), which works very well if \(x\) is much less than one. But if \(x\) happens to exceed one, it does not work and one needs non-perturbative methods, or perhaps a different perturbation parameter.

This is one of the reasons I prefer to decouple the NEGF equations (Eqs.(19.1) through (19.4)) from the MBPT-based methods used to evaluate the self-energy functions. The latter may well evolve and get supplemented as people find better approximations that capture the physics in specific situations.

With equilibrium problems, for example, density functional theory (DFT)-based techniques have proven to be very successful and are often used in quantum chemistry in place of MBPT. I believe one should be cautious about expecting the same success with non-equilibrium problems where a far greater spectrum of many body states are made accessible and can be manipulated through a judicious choice of contacts, but it is quite likely that people will find insightful approaches that capture the essential physics in specific problems.
Like the BTE for semiclassical transport, NEGF-based methods in their simplest form, seem to provide a good description of problems where electron-electron interactions can be treated within a mean field theory based on the widely used picture of quasi-independent electrons moving in a self-consistent potential $U$ due to the other electrons (Section 18.2).

As we saw in Lecture 8, for low bias calculations one needs to consider only the equilibrium potential which is already included in the semi-empirical tight-binding (TB) parameters used to construct our Hamiltonian $[H]$. For real devices operating at high bias, the change in the potential due to any changes in the electron occupation in the channel are routinely included using the Poisson equation which is the simplest approximation to the very difficult problem of electron-electron interactions and there have been extensive discussions of how the self-consistent field (scf) can be corrected to obtain better agreement with experimental results.

However, there are examples where the self-consistent field approach itself seems to fail and some of the most intriguing properties arise from a failure of this simple picture. The purpose of this Lecture is to alert the reader that a straightforward application of NEGF may well miss these important experimentally observable effects. Future challenges and opportunities may well involve effects of this type, requiring insightful choices for $\Sigma$, $\Sigma^{\text{in}}$ if we wish to use the NEGF method.

### 23.1. Coulomb Blockade

In the spirit of the bottom-up approach let us consider the simplest resistor that will show this effect, one that is only slightly more complicated than the one-level resistor we started with (Fig.3.1). We assume two levels, a spin up and a spin down, having the same energy $\epsilon$, with the equilibrium chemical potential $\mu$ located right at $\epsilon$, so that each level is half-filled since the Fermi function $f_0(E=\mu)$ equals 0.5. Based on what we have discussed so far we would expect a high conductance since
the electrochemical potential lies right in the middle of each broadened level as shown in the upper sketch in Fig.23.1.

However, if the single electron charging energy $U_0$ is large then the picture could change to the lower one where one level floats up by $U_0$ due to the electron occupying the other level. Why doesn’t the other level float up as well? Because no level feels any potential due to itself. This self-interaction correction is missed in the self-consistent field (SCF) model discussed in Lecture 8 where we wrote $U=U_0N$. Instead we need an unrestricted SCF where each level $i$ is not restricted to feeling the same potential. Instead it feels a potential $U_i$ that depends on the change in the number of electrons occupying all levels except for $i$:

$$U_i = U_0 (N - N_i)$$ (23.1)

If we were to use Eq.(23.1) instead of $U=U_0N$ we would obtain a picture like the lower one in Fig.25.4, assuming that $\mu$ is adjusted to have approximately one electron inside the channel. We would find a self-consistent solution with

$$N_{dn} = 1, U_{up} = 0, N_{up} = 0, U_{dn} = 0$$
The down level will be occupied ($N_{dn} = 1$) and the resulting potential ($U_{up} = U_0$) will cause the up level to float up and be unoccupied ($N_{up} = 0$). Because it is unoccupied, the potential felt by the down level is zero ($U_{dn} = 0$), so that it does not float up, consistent with what we assumed to start with.

Of course, the solution with up and down interchanged

$$N_{up} = 1, U_{dn} = 0, N_{dn} = 0, U_{up} = 0$$

is also an equally valid solution. Numerically we will converge to one or the other depending on whether we start with an initial guess that has more $N_{up}$ or $N_{dn}$. Experimentally the system will fluctuate between the two solutions randomly over time.

Why have we not worried about this before? Because it is not observable unless the charging energy $U_0$ is well in excess of both $kT$ and the broadening. $U_0/q$ is the potential the channel would float to if one electron were added to it. For a large conductor this potential is microvolts or smaller and is unobservable even at the lowest of temperatures. After all, any feature in energy is spread out by $kT$ which is $\sim 25$ meV at room temperature and $\sim 200$ µeV at $\sim 1$K. The single electron charging effect that we are talking about, becomes observable at least at low temperatures, once the conductor is small enough to make $U_0$ of the order of a meV. For molecular sized conductors, $U_0$ can be hundreds of meV making these effects observable even at room temperature.

Fig.23.2. The single electron charging energy $U_0$ is the electrostatic energy associated with one extra electron in the channel.

However, there is a second factor that also limits the observability of this effect. We saw in Lecture 19 that in addition to the temperature
broadening $\sim kT$, there is a second and more fundamental broadening, 
$\gamma \sim \hbar/t$ related to the transfer time. Single electron charging effects will 
be observed only if the Coulomb gap $U_0$ exceeds this broadening: 
$U_0 > \hbar/t$. For this reason we would not expect to see this effect even 
in the smallest conductors, as long as it has good contacts.

24.1.1. Current versus voltage

Let us now move on from the low bias conductance to the full current-
voltage characteristics of the two-level resistor. For simplicity we will 
assume that the levels remain fixed with respect to the source and are 
unaffected by the drain voltage, so that we do not have to worry about 
the kind of issues related to simple electrostatics that we discussed in 
Lecture 8.

A simple treatment ignoring electron-electron interactions then gives the 
curve marked "non-interacting" in Fig.23.3. Once the electrochemical 
potential $\mu_2$ crosses the levels at $\epsilon$, the current steps up to its maximum 
value.

If we now include charging effects through a self-consistent potential 
$U=U_0$, the current step stretches out over a voltage range of $\sim U_0/q$, since 
the charging of the levels makes them float up and it takes more voltage 
to cross them completely.

Fig.23.3. 
Current-voltage characteristic of a two-
level resistor with $U=0$ and with $U= 
U_0 N$
But if we include an SCF with self-interaction correction (Eq.(23.1)) we calculate a current-voltage characteristic with an intermediate plateau as shown in Fig.23.4 which can be understood in terms of the energy level diagrams shown. At first only the lower level conducts giving only half the maximum current and only when the voltage is large enough for $\mu_2$ to cross $\varepsilon + U_0$ that we get the full current.

Such intermediate plateaus in the $I-V$ characteristics have indeed been observed but the details are not quite right. The correct plateau current is believed to be $2/3$ and not $1/2$ of the total current of $2q/t$. This represents an effect that is difficult to capture within a one-electron picture, though it can be understood clearly if we adopt a different approach altogether, which we will now describe.
23.2. Fock Space Description

This approach is based on the Fock space picture introduced in Lecture 16. As we discussed earlier, in this new picture we do not think in terms of one-electron levels that get filled or emptied from the contacts. Instead we think in terms of the system being driven from one state to another.

For example Fig. 23.5 shows how we would view the one-level resistor in this Fock space picture where the system can be one of two states: 0 representing an empty state, and 1 representing a full state. Fig. 23.6 shows the two pictures for a two-level resistor. In general a $N$-level resistor will have $2^N$ Fock space states.

23.2.1. Equilibrium in Fock space

As we discussed in Lecture 16, there is a well-defined procedure for finding the probabilities of finding the system in a given eigenstate $i$ at equilibrium.
\[ p_i = \frac{1}{Z} e^{-(E_i - \mu N_i)/kT} \]  
(Same as Eq.(16.14))

We could use this to calculate any equilibrium property. For example suppose we want to find the number of electrons, \( n \) occupying the two-level channel shown in Fig.23.6 if it is in equilibrium with an electrochemical potential \( \mu \).

Fig.23.7 shows the result obtained by plotting \( n \) versus \( \mu \) from the equation

\[ n = \sum_i N_i p_i = p_{01} + p_{10} + 2p_{11} \]

using the equilibrium probabilities from Eq.(16.14) cited above. Note how the electron number changes by one as \( \mu \) crosses \( \epsilon \) and then again when \( \mu \) crosses \( \epsilon + U_0 \) in keeping with the lower picture in Fig.23.1.

Note, however, that we did not assume the picture from Fig.23.1 with two one-electron states at different energies. We assumed two one-electron states with the same energy (Fig.23.6) but having an interaction energy that is included in the Fock space picture.

If we are interested in the low bias conductance \( G \) as a function of \( \mu \), we could deduce it from the \( n(\mu) \) plot in Fig.23.7. As we discussed in Lecture 2, current flow is essentially because the two contacts with different \( \mu \)'s have different agendas, since one likes to see more electrons in the channel than the other. From this point of view one could argue that the conductance should be proportional to \( dn/d\mu \) and show peaks at

\[ \mu = \epsilon \quad \text{and at} \quad \mu = \epsilon + U_0 \]

as shown. This is indeed what has been observed experimentally for the low bias conductance of small conductors in the single-electron charging regime where \( U_0 \) exceeds both the thermal energy \( kT \) and the energy broadening due to contacts.
Fig. 23.7. Equilibrium number of electrons, \( n \) in the two-level channel shown in Fig. 23.6 as a function of \( \mu \), assuming \( \epsilon = 10kT \), \( U_0 = 20kT \). The conductance can be argued to be proportional to the derivative \( dn/\mu \) showing peaks when \( \mu \) equals \( \epsilon \) and \( \epsilon + U_0 \).

As we saw in Lecture 15, low bias conductance is an equilibrium property that can be deduced using the principles of equilibrium statistical mechanics. Current flow at higher voltages on the other hand requires the methods of non-equilibrium statistical mechanics. Let me explain briefly how one could understand the 2/3 plateau shown in Fig. 23.4 by calculating the current at high bias in the Fock space picture.

### 23.2.2. Current in the Fock space picture

To calculate the current we write an equation for the probability that the system will be found in one of its available states, which must all add up to one. For example for the one level resistor we could write

\[
v_1p_0 = v_2p_1 \quad \rightarrow \quad \frac{p_1}{p_0} = \frac{v_1}{v_2} \quad \rightarrow \quad p_1 = \frac{v_1}{v_1 + v_2}
\]
assuming that the left contact sends the system from the 0 state to the 1 state at a rate $v_1$, while the right contact takes it in the reverse direction at a rate $v_2$ and at steady-state the two must balance. The current is given by

$$I = qv_2p_1 = q \frac{v_1v_2}{v_1 + v_2} \quad (23.2)$$

in agreement with our earlier result in Lecture 19 (see Eq.(19.10b)) obtained from a one-electron picture.

But the real power of this approach is evident when we consider levels with multiple interacting levels. Consider for example the two-level resistor biased such that electrons can come in from the left contact and transfer the system from $00$ to $01$ or to $10$, but not to the $11$ state because of the high charging energy $U_{00}$. This is the biasing condition that leads to a plateau at $2/3$ the maximum value (Fig.23.4) that we mentioned earlier.

In this biasing condition, the system can only come out of the $11$ state, but never transfer into it, and so the steady-state condition can be calculated simply by considering the kinetics of the three remaining states in Fock space, namely $00$, $01$ and $10$:

$$2v_1p_{00} = v_2(p_{01} + p_{10})$$

$$\Rightarrow \quad \frac{p_{01} + p_{10}}{p_{00}} = \frac{2v_1}{v_2}$$

$$\Rightarrow \quad p_{01} + p_{10} = \frac{2v_1}{2v_1 + v_2}$$

where we have made use of the requirement that all three probabilities must add up to one. Hence
Does NEGF Include “Everything?”

\[ I = qv_2(p_{01} + p_{10}) = \frac{q(2v_1v_2)}{2v_1 + v_2} \]

With \( v_1 = v_2 \rightarrow I = \frac{2}{3}qv_1 \)

which is 2/3 the maximum current as stated earlier.

It is important to note the very special nature of the solution we just obtained which makes it hard to picture within a one-electron picture. We showed that the system is equally likely to be in the states 00, 01 and the 10 states, but zero probability of being in the 11 state.

![Fig. 23.8](image)

The intermediate plateau in the current corresponds to the channel being in a strongly correlated state.

In other words, if we looked at the up-spin or the down-spin state (in the one-electron picture) we would find them occupied with 1/3 probability. If electrons were independent then we would expect the probability for both to be occupied to be the product = 1/9.

Instead it is zero, showing that the electrons are correlated and cannot be described with a one-electron occupation factor \( f \) of the type we have been using throughout these lectures. Even with quantum transport we replaced the \( f \)'s with a matrix \( G^0 \) obtained by summing the \( \psi \psi^+ \) for individual electrons. This adds sophistication to our understanding of the one-electron state, but it still does not tell us anything about two-electron correlations.
23.3. Entangled states

What we just saw with one quantum dot is actually just the proverbial tip of the iceberg. Things get more interesting if we consider two or more quantum dots.

For example, with two coupled quantum dots we could write the one-electron Hamiltonian matrix as a 4x4 matrix using the up and down states in dots 1 and 2 as the basis functions as follows:

$$[H] = \begin{pmatrix}
  u_1 & u_2 & d_1 & d_2 \\
  u_1^* & \varepsilon_1 & t & 0 \\
  u_2 & t & \varepsilon_2 & 0 \\
  d_1 & 0 & 0 & \varepsilon_1 \\
  d_2 & 0 & t & \varepsilon_2
\end{pmatrix}$$

But what are the Fock space states? With four one-electron states we expect a total of $2^4 = 16$ Fock space states, containing 0, 1, 2, 3 or 4 electrons. The number of $n$-electron states in Fock space is given by $^4C_n$: one with $n=0$, four with $n=1$, six with $n=2$, four with $n=3$ and one with $n=4$. 
If there were no inter-dot coupling then these sixteen states would be the eigenstates and we could analyze their dynamics in Fock space just as we did for one dot. But in the presence of inter-dot coupling the true eigenstates are linear combinations of these states and these entangled states can lead to novel physics and make it much more interesting.

The 0-electron and 4-electron states are trivially composed of just one Fock space state, while the 1-electron state is essentially the same as the states in a one-electron picture. Indeed the 3-electron state also has a structure similar to the one-electron state and could be viewed as a 1-hole state.

The 2-electron states, however, have an interesting non-trivial structure. Consider the six 2-electron states which we label in terms of the two states that are occupied: $u_1d_1$, $u_1d_2$, $u_2d_1$, $u_2d_2$, $u_1u_2$, $d_1d_2$. Using these we can write the Fock space Hamiltonian $[HH]$ as explained below.

The **diagonal** elements of $[HH]$ are written straightforwardly by adding the one-electron energies plus an interaction energy $U_0$ if the two basis functions happen to be on the same dot making their Coulomb repulsion much stronger than what it is for two states on neighboring dots.

<table>
<thead>
<tr>
<th>Electron</th>
<th>States</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-electron</td>
<td>![1111]</td>
</tr>
<tr>
<td>3-electron</td>
<td>![1110, 1101, 1011, 0111]</td>
</tr>
<tr>
<td>2-electron</td>
<td>![1100, 1010, 1001, 0110, 0101, 0011]</td>
</tr>
<tr>
<td>1-electron</td>
<td>![1000, 0100, 0010, 0001]</td>
</tr>
<tr>
<td>0-electron</td>
<td>![0000]</td>
</tr>
</tbody>
</table>
The off-diagonal entries $t$ are obtained by noting that this quantity couples the one electron states $u_1$ to $u_2$ and $d_1$ to $d_2$. With two electron states we have inserted $t$ for non-diagonal elements that couples those states for which one state remains unchanged while the other changes from $u_1$ to $u_2$ or from $d_1$ to $d_2$.

The lowest eigenstate obtained from the two-electron Hamiltonian in Eq.(23.4) is with a wavefunction of the form $(s_1, s_2 < 1)$

$$S : S\left(\{u_1d_2\} + \{u_2d_1\}\right) + s_1\{u_1d_1\} + s_2\{u_2d_2\} \quad (23.5)$$

is called the singlet state. Next comes a set of three states (called the triplets) that are higher in energy. These have the form

$$T1 : \frac{1}{\sqrt{2}}\left(\{u_1d_2\} - \{u_2d_1\}\right)$$

$$T2 : \{u_1u_2\}$$

$$T3 : \{d_1d_2\} \quad (23.6)$$
A system with two electrons is normally viewed as occupying two one-electron states. The states T2, T3 permit such a simple visualization. But the states S and T1 do not.

For example, each term in the state

\[ T1 : \frac{1}{\sqrt{2}} \left( \{u_1d_2\} - \{u_2d_1\} \right) \]

permits a simple visualization: \( \{u_1d_2\} \) stands for an upspin electron in 1 and a downspin electron in 2 while \( \{u_2d_1\} \) represents an upspin in 2 and a downspin in 1. But the real state is a superposition of these two “simple” or unentangled states and there is no way to define two one-electron states \( a \) and \( b \) such that the two-electron state could be viewed as \( \{ab\} \).

Such states are called entangled states which comprise the key entity in the emerging new field of quantum information and computing.

How would we compute the properties of such systems? The equilibrium properties are still described by the general law of equilibrium stated earlier

\[ p_i = \frac{1}{Z} e^{-(E_i - \mu N_i)/kT} \]  
(Same as Eq.(16.14))

and using the equilibrium properties to evaluate the average number of electrons.

\[ n = \sum_i N_i p_i \]

The energies \( E_i \) are obtained by diagonalizing the Fock space Hamiltonian \( HHJ \) that we just discussed. Fig.22. Fig.23.9 shows the plot of \( n \) versus \( \mu \) which looks like Fig.23.7, but the middle plateau now involves the entangled singlet state just discussed. There is also some additional structure that we will not get into. The main point we wanted to make is that the law of equilibrium statistical mechanics is quite general and can be used in this case.
Lessons from Nanoelectronics

But the calculation of current at high bias is a non-equilibrium problem that is not as straightforward. Using the entangled states one could set up a rate equation as we did in the last Section and understand some of the interesting effects that have been observed experimentally including negative differential resistance (NDR), that is a decrease in current with increasing voltage (see for example Muralidharan et al. 2007). More generally one needs quantum rate equations to go beyond the simple rate equations we discussed and handle coherences (Braun et al. 2004, Braig and Brouwer 2005).

Can we model transport involving correlated and/or entangled states exactly if we use a Fock space picture instead of using NEGF and including interactions only approximately through self-energies? Sort of, but not quite.

There are two problems. The first is practical. A $N$-level problem in the one-electron picture escalates into a $2^N$ level problem in the Fock space picture. The second is conceptual.

We saw in Lecture 19 how the NEGF method allows us to include quantum broadening in the one-electron Schrödinger equation. To our knowledge there is no comparable accepted method for including
broadening in the Fock space picture. So the rate equation approach from
the last Section works fine for weakly coupled contacts where the
resulting broadening is negligible, but the regime with broadening
comparable to the charging energy stands out as a major challenge in
transport theory. Even the system with two levels (Fig.23.7) shows
interesting structure in n(μ) in this regime ("Kondo peak") that has
occupied condensed matter physicists for many decades.

One could view Coulomb blockade as the bottom-up version of the Mott
transition a well-studied phenomenon in condensed matter physics. In a
long chain of atoms, the levels ε and ε+U₀ (Fig.23.1) will each broaden
into a band of width ~2t₀, t₀/ℏ being the rate at which electrons move
from one atomic site to the next. These are known as the lower and upper
Hubbard bands. If their separation U₀ exceeds the width 2t₀ of each band
we will have a Mott insulator where the electrochemical potential lies in
the middle of the two bands with very low density of states and hence
very low conductance. But if U₀ is small, then the two bands form a
single half-filled band with a high density of states at E = μ₀ and hence a
high conductance.

Needless to say, the full theory of the Hubbard bands is far more
complicated than this oversimplified description might imply and it is
one of the topics that has occupied condensed matter theorists for over
half a century. Since the late 1980's it has acquired an added significance
with the discovery of a new class of superconductors operating at relatively
high temperatures above 100K, whose
mechanism continues to be
controversial and hotly debated.

This problem remains one of the outstanding problems of condensed
matter theory, but there seems to be general agreement that the essential
physics involves a two-dimensional array of quantum dots with an inter-
dot coupling that is comparable to the single dot charging energy.
Lecture 24

The Quantum and the Classical

24.1. Spin coherence
24.2. Pseudo-spins
24.3. Quantum Entropy
24.4. Does Interaction Increase the Entropy?
24.5. Spins and magnets

Intel has a presentation, *From Sand to Circuits*, on their website [http://www.intel.com/about/companyinfo/museum/exhibits/sandtocircuits/index.htm](http://www.intel.com/about/companyinfo/museum/exhibits/sandtocircuits/index.htm) describing the amazing process that turns grains of sand into the chips that have enabled the modern information age. As I explained at the outset, these Lectures were about the physics that these “grains of sand” and the associated technology have helped illuminate in the last 25 years, the physics that helps validate the concept of an elastic resistor with a clean separation of entropy-driven processes from the force-driven ones.

Interestingly much of this physics does not involve the quantum aspects and can be understood within a semiclassical picture. Nearly all of what we discussed in Parts 1 and 2 of these lectures follow from the Boltzmann equation. Even though the modern nanotransistor that powers today’s laptops is only a few hundred atoms long, it remains in essence a classical device controlled more by electrostatics than by quantum subtleties. Indeed a recent paper reports classical behavior down to atomic scale conductors (Weber et al. 2012).

Is this inevitable? Or could we harness our impressive progress in nanofabrication and understanding to create a fundamentally different class of quantum devices that will take us beyond today’s charge-based information processing paradigms. Let me end these lectures with a few rambling thoughts (not answers!) on these lines.
24.1. Spin Coherence

The difference between quantum and classical is probably best exemplified by the spin of the electron. Consider for example, the experiment that we discussed in Lectures 14 and 22. An input magnet injects spins into the channel which produce a voltage on the output magnet given by

\[
\mu_p = \mu + \frac{\hat{P} \cdot \hat{n}}{2} \tag{24.1}
\]

Fig.24.1. An input magnet injects spins into the channel which produce a voltage on the output magnet that depends on the cosine of the angle between the two magnets.

In Lecture 22 we saw how the NEGF model gives a complex matrix $[G^r]$ whose (2x2) components at each point along the diagonal can be expressed in terms of the physically transparent concepts of electron and spin density, $N$ and $S$:

\[
\begin{bmatrix}
1up & 1dn & 2up & 2dn & 3up & 3dn \\
1up & 1dn & 2up & 2dn & 3up & 3dn \\
2up & 2dn & N_2, \vec{S}_2 \\
3up & 3dn & N_3, \vec{S}_3 \\
\end{bmatrix}
\]

\[
\rightarrow z \quad 1 \quad 2 \quad 3
\]
The NEGF equation for current then predicts the result in Eq. (24.1), with \( \mu \) and \( \bar{\rho} \), related to \( N \) and \( \bar{S} \) respectively.

This simple example illustrates the relation between quantum and classical descriptions. If the transverse components of spin are negligible then we can describe the physics in terms of \( N \) and \( S_z \) only. We could interpret the non-zero components on the diagonal

\[
\begin{bmatrix}
N + S_z & S_x - i S_y \\
S_x + i S_y & N - S_z
\end{bmatrix}
\]

(Same as Eq.(22.30))

\[
\frac{[G^n]}{2\pi} =
\begin{bmatrix}
N + S_z & S_x - i S_y \\
S_x + i S_y & N - S_z
\end{bmatrix}
\]

\( (N + S_z) \) as number of up electrons

\( (N - S_z) \) as number of down electrons

(per unit energy) and then write semiclassical equations for the two types of electrons. That is essentially what we did in Lecture 14.

When does this work? One possibility is that the magnets are all collinear and there is no spin-orbit coupling so that we are restricted to angles \( \theta \) that are multiples of 180 degrees (Fig.24.1). Another possibility is that various spin dephasing processes are strong enough to reduce transverse components of spin to negligible proportions. And if the z-components are reduced too, then we would not have to worry about spin at all.

What if we had collinear magnets but they point not along \( z \), but along \( x \)? Now the \( [G^n] \) matrix is not diagonal

\[
\frac{[G^n]}{2\pi} \rightarrow
\begin{bmatrix}
N & S_x \\
S_x & N
\end{bmatrix}
\]

and it might appear that a semiclassical description is not possible. The trick is to choose the coordinates or more generally the “basis” correctly. What we should do is to select a basis in which up and down point along \( +x \) and \( -x \) respectively so that in this basis \( [G^n] \) is diagonal.
In a word, we should call the direction of the magnet $z$ instead of $x$! This sounds like a trivial observation, but can provide interesting insights that may not be quite so obvious. Let me explain.

### 24.2. Pseudo-spin

One of the nice things about the formalism of spin matrices (Lecture 22) is that it goes way beyond spins; it applies to any two-component complex quantity. For example in Lectures 18 and 19 we talked about the graphene lattice where the unit cell has an “A” atom (on the lattice sites marked with a red circle) and a “B” atom (on the unmarked lattice sites).

The wavefunction in a unit cell is described by a two component complex quantity:

$$\{\psi\} = \begin{pmatrix} \psi_A \\ \psi_B \end{pmatrix}$$

and we could look at the corresponding $[G^n]$ and use our old relation from Eq.(22.30) to define a pseudo-spin

$$\frac{[G^n]}{2\pi} \rightarrow \begin{bmatrix} N + S_{up} & 0 \\ 0 & N + S_{dn} \end{bmatrix}$$

This has nothing to do with the real spin, just that they share the same mathematical framework. Once you have mastered the framework, there is no need to re-learn it, you can focus on the physics. In the literature on graphene, there are many references to pseudo spin and what direction it points in.
Lessons from Nanoelectronics

Let me point out a less familiar example of pseudospin involving an example we have already discussed. In Lecture 20, we discussed the potential variation across a single scatterer with transmission equal to $T$ (Fig.20.7). Let us just look at the diagonal elements of $\{G^\alpha\}$ for the same problem. There are oscillations on the left of the barrier with a constant density on the right. The reason Fig.20.7 shows oscillations on the right as well is that we were looking at the occupation obtained from $G^\alpha/A$ and $A$ has oscillations on the right. But let us not worry about that.

Let us see how we can use pseudospins to understand the spatial variation of the diagonal elements of $\{G^\alpha\}$.

Let us view positive and negative going states as the up and down components of a pseudospin. The pseudospinor wavefunction on the left and right of the barrier have the form

$$\left\{ \begin{array}{c} \psi \\ \bar{\psi} \end{array} \right\} \rightarrow \left\{ \begin{array}{c} e^{ikz} \\ re^{-ikz} \end{array} \right\}, \quad \left\{ \begin{array}{c} t e^{ikz} \\ 0 \end{array} \right\}$$

$$\left\{ \psi \right\} \left\{ \psi \right\}^+ \rightarrow \begin{bmatrix} 1 & r e^{i2kz} \\ r e^{-2ikz} & rr^* \end{bmatrix}, \quad \begin{bmatrix} t^* & 0 \\ 0 & 0 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} N+S_z & S_x - iS_y \\ S_x + iS_y & N-S_z \end{bmatrix}$$
This suggests that the pseudospins to the left of the barrier are described by (assuming \( r, t \) are real)

\[
\begin{align*}
N_{\text{Left}} &= \frac{(1+r^2)}{2} \\
N_{\text{Right}} &= \frac{t^2}{2} \\
S_z_{\text{Left}} &= \frac{(1-r^2)}{2} \\
S_z_{\text{Right}} &= \frac{t^2}{2} \\
S_x &= r \cos 2kz \\
S_y &= -r \sin 2kz
\end{align*}
\]

In other words, on the left of the barrier, the pseudospin is rotating round and round in the \( x-y \) plane. When we plot \( G^n(z,z) \), we are looking at the sum of the two pseudospin components and squaring the sum, which amounts to

\[
\text{Trace}\left\{1 \ 1\right\}\{\psi\}\{\psi\}^+ \left\{1 \ 1\right\} = \text{Trace}\left\{1 \ 1\right\}\{\psi\}\{\psi\}^+
\]

In effect we are using a pseudomagnet with \( \Gamma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \) which corresponds to one polarized 100\% along \( x \). So from Eq.(24.1), the measured potential should be proportional to

\[
N + \hat{x} \cdot \vec{S} \rightarrow \frac{1+r^2}{2} + r \cos 2kz, \quad \frac{t^2}{2}
\]

which describes the numerical results quite well.

This is a relatively familiar problem where the concept of pseudospin probably does not add much to our undergraduate understanding of one-dimensional standing waves. The purpose was really to add our understanding of pseudospins!
Lessons from Nanoelectronics

24.3 Quantum Entropy

Now that we have seen how “spins” appear everywhere, let us talk briefly about the information content of a single spin which as we discussed in Lecture 17 is related to the thermodynamic entropy. We talked about the entropy of two examples of a collection of $N$ spins obtained from the expression

$$ S = - \sum p_i \ln p_i $$

From a quantum mechanical point of view we could write the wavefunction of a single spin in collection A as

$$ \psi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \rightarrow \psi\psi^+ = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} $$

and interpret the diagonal elements of $\psi\psi^+$ ($1$ and $0$) as the $\{p_i\}$’s to use in Eq.(24.2). Writing $\psi\psi^+$ for a spin in collection B requires us to take a sum of two equally likely possibilities:

$$ \psi\psi^+ = 0.5 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 0.5 \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} $$

Once again we can interpret the diagonal elements of $\psi\psi^+$ (both $0.5$) as the $\{p_i\}$’s to use in Eq.(24.2) and get our semiclassical answers.

What if we have collection C, which looks just like collection A, but the spins all pointing along $x$ and not $z$, we then have

$$ \psi = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \rightarrow \psi\psi^+ = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} $$

If we just took the diagonal elements of $\psi\psi^+$ (both $0.5$) we
obtain the same answer as we got for collection B which is obviously wrong. A collection with all spins pointing along $x$ (C) should have the same entropy as a collection pointing along $z$ (A) rather than a random collection (B).

The correct answer is obtained if we first diagonalize $\psi\psi^+$ and then use its diagonal elements (which are the eigenvalues) as the $\{p_i\}$'s in Eq.(24.2). This is accomplished if we generalize Eq.(24.2) to write

$$\frac{S}{k} = -\text{Trace}[\rho \ln \rho] \quad (24.3)$$

where $\rho = \psi\psi^+$ is a 2x2 matrix (called the density matrix).

**24.3.1. How much information can one spin carry?**

Suppose we decide to use the spin of the electron, that is the direction of the input magnet in Fig.24.1 to convey information. It would seem that we could send large amounts of information, since there are now many possibilities. For example, suppose we choose a set of say 64 directions of the magnetization to convey information, it would seem that the entropy would be

$$\frac{S}{k} = \ln 64$$

Note that we are using 64 figuratively to represent the number of magnetization directions we use, which could just as well be 10 or 100.

We have seen in Lectures 14, 22 that a magnetic voltage probe making an angle $\theta$ with the injected spins measures a voltage proportional that depends on $\theta$ (Fig.24.1) and it would seem that we could measure the direction of spin simply by measuring the voltage. This would allow us to encode 64 possible values of $\theta$ thereby transmitting $\ln 64$ rather than $\ln 2$.

But how can this be correct? Didn’t we argue earlier that for one spin $\frac{S}{k} = \ln 2$ rather than $\ln 64$? These two arguments can be reconciled.
by noting that in order to measure a voltage that depends on $\theta$ we need many many electrons so that we can take their average. An individual electron would either transmit or not transmit into the magnet with specific probabilities that depend on $\theta$. Only by averaging over many electrons would we get the average values that we have discussed. This means that we could send $\ln 64$ worth of information, but only if we send many identically prepared electrons, so that the receiver can average over many measurements.

But couldn’t we take one electron that we receive and create many electrons with the same wavefunction? After all, we can always copy a classical bit of information. There is a “no cloning theorem” that says we cannot copy the quantum state of an electron. The sender has to send us identically prepared electrons if we want to make many measurements and average.

These concepts are of course part of the emerging field of quantum information on which much has been written and will be written. My point here is simply that compared to charge, spin represents a qualitatively different state variable that could be the basis for fundamentally different approaches to information processing.

24.4. Does interaction increase the entropy?

Back in Lecture 17 we discussed how a perfect anti-parallel (AP) spin valve could function like an info-battery (Fig. 17.3) that extracts energy from a collection of spins as it goes from the low entropy state A to the high entropy state B.
But exactly how does this increase in entropy occur? In Lecture 17 we described the interaction as a “chemical reaction”

\[ u + D \iff U + d \] (Same as Eq.(17.7))

where \( u, d \) represent up and down channel electrons, while \( U, D \) represent up and down localized spins.

From a microscopic point of view the exchange interaction creates a superposition of wavefunctions as sketched below:

\[
u \times D \xrightarrow{\text{\[ \begin{array}{c|c|c} \uparrow & \uparrow & \downarrow \\
\hline
\end{array}\]} \frac{1}{\sqrt{2}} u \times D + \frac{1}{\sqrt{2}} d \times U
\]

We have shown equal superposition of the two possibilities for simplicity, but in general the coefficients could be any two complex numbers whose squared magnitude adds up to one.

Now the point is that the superposition state

\[
\frac{1}{\sqrt{2}} u \times D + \frac{1}{\sqrt{2}} d \times U
\]

has zero entropy just like the original state \( u \times D \). Indeed we could picture a pseudo-spin whose up and down components are \( u \times D \) and \( d \times U \). The interaction merely rotates the pseudo-spin from the \( z \) to the \( x \) direction and as we discussed in the last Section, mere rotation of spins or pseudo-spins causes no generation of entropy.

So how does the increase in entropy occur? The itinerant electron eventually gets extracted from the channel. At that moment there is a
“collapse of the wavefunction” either into a $u \times D$ or a $d \times U$ depending on whether the channel electron is extracted by the source as an up electron or by the drain as a down electron. The localized spin is left behind in a down or an up state with 50% probability each. This is when the entropy increases by $k \ln 2$.

24.5. Spins and Magnets

Spins have long been viewed as the quintessential quantum object and we have seen how we are learning to use solid state devices to control and manipulate not just single spins but even more sophisticated quantum objects like entangled spins. We have also seen that diverse quantum objects can all be viewed as two-component spins.

Actually this analogy need not be limited to two-component objects. As we mentioned earlier, the NEGF model gives a complex matrix $[G^\alpha]$ whose (2x2) components at each point along the diagonal can be expressed in terms of the physically transparent concepts of electron and spin density, $N$ and $\vec{S}$:

$$
\begin{array}{cccc}
1_{up} & 1_{dn} & 2_{up} & 2_{dn} & 3_{up} & 3_{dn} \\
1_{up} & \begin{bmatrix} N_1, \vec{S}_1 \end{bmatrix} & 2_{up} & 2_{dn} & 3_{up} & 3_{dn} \\
2_{up} & 2_{dn} & \begin{bmatrix} N_2, \vec{S}_2 \end{bmatrix} & 1 & 2 \rightarrow \vec{z} \\
3_{up} & 3_{dn} & \begin{bmatrix} N_3, \vec{S}_3 \end{bmatrix} & \end{array}
$$

We could extend this viewpoint beyond the spin component. One could envision say a 10x10 block involving say multiple orbitals that retains its phase coherence for relatively long lengths of time and treat it as a kind of giant quasi spin-like object. Of course we could not describe it with just four numbers $N, S_x, S_y, S_z$. We would need 100 real components to
represent all the information contained in a 10x10 Hermitian matrix and could use the NEGF equations to come up with a diffusion equation for this 100-component object just as we discussed earlier for spin diffusion (Eq.(22.37)).

Could we use these giant quasi spins to encode and transmit information? Perhaps, but there is an important lesson the last twenty years have taught us. I am talking about the integration of what used to be two distinct fields, namely spintronics and magnetics.

Even ten years ago there was little overlap between these two fields. Spintronics was about the manipulation of individual spins motivated largely by basic low temperature physics, while magnetics was motivated largely by practical room temperature applications involving the manipulation of classical magnets whose magnetism is a result of enormous internal interactions that make all the elementary magnets (or spins) line up.

The first step in the integration of these two fields was the use of magnetic contacts to inject spins leading to the spin valve and eventually the magnetic tunnel junctions (MTJ’s) whose change in resistance ($R_p$ to $R_{AP}$, See Fig.14.1) is now routinely used to read information stored in magnets.

The second step was the demonstration of spin-torque which allows injected spins to turn magnets provided they are not more than a few atomic layers in thickness. This effect could be used to write information into nanomagnets. Memory devices utilizing spin for both reading and writing are now being actively developed.

The point I want to make is that for information processing it is not enough to have a spin, we also need a magnet. In standard charge-based architectures information is stored in capacitors and transmitted from capacitor to capacitor. Similarly we need a magnet to implement a spin capacitor and devices to transmit the information from magnet to
magnet. Developments in the last decade have given us the basic ingredients. Whether we can build a information processing technology around it, remains to be seen.

It is tempting to go beyond simple spins and look at all kinds of exotic two-component pseudo-spins or multi-component quasi-spins that maintain phase coherence over useful lengths of time. But it seems to me that a key question one should ask is, “do we have a quasi-magnet to generate and detect the quasi-spin?”

These are the kinds of questions that will probably be explored and answered in the coming years. The purpose of these notes was to try to convey the insights from the last thirty years drawn from the remarkable experiments made possible by the advances in nanoelectronics. These experiments have helped establish the elastic resistor as a useful approximation to short conductors, and we have tried to show that it can also be used to understand long conductors by viewing them approximately as combinations of elastic resistors.

As I stated in the introductory lecture, I believe that the results we have discussed all follow directly from the formal theories of semiclassical or quantum transport. The real value of the bottom-up approach based on elastic resistors is in improving our physical understanding, and this will hopefully facilitate the insights needed to take us to the next level of understanding, discovery and innovation.

The last thirty years have taught us the significance of contacts, even the most mundane ones. We have also seen some of the interesting things that can be achieved with “smarter” contacts like the magnetic ones. Perhaps in the coming years we will learn to manipulate and control them creatively so that they become an active part of information processors.