Quantum Matter: Coherence And Correlations

Henrik Johannesson   Jari Kinaret

March 26, 2007

Contents

Preface

This is work in preparation. There are undoubtedly many typos, and the notations between the different chapters may not be fully uniform, but hopefully the notes will improve during the course and the coming years. The document will be updated as errors are detected and corrected.

We welcome your comments on explanations that you find confusing or particularly clear, trivial or incomprehensible.

Chapter 1

Quantum Coherence in Condensed Matter

1.1 Effects of Phase Coherence

1.1.1 Resonant Tunneling

Consider a double barrier structure comprising two scattering centers in an otherwise clean one-dimensional structure as shown in Figure 1.1. Classically the transmission probability through the structure can be obtained from the probability of transmission (T_j) and reflection (R_j) for each barrier ($T_j + R_j = 1$) by accounting for multiple reflections between barriers as

$$T = T_1(1 + R_2R_1 + R_2R_1R_2R_1 + \ldots)T_2 = \frac{T_1T_2}{1 - R_1R_2} = \frac{T_1T_2}{T_1 + T_2 - T_1T_2} \rightarrow_{T_2=T_1} \frac{T_1}{2 - T_1}$$

which is always less than T_1, *i.e.* it is harder to get through two barriers than one. Hardly surprising. What may be surprising at first is that the total transmission probability is not proportional to T_1^2 but to only T_1 for small transmission. The reason is that if the particle gets through the first barrier, it bounces several times between the barriers and hence has many attempts to get through the second barrier; in the limit of small T_1 a particle is almost as likely to exit to the right as it is to exit to the left (once it has passed the first barrier), hence $T \approx T_1/2$ as our calculation shows.

In quantum mechanics we cannot sum the probabilities for different trajectories but, instead, we have to sum the transmission amplitudes. During each passage between the barriers

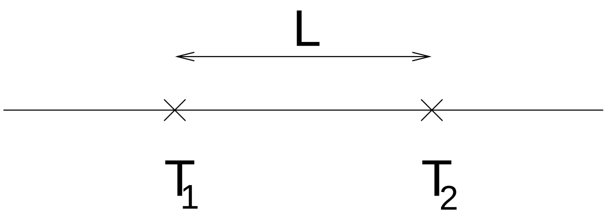

Figure 1.1: A double barrier structure comprising two barriers with transmission probabilities T_1 and T_2, separated by distance L.

the wave function accumulates phase kL, which yields the total transmission amplitude

$$t = t_1 e^{ikL}(1 + e^{2ikL} r_2 r_1 + e^{4ikL} r_2 r_1 r_2 r_1 + \ldots) t_2 = \frac{e^{ikL} t_1 t_2}{1 - e^{2ikL} r_1 r_2}$$

so that the transmission probability is given by

$$T = \frac{|t_1 t_2|^2}{|1 - e^{2ikL} r_1 r_2|^2} \rightarrow_{r_2 = r_1, t_2 = t_1} \frac{|t_1|^4}{|1 - e^{2ikL} r_1^2|^2}.$$

We now write $r_1 = \sqrt{R_1} e^{i\phi_R}$ and $t_1 = \sqrt{T_1} e^{i\phi_T}$ where $T_1 + R_1 = 1$ to get

$$T = \frac{T_1^2}{(1 - \cos(2kL + 2\phi_R) R_1)^2 + \sin^2(2kL + 2\phi_R) R_1^2} = \frac{T_1^2}{T_1^2 + 4(1 - T_1)\sin^2(kL + \phi_R)}.$$

Now it is not true that $T < T_1$, and in particular if $kL + \phi_R = n\pi$, $n \in \mathbf{Z}$, $T = 1$ regardless of the transparency of an individual barrier: the particle is transmitted with unit probability regardless of how reflective the individual barriers are. This is the well-known phenomenon of resonant tunneling. Near a resonance, i.e. for $kL \approx n\pi - \phi_R$, we get

$$T \approx \frac{T_1^2}{T_1^2 + 4(1 - T_1)[kL - (n\pi - \phi_R)]^2} = \frac{1}{1 + 4\frac{1 - T_1}{T_1^2}[kL - (n\pi - \phi_R)]^2}$$

showing the Lorentzian line shape of the resonance. It is important to notice that the resonance width is given by $1 = 4\frac{1-T_1}{T_1^2}[kL - (n\pi - \phi_R)]^2$ or $[kL - (n\pi - \phi_R)] = \pm\frac{T_1}{2\sqrt{1-T_1}}$ implying that for low-transparency barriers the resonances are very narrow, and in the limit of vanishing barrier transparency the resonance becomes a δ-peak.

Two features are responsible for the difference between the classical and the quantum results: (i) in quantum mechanics we had to add the amplitudes of different ways of arriving at the same final state, and (ii) the phase accumulated during propagation between the barriers is always kL. The first of these is carved in the stone that Bohr & Co. brought down the mountain, but the second is an assumption.

Let us now relax the assumption and postulate that, due to some unknown process, the phase changes by an amount θ_j during the j^{th} round trip between the barriers. We then have

$$t = t_1 e^{ikL + \theta_0/2}(1 + e^{2ikL + \theta_1} r_2 r_1 + e^{4ikL + \theta_1 + \theta_2} r_2 r_1 r_2 r_1 + \ldots) t_2$$

so that

$$\begin{aligned}
& T(\{\theta_j\}_{j=1}^{\infty}) \\
= \; & |t_1|^4 \left|1 + e^{2ikL + \theta_1} r_1^2 + e^{4ikL + \theta_1 + \theta_2} r_1^4 + \ldots\right|^2 \\
= \; & T_1^2 \left|1 + e^{2ikL + \theta_1 + 2\phi_R} R_1 + e^{4ikL + \theta_1 + \theta_2 + 4\phi_R} R_1^2 + \ldots\right|^2
\end{aligned}$$

or

$$\begin{aligned}
T = \; & T_1^2 \left|\textstyle\sum_{n=0}^{\infty} e^{i2n(kL + \phi_R) + i\sum_{j=1}^{n}\theta_j} R_1^n\right|^2 \\
= \; & T_1^2 \textstyle\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} e^{i2(n-m)(kL - \phi_R) + i\sum_{j=1}^{n}\theta_j - i\sum_{j=1}^{m}\theta_j} R_1^{n+m} \\
= \; & T_1^2 \left[\textstyle\sum_{n=0}^{\infty} R_1^{2n} + 2\mathrm{Re}\sum_{n=0}^{\infty}\sum_{m=0}^{n-1} e^{i2(n-m)(kL - \phi_R) + i\sum_{j=m+1}^{n}\theta_j} R_1^{n+m}\right]
\end{aligned}$$

where the second term describes interference between different paths. Let us now assume that all the phase changes θ_j are independent random variables that follow a Gaussian distribution

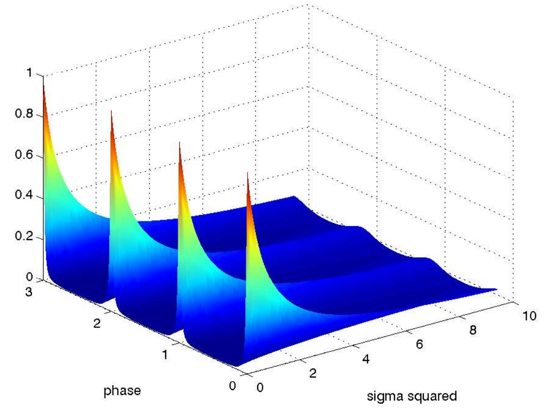

Figure 1.2: Transmission as a function of $2(kL - \phi_R)$ and σ^2 for $T_1 = 0.1$.

$P(\theta)$ with standard deviation σ. We can then obtain the average transmission probability $\langle T \rangle$ for a collection (ensemble) of such systems by

$$\langle T \rangle = \int_{-\infty}^{\infty} \prod_{j=1}^{\infty} d\theta_j P(\theta_j) T(\{\theta_j\}_{j=1}^{\infty}) = T_1^2 \left[\sum_{n=0}^{\infty} R_1^{2n} + 2\mathrm{Re} \sum_{n=0}^{\infty} \sum_{m=0}^{n-1} e^{i2(n-m)(kL-\phi_R)-\frac{1}{2}(n-m)\sigma^2} R_1^{n+m} \right]$$

where I used $\int_{-\infty}^{\infty} d\theta\, P(\theta) e^{i\theta} = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} d\theta\, e^{-\theta^2/(2\sigma^2)} e^{i\theta} = e^{-\sigma^2/2}$. This result shows that the interference terms are reduced by the phase fluctuations: the exponent has acquired the part $-(n-m)\sigma^2/2$ the dampens the off-diagonal $n \neq m$ terms. Proceeding further by evaluating the geometric sums we get

$$\begin{aligned} \langle T \rangle &= T_1^2 \left[\frac{1}{1-R_1^2} + 2\mathrm{Re} \sum_{n=0}^{\infty} e^{i2n(kL-\phi_R)-n\frac{1}{2}\sigma^2} R_1^n \frac{1-e^{-i2n(kL-\phi_R)+n\frac{1}{2}\sigma^2} R_1^n}{1-e^{-i2(kL-\phi_R)+\frac{1}{2}\sigma^2} R_1} \right] \\ &= T_1^2 \left[\frac{1}{1-R_1^2} \frac{e^{\sigma^2} R_1^2 - 1}{e^{\sigma^2} R_1^2 + 1 - 2e^{\frac{1}{2}\sigma^2} \cos(2(kL-\phi_R)) R_1} + 2\mathrm{Re} \frac{1}{1+R_1^2 - e^{-i2(kL-\phi_R)+\frac{1}{2}\sigma^2} R_1 - e^{i2(kL-\phi_R)-\frac{1}{2}\sigma^2} R_1} \right] \end{aligned}$$

Since there were ample possibilities for errors in this lengthy calculation, let us check the two limits we know: the classical result and deterministic quantum case. The classical limit is obtained by completely smearing out the phase information by setting $\sigma^2 \to \infty$, which yields

$$\langle T \rangle_{\sigma \to \infty} = \frac{T_1^2}{1 - R_1^2} = \frac{T_1}{2 - T_1}$$

as above, and the deterministic phase evolution is obtained by setting $\sigma^2 = 0$, which yields

$$\langle T \rangle_{\sigma=0} = \frac{T_1^2}{1 + R_1^2 - 2\cos(2(kL - \phi_R)) R_1}$$

also in agreement with the previous result. For intermediate σ^2 the result is between the classical and quantum limits as shown in Figure 1.2.

Thus, we have concluded that if the phase accumulated during propagation fluctuates randomly, resonant tunneling (and other interference phenomena *e.g.* in optics) are suppressed. How much phase fluctuations can be tolerated, then? Let us consider the resonant case $kL = \phi_R$ so that

$$\begin{aligned}\langle T\rangle = &\ T_1^2\left[\frac{1}{1-R_1^2}\frac{e^{\sigma^2}R_1^2-1}{e^{\sigma^2}R_1^2+1-2e^{\frac{1}{2}\sigma^2}R_1}+2\frac{1}{1+R_1^2-e^{\frac{1}{2}\sigma^2}R_1-e^{-\frac{1}{2}\sigma^2}R_1}\right]\\ = &\ \frac{e^{\frac{1}{2}\sigma^2}+R_1}{e^{\frac{1}{2}\sigma^2}-R_1}\frac{1-R_1}{1+R_1}\end{aligned}$$

The relevant scale for importance of phase fluctuations is seen by solving $T = \frac{1}{R_1+1}$ which is the average value of the classical ($T = \frac{1-R_1}{1+R_1}$) and resonant ($T = 1$) transmissions. This yields $\frac{1}{2}\sigma^2 = \ln(1+T_1)$ For small barrier transparency this gives $\frac{1}{2}\sigma^2 \approx T_1$, which can be understood by noticing that, classically, the average number of times the particle must impinge upon the second barrier before it succeeds in getting through is $1/T_1$, so that the total variance of its phase upon exit is $(1/T_1)\sigma^2$; hence, if $\sigma^2 \sim T_1$, the total variance upon exit is a number independent of T_1, making it reasonable that resonant transmission can be seen if $\sigma^2 \ll T_1$ while if $\sigma^2 \gg T_1$, resonant transmission is destroyed.

Phase fluctuations result in a randomization of the phase of the wave function. If a state that has energy ϵ and a well-defined phase at time $t = 0$ undergoes phase fluctuations, the phase measured after time δt has a probability distribution centered around $\phi = \epsilon t/\hbar$. The width of the distribution increases with δt, and after some time the width has increased to 2π, meaning that the phase has become completely uncertain (phases that differ by a multiple of 2π are indistinguishable). The time it takes to reach this situation is called the phase breaking time τ_ϕ. An alternative definition of the phase breaking time is based on evaluating the average phase factor $\langle e^{i\phi(t)}\rangle$ as a function of time: at $t > 0$ the magnitude of the average phase factor decreases exponentially with time, and the phase breaking time can be defined through $|\langle e^{i\phi(t)}\rangle| \sim e^{-t/\tau_\phi}$. Both definitions yield the same result within factors of order unity.[1] In order for resonant transmission to survive phase fluctuations, the above analysis shows that, at small barrier transparencies, phase breaking time must be at least $\tau_\phi \gtrsim \frac{2L}{vT_1}$ where v is the velocity of the particle between the barriers; faster de-phasing results in a transmission that is far below unity.

1.1.2 Persistent Currents

We will now consider a collection of independent electrons confined to a small ring-shaped conductor of radius R. The ring is pierced by a magnetic field that is entirely confined in the inside of the ring and does not penetrate into the conductor at all. Hence, classically the electrons are completely unaffected by the magnetic field since the field strength $\mathbf{B}$ vanishes inside the conductor. Quantum mechanically, however, the fundamental quantity is the vector potential $\mathbf{A}$, which does not vanish inside the conductor, and may therefore influence the electrons.

[1] Technically, one should distinguish between the energy relaxation time τ_E and the phase breaking time τ_ϕ. The former describes how quickly an excited state such as an electron above the Fermi surface loses its energy, while the latter describes how quickly the phase of the wave function describing the excitation becomes randomized. Usually the two times are roughly equal but if energy relaxation occurs through a series of scattering events all involving small energy transfers, the phase memory may be lost before energy is fully relaxed.

The system is described by the Schrödinger equation

$$\left[\frac{1}{2m}\left[-i\hbar\nabla + e\mathbf{A}(\mathbf{r})\right]^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r})$$

where $\nabla\times\mathbf{A} = \mathbf{B}$ and $V(\mathbf{r})$ is the potential that confines electrons in the ring. In the simplest case the ring is infinitesimally thin both in the radial direction and the z-direction so that $V(\mathbf{r}) = V_0\delta(r-R)\delta(z)$, $V_0 < 0$. In this case the only degree of freedom is the angular position φ along the ring, and the Schrödinger equation simplifies to

$$\frac{1}{2m}\left[-i\frac{\hbar}{R}\frac{d}{d\varphi} + eA(\varphi)\right]^2\psi(\varphi) = \epsilon\psi(\varphi)$$

where we assumed that the magnetic field is in the z-direction and cylindrically symmetric about the origin $r = 0$. To be specific, let us consider a magnetic field that is non-zero only for $r = 0$ and has a total magnetic flux Φ. We then have $\Phi = \int_\Omega d^2r\nabla\times\mathbf{A}(\mathbf{r}) = \oint_{\partial\Omega} d\ell\cdot\mathbf{A}(\mathbf{r})$ where Ω is an arbitrary region surrounding the origin and $\partial\Omega$ is its boundary. Let us specialize to a circular region with radius ρ so that $\Phi = \rho\int_0^{2\pi} d\varphi\hat{\varphi}\cdot\mathbf{A}(\mathbf{r})$. By cylindrical symmetry the integrand must be independent of φ so that we have $\Phi = 2\pi\rho\hat{\varphi}\cdot\mathbf{A}(\rho,\varphi)$ or $\mathbf{A}(\rho,\varphi) = \frac{\Phi}{2\pi\rho}\hat{\varphi}$. Hence, within the conductor the vector potential is given by $\mathbf{A}(R,\varphi) = \frac{\Phi}{2\pi R}\hat{\varphi}$, and the Schrödinger equation becomes

$$\frac{\hbar^2}{2mR^2}\left[-i\frac{d}{d\varphi} + \frac{\Phi}{\Phi_0}\right]^2\psi(\varphi) = \epsilon\psi(\varphi)$$

where $\Phi_0 = h/e$ is the magnetic flux quantum. Since the variable φ does not appear in the Hamiltonian, the Hamiltonian commutes with the angular momentum operator, and the eigenfunction $\psi(\varphi)$ can be written as $\frac{1}{\sqrt{2\pi}}e^{i\ell\varphi}$. Substituting in the equation we get $\epsilon_\ell = \frac{\hbar^2}{2mR^2}\left[\ell + \frac{\Phi}{\Phi_0}\right]^2$. Since the wave function must be single valued as $\varphi \to \varphi + 2\pi$, the angular momentum quantum number ℓ must be an integer.

Thus, the single particle levels are characterized by an integer valued quantum number ℓ, which is connected to the angular momentum in the z-direction through $L_z = \hbar\ell$, and have energies $\epsilon_\ell = \frac{\hbar^2}{2mR^2}\left[\ell + \frac{\Phi}{\Phi_0}\right]^2$. At temperature $T = 0$ sufficiently many states with lowest energies are occupied that all electrons can be accommodated. To see what implications this has let us first consider spinless electrons (fictitious particles that are identical to electrons except that they have no spin). Then each ℓ-state can accommodate one electron. If the number of electrons in the ring is odd, $N = 2M + 1$, then at zero penetrating flux $\Phi = 0$ the states with $\ell = -M$ to $\ell = M$ are occupied, and the two highest occupied states are degenerate. If Φ is now increased from zero, the energies of states with negative ℓ are reduced while those of states with positive ℓ are increased, and for $\Phi = \Phi_0/2$ it becomes energetically favorable to occupy the state $\ell = -(M+1)$ rather than the state $\ell = +M$, meaning that at this value of the magnetic flux the total angular momentum increases from $L_z = 0$ to $L_z = -(2M+1)\hbar = -N\hbar$. At $\Phi = 3\Phi_0/2$ another increase takes place and L_z becomes $-2N\hbar$ so that in general $L_z = -[\Phi/\Phi_0 + 1/2]N\hbar$ where $[z]$ is the least integer not greater than z. For an even number of particles $N = 2M$ at $\Phi = 0$ the states from $\ell = -(M-1)$ to $\ell = (M-1)$ are occupied, plus either one of the states $\ell = \pm M$, and the total angular momentum is seen to be $L_z = -M\hbar = -(N/2)\hbar$ for $0 < \Phi < \Phi_0$, so that $L_z = -[\Phi/\Phi_0]N\hbar - (N/2)\hbar$.

The total angular momentum is not directly observable. However, the velocity associated with the state $|\ell\rangle$ is $\hbar^{-1}R\partial_\ell\epsilon_\ell = \frac{\hbar}{mR}\left[\ell + \frac{\Phi}{\Phi_0}\right]$, and the current carried by a single electron traveling with velocity v is $-ev/(2\pi R)$, so that the total current carried by a collection of electrons with the total angular momentum L_z is

$$I = -e\frac{1}{2\pi R}\frac{\hbar}{mR}\left(\hbar^{-1}L_z + N\frac{\Phi}{\Phi_0}\right) = -Ne\frac{\hbar}{2\pi mR^2}\left(\frac{L_z}{N\hbar} + \frac{\Phi}{\Phi_0}\right).$$

This current can, in principle, be measured: the current in the ring causes a magnetic flux through the loop (*cf.* an electromagnet), which slightly changes the total magnetic field from the externally applied one. In practice, however, the change is so small that the measurements are extremely challenging. Substituting the total angular momentum L_z obtained above shows that the quantity in the parentheses varies between $-\frac{1}{2}$ and $+\frac{1}{2}$, so that the maximum persistent current is $\frac{N}{2\pi R}\frac{e\hbar}{mR}$. This can be written as ev_F/L where v_F is the Fermi velocity (velocity of highest occupied state at $\mathbf{B} = 0$) and $L = 2\pi R$ is the circumference of the ring. Hence, the total current is effectively given by the last occupied state as contributions from the other states cancel out.

Persistent currents are purely a quantum mechanical phenomenon even though they resemble ordinary diamagnetic or paramagnetic response — even in the incoherent regime surface currents arise as a respond to an external magnetic field. The difference is two-fold: persistent currents emerge even if the magnetic field inside the conductor vanishes as long as the electronic states encircle a magnetic flux (classical response is proportional to magnetic field in the conductor), and the magnitude of the current is inversely proportional to the system size L. Persistent currents are also fundamentally different from currents that appear as a response to external electric field. The conductive currents represent a balance between the external fields that tend to excite electrons to states with higher energies, and scattering mechanisms that tend to relax the electron distribution towards a thermal equilibrium. Persistent currents, in contrast, are a property of the ground state and exist even in a thermal equilibrium. Equilibrium currents can only exist if the time reversal invariance is broken: time reversal invariance implies that time-reversed states are degenerate and hence occupied equally in an equilibrium, and since time-reversed states carry opposite currents, the net current in equilibrium vanishes.

A crucial requirement for the appearance of persistent currents is that the phase coherence of the wave function is maintained around the loop — if the state acquires an undetermined phase during propagation around the loop, there is no reason for ℓ to be integer, and there is no response to the external flux. Consequently, the effect can only be seen in relatively small rings and at low temperatures.

Home problem 1: Persistent currents
Consider a system of noninteracting electrons, *i.e.* spin-$\frac{1}{2}$ particles with charge $-e$.

1. Sketch the persistent current at zero temperature as a function of the magnetic flux for systems with $4M$, $4M+1$, $4M+2$ and fbox $4M+3$ electrons.

2. Estimate the temperature requirement to observe persistent current. Can you say something about how the current behaves as a function of temperature?

1.2 Coherent transport

Let us now investigate transport in the phase-coherent regime in more detail, and pay particular attention to low-dimensional systems.

1.2.1 Landauer-Büttiker formalism

Consider an impurity-free one-dimensional wire, *i.e.* a wire that is so narrow that only one transverse mode is occupied: current between two reservoirs with chemical potentials μ_L and μ_R is given by (charge) × (density) × (velocity), which yields

$$\begin{aligned}
I &= -e\int_{-\infty}^{\infty} d\epsilon\, D(\epsilon)[f(\epsilon-\mu_L)-f(\epsilon-\mu_R)]v(\epsilon) \\
&= -e\int_{-\infty}^{\infty} d\epsilon\, \tfrac{1}{2\pi}\tfrac{dk}{d\epsilon}[f(\epsilon-\mu_L)-f(\epsilon-\mu_R)]\tfrac{1}{\hbar}\tfrac{d\epsilon}{dk} \\
&= -e\tfrac{1}{h}\int_{-\infty}^{\infty} d\epsilon\, [f(\epsilon-\mu_L)-f(\epsilon-\mu_R)] \\
&= -e\tfrac{1}{h}\int_{-\infty}^{\infty} d\epsilon\, [f(\epsilon-\mu-\tfrac{1}{2}eV)-f(\epsilon-\mu+\tfrac{1}{2}eV)]
\end{aligned}$$

and at low temperatures $f(\epsilon) \approx \Theta(-\epsilon)$ giving

$$\frac{dI}{dV} = \frac{e^2}{h}$$

which is the quantum conductance. The inverse quantum conductance $R_K = h/e^2$ is known as the von Klitzing constant, or quantum resistance, and roughly equal to 26 kΩ. This conductance value plays an important role in many small devices that often behave qualitatively differently depending on whether some device resistances are smaller or larger than the quantum value.

A crucial ingredient of the derivation was the cancelation between the (directional) density of states $D(\epsilon) = \frac{1}{2\pi}\frac{dk}{d\epsilon}$ and the velocity $\frac{1}{\hbar}\frac{d\epsilon}{dk}$. If the wire has width W, the energy of the n^{th} transverse mode is $\frac{\hbar^2}{2m}\frac{\pi^2}{W^2}n^2$, so if the Fermi energy exceeds $4\frac{\hbar^2}{2m}\frac{\pi^2}{W^2}$, two transverse modes are occupied *etc.*. In the absence of scattering, each transverse mode contributes independently to the current, and if N transverse modes are occupied, the conductance is therefore $G_N = N\frac{e^2}{h}$. Allowing for two spin states, the conductance becomes $G_N = 2N\frac{e^2}{h}$.

Hence, we find that a one-dimensional conductor has a finite conductance at zero temperature even if there are no impurities. This is in a remarkable contradiction with conventional wisdom that states that the conductance of a pure metal diverges at zero temperature! What about disordered wires where an electron entering the wire has transmission amplitude t, $|t|^2 \leq 1$, to get through the wire? We can analyze this case quite easily by noticing that those electrons whose energies lie below both Fermi levels μ_L and μ_R do not contribute to a net current through the wire — there are equally many electrons moving in both directions — and for energies $\mu_L > \epsilon > \mu_R$ there are only electrons entering from the left; of these, only the ones that emerge out of the wire on the right contribute to the net current, implying that

$$I = -e\frac{1}{h}\int_{\mu_R}^{\mu_L} d\epsilon\, |t(\epsilon)|^2 = |t|^2\frac{e^2}{h}V$$

where I assumed that the transmission amplitude is only weakly dependent on the energy on the relevant energy range. The result is intuitively appealing: a reduced probability for transmission results in a reduced conductance. This result is known as the Landauer formula for conductance after the late Rolf Landauer, and is often used to calculate conductances of quantum mechanically coherent structures.

The expression can be generalized, firstly, to the case of many modes in a two-terminal wire. Then electrons entering the wire in mode n may be scattered into mode n', and either be transmitted through the wire with amplitude $t_{n'n}$ or be reflected with amplitude $r_{n'n}$. The resulting conductance for the wire can be written as

$$G = \mathrm{T}r(t^\dagger t)\frac{e^2}{h}$$

(proof left as a home problem). A generalization to a more complicated conductor with many current terminals and voltage probes is also straightforward: At current terminals we specify the chemical potentials, and at voltage probes the chemical potential is adjusted so that the net current through the voltage probe is zero (ideal volt meter). The resulting expressions for current and voltages, which only require knowledge of the different transmission and reflection amplitudes, are occasionally lengthy but the underlying physics is quite simple. These generalizations of the Landauer formula were first considered by Markus Büttiker in the 1980s, and are known as Landauer-Büttiker formalism.

While the physics of the Landauer-Büttiker formalism looks simple in the formalism, there are a number of subtleties. Firstly, the result goes against "common knowledge", and initially the formalism was met with a great deal of scepticism. Alternative derivations using slightly different arguments were presented, with a final result that assumed the form $G = \frac{|t|^2}{1-|t|^2}\frac{e^2}{h}$ that agrees with the above result for small transmissions but reproduces the classical divergent result as $|t|^2 \to 1$. The debate was only resolved once Daniel Fisher and Patrick Lee in 1981 derived the Landauer result using a standard field theoretical technique that was significantly more complicated but also more acceptable to physicists at large. Somewhat later it was realized that the difference between $|t|^2$ and $|t|^2/(1-|t|^2)$ corresponds to two different experiments: the first result is obtained if the wire is connected to two reservoirs with fixed chemical potentials, and the second result corresponds to a contactless measurement. Since one typically measures conductance with the first method, the counterintuitive result is usually the appropriate one: it essentially states that the smallest theoretically achievable contact resistance between a single mode quantum wire is $e^2/(2h)$, so a wire with two ends has a minimal resistance of e^2/h.

The second subtlety has to do with dissipation. Resistance implies dissipation, *i.e.* that energy is transferred from the electron system to something else, and it is hard to see where and how such a dissipation can occur in a clean wire. This problem can be resolved by considering a three terminal configuration where a clean wire connects the left terminal L to a voltage probe VP and a clean wire connects the voltage probe to the right terminal R. If the transmission probabilities between L and VP, and between VP and R, are equal to one, then the condition that the net current in the voltage probe vanishes implies that the the chemical potential of the voltage probe must equal the average of the chemical potentials μ_L and μ_R, independent of where the voltage probe is located. Consequently, the interior of the wire is at a constant potential, the internal resistance of the wire is zero, and the entire voltage drop occurs at the end of the wire, in accordance with the contact resistance interpretation given above. Hence, dissipation occurs at the contacts between the wire and the reservoir (typically slightly on the reservoir side). Since in the reservoirs there are many degrees of freedom with a dense energy spectrum, energy can easily be redistributed between them, and the conceptual problem with dissipation disappears.

1.2.2 Integer Quantum Hall Effect

A particular system where the Landauer-Büttiker formalism can be applied very easily and successfully is a two-dimensional electron gas in a strong perpendicular magnetic field. This system has proven to exhibit very rich physics, and has thus far resulted in two Noble prizes in Physics. For now we will focus on the Integer Quantum Hall Effect (IQHE) that can be understood without considering the effects of electron-electron interactions (hence "two-dimensional electron *gas*" — technically, *gas* implies a non-interacting system). The Fractional Quantum Hall Effect (FQHE) that takes place in a two-dimensional electron *liquid* at higher magnetic fields is discussed in the chapter on the joined effects of coherence and interactions.[2]

Our starting point is the Schrödinger equation for electrons confined to a plane and subjected to a magnetic field. The equation reads

$$\frac{1}{2m}(-i\hbar\nabla + e\mathbf{A})^2\psi(\mathbf{r}) = E\psi(\mathbf{r})$$

where $\mathbf{A}$ is a vector potential associated with the magnetic field $\mathbf{B} = B\hat{\mathbf{z}}$ through $\mathbf{B} = \nabla \times \mathbf{A}$. There are many choices of the vector potential that give rise to the same magnetic field (many gauges), and for our present purposes the choice $\mathbf{A} = Bx\hat{\mathbf{y}}$ is the most convenient (transverse gauge). Inserting this to the Schrödinger equation yields

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + 2i\frac{eB}{\hbar}x\frac{\partial}{\partial y}\right]\psi(x,y) + \frac{1}{2}m\omega_c^2 x^2\psi(x,y) = E\psi(x,y)$$

where $\omega_c = \frac{eB}{m}$ is the cyclotron frequency. The Hamiltonian is seen to commute with ∂_y, implying that the wave functions can be chosen to be eigenfunctions of the momentum in the y-direction, and written in the form $\psi(x,y) = e^{iky}u_k(x)$. The remaining function $u_k(x)$ can be solved from

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}u_k(x) + \frac{1}{2}m\omega_c^2(x + \frac{\hbar k}{eB})^2 u_k(x) = E u_k(x).$$

This is recognized as a Schrödinger equation for a harmonic oscillator with frequency ω_c and center at $x_0(k) = -\frac{\hbar k}{eB} = -k\ell_c^2$, where $\ell_c = \sqrt{\frac{\hbar}{eB}}$ is the magnetic length. Hence, the energy

[2] Experimentally the two effects are seen in similar, often the same, system. The distinction between electron gas or electron liquid only refers to the importance of electron-electron interactions in the explanations of the experimentally observed effects.

spectrum of electrons in a magnetic field is given by $E_{nk} = (n + \frac{1}{2})\hbar\omega_c$ which is independent of the quantum number k that determines both the momentum in the y-direction and the center of the wave function in the x-direction. The wave functions may be visualized as spaghetti centered at different points in the x-direction and running along the y-direction. The energy of each strand of spaghetti is independent of its position in the sample and is entirely given by the quantum number n. States with the same n are degenerate and form a so-called Landau level. The number of states in a Landau level can be determined by considering periodic boundary conditions in the y-direction, which implies a finite spacing $\Delta k = 2\pi/L$ between the allowed k-values, and requiring that the center of the state falls within the sample in the x-direction. The end result is that the degeneracy of a Landau level is AB/Φ_0 where AB is the magnetic flux through a sample of area A and $\Phi_0 = h/e$ is the magnetic flux quantum. For a typical experimental sample the area is about $10^{-4}m^2$ so that at a magnetic field of one tesla the Landau level degeneracy is roughly 2.5×10^{10}.

When the chemical potential lies in the gap between two Landau levels, those levels that are below μ are full and levels above μ are empty. The density of electrons in the sample is therefore NB/Φ_0. We know from simple electron transport theories that the Hall conductance of an electron system with areal density ρ is given by $\rho e/B$ so that if $\rho = NB/\Phi_0$, the Hall conductance equals $\sigma_{xy} = Ne/\Phi_0 = N\frac{e^2}{h}$. Hence, if an integer number of Landau levels is occupied, the Hall conductance is quantized to an integer times the von Klitzing conductance. For the experimental discovery of this Integer Quantum Hall Effect, Klaus von Klitzing was awarded the Nobel prize in Physics in 1985. The degree of accuracy of the quantization is such (roughly 10^{-10}) that it has been adopted as a resistance standard, effectively replacing the old definition of an ampere: the ohm is defined so that the Hall resistance of a certain type of device equals 25812.807 Ω. Experimentally it is also seen that when the Hall conductance is quantized, the longitudinal conductance (which is the usual dissipative conductance) vanishes. This can be understood as a direct consequence of the fact that when some Landau levels are completely full and others completely empty, the only scattering mechanisms that could lead to resistance require exciting electron from one Landau level to another, which requires energy $\hbar\omega_c$; for a typical experiment in GaAs, this energy at a field of one tesla equals 17 meV which is quite large, comparable to thermal energy at about 200 K.[3]

While the above explanation of the IQHE is at first sight appealing, it does not fare well under a more careful analysis. Firstly, all experimental samples are dirty to varying degrees, and it seems unreasonable from the above arguments that the Hall conductance should be quantized to the observed degree of accuracy. Secondly, in a typical experiment the electron density is fixed by charge neutrality — deviating from charge neutrality is far too costly energetically — and only for isolated, individual points on the magnetic field axis does the fixed electron density correspond to a chemical potential in the gap between Landau levels; yet, the Hall conductance assumes its quantized value over wide ranges of applied fields. Thirdly, it appears absurd that a low-energy measurement, which typically only probes electrons near the Fermi level, should be sensitive to electrons far below the Fermi surface. Fourthly, a more detailed version of the above analysis, which allows for scattering by impurities, involves the only existence proof in condensed matter physics: regardless of the level of disorder, there exists at least one electron state per Landau level that can propagate through the sample;

[3] It may be surprising that a vanishing longitudinal conductance implies vanishing longitudinal resistance. The explanation is that both conductance and resistance are 2-by-2 matrices and $\hat{\rho} = \hat{\sigma}^{-1}$. If $\sigma_{xx} = 0$ in $\hat{\sigma} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ -\sigma_{xy} & \sigma_{xx} \end{pmatrix}$, also the diagonal element of the corresponding resistance tensor vanishes.

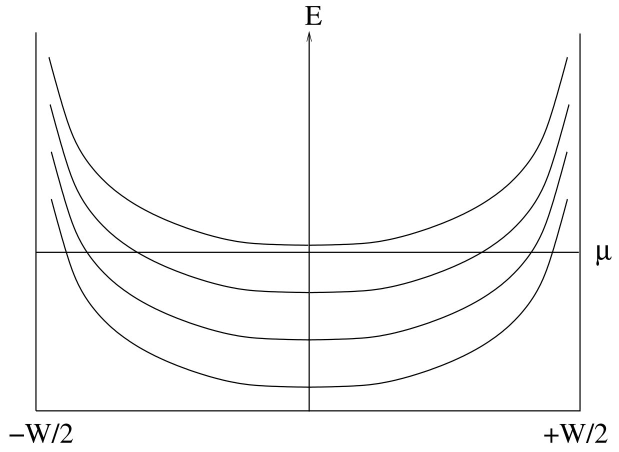

Figure 1.3: *Edge states in the integer quantum Hall regime.* Energy of a state in a sample of width W *vs.* the expectation value of the position. Note that each Landau level contributes one state at the chemical potential near both sample edges.

typically, in physics proofs are constructive, saying something about the solution (in addition to its existence).

All the above objections can be dealt with using the Landauer-Büttiker formalism as we will now see. To begin with, let us introduce a potential $V(x, y)$ that confines the electrons to the sample. A general potential renders the Schrödinger equation unsolvable analytically — essentially only a parabolic confinement is analytically tractable in the x-direction — but the qualitative consequences can easily be inferred. As a result of the confinement in x-direction, the energy of a state becomes dependent on where the state is localized in the transverse (x) direction, hence, the energy E_{nk} acquires a k-dependence. Also, regardless of the value of k, the state cannot be located outside the sample as defined by $V(x, y)$; hence, for any k, $\langle x\rangle_{nk} = \int_{-\infty}^{\infty} dx\, |u_{nk}(x)|^2 x \in [-W/2, W/2]$ where $\pm W/2$ are the edges of the sample. Consequently, the energy E_{nk} as a function of the expectation value $\langle x\rangle_{nk}$ is qualitatively given by Figure 1.3. The figure shows, firstly, that there are states at the Fermi level regardless of the value of the chemical potential, secondly, the states at the Fermi level are located near the edges of the sample, and thirdly, each (even partially) occupied Landau level contributes one state at the Fermi level near both sample edges. Additionally, since $\partial_k \langle x\rangle_{nk} < 0$, states near the right edge of the sample have negative velocities in the longitudinal (y) direction, while states near the opposite edge have positive velocities.

Now the mysteries noted above disappear one by one: according to Landauer-Büttiker formalism, the conductances only depend on the probabilities of electrons being transported from one electrode to another, and since electrons on the Fermi level are confined near sample edges, they are inevitably transported from one electrode to the next one in the clockwise direction — transmission probability to the next electrode in the clockwise direction is 1 and in the counterclockwise direction it is 0; the number of states at the Fermi level stays constant over wide ranges of the magnetic field, implying wide quantization plateuax as is observed; since each even partially occupied Landau level contributes states to the Fermi level, their contribution to low energy experiments is natural; finally, the nature of the propagating modes is now clear: they are edges modes propagating along the edges of the sample. We can now also see that sufficient disorder would allow electrons to scatter all the way across the sample and eventually reverse their direction of propagation, thereby destroying the quantization. Careful consideration along these lines can be used to predict the widths of the Hall plateaux and how they depend on sample width, temperature, disorder potential *etc.*.

1.3 Localization: coherence effects in disordered systems

All naturally occurring systems contain impurities, defects, or other imperfections to varying degrees. These result in a potential landscape that has a random potential superimposed on a regular one such as the potential arising from a periodic crystal. This random potential gives rise to many different effects such as a transition from ballistic transport (transport without scattering) to diffusive transport that is described by frequent scattering by imperfections. The nature of these two transport regimes is most clear if we consider how far a particle moves in time δt: in the ballistic case, if the particle was at position $\mathbf{r}_1$ at time t_1, then at time t_2 it will be in position $\mathbf{r}_2$ such that $\langle(\mathbf{r}_1-\mathbf{r}_2)^2\rangle^{1/2} = v|t_2-t_1|$ while in the diffusive case the root mean square displacement increases as $\langle(\mathbf{r}_1-\mathbf{r}_2)^2\rangle^{1/2} = 2D\sqrt{|t_2-t_1|}$ where v is the speed and D the diffusion constant.

The two transport regimes are separated by the mean free path ℓ, or the elastic scattering time τ_{el}: transport is ballistic in length scales $L < \ell$ or time scales $t < \tau_{\text{el}}$, and diffusive on larger scales. Phase-breaking phenomena introduce a new time scale τ_ϕ or corresponding length L_ϕ. In practice phase breaking occurs at longer length and time scales than impurity scattering (this need not be true in the cleanest semiconductor systems) so that the system size L may be either in the ballistic, phase-coherent regime $L < \ell$, in the diffusive, phase coherent regime $\ell < L < L_\phi$, or in the classical regime $L_\phi < L$.

Somewhat surprisingly, quantum phenomena may be seen even in the classical regime when the system size is larger than either ℓ or L_ϕ. One of these phenomena is strong localization, or simply localization, or more specifically Anderson localization after Philip W. Anderson, one of the most influential condensed matter physicists of this and the last century and the winner of the 1977 Nobel prize in Physics (for his work on localization). Strong localization is an effect that arises from similar considerations that result in the concept of the mean free path, and in some cases it results in a conductance reduction by 100%. Weak localization, in contrast, is much weaker effect, mostly associated with phase breaking, and discussed first in the late 1970s by a school of Russian physicists including Boris Altshuler, Arkadi Aranov, Dmitri Khmelnitskii, Anatoly Larkin and Boris Spivak.

Strong localization

Consider a d-dimensional cubical sample with side length L — how does the conductance between two opposite faces change with changing L? The classical answer is easy: the conductance is directly proportional to the cross section of the conductor, which is L^{d-1}, and inversely proportional to the length of the conductor, so that $G \sim L^{d-2}$. This result is usually given in the form $\frac{d\ln G}{d\ln L} = \beta(G)$ where $\beta(G)$ is called the scaling function, and has classically the form $\beta(G) = (d-2)$. Hence, in three dimensions the conductance increases with system size, in one dimension it decreases, and in two dimensions the classical conductance is size-independent.

Quantum mechanical effects lead to corrections from this classical behavior. If the system is very disordered, particles tend to be localized near the potential minima, and transport through the system becomes difficult. The eigenstates $\psi_\alpha(\mathbf{r})$ of a random potential have a typical size ξ so that $|\psi_\alpha(\mathbf{r}-\mathbf{r}_\alpha)|^2 \sim e^{-|\mathbf{r}-\mathbf{r}_\alpha|/\xi}$ for $|\mathbf{r}-\mathbf{r}_\alpha| \gtrsim \xi$. At zero temperature there is only one state at the Fermi level, and all transport involves particles entering and leaving that state. Transport hence involves a factor that is the amplitude of an electron at one edge of the sample being in the state on the Fermi level, and a factor that an electron in the state on the Fermi level is on the opposite edge of the sample. The product of these two factors is independent of where in the sample the state on the Fermi level is located, and yields a conductance that decreases exponentially with the sample size as $e^{-L/\xi}$. This simple analysis excludes transport through extended states (such as $\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}$) which occur in cleaner systems and have large amplitudes on opposite sample edges, thereby avoiding the exponential decay. If the disorder is not strong, the classical result is only slightly modified, and the end result is

$$\beta(G) = \begin{cases} (d-2) - \frac{a}{G}, & G \gg \frac{e^2}{h} \\ \ln G, & G \ll \frac{e^2}{h} \end{cases}$$

Here a is a constant that is usually positive.

An important feature of the scaling function is that $\beta(G)$ is negative for all G in one and two dimensions (in the two-dimensional case $\beta(G)$ may become positive for large G in the presence of an external magnetic field), while in three dimensions $\beta(G)$ is negative for small G and positive for large G. The sign of β is of substantial importance (indeed, discovering a sign change of the corresponding scaling function in QCD gave the Nobel prize in Physics in 2004). A positive $\beta(G)$ means that the conductance increases with increasing system size, a behavior usually associated with metals, and a negative $\beta(G)$ implies that conductance decreases with increasing system size, a behavior typical of insulators. The point separating these two behaviors, $\beta(G_c) = 0$, is associated with a metal-insulator transition. In the absence of magnetic fields the metal-insulator transition is only seen in three dimensions where sufficiently clean systems behave as metals and sufficiently dirty systems as insulators.

From the above discussion it may not be completely clear why strong localization is a quantum phenomenon: very little quantum mechanics was visible in the analysis. Firstly, in three dimensions the existence of a metal-insulator-transition is clearly of quantum mechanical origin as the classical analysis predicts always a positive $\beta(G)$, and the transition appears as a result of two asymptotic behaviors with different signs of the scaling function. Secondly, you can understand the localization as arising from transport through a random potential landscape with many barriers and potential wells. The total transmission through this complicated potential is in general quite low as we saw in our analysis of resonant tunneling, and only at certain resonant energies may the transmission be substantial. Since the resonant energies for different potential wells in the random potential are different, it is quite difficult to get a resonant transmission through the whole structure.[4]

There are also other mechanisms that may lead to a metal-insulator transition, typically associated with electron-electron interactions. They will be discussed in the Chapter on correlation effects. Conventionally, a metal-insulator transition that is associated with disorder in the system is known as Anderson transition, and a transition associated with electron-electron interactions is known as Mott transition after Sir Neville Mott who shared the 1977 Nobel prize with Phil Anderson and John van Vleck.

[4] Such resonances, known as stochastic or Azbel resonances, are predicted to exist even for random potentials, but they are typically very narrow and occur at random energies.

At $T = 0$ localization implies that conductance decreases exponentially with increasing sample size. What about finite temperatures? At a finite temperature transport takes place by electrons hopping between different localized states. If an electron hops between two states that have energies E_i and E_f, the energy difference must be supplied by the thermal bath, which implies that the hopping rate will be proportional to $e^{-(E_f-E_i)/k_BT}$. Another factor arises from the overlap between the initial and final states: if they are centered at positions $\mathbf{R}_i$ and $\mathbf{R}_f$, this results in the factor $e^{-|\mathbf{R}_i-\mathbf{R}_f|/\xi}$ where ξ is a length scale describing the size of the localized states. The first, thermal factor favors long hops since if you search over a sphere of radius R, there are typically $\frac{4\pi}{3}R^3D(\epsilon_F)$ states per unit energy near the Fermi level. Here $D(\epsilon_F)$ is the density of states. Consequently, searching over a sphere of this size centered at $\mathbf{R}_i$, one typically finds a final state whose energy deviates from E_i by $\frac{3}{4\pi R^3 D}$. Thus, the rate for hopping a distance R is roughly given by

$$\Gamma(R) \sim e^{-\frac{R}{\xi}-\beta\frac{3}{4\pi R^3 D}}$$

where $\beta = (k_BT)^{-1}$. The hopping rate is maximized for hops of the optimal length $R_{\text{opt}} = \left[\frac{9}{4\pi}\frac{\xi\beta}{D}\right]^{1/4} \sim T^{-1/4}$, that is, at low temperatures long hops are favored and at high temperatures hop length decreases. Typically the hops with optimal length dominate conductance so that the conductance is obtained by substituting R_{opt} to $\Gamma(R)$, which yields a conductance whose temperature dependence is given by

$$G \sim e^{-A/T^{1/4}}.$$

This is known as Mott's $T^{1/4}$ law, and has been verified by several experiments.

The optimal hop length cannot be shorter than the lattice spacing, which implies that the above analysis breaks down at sufficiently high temperatures when the hopping distance becomes a temperature-independent constant, and the hopping rate only depends on temperature in the usual activated manner $G \sim e^{-T_A/T}$.

Weak localization

Weak localization is a general term that refers to many quantum effects in conductance. These effects originate from the phase coherence of the quantum states, and therefore typically depend on the phase breaking time τ_ϕ or, equivalently, on the phase coherence length L_ϕ. As a matter of fact, weak localization corrections to conductance are the most common way of measuring the phase breaking time.

Weak localization results are derived rigorously using quantum field theoretical techniques to analyze the effects of impurities in quantum mechanical systems, and the techniques are beyond the scope of this course. However, many of the results can be understood based on a simple physical picture which we will consider in the following.

The conductance is a measure of how easily electrons can move from one place to another. Large conductance implies that it is easy for an electron to leave its original position and end up somewhere else, while low conductance implies that an electron is likely to stay near its original position. In diffusive systems an electron moves along a rugged trajectory, bouncing off impurities, and after having bounced off many impurities, it may end up near its initial position — hence, some of the possible trajectories of the electron form closed loops. In classical physics this connection is formulated quantitatively as the Einstein relation (derived by Einstein in his most cited article as a connection between viscosity [the equivalent of resistance in fluid dynamics] and diffusion constant), which relates the diffusion constant D to the conductivity σ and the density of states $\frac{dn}{d\mu}$ as $\sigma = e^2 D \frac{dn}{d\mu}$.[5] Let the quantum mechanical amplitude for traveling the loop C_j be A_j. Classically, the probability of the electron returning to the vicinity of its initial position is then the sum of probabilities over all loops, or

$$P_{\mathrm{cl}} = \sum_j |A_j|^2$$

where $P_j = |A_j|^2$ is the probability for traveling one loop. Quantum mechanically, however, we need to sum the amplitudes first and only then take the square, so the quantum return probability is

$$P_{\mathrm{quantum}} = |\sum_j A_j|^2 = \sum_j |A_j|^2 + \sum_{j \neq j'} A_j A_{j'}^*$$

The last sum describes interference between different loops, and each term can be written as $|A_j||A_{j'}| \cos[\theta_j - \theta_{j'}]$ in terms of the magnitudes and phase of the individual amplitudes. If the phases are random, the cosine averages to zero, and the classical and quantum results coincide. For different loops the phases are usually unrelated so the interference effects can be assumed to be small.

However, each loop can be traversed in two directions, clockwise and anticlockwise. Let us call these loops C_j^- and C_j^+, and separate their contributions to the quantum probability. We have then $|A(C_j^-)|^2 + |A(C_j^+)|^2 + A(C_j^-)A(C_j^+)^* + A(C_j^-)^* A(C_j^+)$. The phases of these two trajectories are not unrelated: if an electron has wave vector $\mathbf{k}$ during its passage between point $\mathbf{R}_1$ and $\mathbf{R}_2$ on the counterclockwise path, it accumulates phase $\mathbf{k}\cdot(\mathbf{R}_2 - \mathbf{R}_1)$ during that part of the loop; on the clockwise path the electron propagates from $\mathbf{R}_2$ to $\mathbf{R}_1$ with wave vector $-\mathbf{k}$, and accumulates phase $-\mathbf{k}\cdot(\mathbf{R}_1 - \mathbf{R}_2)$ — exactly the same as on the counterclockwise path, implying that $A(C_j^+) = A(C_j^-)$, and the two countertraversed paths yield a contribution $4|A_j|^2$ to the quantum mechanical return probability but only contribution $2|A_j|^2$ to the classical return probability. Consequently, a quantum mechanical particle is more likely to return to its original location and less likely to move away, which results in a lower conductance than what would be expected classically.

The above argument relies entirely on the special phase relation between the two countertraversed trajectories. This special relationship ceases to be valid if the time it takes for the electron to traverse the loop exceeds the phase breaking time, or if the loop size exceeds the phase breaking length. The minimum propagation time for a loop is roughly given by the elastic scattering time τ since for times shorter than τ the electrons move ballistically along straight trajectories. Hence, only loops with traversal times $\tau < t < \tau_\phi$ contribute to quantum corrections in conductance. On this time scale the motion of an electron is diffusive and the probability distribution of finding the electron at distance r from its initial position is roughly Gaussian with a variance that increases as t, $P(r) \sim t^{-d/2} e^{-r^2/(2Dt)}$, so that the probability of finding it near its initial position ($r = 0$) decreases as $t^{-d/2}$. Hence, the total number of loops with traversal times in the required range is proportional to $\int_\tau^{\tau_\phi} dt\, t^{-d/2}$.

[5] If there is an electric field E across a sample, then the chemical potential of particles with charge q obeys $\frac{d\mu}{dx} = qE$. The diffusive particle current due to a concentration gradient is given by $j_D^{(p)} = -D\frac{dn}{dx} = -D\frac{dn}{d\mu}\frac{d\mu}{dx}$ where the second equation holds under the assumption of local equilibrium. The associated diffusive charge current is hence $j_D^{(c)} = -q^2 D \frac{dn}{d\mu} E$. If the total current vanishes, as is the case in equilibrium, this diffusive current must be canceled by the drift current $j = \sigma E$, which yields the Einstein relation.

Since each of these loops gives a similar relative correction to the conductance, the overall quantum correction is

$$\frac{\delta\sigma}{\sigma} \sim -\kappa \begin{cases} (\tau_\phi/\tau)^{1/2}, & d=1 \\ \ln(\tau_\phi/\tau), & d=2 \\ (\tau_\phi/\tau)^{-1/2}, & d=3 \end{cases}$$

where κ is a constant that depends on τ and hence on the amount of disorder.

Hence, since the phase breaking time depends on temperature, one way to investigate the quantum corrections to conductance in experiments is to look for temperature dependent contributions. This is, however, not very practical as many classical effects also depend on temperature. A better way is to realize that the special phase relationship between the countertraversed paths can be removed by breaking time reversal invariance by introducing a magnetic field: a high magnetic fields changes the trajectories of the electrons, but at low magnetic fields the trajectories are almost unaffected while the relative phases of the two orientations of the loops differ by $2\Phi/\Phi_0$ as we determined in the persistent current analysis. Here Φ is the magnetic flux through the loop, which equals BS where S is the cross-sectional area of the loop in the direction perpendicular to the magnetic field. If the transport is diffusive, the cross sectional area increases with time as $2Dt$. If the phase difference due to flux exceeds roughly 2π, the quantum effects are washed out. This happens if $t \gtrsim \tau_B = \pi\Phi_0/(2DB)$, and we must replace the upper limit of the integral yielding the quantum corrections be the smaller of τ_B and τ_ϕ. This means that as the magnetic field is increased, fewer and fewer loops contribute to the quantum corrections to conductance, making them less and less important. Since the quantum corrections reduce conductance, we conclude that the conductance of a disordered system should increase as the magnetic field is increased. This phenomenon, known as negative magnetoresistance, has been confirmed in numerous experiments, and is used as a standard tool to measure phase breaking times. It is particularly well suited for the task since all classical effects lead to positive magnetoresistance and occur at higher magnetic fields than the quantum corrections.

Universal conductance fluctuations

Another phenomenon associated with phase coherence is known as universal conductance fluctuation. It emerges as an answer to the question *How much does the conductance vary between nominally similar conductors?* By nominally similar we mean that the conductors have same dimensions, same impurity concentrations *etc.* — essentially, we consider an ensemble of conductors and investigate the sample-to-sample variations of the conductances. The variations arise because the impurities occupy different positions and the associated potential variations have different magnitudes in different samples; in effect, the variations reflect the varying potential landscapes in different samples (the potential landscape can be thought of as the sea floor, and the charge carriers as a fluid filling the sea: the conductivity depends on the local sea depth (carrier density) and the location and shape of islands (potential barriers)). An ensemble can be created either by having several physically different samples, or by subjecting one sample to external controls (such as electric or magnetic fields) that change the potential landscape.

Let us again consider what happens in a classical system. Classically we can divide a sample of length L into a series of thin slices. The slices are roughly independent of each other if their thickness exceeds the scattering length ℓ, so the number of slices is L/ℓ. If each slice is nominally similar, they all have average resistances R_0, and the slice-to-slice variance is δR_0^2. The average resistance of the full sample is then a sum of the resistances of individual slices, $R_{\text{avg}} = (L/\ell)R_0$, and the total variance $\delta R^2 = (L/\ell)\delta R_0^2$ so that the relative standard deviation decreases with increasing length as $L^{-1/2}$ as usual.

A key assumption in the above analysis was that the slices were assumed to be independent. If the system is phase coherent, this assumption fails, and we need to reconsider. The exact analysis is rather complicated, but a simple physical argument can be constructed by noticing that in a sample with many transverse modes, the total reflectance $\tilde{R}$ (not resistance) is given by the sum over transverse channels as

$$\tilde{R} = \sum_{\alpha\beta} |r_{\alpha\beta}|^2$$

where $r_{\alpha\beta}$ is the reflection amplitude from transverse mode β to mode α (note that the total reflectance and the total transmittance T are related by $\tilde{R}+T = N$ where N is the number of transverse modes. Since the conductance is $G = \frac{e^2}{h}T$ according to Landauer, knowing $\tilde{R}$ will yield conductance.) Now assume that the N^2 different contributions $\alpha\beta$ to the reflectance are independent so that the variance of the reflectance is given by $N^2[\langle|r_{\alpha\beta}|^4\rangle - \langle|r_{\alpha\beta}|^2\rangle]$. We can evaluate this by dividing the reflection amplitude $r_{\alpha\beta}$ into terms that come from different particle trajectories — Feynman paths, if you are familiar with the path integral formulation of quantum mechanics. This gives $r_{\alpha\beta} = \sum_i r_{\alpha\beta}(i)$ where $r_{\alpha\beta}(i)$ is the contribution of the trajectory i so that the first term of $\delta\tilde{R}^2$ can be written as

$$\begin{aligned}
&\langle|r_{\alpha\beta}|^4\rangle \\
&= \textstyle\sum_{ijkl}\langle r_{\alpha\beta}(i)^* r_{\alpha\beta}(j)^* r_{\alpha\beta}(k) r_{\alpha\beta}(l)\rangle \\
&\approx 2\langle\textstyle\sum_i r_{\alpha\beta}(i)^* r_{\alpha\beta}(i)\rangle^2 \\
&= 2\langle|r_{\alpha\beta}|^2\rangle^2
\end{aligned}$$

where we only included the diagonal (manifestly real) terms $(i = k, j = l)$ and $(i = l, j = k)$ under the assumption that the off-diagonal terms depend on phases that average to zero. From $\langle G\rangle = N\frac{e^2}{h} - \frac{e^2}{h}\sum_{\alpha\beta}\langle|r_{\alpha\beta}|^2\rangle$ and the fact that $G \propto N\ell/L$ (conductance is linearly proportional to wire width and inversely proportional to wire length), we conclude that $\langle|r_{\alpha\beta}|^2\rangle \propto (1/N)(1 - \ell/L)$ so that $\delta\tilde{R}^2 \propto N^2[(1/N)(1 - \ell/L)]^2 = (1 - \ell/L)^2$, that is, the variance of the reflectance is to leading order independent of the size of the sample. Since $T = N - \tilde{R}$, we have $\delta T^2 = \delta\tilde{R}^2 \approx 1$ implying that $\delta G^2 = \left(\frac{e^2}{h}\right)^2$ so that the standard deviation of conductance is equal to the conductance quantum, and independent of sample size or the conductance: universal conductance fluctuations.[6] This is in blatant contradiction with the classical result that would imply sample size dependence (as we saw above) but also typically yield a result that conductance fluctuation depends on the average conductance.

The above seems like a de-tour — we really wanted the conductance, or total transmittance, but we started by analyzing the reflectance — was it really necessary? It turns out that it was. The key assumption above was that the reflection amplitudes $r_{\alpha\beta}$ are statistically independent so that it is possible to add averages and variances. This turns out to be a good assumption. In contrast, the transmission amplitudes $t_{\alpha\beta}$ are not statistically independent, and carrying out the analysis in those terms would either have been more complicated or,

[6] A more careful analysis reveals that the fluctuation are $a\frac{e^2}{h}$ where the constant a depends on the shape of the sample. The constant does not, however, depend on the size of the sample or the impurity concentration (as long as transport is diffusive).

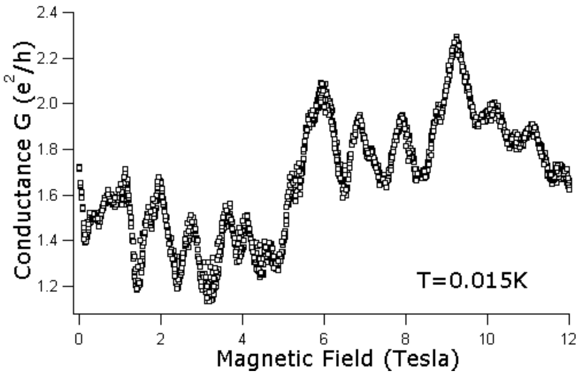

Figure 1.4: A typical magnetofingerprint of a small sample, taken and Georgia University of Technology and showing the conductance as a function of an applied magnetic field.

more likely, yielded an erroneous result. The fact that the transmission amplitudes are not independent can be understood using the trajectory (Feynman path) picture: in a disordered sample, there are relatively few preferred paths through the sample, and the main contribution to transmission comes from these paths (sort of like numerous small mountain streams merging into large rivers before reaching the ocean). The reflected paths, in contrast, typically never penetrate very far into the sample, and one reflection event is therefore quite independent of other events. If you are not satisfied by these heuristic arguments, the real way of obtaining universal conductance fluctuations is explained by Patrick Lee and Douglas Stone in Phys. Rev. Lett. **55**, 1622 (1985). The heuristic argument was presented by Patrick Lee a year later.

Experimentally UCF has been seen in numerous experiments. The typical experiment measures conductance variations as a function of an external weak magnetic field as shown in Fig. 1.4, or as a function of carrier density. The resulting $G(B)$ plot looks like noisy data, but the plot is perfectly reproducible for a given sample (as it should be, it is determined by the exact positions of impurities), and is commonly known as the magnetofingerprint. When the sample is heated up, the impurities can move around, which results in a new magnetofingerprint.

1.4 Origins of decoherence

Wherefore decoherence? The phase of a wave function evolves in time as $e^{iEt/\hbar}$ where E is the energy: hence, phase fluctuations are related to energy fluctuations. To analyze the origin of phase fluctuations, we divide the universe into a 'system' — what we are interested in — and 'environment' — the rest. While the universe as a whole is, presumably, described by quantum mechanical equations of motion and possesses a phase that evolves in a deterministic fashion, any small part of it (the system) is coupled to its environment more or less strongly and therefore does not have a constant energy and, consequently, not entirely deterministic phase evolution. Hence, phase coherence is lost due to interactions between the system and its environment.

The distinction between the system and the environment is quite an abstract one. Sometimes they are two spatially separate parts — *e.g.* a single hydrogen molecule (the system) in hydrogen gas (the environment) — but more often the environment refers to those degrees of freedom that are not explicitly accounted for in the Hamiltonian of the system, as is for instance the case when we describe metals as a collection of independent electrons in a static lattice: the environment of an electron contains both other electrons and lattice ions, and the coupling between the system and the environment is in the form of electron-electron and electron-phonon interactions.

For now, let us denote those degrees of freedom that we are interested in by $\{x_{\rm system}\}$ and the rest by $\{X_{\rm env}\}$. The Hamiltonian can then be written as

$$H = H_{\rm system}(\{x_{\rm system}\}) + H_{\rm env}(\{X_{\rm env}\}) + H_{\rm coupling}(\{x_{\rm system}, X_{\rm env}\})$$

where the first two terms describe the isolated system and environment, respectively, and the last term describe a coupling between the two parts. The isolated systems can be described by wave functions that only depend on one set of variables and whose energies are entirely determined by $H_{\rm system}$ or $H_{\rm env}$,

$$\begin{array}{lcl} H_{\rm system}(\{x_{\rm system}\})\psi_\alpha(\{x_{\rm system}\}) & = & \epsilon_\alpha\psi_\alpha(\{x_{\rm system}\}) \\ H_{\rm env}(\{X_{\rm env}\})\phi_\beta(\{X_{\rm env}\}) & = & \epsilon_\beta^{\rm env}\phi_\beta(\{X_{\rm env}\}) \end{array}$$

The separation system-environment is only useful if the coupling between those two is so weak that to a first approximation we may entirely neglect it. Then the starting point of our description of the system is in terms of the eigenstates of an isolated system,

$$\begin{array}{l} [H_{\rm system}(\{x_{\rm system}\}) + H_{\rm env}(\{X_{\rm env}\})]\,[\psi_A(\{x_{\rm system}\})\phi_B(\{X_{\rm env}\})] \\ = (\epsilon_A + \epsilon_B^{\rm env})\,[\psi_A(\{x_{\rm system}\})\phi_B(\{X_{\rm env}\})]\,. \end{array}$$

The impact of the coupling to the environment is perturbative, resulting in mixing of eigenstates whose energies are close to each other and slight shifts of the eigenenergies. More importantly, the joint eigenfunctions of the system-environment complex are not simply given by products of a system eigenfunction and an environment eigenfunction, as would be the case if the two parts were completely decoupled, but are of the more general form

$$\begin{array}{l} H\Psi_\gamma(\{x_{\rm system}, X_{\rm env}\}) = E_\gamma\Psi_\gamma(\{x_{\rm system}, X_{\rm env}\}) \\ \Psi_\gamma(\{x_{\rm system}, X_{\rm env}\}) = \sum_{\alpha,\beta} C_{\gamma\alpha\beta}\psi_\alpha(\{x_{\rm system})\phi_\beta(\{X_{\rm env}\}) \neq \psi_A(\{x_{\rm system})\phi_B(\{X_{\rm env}\}) \end{array}$$

To see explicitly how coupling to an environment leads to de-phasing, consider a situation in which the system and the environment are decoupled until time $t = 0$, and then a coupling is switched on. At time $t = 0^-$ the system was in its eigenstate $|k\rangle$ with energy ϵ but because of the coupling, at times $t > 0$ the state $|k\rangle$ is no longer an eigenstate with a definite energy, and therefore its phase does not evolve in a deterministic fashion: $|k\rangle$ splits into many components with different energies and different phase evolutions. Because a perturbation such as the system–environment coupling predominantly couples unperturbed states whose energies are close to each other, the distribution of the phase of the state at $t > 0$ is initially quite narrow and increases with time. In a simplest model the phase performs a random walk with both an average and variance that increase roughly linearly with time, $\langle\phi(t)\rangle \sim \bar{\epsilon}t$ and $\langle[\phi(t) - \bar{\epsilon}t]^2\rangle \sim D_\phi t$. If the coupling with the environment is symmetric in the energy space, we have $\bar{\epsilon} \approx \epsilon$, but in general this need not be the case.

One common way of describing the system-environment problems is to use density matrices. Let us first ignore the division between the degrees of freedom, and simply consider a physical system with states $|\Psi_\gamma\rangle$. If the probability of the system being in state $|\Psi_\gamma\rangle$ is w_γ, the expected result of a measurement that is described by operator $\hat{A}$ is

$$\langle\langle A\rangle\rangle = \sum_\gamma w_\gamma \langle\Psi_\gamma|\hat{A}|\Psi_\gamma\rangle$$

where we introduced the notation $\langle\langle\hat{A}\rangle\rangle$ to denote the ensemble average. We can re-write the ensemble average in terms of another basis $|\Phi\rangle$ as

$$\langle\langle A\rangle\rangle = \sum_\gamma w_\gamma \sum_{\Phi,\Phi'} \langle\Psi_\gamma|\Phi\rangle\langle\Phi|\hat{A}|\Phi'\rangle\langle\Phi'|\Psi_\gamma\rangle = \sum_{\Phi,\Phi'}\left(\sum_\gamma w_\gamma\langle\Phi'|\Psi_\gamma\rangle\langle\Psi_\gamma|\Phi\rangle\right)\langle\Phi|\hat{A}|\Phi'\rangle.$$

Defining a new operator

$$\hat{\rho} = \sum_\gamma w_\gamma |\Psi_\gamma\rangle\langle\Psi_\gamma|$$

we can write the ensemble average as

$$\langle\langle A\rangle\rangle = \sum_{\Phi,\Phi'}\langle\Phi'|\hat{\rho}|\Phi\rangle\langle\Phi|\hat{A}|\Phi'\rangle = \mathrm{Tr}(\hat{\rho}\hat{A}).$$

The operator $\hat{\rho}$ is known as a density matrix and it is quite a convenient tool in describing coupled quantum systems.

By substituting $\hat{A} = 1$ we see that $\mathrm{Tr}\hat{\rho} = 1$ which corresponds to normalization of probabilities, $\sum_\gamma w_\gamma = 1$. In the special case that one of the weights w_γ equals unity, meaning that the system is known to be in a particular state, we even have $\mathrm{Tr}\hat{\rho}^2 = 1$.[7]

Applying the density matrix formalism to the system-environment complex, it is most convenient to use the direct product basis $\psi_\alpha\phi_\beta$, and write the density matrix as

$$\rho = \sum_{\alpha,\beta} w_{\alpha\beta}|\psi_\alpha\rangle|\phi_\beta\rangle\langle\phi_\beta|\langle\psi_\alpha|.$$

If we now consider a measurement that is only sensitive to the system degrees of freedom, which are included in ψ's, we have the ensemble average

$$\begin{aligned}\langle\langle A\rangle\rangle &= \textstyle\sum_{\alpha,\beta} w_{\alpha\beta}\langle\phi_\beta|\phi_\beta\rangle\langle\psi_\alpha|\hat{A}|\psi_\alpha\rangle\\ &= \textstyle\sum_{\phi,\phi',\psi,\psi'}\left(\sum_{\alpha,\beta} w_{\alpha\beta}\langle\phi'|\phi_\beta\rangle\langle\psi'|\psi_\alpha\rangle\langle\phi_\beta|\phi\rangle\langle\psi_\alpha|\psi\rangle\right)\langle\phi|\phi'\rangle\langle\psi|\hat{A}|\psi'\rangle\\ &= \textstyle\sum_{\psi,\psi'}\left(\sum_{\phi,\alpha,\beta} w_{\alpha\beta}|\langle\phi|\phi_\beta\rangle|^2\langle\psi'|\psi_\alpha\rangle\langle\psi_\alpha|\psi\rangle\right)\langle\psi|\hat{A}|\psi'\rangle.\end{aligned}$$

Here we can identify the quantity inside the parentheses as a reduced density matrix for the system

$$\hat{\rho}_S = \sum_\phi\sum_{\alpha,\beta} w_{\alpha\beta}\langle\phi|\phi_\beta\rangle\langle\psi'|\psi_\alpha\rangle\langle\psi_\alpha|\psi\rangle\langle\phi_\beta|\phi\rangle \equiv \mathrm{Tr}_\phi\hat{\rho}$$

which is a partial trace of the full density matrix ρ. The density matrix ρ_S only depends on the system degrees of freedom, and has matrix elements $\langle\psi|\hat{\rho}_S|\psi'\rangle$. The system expectation values can now be written as

$$\langle\langle A\rangle\rangle = \mathrm{Tr}(\hat{\rho}_S\hat{A}).$$

[7] This special situation is known as a pure state, and the more general situation is referred to as a mixed state.

The reduced density matrix $\hat{\rho}_S$ is a much more convenient quantity than the full density matrix $\hat{\rho}$. The diagonal elements of the reduced density matrix are given by

$$\langle\psi|\hat{\rho}_S|\psi\rangle = \sum_{\phi}\sum_{\alpha,beta} w_{\alpha\beta}\langle\phi|\phi_\beta\rangle\langle\psi|\psi_\alpha\rangle\langle\psi_\alpha|\psi\rangle\langle\phi_\beta|\phi\rangle = \sum_{\phi}\sum_{\alpha,\beta} w_{\alpha\beta}|\langle\phi|\phi_\beta\rangle|^2|\langle\psi|\psi_\alpha\rangle|^2$$

which is manifestly real and positive. Physically, the diagonal matrix elements of the density matrix given the probability of finding the system in a particular state. The off-diagonal matrix elements of $\hat{\rho}_S$ are, in contrast, given by sums of complex numbers. They represent coupling, or coherence, between the different states of the system.

While the reduced density matrix is more manageable than the full density matrix in terms of degrees of freedom, its dynamics is complicated to describe. In principle the time dependence can be obtained by first considering the full density matrix $\hat{\rho}(t) = \sum_\Psi |\Psi(t)\rangle\langle\Psi(t)|$ and taking the partial trace over the environmental degrees of freedom. In practice, however, the calculations tend to get quite complicated, and are usually carried out using the path integral formalism of quantum field theory. Typically, however, the off-diagonal matrix elements of the reduced density matrix decrease as a function of time, reflecting the reduction of coherence due to coupling to the environmental modes, The diagonal elements remain non-zero as required by the probability conservation $\mathrm{Tr}\hat{\rho} = 1$. Hence, the diagonal elements of ρ_S have a simple classical interpretation as probabilities while the off-diagonal elements are inherently quantum mechanical. The exact dynamics of the density matrix typically couples the diagonal and off-diagonal elements (time evolution of the diagonal elements is connected to the off-diagonal elements and *vice versa*). Often, however, one can show that the off-diagonal elements are of lesser importance and one can express the system's time evolution entirely in terms of probabilities of the system being in a particular state. This description is known as a master or rate equation, and is often used as a starting point for dynamic analyses in the classical or semi-classical regime. We will employ the master equation formalism in the discussion of Coulomb blockade systems in the next chapter.

Dissipation, or equilibration in general, is not straightforward to describe in quantum mechanics. In classical physics dissipation is often accounted for by viscous damping terms in the equations of motion along the lines

$$m\partial_t^2 x(t) = F(x(t)) - \gamma\partial_t x(t)$$

where the two first terms constitute the Newtonian equation of motion for a particle of mass m in a force field that depends on the particle's position. The second term on the right hand side results in an acceleration that is in opposite direction as the velocity ($\gamma > 0$), in other words, it describes dissipation. Multiplying the equation by $\partial_t x$, using $F(x) = -\partial_x V(x)$ and $E = (m/2)[\partial_t x]^2 + V(x)$ we see that the energy of the particle decreases according to

$$\partial_t E(t) = -\gamma[\partial_t x(t)]^2 = -\frac{2\gamma}{m}E_{\mathrm{k}in}(t).$$

In ordinary quantum mechanical treatments the energy is an eigenvalue of the Hamilton operator, and as long as the Hamiltonian is time independent, the energy is also time independent: quantum mechanics cannot describe dissipation. However, even in quantum mechanics, energy can flow between two coupled subsystems, and the energy any one subsystem need not be conserved. Using this idea one has introduced quantum descriptions coupling the interesting degrees of freedom (the system) to uninteresting ones (the environment), thereby allowing some dissipation in *the system*. The most common implementation of this idea is due to Caldeira and Leggett, who described the environment as a collection of independent harmonic oscillators, $H_{\mathrm{env}}(\{X_j\}) = \sum_{j=1}^{\infty}\left[-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial X_j^2} + \frac{1}{2}m\omega_j^2 X_j^2\right]$ that are linearly coupled to the system degree of freedom x by $H_{\mathrm{coupling}} = \sum_j C_j x X_j$. This method is quite general as the spectrum of environmental frequencies $\{\omega_j\}$ and the coupling to the different environmental modes C_j can be chosen to meet the requirement of the specific problem at hand. In particular, choosing a set of linearly spaced oscillator frequencies and coupling constants C_j that increases linearly with the frequency ω_j of the environmental mode, results in the classical viscous (ohmic) damping.[8]

In the usual description of condensed matter systems the starting point — the system — is to treat electrons as a collection of independent particles and the underlying lattice as a static structure only giving rise to a periodic potential. Within this assumption, the electrons are described as Bloch waves with well-defined band indices and quasimomenta. This description is only valid over time scales in which it is reasonable to neglect the coupling between the electrons themselves (electron-electron interaction) and coupling between electrons and lattice vibrations (electron-phonon coupling), or any other perturbation that changes the energy of the electron. These two scattering mechanisms have been considered in great detail; however, there is only limited consensus of what the resulting phase breaking time is: for the electron-phonon scattering it has been shown that $\tau_\phi^{-1} \sim T^p$ where the exponent p is either 4 (for dirty 3-dimensional metals), 3 (clean metals), or 2 (dirty metals with some impurities that do not vibrate together with the host lattice). Experimentally, values of p ranging from roughly 1.4 (outside the wide theoretical range!) to about 4 have been observed. Typically, the phase breaking time associated with electron-phonon scattering is of the order of 10^{-11} - 10^{-7} seconds at the temperature of one kelvin depending on, *e.g.*, the sample dimensionality and amount of disorder.

The issue with the phase breaking time arising from electron-electron interactions is rather similar. Theoretically, in clean 3-dimensional metals the inverse phase breaking time due to electron-electron scattering is predicted to vary as the second power of temperature ($p = 2$) while in disordered metals it is expected to obey a $p = 3/2$ law.[9] Experimentally, the first prediction is confirmed (with the reservations at the lowest temperatures that we are soon to discuss) while in disordered samples the data suggests $p \approx 1$ rather than $p = 3/2$. For two-dimensional electron systems t_ϕ^{-1} is expected to vary as $T^2 \ln T$ at relatively high temperatures and as $T \ln T$ at low temperatures; experimentally, the data suggests $AT^2 + BT^{2/3}$ where the second term arises from the scattering processes involving very small energy transfers — this term would be absent in the expression for the energy relaxation time τ_E. The low-energy processes dominate at temperatures below 1 K.

Often we talk about phase breaking length L_ϕ instead of phase breaking time. The two are in one-to-one correspondence to each other: phase breaking time is the distance that an electron, on the average, travels in time τ_ϕ. If the system is clean, electron transport is ballistic and we have $L_\phi = v_F \tau_\phi$, while if the system is disordered, electrons are scattered by impurities and travel in a diffusive manner, leading to $L_\phi = \sqrt{D\tau_\phi}$ where D is the diffusion constant.

[8] It turns out that the ohmic limit is slightly pathological and a more reasonable model is obtained if the linear increase of C_j is cut off at some high frequency ω_D.

[9] The first of these results will be derived later in the discussion on Fermi liquids and electron-electron interactions.

Another time describing decoherence effects is the Thouless time τ_{Th} or the corresponding Thouless energy $E_{\mathrm{Th}} = \hbar/\tau_{\mathrm{Th}}$ and the Thouless length $L_{\mathrm{Th}} = \sqrt{D\tau_{\mathrm{Th}}}$. The Thouless energy measures the system's sensitivity to boundary conditions: it is the change in the energy eigenvalue when the periodic boundary conditions $\psi(x+L) = \psi(x)$ are replaced by the antiperiodic ones, $\psi(x+L) = -\psi(x)$. If the phase of the state gets randomized between the boundaries of the sample, or if the amplitude of the wave function vanishes on the opposite side of the sample, the changing boundary conditions have no implications, and $E_{\mathrm{Th}} = 0$. Since E_{Th} may vary from one quantum state to another, the value that characterizes an entire sample is obtained by averaging, and the average value can be approximately related to the system size by the classical diffusion law $E_{\mathrm{Th}} = \hbar D/L^2$. The Thouless energy appears in many phase-sensitive results, for instance the magnitude of the persistent currents is proportional to $\sqrt{\Delta E_{\mathrm{Th}}}$ where Δ is the typical level spacing; the conductance of a quantum wire is roughly $\frac{\Delta}{E_{\mathrm{Th}}}\frac{e^2}{h}$, *etc.*. The Thouless time is not quite the same as the de-phasing time since insensitivity to boundary conditions arises also as a consequence of strong localization.

Fact or Fiction?

As is probably evident from the above discussion, the temperature dependences of the different contributions to the phase breaking time are not completely understood: the theoretical analysis is complicated by the interplay of many physical mechanisms, and experimentally it is very hard to isolate or even identify the different contributing scattering mechanisms.

Quite recently (1998), a group of experimentalists carried out a series of experiments indicating that the phase breaking time saturates at low temperatures as shown in Figure 1.5, instead of diverging as all theoretical analyses predict. This resulted in a heated debate where the theorists argued that the experiments were wrong in the sense that either the electrons were heated up by some mechanism, or that there was a scattering mechanism that was not frozen out at the experimental temperatures (for example, magnetic impurity scattering, or time-dependent noise that leaked into the measurement system due to insufficient filtering). The experimentalists maintained the integrity of their data and how it was obtained.

Subsequently, many theoretical explanations were put forward as possible explanations, but most of them could be discounted on various grounds. Recently, the discussion has essentially died out — the issue has become very much of a hot potato, best not to touched — and no true consensus has been reached. Most physicist, I believe, continue to think that the phase breaking time increases beyond bounds as the temperature approaches absolute zero.[a] Were the opposite true, this would have major consequences and invalidate much of the theoretical foundation of condensed matter physics.

[a] The leading candidate for an explanation is magnetic impurities, which have been shown to affect the low-temperature de-phasing rate even in concentration below 1 ppm, and result in a saturation through a compensation between the electron-electron scattering and the Kondo effect, both discussed in the next chapter.

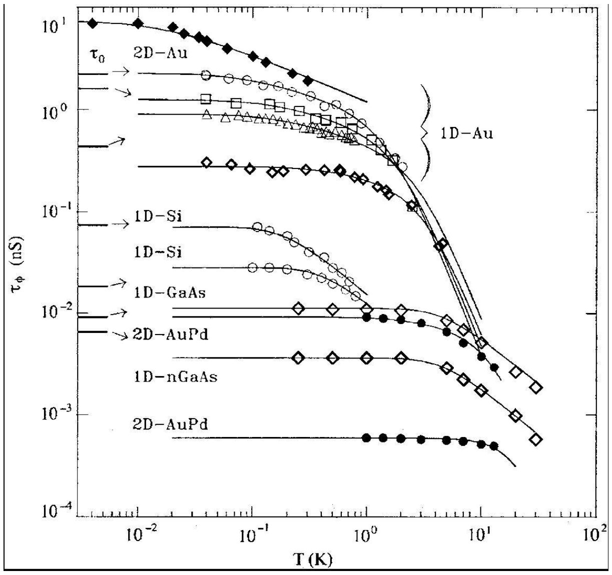

Figure 1.5: Temperature dependence of the electron dephasing time τ_ϕ in a number of systems (from R. A. Webb, P. Mohanty, and E. M. Q. Jariwala, *Fortschr. Phys.* **46**, 779 (1998)).

Chapter 2

Correlation Effects

2.1 Correlations in classical systems: Coulomb blockade

2.1.1 Double Barrier Structure: Single Electron Transistor

Qualitative discussion

How much energy is required to add one electron to an electric conductor? The precise answer to this question depends both on the internal structure of the conductor through electron-ion interactions, and on the presence of other metallic bodies through the creation of image charges. For simplicity, let us ignore all complications of the first type, assume a spherical conductor of radius R, ground all other metallic bodies, and place them infinitely far from the conductor. The electrostatic potential of the conductor then depends on its charge through $V(Q) = \frac{Q}{4\pi\epsilon R}$ so that the energy cost of charging up the conductor with charge Q is

$$E(Q) = \int_0^Q dq \frac{q}{4\pi\epsilon R} = \frac{Q^2}{8\pi\epsilon R}$$

Writing this in the usual capacitance form $Q^2/(2C_0)$ shows that the ground capacitance of the conductor is $C_0 = 4\pi\epsilon R$, which is directly proportional to the size of the conductor (for the older generation this has no news value: in older systems of units the unit of capacitance is centimeter).

Consider now a very small grain, connected to two external electrodes by means of tunnel junctions, and to a third electrode by a capacitive coupling as indicated in Figure 2.1. We will refer to the first two electrodes as source and drain, and to the last one as gate. Now the electrostatic situation is more complicated, and we need to analyze the circuit more carefully to determine the energy $E(Q)$. Let us indicate the potentials on the external electrodes V_j and the capacitances between the grain and the electrodes by C_j $(j = s, d, g)$. If we denote the image charges on the electrodes by $-Q_j$ and on the ground plate by $-Q_0$,[1] we have $Q = Q_0 + \sum_j Q_j$, $V = Q_0/C_0$, and $V = V_j + Q_j/C_j$ where V is the potential on the grain. By solving this system of equations we get $V(Q) = [Q + \sum_j C_j V_j]/C_\Sigma$ and integration yields

$$E(Q) = \frac{Q^2}{2C_\Sigma} + \frac{1}{C_\Sigma} Q \sum_j C_j V_j = \frac{1}{2C_\Sigma}(Q - \tilde{Q})^2 - \frac{1}{2C_\Sigma}\tilde{Q}^2$$

[1] In practice, "ground plate" refers to all metallic bodies other s, d, and g. They are usually far from the grain so we can often ignore the small capacitance C_0.

where $C_\Sigma = C_0 + \sum_j C_j$ is the total capacitance between the grain and other conducting bodies kept at constant potential, and $\tilde{Q} = -\sum_j C_j V_j$.

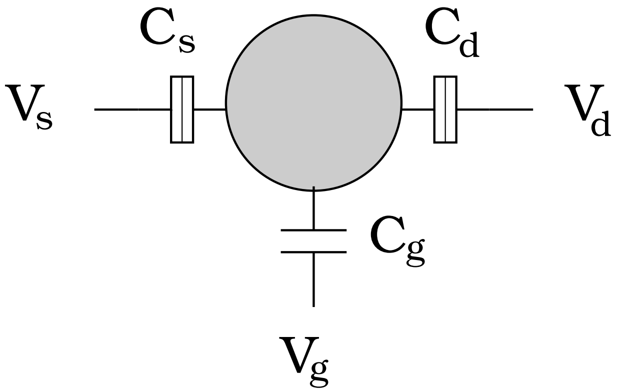

Figure 2.1: Schematic picture of a single electron transistor (SET)

The first issue we wish to address is to determine how much the charge on the grain varies. Let us start by considering the equilibrium case $V_s = V_d$. The grain charge changes only if an electron tunnels between the grain and either the source or the drain. Let us take as a reference state a configuration where the charge on the grain is zero, so that if M electrons of charge $-e$ tunnel into it, the grain charge will become $-Me$, and the electrostatic energy of the grain becomes $E(-Me)$. The total change in electrostatic energy due to the tunneling processes is

$$\delta E(M) = [E(-Me) + MeV_s] - E(0) = M^2e^2/(2C_\Sigma) - Me[-V_s - \frac{1}{C_\Sigma}\tilde{Q}],$$

where we have accounted for the change in the electrostatic energy in the source (or drain) electrode. We can simplify the expression to get $\delta E(M) = E_C[M - (Q_0^{\text{eff}}/(-e))]^2 - \frac{(Q_0^{\text{eff}})^2}{2C_\Sigma}$ where $Q_0^{\text{eff}} = -\sum_j C_j V_j + C_\Sigma V_s$ is an offset charge and $E_C = e^2/(2C_\Sigma)$ is the charging energy. Since $\delta E(M)$ has its minimum at $M \approx Q_0^{\text{eff}}/(-e)$, at low temperatures charge tends to flow between the grain and the electrodes so that the grain charge becomes approximately Q_0^{eff}.

The fluctuations in the grain charge are given by the variance $\delta Q^2 = \langle Q^2\rangle - \langle Q\rangle^2$, where the angular brackets $\langle\ldots\rangle$ denote average over different charge states of the grain. In equilibrium the probabilities of different charge states of the grain are proportional to $e^{-\beta\delta E(M)}$ and we have

$$\langle Q\rangle = Z^{-1}\sum_M (-Me)e^{-\beta\delta E(M)}$$

and

$$\langle Q^2\rangle = Z^{-1}\sum_M (Me)^2 e^{-\beta\delta E(M)}$$

where $Z = \sum_M e^{-\beta\delta E(M)}$. At high temperatures the granularity of charge can be ignored and we get by symmetry $\langle Q\rangle = Q_0^{\text{eff}}$, and from the equipartition theorem $[1/(2C_\Sigma)]\langle(Q - \langle Q\rangle)^2\rangle = \frac{1}{2}k_BT$ or $\delta Q^2_{\text{eq}} = C_\Sigma k_B T$. At low temperatures, in contrast, the granularity of charge cannot be ignored, and we must do the sums numerically to obtain δQ_{eq}^2.[2] The fluctuations are smallest when the minimum of $\delta E(z)$ occurs at an integer value of z, and they are largest when the minimum occurs at a half-integer value of z. In terms of the offset charge Q_0^{eff} these two cases correspond to an integer and half-integer multiples of e. In Figure 2.2 we have plotted the charge fluctuations in units of e^2 as a function of βE_c and Q_0^{eff}/e. The behavior of the system is periodic in Q_0^{eff} with period e, and we only need to consider offset charges in the range $Q_0^{\mathrm{eff}} \in]-e/2, e/2]$.[3] The essential physics behind the fluctuation results is now easy to understand: the tunneling electrons try to screen out the offset charge the best they can; for $Q_0^{\mathrm{eff}} = 0$ the screening is optimal if no extra electrons tunnel to or from the grain, whereas for $Q_0^{\mathrm{eff}} = e/2$ perfect screening is not possible due to charge granularity, and the optimal solution is to have the number of extra electrons fluctuate equally between 0 and 1 so that the grain is charge neutral in the average.

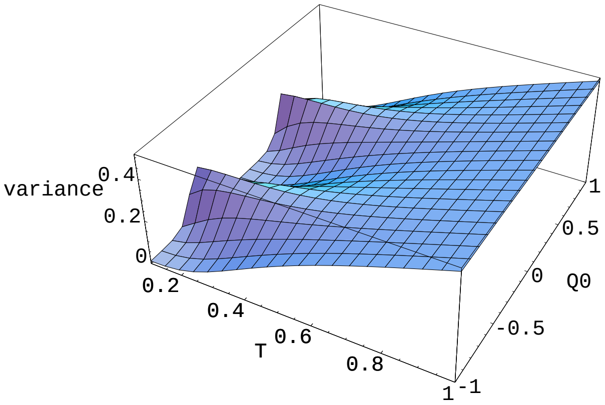

Figure 2.2: Fluctuations of the grain charge as a function of temperature and the offset charge.

So far we have considered the equilibrium case with $V_s = V_d$, and we have found that the amplitude of charge fluctuations in the grain depend on Q_0^{eff} and therefore on V_g. The charge fluctuations do not result in a net current through the grain since charges are equally likely to enter and exit the grain through either source or drain junction. Let us now introduce a small source-drain voltage V_{sd} such that $V_s = V_d + V_{sd}$. The structure is no longer symmetric, and there is a net current flowing from source to drain through the grain. The amount of this current is determined by the amount of asymmetry, *i.e.* by V_{sd}, and by the ease with which charge in the grain can fluctuate, which is related to δQ_{eq}^2.

For simplicity, let us first consider the zero temperature case so that at $V_{sd} = 0$ charge fluctuations are completely suppressed except if Q_0^{eff} is a half-integer multiple of e. Then charge can only flow through the grain if each step in the current-carrying process reduces the total electrostatic energy — that is, if it is profitable from an energy point of view for an electron to jump into the grain from one either the source or drain, and then continue to the other electrode. Thus, charge can only flow through the system if (now we denote the electric

[2] Actually, the sums can be related to the Jacobi elliptic theta function of the third type, $\vartheta_3[i\beta E_c(Q_0^{\mathrm{eff}}/(-e)), e^{-\beta E_c}]$ and its derivatives, see *e.g.* I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, And Products*, (Academic, Orlando, 1980).

[3] The offset charge Q_0^{eff} need not be an integer multiple of e since it depends on the continuously varying voltages V_s, V_d, and V_g.

charge of the tunneling particle by q rather than $-e$)

$$q[V^{(0)} + \frac{1}{2}V_{sd}] > E(Q+q) - E(Q) > q[V^{(0)} - \frac{1}{2}V_{sd}]$$

where $V^{(0)} = (V_s + V_d)/2$. For a symmetric structure with $C_s = C_d$ we get that charge can only flow if

$$|V_{sd}| > \frac{2Q_0^{\text{eff}} + |q|}{C_\Sigma}$$

Hence, if Q_0^{eff} vanishes, current can only flow if $V_{sd} > eC_\Sigma$, which is the effect known as *Coulomb blockade: at low source drain voltages the charging energy cost blocks current flow through the grain*; at a finite temperature thermal fluctuations smoothen out the blockade and some current can flow even at smaller voltages.[4] Since the offset charge can be controlled with V_g, the current blockade may be lifted if V_g is adjusted so that Q_0^{eff} is a half-integer multiple of e. This can be used to construct a switch in which the source-drain current is controlled by the gate voltage. The switch is known as the *single electron transistor* since its operation is based on changing Q_0^{eff} by less than the charge of a single electron.

We made two implicit assumptions in the above analysis: we assumed that the number of electrons in the grain was an integer, and we ignored any charge redistribution effects. The former assumption is justified if the grain is sufficiently well isolated from its surroundings so that quantum fluctuations do not destroy charge quantization. We may estimate the typical time scale of charge variations using classical circuit theory and an equivalent circuit for the double barrier structure (Figure 2.3) where we have included the resistances of the tunnel barriers. A net charge Q on the grain relaxes towards zero exponentially as $Q(t) = Q(0)e^{-t/(R_{\text{eff}}C_\Sigma)}$ where $C_\Sigma = C_s + C_d + C_g$ and $R_{\text{eff}} = R_s \parallel R_d \equiv R_sR_d/(R_s + R_d)$ is the total resistance between the grain and ground. Hence, the relevant time scale for charge variations is the RC-time $\tau_{RC} = R_{\text{eff}}C_\Sigma$. The quantum mechanical lifetime of a spontaneous charge fluctuation is $\tau_Q = \hbar/E(e) = 2C_\Sigma\hbar/e^2$. If this time is much shorter than the time needed to create a charge variation on the grain, $\tau_Q \ll \tau_{RC}$ or, equivalently, $R_{\text{eff}} \gg \hbar/e^2$, spontaneous charge fluctuations are greatly suppressed, and the assumption of integer charge on the grain is justified.

Our other implicit assumption, ignoring charge redistribution effects within the grain, is valid if we only consider frequencies that are much lower than the frequency of charge motion inside the grain, that is, plasma frequency. In the case of a three-dimensional grain this condition is not particularly restrictive since 3D plasma frequency is very high, approximately 1THz. In 2D the situation is somewhat more difficult since the plasma frequency for two-dimensional charge oscillations depends on the wavelength, but plasma frequencies in two-dimensional grains that are small enough to exhibit Coulomb blockade are also in the terahertz range.

Apart from these approximations charging phenomena are very robust. Since Coulomb blockade is effectively a classical phenomenon, its existence does not depend on subtleties like phase coherence. In is also quite insensitive to impurities — they may affect the precise gate voltage values at which a transmission resonance occurs, but they do not change the spacing of the resonance peaks. Consequently, Coulomb blockade structures are quite attractive

[4] The characteristic temperature scale T_C, which is given by $k_BT_C \approx \frac{e^2}{2C_\Sigma}$, increases with decreasing device size, implying that practical Coulomb blockade devices must be quite small. Using the spherical grain approximation and relative dielectric constant $\epsilon_r \approx 10$ we have $T_C \approx \frac{1\mu m}{R}K$.

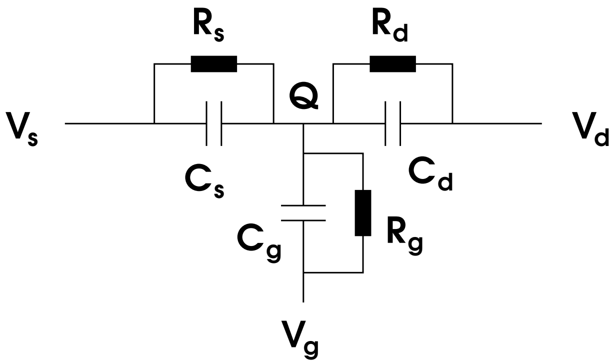

Figure 2.3: Equivalent circuit for a double barrier structure. The gate resistance R_g is usually much larger than the two other tunnel resistances and therefore we can set $R_g = \infty$.

from an application point of view, and indeed some applications based on them are already approaching commercialization. The major difficulty is the fabrication of sufficiently small structures that can operate at practical temperatures (77K or higher).

Quantitative analysis

A more complete description of Coulomb blockade phenomena can be obtained from a master (or rate) equation which describes how the probability $P(N,t)$ of finding net charge $-Ne$ on the grain evolves with time. The charge on the grain may change by tunneling across either junction, which leads to the set of equations

$$\begin{array}{rcl} \frac{dP(N,t)}{dt} &=& \Gamma_s[\Delta E_s^+(N-1)]P(N-1,t) - \Gamma_s[\Delta E_s^-(N)]P(N,t) \\ &+& \Gamma_s[\Delta E_s^-(N+1)]P(N+1,t) - \Gamma_s[\Delta E_s^+(N)]P(N,t) \\ &+& \Gamma_d[\Delta E_d^+(N-1)]P(N-1,t) - \Gamma_d[\Delta E_d^-(N)]P(N,t) \\ &+& \Gamma_d[\Delta E_d^-(N+1)]P(N+1,t) - \Gamma_d[\Delta E_d^+(N)]P(N,t) \end{array} \qquad (2.1)$$

where $\Delta E_s^+(N)$ is the energy change due to an additional electron entering the grain across the left junction if the grain initially has net charge $-Ne$, $\Delta E_s^-(N)$ is the energy change due an electron leaving the grain across the left junction, and $\Delta E_d^{\pm}(N)$ are the corresponding energy changes associated with tunneling events across the right junction. These energy changes are determined by electrostatic energies; for example, for a symmetric structure with $V_d = -V_s$ and $C_d = C_s$ we get

$$\Delta E_s^+(N) = \frac{e^2}{2C_\Sigma}[(N+1)^2 - N^2] - e(\frac{C_g}{C_\Sigma}V_g - V_s).$$

If the gate capacitance is much larger than the junction capacitances, this is approximately $\Delta E_s^+(N) \approx (N+\frac{1}{2})\frac{e^2}{C_g} - e(V_g - V_s)$.

The coefficients $\Gamma(\Delta E)$ give the tunneling rates across the barriers, and are determined by barrier properties and energy considerations. Probably the simplest way to obtain them is to assume that (i) density of states is constant on both sides of the junction,[5] (ii) tunneling

[5] Constant density of states is an approximation that can be satisfied only if the spacing between different energy levels is much small than the other relevant energy scales k_BT and $|eV_{sd}|$.

matrix element is the same for each tunneling process regardless of the quantum mechanical states involved in the process, (iii) all dissipative mechanisms can be ignored so that energy is conserved in each tunneling process, and (iv) chemical potentials between the two sides of the junction differ by $\Delta\mu$. Using these assumptions, the tunneling rate Γ is proportional to $\int_{-\infty}^{\infty} d\epsilon\, n_F(\epsilon - \mu - \Delta\mu)[1 - n_F(\epsilon - \mu)]$ where $n_F(\epsilon)$ is the Fermi function and the two factors give the proportions of full and empty states on the two sides of the barrier, respectively. Changing integration variable to $z = e^{\beta(\epsilon-\mu)}$ yields $\Gamma(\Delta\mu) = A\Delta\mu/(1 - e^{-\beta\Delta\mu})$ where A is a proportionality constant. The constant A can be determined by calculating the current across the junction, which is on one hand given by $-e[\Gamma(\Delta\mu) - \Gamma(-\Delta\mu)] = -eA\Delta\mu$, but on the other hand it is also given by U/R_T where $U = \Delta\mu/(-e)$ and R_T is the tunneling resistance. This yields $A = 1/(e^2 R_T)$ and hence $\Gamma(\Delta\mu) = \frac{1}{e^2 R_T}\frac{\Delta\mu}{1-e^{-\beta\Delta\mu}}$. The chemical potential difference $\Delta\mu$ can also be viewed as a energy difference ΔE between the states of the system before and after a tunneling event, so that $\Gamma(\Delta E) = \frac{1}{e^2 R_T}\frac{\Delta E}{1-e^{-\beta\Delta E}}$. Note that the $T = 0$ limit of these tunneling rates is particularly simple: $\Gamma(\Delta E) = \Theta(\Delta E)\frac{\Delta E}{e^2 R_T}$, which vanishes if the energy of the system would increase as a result of a tunneling process.

Most of the time we are interested in the steady state behavior of the device so that $\dot{P}\,(N) = 0$. In this case the master equation reduces to a matrix equation the solution of which yields the steady state probabilities $P(N)$. As an example of the master equation approach we now apply (2.1) to a symmetric structure $R_s = R_d$, $C_s = C_d$, and $V_s = -V_d = V_{sd}/2$, and determine the steady state probabilities $P(N)$ and the current through the structure as a function of the source-drain voltage.

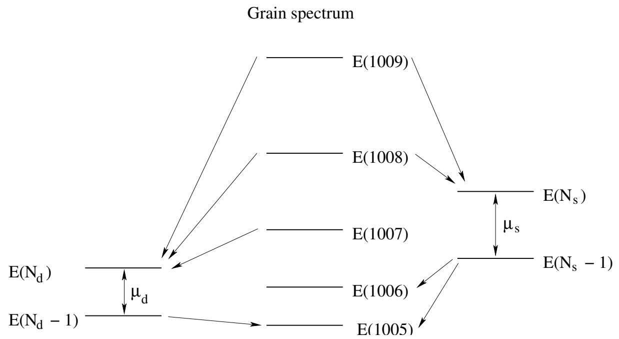

Figure 2.4: Definition of the continuous parameters N_s and N_d. In this particular case, $N_s \approx 1007.5$ and $N_d \approx 1006.5$, which implies that in steady state ($T = 0$) only the probabilities $P(1006)$ and $P(1007)$ are non-zero. The arrows indicate possible transitions involving configurations with the indicated number of electrons in the grain before the transition.

For simplicity we start with the zero temperature case and set $V_g = 0$. It is useful to define continuous parameters N_s and N_d such that $\mu_s + E(N_s - 1) = E(N_s)$ and similarly for N_d. Hence, electrons can enter the grain with N electrons from lead δ ($\delta = s, d$) if $N_\delta - 1 > N$, and they can leave the grain with N electrons by tunneling into lead δ if $N > N_\delta$, see Figure 2.4. For concreteness, let us assume that the polarity of the applied voltage is such that $\mu_s > \mu_d$ and hence $N_s > N_d$. We can distinguish three different ranges of grain charge N: (i) $N > N_s - 1$: all rates for tunneling into the grain are zero, (ii) $N_d > N$: all rates for tunneling out of the grain are zero, and (iii) net charges not covered by cases (i) or (ii): at least some in-tunneling and out-tunneling rates are nonzero. The states corresponding to cases (i) and (ii) can occur in steady state only if they can be reached from some state with lower (case (i)) or higher (case (ii)) charge N, consequently, $P(N)$ must vanish for $N > N_s$ and for $N_d - 1 > N$. Let us now define $N_0 = [N_d]$ and $m = [N_s] - [N_d]$ where $[x]$ is the largest integer not larger than x. Physically, N_0 corresponds the smallest number of electrons the grain may contain in steady state, and $N_0 + m$ corresponds to the largest number of electrons in the grain in steady state, *i.e.*, $P(N_0 + \alpha)$ is non-zero only for $\alpha = 0, \dots, m$. At $T = 0$ the tunneling rates are given by $\Gamma_s^+(N) = \frac{1}{R_s C_\Sigma}\mathrm{pos(N_s - N - 1)}$ and $\Gamma_d^-(N) = \frac{1}{R_s C_\Sigma}\mathrm{pos(N - N_d)}$, where $\mathrm{pos(x) = x\Theta(x)}$ is x if x is positive and zero otherwise.

We can now solve the master equation recursively starting from $\alpha = 0$ and obtain

$$\frac{P(N_0+\alpha)}{P(N_0)} = \prod_{j=0}^{\alpha-1} \frac{N_s - N_0 - j - 1}{N_0 - N_d + j + 1} = \frac{\Gamma(N_s - N_0)\Gamma(N_0 - N_d + 1)}{\Gamma(N_s - N_0 - \alpha)\Gamma(N_0 - N_d + 1 + \alpha)}$$

where $\Gamma(z)$ is the gamma function and we used $z = \Gamma(z+1)/\Gamma(z)$. For small voltages we have $[N_s] = [N_d]$, all rates are zero, hence only $P(N_0)$ is non-zero ($P(N_0) = 1$), and the current vanishes. The current can only start to flow when $[N_s] - [N_d] = 1$, which is exactly the condition we obtained earlier with a simple energetics argument. For large voltages $|V_s - V_d|$ the width of the allowed charge interval, m, is large, and we can approximate the distribution $P(N)$ by a Gaussian — note, for example, that if N_s and N_d are integers, the distribution is binomial, which may be approximated by a Gaussian for large m. We write $P(N) \propto e^{-\gamma(N-\overline{N})^2}$ which has its maximum for $N = \overline{N}$. We can determine $\overline{N}$ by symmetry: the only α-dependent terms are the two Γ-functions in the denominator, and their product is minimized if the arguments of the functions are equal, which yields $\overline{N} = \frac{N_s + N_d - 1}{2}$. The width of the distribution we obtain by writing $P(N_0+\alpha) \propto \exp[-\ln\Gamma(N_s - N_0 - \alpha) - \ln\Gamma(N_0 - N_d + 1 + \alpha)]$ and expanding the exponent to second order in α near $N_0 + \alpha = \overline{N}$. This yields $\gamma = \psi'(\frac{N_s - N_d + 1}{2})$ where $\psi(z)$ is the logarithmic derivative of the Γ-function, $\psi(z) = \frac{d}{dz}\ln\Gamma(z)$. For large voltages, *i.e.* for large $(N_s - N_d)$, we may approximate $\Gamma(z) \approx e^{z(\ln z - 1)}$ (Stirling's formula) so that $\psi(z) \approx \ln(z)$ and $\psi'(z) \approx 1/z$, for details see *e.g.* I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, And Products*, (Academic, Orlando, 1980). Hence, we have

$$P(N) \approx \frac{1}{\sqrt{\pi/\gamma}} e^{-\gamma(N-\overline{N})^2}$$

where $\overline{N} = \frac{N_s + N_d - 1}{2}$ and $\gamma = \frac{2}{N_s - N_d}$, and the prefactor was determined by normalization. Thus, the probability $P(N)$ is largest for those charge states N which are strongly coupled to both leads, and decreases to zero near the edges of the allowed interval — roughly speaking, states with large N are rapidly emptied into the right lead but only slowly replenished from the left lead, while the opposite is true for states with small N. In the middle of the allowed range tunneling-in and tunneling-out are equally fast.

Inserting this into the expression for the current at the left junction $I = \sum_N \frac{e}{R_s C_\Sigma}(N_s - N - 1)P(N)$ yields

$$I = \frac{e}{R_s C_\Sigma}\frac{N_s - N_d - 1}{2} = \frac{1}{R_s}\left(V_s - \frac{e}{2C_\Sigma}\right) = \frac{1}{2R_s}\left(V_{sd} - \frac{e}{C_\Sigma}\right).$$

Thus, the large-voltage IV-curve has the same slope as the IV-curve in the absence of Coulomb blockade, but it is shifted to larger voltages by amount $e/C_\Sigma = 2E_C/e$.

The finite temperature case is harder to analyze except in the limit $E_C \gg k_B T \gg |eV_{sd}|$ when at most two charge states are occupied with significant probability. The resulting two-state system can be solved straightforwardly to obtain

$$\begin{aligned} I = \quad & \frac{\Delta E_s \Delta E_d}{eR_s} \frac{\sinh[\frac{1}{2}\beta(\Delta E_s - \Delta E_d)]}{\cosh[\frac{1}{2}\beta(\Delta E_s + \Delta E_d)] - \cosh[\frac{1}{2}\beta(\Delta E_s - \Delta E_d)]} \\ \times \quad & \frac{1}{\Delta E_s \coth[\frac{1}{2}\beta\Delta E_s] + \Delta E_d \coth[\frac{1}{2}\beta\Delta E_s]} \end{aligned}$$

where $\Delta E_s = E(N-1) - E(N) + eV_s$ and $\Delta E_d = E(N-1) - E(N) - eV_s$. The differential conductance $G(V) = \frac{dI}{dV_{sd}}$ at zero bias is now

$$G(V_{sd} = 0) = \frac{1}{4R_s} \frac{\beta \Delta E}{\sinh(\beta \Delta E)} \tag{2.2}$$

where $\Delta E = E(N-1) - E(N)$. As a function of gate voltage we have $G(V_g) = \frac{1}{4R_s} \frac{\beta e(V_g - V_g^{(N)})}{\sinh[\beta e(V_g - V_g^{(N)})]}$ where $V_g^{(N)}$ is the gate voltage value for which $E(N) = E(N-1)$. The differential conductance at zero applied bias voltage is plotted in Figure 2.5.

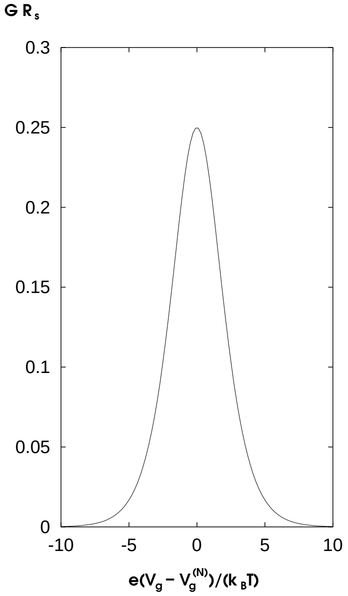

Figure 2.5: Differential conductance at zero bias voltage as a function of the gate voltage. The full width at half maximum is $\Delta V_g \approx 4.355 \frac{k_B T}{e}$.

Hence, the zero bias conductance exhibits a series of peaks and valleys as the gate voltage is varied, corresponding to different values of N. All peaks have the same, temperature independent height $\frac{1}{4R_s}$ which is exactly half of the large voltage conductance, whereas the peak widths are linearly proportional to temperature T. The peak width is directly related to the number of single particle states that contribute to G, which increases as T, but since the total conductance of the peak is constant, the contribution from each single particle state decreases with increasing temperature as T^{-1}. This behavior can be verified experimentally in semiconducting quantum dots, where the density of states is very small so that $D(\epsilon_F)k_BT \ll 1$ and the number of contributing single particle states does not increase smoothly with T. In this case the T-dependence of the conductance peaks follows the T^{-1}-law as has been seen experimentally.

In the appendix A we discuss the role if the electromagnetic environment on Coulomb blockade. It is clear from the beginning that the electromagnetic environment — that is, the impedance of the equipment connected to the tunnel junctions — plays an important role in CB structures: we have thus far considered blockade in a system with two junctions and concluded that the resistivities must exceed the quantum resistance for the blockade to appear. In the appendix we analyze a more general case where a single junction is connected to an impedance $Z(\omega)$, and we will find that a conventional blockade emerges if the low-frequency impedance exceeds 26 kΩ, while for lower impedances the low-frequency current follows a power law $I \sim V^{1+\alpha}$ where $\alpha \approx Z(0)/26k\Omega$. The importance of this result will become more clear in the section on Luttinger liquids, where we discuss another origin of power laws in conductance. Experimentally, it can be quite difficult to infer the origin of this type of simple behavior that can be attributed to different physical reasons.

Home problem 2: Coulomb staircase
Consider a strongly asymmetric double barrier structure such that $R_R \gg R_L$ meaning that the central grain is much more strongly coupled to the left lead. A voltage V_L is applied to the left lead and voltage $V_R = -V_L$ is applied to the right lead. Both leads are assumed to be internally in equilibrium at all times. For simplicity, consider the zero temperature limit only.

1. Without doing any calculations but simply relying on physical argumentation, determine qualitatively the N-dependence of $P(N; V_L)$.

2. Using the master equation (2.1) show that your result in part 1 is correct.

3. Using the result of part 2, show that the current-voltage characteristics of this asymmetric structure are step-like, *i.e.* the current is nearly constant for a range of voltages V_L, and then increases steeply to new plateau. What is the step height ΔI? What is the physical reason for the current steps?

2.2 Correlations in quantum matter

The electron-electron interaction has also more subtle effects than those describable within the capacitance approach used in the discussion of the Coulomb blockade. Interactions between charge carriers result in correlations in particle motion, *i.e.* the motion of one particle affects that of another one. For the mesoscopic phenomena discussed in Chapter 1 these correlation effects are of little qualitative importance, and, if at all necessary, can be accounted for by adjusting the phase breaking time (see the discussion in Sec. 1.2). The fact that we can "absorb" the effect of the electron interaction by simply renormalizing a parameter may appear rather surprising. In fact, this state of affair is quite common for interacting fermions, and applies to a variety of systems, spanning very different energy and length scales: Conduction electrons in ordinary metals and semiconductors, liquid He^3, the interior of neutron stars, nuclear matter, and quark-gluon plasmas all belong to this class. To a first (and often very good!) approximation, properties of these systems can be obtained by simply treating the fermions as independent particles with adjusted, or *renormalized* parameters. In the elementary text book treatment of conduction electrons in metals one does not even bother to carry out this renormalization (c.f. Sommerfeld's *independent electron model)*, and still the theory gets most of the physics right.

How can such a simple-minded approach possibly work? To answer this question, one must include the electron-electron interaction (or, in a more general setting, the interaction among whatever fermions that make up the system) and study its effect (or "non-effect"!) on the physics. There are many ways of doing this, the (historically) most important being *perturbation theory*. Other approaches include *variational calculations* and *renormalization group methods*. The picture that emerges from all these different methods may be interpreted within *Landau Fermi liquid theory*, one of the high points of theoretical physics from the last century (Lev Landau, 1956). Landau's theory explains why the independent electron approximation works so well in many condensed matter applications (like the study of electronic properties of mesoscopic systems or bulk metals). Equally important, the theory points to its own demise for electrons that organize to form collective phases of matter: Already the same year as Landau worked out his theory, Leon Cooper pointed out that a weak attractive interaction among electrons in a metal − mediated by phonons − would cause an "instability" of the electron liquid, leading to superconductivity. This phase of electronic quantum matter cannot be described by Landau Fermi liquid theory but requires a very different type of theory. Since the late 70s an increasing number of condensed matter systems have been discovered which do not conform to Fermi liquid theory. Examples include fractional quantum Hall systems, high-temperature superconductors and other complex oxides, "heavy fermion" materials, quasi-one-dimensional organic conductors and carbon nanotubes, trapped ultra cold Fermi gases, and a rapidly growing number of specially designed semiconductor-based nanoscale devices. Our theoretical understanding of these systems remain scattered and incomplete. Despite a concerted effort by many researchers, real progress is coming only slow and piecewise. In most cases we understand *why* Landau Fermi liquid theory fails, but there is yet no consensus that the proposed alternative theories properly "do the job" [6]. In other cases the very breakdown of Fermi liquid theory remains somewhat of a mystery. A case in point is the so called "optimally doped" metallic phase of the Cu-O based high-temperature superconductors where *every* experimentally measured non-equilibrium property (resistivity, Raman scattering, nuclear relaxation rate,...) is in conflict with the Fermi liquid picture of how a metal should behave. A recent count estimates that more than 10^5 scientific articles have been published about this problem since the discovery of the high-temperature superconductors in 1987. Yet, there is no consensus about what mechanism is responsible for their

[6] There is one striking exception to this statement: The theory of the fractional quantum Hall effect (FQHE) pioneered by Robert Laughlin. For his achievement, Laughlin shared the 1998 Nobel Prize in Physics with Horst Störmer and Daniel Tsui, the two experimentalists who discovered the effect. More about this fascinating physics in the next chapter!

violation of "text-book" Fermi liquid physics.

The importance and depth of the problems involved – their common denominator being the presence of strong correlations – has turned the field of strongly correlated quantum matter into one of the most active in Physics today. The impetus is two-fold: On the practical side, the high sensitivity of many correlated-electron materials to changes in external parameters holds promise for the development of new technologies (sensors, elements for control and diagnostics,...), while in other applications this same property may interfere with desired device operation, in particular at the nano or mesoscopic scale. Either way, it is crucial to develop the theoretical tools required for an understanding of the underlying physics. At a more "fundamental" level, the very existence of strongly correlated quantum matter raises difficult conceptual questions about how to understand the emergence of the defining properties at the mesoscopic or macroscopic scale from those at the atomic level.

The problem is hard on two counts. First, given a model of a many-particle system where interaction and correlation effects are built in from "scratch" (and not simply added as a perturbation) we lack the conceptual, mathematical, and computational tools to efficiently carry out a reliable analysis, except in a few fortuitous circumstances.[7] Secondly, materials that exhibit effects from strong electron correlations – such as metallic oxides, intermetallic compounds, or organic conductors – often have a complex structure that is difficult to characterize in detail. Also, the typical hypersensitivity of these materials to atomic imperfections and sample preparation methods make experiments hard to reproduce. All is not gloom, however. Progress in nanotechnology has made possible the manufacturing of nanoscale structures that show strong correlation effects, and which are much easier to control experimentally than "traditional" bulk materials. An important example is the *Kondo effect* in quantum dots, where high-precision measurements have allowed tests of theory at an unprecedented level. Another example is the *Mott transition* in ultra-cold fermionic gases trapped in one-dimensional optical lattices. The study of these and related phenomena – to be discussed in Sec. III – has opened up a vista on problems that for a long time were considered to be out-of-reach for experiments. Indeed, the study of correlation effects in nanoscale structures has boomed in the last few years, both in experiment and theory. There is a perception among many researchers that this is the "way to go" to make progress on the notoriously difficult problem of strongly correlated quantum matter. Yet, many questions remain unanswered, and new questions – specific to the nano- or mesoscale – have appeared. One such very basic question is how to understand the interplay between coherence and correlation effects. In the single-particle picture of mesoscopic physics each electron – or *quasiparticle*, to use the language of Landau Fermi liquid theory – is associated with a wave function that carries a phase that is well-defined on time scales shorter than the phase breaking time. This leads to all kinds of strange and beautiful coherence phenomena, some of which have been discussed in the previous chapter. If interactions among electrons become more pronounced, as a result of changing temperature, magnetic field, or some other experimental parameter, the single-particle physics may break down and give way to a *strongly correlated system* where the electrons organize into collective states. Given this, how do we account for coherence effects at the nanoscale? Are they lost, or do they reappear at the new collective level? This is a difficult and fascinating question that defines much of current

[7] The theory of the FQHE, mentioned in the previous footnote, is one example. Another class of problems where there has been substantial theoretical progress are those that can be reduced to a one-dimensional geometry for which there exist a number of powerful analytical and computational techniques *(Bethe Ansatz,* Density Matrix Renormalization Group,...). More about that in Sec III.

research in the field. We shall touch on it as we go along, but an answer must await progress in experiments as well as in theory.

The rest of the chapter is organized as follows: In section 2.2.1 we give a pedestrian view of perturbation theory applied to interacting electrons. This section serves a two-fold purpose: First, to introduce some standard concepts and methods from many-particle physics (second quantization, Feynman diagrams, ...), and secondly, to provide a "microscopic" underpinning of Fermi liquid theory. This theory is the topic of Sec 2.2.2. After a general introduction we look at a few simple applications of Landau's theory to liquid ^{3}He, and then discuss how to extend it to electrons in metals. Most of the material follows Landau's original approach, but we shall also look at the more "modern" view of a Fermi liquid as a fixed point theory under the renormalization group. Sec 2.2.3 deals with failures of Fermi liquid theory, taking us into the exotic and sometimes bewildering realm of *non-Fermi liquid physics*. We shall here focus on one particular paradigm for thinking about non-Fermi liquids, that of *quantum criticality*, and briefly discuss its application to *heavy fermion materials*. In Sec 2.2.4 finally, we give a fairly detailed exposition of another paradigm that replaces that of Landau's whenever electrons are confined to one dimension (as in quantum wires or in carbon nanotubes): the *Luttinger liquid*. This section also contains an introduction to the powerful method of *bosonization*.

2.2.1 Interacting electrons: perturbative approach

As our case study, let us take a metal and write down its Hamiltonian H:

$$H = H_{el} + H_{ion} + H_{el-ion} + H_{el-imp} \tag{2.3}$$

where

$$\begin{aligned} H_{el} &= H_0 + H_{int} && (2.4) \\ H_{ion} &= H_{ion,kin} + H^0_{ion-ion} + H_{phonon} && (2.5) \\ H_{el-ion} &= H^0_{el-ion} + H_{el-phonon} && (2.6) \\ & && (2.7) \end{aligned}$$

The term H_{el} represents the part of H that contains only the conduction electrons: H_0 is the kinetic part, while H_{int} is the Coulomb interaction among the conduction electrons. H_{ion} in turn is the part that contains only the ions that make up the crystal lattice. $H_{ion,kin}$ is the kinetic energy of the ions, $H^0_{ion-ion}$ describes the interaction between the ions in their equilibrium positions, and H_{phonon} is the correction to this interaction from vibrations of the ions around their equilibrium positions. H_{el-ion} in turn controls the interaction between the conduction electrons and the ions, with the term H^0_{el-ion} representing the interaction with the ions in their equilibrium positions, and $H_{el-phonon}$ the electron-phonon term that encodes the effect of the lattice vibrations on the conduction electrons. Finally, H_{el-imp} is an interaction between conduction electrons and impurities or lattice defects.

Elementary text books in condensed matter physics start with

$$H_0 + H^0_{el-ion} = \sum_j \frac{\boldsymbol{p}_j^2}{2m} + \frac{1}{2}\sum_{j,\ell} V_{el-ion}(\boldsymbol{r}_j - \boldsymbol{R}_\ell), \tag{2.8}$$

where the index j labels the electron momenta $\boldsymbol{p}_j$ and coordinates $\boldsymbol{r}_j$, with ℓ labeling the ions with coordinates $\boldsymbol{R}_\ell$. V_{el-ion} is the electron-ion potential. Assuming that the lattice is periodic, one shows that the effect of the potential term can be encoded by replacing the free electron wave functions (plane waves) by Bloch wave functions. One then adds the electron-impurity term H_{el-imp} and studies its effect on the Bloch electrons, most often using a relaxation time approximation. Finally, one cranks up the temperature, and includes also the electron-phonon term $H_{el-phonon}$. (*Nota bene*: The Hamiltonian knows nothing about temperature. It is only the *effect* of the electron-phonon term $H_{el-phonon}$ that becomes important at finite temperature, and therefore has to be included.) But what about the Coulomb interaction H_{int} among the conduction electrons? In fact, it is usually neglected! Since all interaction terms in the Hamiltonian in (2.3) are of the same order of magnitude, all being Coulomb interactions, it is *a priori* not obvious that this approach should work. That a separation between electron-electron and electron-ion interactions makes sense is another matter: The motions of the ions are slow while the electrons move fast, hence the two dynamics are effectively decoupled *(Born-Oppenheimer approximation)*. But how can one exclude the electron-electron interaction all together? Here we "turn the table" and focus on the electron Hamiltonian

$$H'_{el} = \sum_i \frac{\boldsymbol{p}_i^2}{2m_e} + \frac{1}{8\pi\epsilon_0} \sum_{i,j} \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} + H_+. \tag{2.9}$$

To satisfy the condition of charge neutrality, we have included the lattice ions as a fixed uniform positive charge background, representing it by the ion *self-energy term* H_+.[8] Smearing out the ion charge uniformly over space and decoupling it from the conduction electrons means that we neglect the presence of the crystal lattice and its effect on the electrons. In particular, it means that we forsake the possibility to have a nontrivial band structure or, for that matter, an interaction between electrons and phonons, or between electrons and whatever defects or impurities are embedded in the lattice. It also means that we take the electron coordinates as continuum variables. The continuum limit approximation introduces some formal problems, which, however, can rather easily be taken care of. The neglect of band structure and interactions with phonons or impurities may appear as a more serious distortion of the facts. However, as long as we are only interested in the *qualitative* aspects of the electron-electron interaction, the omission of the lattice is legitimate.[9] The model in Eq. (2.9) goes under many names: The *jellium model*, the *interacting electron gas*, or, maybe most appropriate – considering the high density of conduction electrons in a metal – the *electron liquid*.

To make progress we shall carry out *second quantization* of H_{el} in (2.9). (See Appendix 3.1.1 for details.) As a first step we pick a single-particle basis with states $|\lambda\rangle \equiv |\boldsymbol{k}_\lambda, \sigma_\lambda\rangle$, with $\boldsymbol{k}_\lambda\,[\sigma_\lambda]$ the single-particle momentum [spin] in the state with label λ. Evaluating the matrix elements

$$\langle\lambda'|\frac{\boldsymbol{p}^2}{2m_e}|\lambda\rangle \;=\; \frac{\boldsymbol{p}^2}{2m_e}\delta_{\lambda\lambda'} \equiv E_\lambda \delta_{\lambda\lambda'} \tag{2.10}$$

[8] As a reminder of having added H_+ to H_{el} in Eq. (2.3) we have put a prime on H_{el}.

[9] This is actually a bit optimistic. There exist a number of situations where lattice and impurity effects feed back dramatically on the electron-electron interaction. A case in point is the *nesting* of the square Fermi surface for electrons hopping on a two-dimensional lattice. "Nesting" means that one section of the Fermi surface can be connected to another section via a constant vector. This feature results in a strong enhancement of interaction effects. Nested Fermi surfaces have been invoked by some theorists trying to explain the strange metallic behavior of complex oxides, like the high-temperature superconductors. We shall come back to the concept of nesting when discussing singular scattering of electrons in one dimension.

$$\frac{e^2}{4\pi\epsilon_0}\langle\lambda'\mu'|\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}|\lambda\mu\rangle \;=\; \frac{e^2}{4\pi\epsilon_0 k_\nu^2}\delta_{\sigma_\lambda\sigma_{\lambda'}}\delta_{\sigma_\mu\sigma_{\mu'}} \equiv V_\nu\delta_{\sigma_\lambda\sigma_{\lambda'}}\delta_{\sigma_\mu\sigma_{\mu'}}, \tag{2.11}$$

with $k_\nu = |\boldsymbol{k}_\nu| \equiv |\boldsymbol{k}_\lambda - \boldsymbol{k}_{\lambda'}| = |\boldsymbol{k}_{\mu'} - \boldsymbol{k}_\mu|$, and following the second-quantization manual in the Appendix, we obtain:

$$H'_{el} = \sum_\lambda E_\lambda c^\dagger_\lambda c_\lambda + \frac{1}{2}\sum_{\lambda\mu\nu} V_\nu c^\dagger_{\lambda-\nu}c^\dagger_{\mu+\nu}c_\mu c_\lambda + H_+. \tag{2.12}$$

Here, and in what follows, we use the short-hand notation $\lambda-\nu \equiv (\boldsymbol{k}_\lambda - \boldsymbol{k}_\nu, \sigma_\lambda)$ and $\mu+\nu \equiv (\boldsymbol{k}_\mu + \boldsymbol{k}_\nu, \sigma_\mu)$, exploiting the fact that the Coulomb interaction, with momentum transfer indexed by ν, is spin independent.

The second-quantized form of H'_{el} in Eq. (2.12) allows us to write down a perturbative expansion of the groundstate energy E on compact form:

$$E = \langle 0|H_0|0\rangle + \langle 0|H_{int}|0\rangle + \sum_i \frac{\langle 0|H_{int}|i\rangle\langle i|H_{int}|0\rangle}{E_i - E_0} + \text{higher order terms} + H_+ \tag{2.13}$$

with $H_0 \equiv \sum_\lambda E_\lambda c^\dagger_\lambda c_\lambda$ and $H_{int} \equiv (1/2)\sum_{\lambda\mu\nu} V_\nu c^\dagger_{\lambda-\nu}c^\dagger_{\mu+\nu}c_\mu c_\lambda$, and where

$$|0\rangle = \prod_{\substack{k<k_F\\ \sigma}} c^\dagger_{\boldsymbol{k}\sigma}|\text{vacuum}\rangle \tag{2.14}$$

is the groundstate of the non-interacting theory H_0.[10] The first-order term, $\langle 0|H_{int}|0\rangle$, is a sum over matrix elements $\langle 0|V_\nu c^\dagger_{\lambda-\nu}c^\dagger_{\mu+\nu}c_\mu c_\lambda|0\rangle$. It is convenient to represent these matrix elements by *Feynman diagrams* (see Appendix C). The basic diagram for a first-order process is depicted in Fig. 2.6.

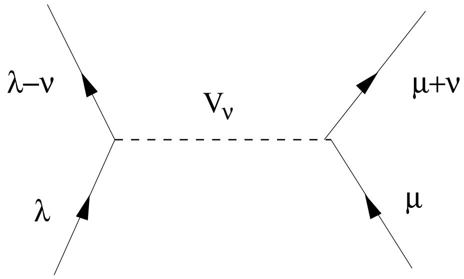

Figure 2.6: Feynman diagram representing the matrix element $\langle out|V_\nu c^\dagger_{\lambda-\nu}c^\dagger_{\mu+\nu}c_\mu c_\lambda|in\rangle$.

Since | in⟩ =| out⟩ =| 0⟩ (using the notation from the Appendix) we have to identify incoming and outgoing electron lines. There are two way of doing this, corresponding to the Feynman diagrams in Figs. 2.7 and 2.8 (b), respectively.

[10] In the electron liquid model employed here, the dispersion relation is spherically symmetric, implying that the ground state $|0\rangle$ in D=3 is a filled Fermi sphere, with the *Fermi surface* the surface of the sphere. (In D=2 [D=1] the "Fermi sphere" is a filled disc [line].) Taking the lattice into account implies the possibility of a nontrivial bandstructure, with a Fermi surface that may take on a complex shape. The best available methods to obtain the groundstate $|0\rangle$ in the presence of the lattice are those method based on density functional theory (pioneered by Walter Kohn who was awarded the 1998 Nobel Prize in Chemistry for his contribution).

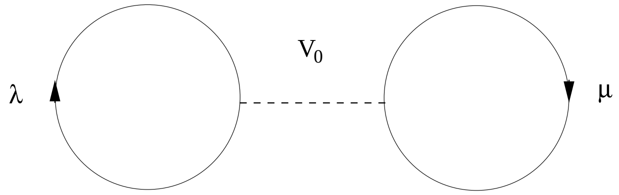

Figure 2.7: Feynman diagram for the matrix element $\langle 0|V_0 c^\dagger_\lambda c^\dagger_\mu c_\mu c_\lambda|0\rangle$.

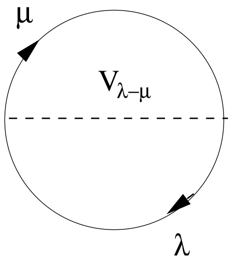

Figure 2.8: Feynman diagram for the matrix element $\langle 0|V_{\lambda-\mu} c^\dagger_\mu c^\dagger_\lambda c_\mu c_\lambda|0\rangle$.

The diagram in Fig. 2.7 is obtained by taking $\lambda = \lambda - \nu$ and $\mu = \mu - \nu$, implying that $\nu = 0$. Anticommuting c_λ through c_μ and $c^\dagger_\mu$ and carrying out the sum in H_{int}, it immediately follows that the diagram in Fig. 2.7 contributes a term to the groundstate energy that is *identical* to the Hartree term in mean field theory:

$$E_{\text{Hartree}} = \frac{V_0}{2} \sum_{\lambda\mu} n_\lambda n_\mu. \tag{2.15}$$

Here n_λ and n_μ are the groundstate densities of electrons with quantum numbers λ and μ, respectively. In the thermodynamic limit (i.e. with a very large number of conduction electrons) one can show that the Hartree term in Eq. (2.15) cancels the ion self-energy term H_+. This simplifying feature is special to the electron liquid, and does not hold on a lattice, where the electrons see a non-uniform background of ions.

Turning to the diagram in Fig. 2.8, we have again identified $|\text{in}\rangle$ and $|\text{out}\rangle$ states, now by taking $\lambda = \mu + \nu$, implying that $\nu = \lambda - \mu$. By again anticommuting the electron operators in H_{int} and performing the sum over λ and μ one obtains the contribution:

$$E_{\text{exchange}} = -\sum_{\lambda\mu} \frac{V_{\lambda-\mu}}{2} n_\lambda n_\mu \delta_{\sigma_\lambda \sigma_\mu}, \tag{2.16}$$

i.e. the well known Hartree-Fock exchange term from mean-field theory (for a review, see the Appendix).

Summarizing, we have obtained the satisfying result that *lowest-order perturbation theory for the electron liquid exactly reproduces mean-field theory.* Strengthened by this success, let us next tackle the second-order term in the perturbative expansion in Eq. (2.13):

$$\begin{aligned}&\sum_i \tfrac{\langle 0|H_{int}|i\rangle\langle i|H_{int}|0\rangle}{E_i-E_0}\\ &= \tfrac{1}{4}\sum_i \sum_{\substack{\lambda\mu\nu\\ \lambda'\mu'\nu'}} V_\nu V_{\nu'} \tfrac{\langle 0|c^\dagger_{\lambda-\mu}c^\dagger_{\mu+\nu}c_\mu c_\lambda|i\rangle\langle i|c^\dagger_{\lambda'-\mu'}c^\dagger_{\mu'+\nu'}c_{\mu'}c_{\lambda'}|0\rangle}{E_i-E_0}\\ &= \text{const.}\times\sum_i \int d^3\boldsymbol{k}_\nu d^3\boldsymbol{k}_{\nu'}\tfrac{1}{k_\nu^2}\tfrac{1}{k_{\nu'}^2}\times[......].\end{aligned} \qquad (2.17)$$

In the third line we have replaced the sums over ν and ν' by integrals, and used that $V_\nu \sim 1/k_\nu^2$. The "[... ...]" is shorthand for "everything else" (which, at this point of our analysis, is not important). To find out about possible constraints on the integration variables k_ν and $k_{\nu'}$ in (2.17), we again pass to diagrammatic language. From the Appendix, we know that the basic second-order Feynman diagram can be drawn as in Fig. 2.9.

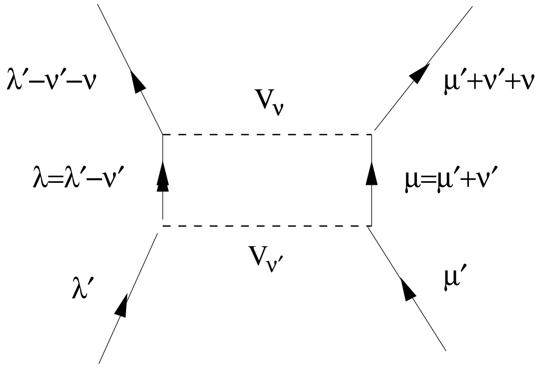

Figure 2.9: Basic second-order Feynman diagram.

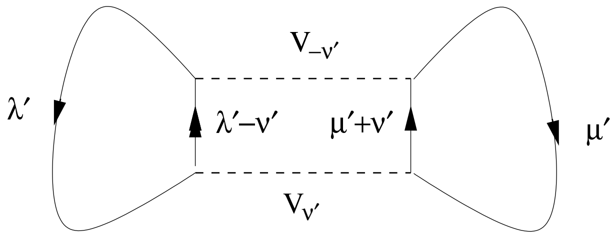

Figure 2.10: Feynman diagram representing the second-order process $\langle 0|V_{-\nu'}c^\dagger_{\lambda'}c^\dagger_{\mu'}c_{\lambda'-\nu'}c_{\mu'+\nu'}|i\rangle\langle i|V_{\nu'}c^\dagger_{\lambda'-\nu'}c^\dagger_{\mu'+\nu'}c_{\lambda'}c_{\mu'}|0\rangle$.

Identifying incoming and outgoing states, $|\text{in}\rangle = |\text{out}\rangle = |0\rangle$, produces two new diagrams, one of which is depicted in Fig. 2.10. This particular diagram is obtained from that in Fig. 2.9, by taking $\lambda' = \lambda' - \nu' - \nu$ and $\mu' = \mu' + \nu' + \nu$, implying that $\nu = -\nu'$, i.e. $\boldsymbol{k}_\nu = -\boldsymbol{k}_{\nu'}$. Carrying out the integrations we thus obtain the second-order contribution

$$\text{const.}\times\sum_i \int d^3\boldsymbol{k}_\nu d^3\boldsymbol{k}_{\nu'}\frac{1}{k_\nu^2}\frac{1}{k_{\nu'}^2}\delta^{(3)}(\boldsymbol{k}_\nu+\boldsymbol{k}_{\nu'})\times[......] \to \infty. \qquad (2.18)$$

The divergence in (2.18) is bad news. What to do?

The standard route is to resort to perturbative quantum field theory which provides a machinery to *re-sum the perturbation series* and obtain a finite answer. Infinite pieces of various sums and integrals (like that coming from the second-order contribution in Eq. (2.18) are isolated and made to cancel each other. For the electron liquid this feat was first achieved by Gell-Mann and Brückner in 1957, building on earlier work by Bohm and Pines. The particular calculational strategy taken by Gell-Mann and Brückner is known as the "random phase approximation" (RPA), and becomes exact in the limit of an infinite density of electrons. Unfortunately, its exposition goes beyond the present course.[11] Here we shall instead "simulate" the result of the RPA calculation by using a few tricks and shortcuts (and sticking with ordinary quantum mechanics).

By inspection of (2.18) it is clear that the heart of the problem is that we have a long-range Coulomb interaction, falling off as $1/r$, with r the distance between two electrons:

$$\sim \int \frac{\mathrm{e}^{i\boldsymbol{k}\cdot\boldsymbol{r}}}{r} d^3\boldsymbol{r} = \frac{4\pi}{k^2} \to \infty \,\text{as}\, |\boldsymbol{k}| \to 0. \tag{2.19}$$

To cure the problem we "regulate" the theory (as it is called in technical jargon). That is, we remove the zero mode $\boldsymbol{k} = 0$, "cutting off" the tail of the Coulomb interaction at large distances. This takes care of the divergence of the second-order term in (2.18). However, the full perturbation series still diverges since the second- and higher-order terms can be made arbitrarily large. So we need at least one more trick: We split the Coulomb interaction in two parts; one that acts over short distances only ($k > k_c$), and one that acts over long distances ($k < k_c$).[12] The short-range part roughly falls off as an exponentially screened (Yakawa) potential, while for sufficiently large values of k_c the long-range part varies very slowly in space. This last property can be taken advantage of. Describing the effect of the long-range Coulomb interaction by a time-dependent electric field $\boldsymbol{E}$, we write:

$$\boldsymbol{E} = -\frac{\partial \boldsymbol{A}}{\partial t} - \nabla \sum_{\substack{i,j \\ i\neq j}} V_{k<k_c}(|\boldsymbol{r}_i - \boldsymbol{r}_j|) \approx -\frac{\partial \boldsymbol{A}}{\partial t}, \tag{2.20}$$

where in the last term we have dropped the gradient contribution, exploiting the fact that $V_{k<k_c}(r)$ varies slowly in space. Writing the vector potential $\boldsymbol{A}$ and the field $\boldsymbol{E}$ as Fourier series,

$$\boldsymbol{A}(\boldsymbol{r}) = \frac{1}{\sqrt{V\epsilon_0}} \sum_{\boldsymbol{k}} \hat{k} Q_k \exp(i\boldsymbol{k}\cdot\boldsymbol{r}), \tag{2.21}$$

$$\boldsymbol{E}(\boldsymbol{r}) = -\frac{1}{\sqrt{V\epsilon_0}} \sum_{\boldsymbol{k}} \hat{k} P_k \exp(i\boldsymbol{k}\cdot\boldsymbol{r}), \tag{2.22}$$

[11] The RPA calculation is a main staple of any Ph.D. level course in many-particle physics, but requires an extensive grounding in quantum field theory techniques and Feynman diagrammatics. The notion of "random phase approximation" comes from the neglect of certain terms in the perturbative expansion that carry randomly varying phases. Physically, the approximation is equivalent to neglecting the coupling between density fluctuations of different wave vectors. The RPA has been widely used in condensed matter and nuclear physics.

[12] What is a "short distance" depends on the choice of k_c. One can show that a good choice is to take $k_c \approx \omega_p/v_F$, where v_F is the Fermi speed and ω_p the "plasma frequency", to be defined below. For our qualitative discussion here, however, we may keep k_c as an arbitrary parameter.

we can take the canonically conjugate variables Q_k and $P_k \equiv \partial Q_k/\partial t$ as *collective coordinates* that describe the effect of the long-range Coulomb interaction of the electrons.[13] To see how this procedure simplifies the problem, let us go back to first-quantized language and express the Coulomb interaction H_{int} between the electrons as a Fourier series (with the zero mode subtracted):

$$\begin{aligned}
H_{int} &= \frac{e^2}{2V\epsilon_0}\sum_{\substack{i,j\\ i\neq j}}\left(\sum_{0<k<k_c}+\sum_{k\geq k_c}\right)\frac{\exp(i\boldsymbol{k}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j))}{k^2}\\
&= \frac{e^2}{2V\epsilon_0}\sum_{i,j}\left(\sum_{0<k<k_c}+\sum_{k\geq k_c}\right)\frac{\exp(i\boldsymbol{k}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j))}{k^2}-\frac{Ne^2}{2V\epsilon_0}\sum_{k\neq 0}\frac{1}{k^2}\\
&\equiv H_{k<k_c}+H_{k>k_c}-\frac{Ne^2}{2V\epsilon_0}\sum_{k\neq 0}\frac{1}{k^2}. \qquad (2.23)
\end{aligned}$$

Quantizing $\boldsymbol{A}$ and $\boldsymbol{E}$, i.e. taking $[Q_{\boldsymbol{k}'}, P_{\boldsymbol{k}}] = i\hbar\delta_{\boldsymbol{k}',\boldsymbol{k}}$, we then replace $H_{k<k_c}$ by an integral over the effective electric field $\boldsymbol{E}$,

$$H_{k<k_c} = \frac{\epsilon_0}{2}\int d^3\boldsymbol{r}\boldsymbol{E}^2 \qquad (2.24)$$

and absorb the vector potential $\boldsymbol{A}$ into the kinetic term of H_{el} by *minimal coupling*

$$\boldsymbol{p}_i \rightarrow \boldsymbol{p}_i + e\boldsymbol{A}. \qquad (2.25)$$

We can then write $H_{el} = H_0 + H_{int}$ as

$$\begin{aligned}
H_{el} &\approx \sum_i \frac{1}{2m_e}\left(\boldsymbol{p}_i+\frac{e}{\sqrt{V\epsilon_0}}\sum_{k<k_c}\hat{k}\,Q_k\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)\right)^2\\
&+ \frac{e^2}{2V\epsilon_0}\sum_{i,j}\sum_{k\geq k_c}\frac{\exp[i\boldsymbol{k}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j]}{k^2}-\frac{Ne^2}{2V\epsilon_0}\sum_{k\neq 0}\frac{1}{k^2}\\
&+ \frac{1}{2V}\sum_{\substack{0<k<k_c\\ 0<k'<k_c}}\hat{k}\cdot\hat{k}'P_kP_{k'}\int d\boldsymbol{r}\exp[i(\boldsymbol{k}+\boldsymbol{k}')\cdot\boldsymbol{r}] \qquad (2.26)
\end{aligned}$$

It is maybe not obvious that we have gained much by this detour. With a few more manipulations, however, we will easily be able to read off the physics! First, recall that $\boldsymbol{p}_i$ and $\boldsymbol{r}_j$ are canonically conjugate variables, i.e. $[r_{i\alpha}, p_{j,\beta}] = i\hbar\delta_{ij}\delta_{\alpha\beta}$, with $\alpha,\beta = x,y,z$. This implies that $\boldsymbol{p}_i\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)+\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)\boldsymbol{p}_i = 2\boldsymbol{p}_i\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)-\hbar k\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i)$. Introducing the ("plasma frequency") parameter $\omega_p^2 \equiv ne^2/m\epsilon_0$ with $n \equiv N/V$, and using that $\int\exp[i(\boldsymbol{k}+\boldsymbol{k}')\cdot\boldsymbol{r}]d\boldsymbol{r} = V\delta(\boldsymbol{k}+\boldsymbol{k}')$, it is then straightforward to mold Eq. (2.27) on the form

$$H_{el} \approx \sum_i\frac{\boldsymbol{p}_i^2}{2m_e}+\frac{e^2}{2V\epsilon_0}\sum_{i,j}\sum_{k\geq k_c}\frac{\exp[i\boldsymbol{k}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j]}{k^2}-\frac{ne^2}{2\epsilon_0}\sum_{k\neq 0}\frac{1}{k^2}$$

[13] A quick recipe for showing that Q_k and P_k are canonically conjugate: Write a Hamiltonian $H \sim \int d^3\boldsymbol{r}\boldsymbol{E}^2$, insert the Fourier series for $\boldsymbol{E}$ from Eq. (2.21), differentiate with respect to P_k, and use Hamilton's equation $\partial H/\partial P_k = \partial Q_k/\partial t$.

$$\begin{aligned}
&+ \quad \frac{1}{2}\sum_{0<k<k_c}(P_k^* P_k + \omega_p^2 Q_k^* Q_k - \frac{ne^2}{\epsilon_0 k^2}) \\
&+ \quad \frac{e}{\sqrt{V\epsilon_0 m_e}}\sum_{k<k_c}\hat{k}Q_k \cdot \sum_i(\boldsymbol{p}_i - \frac{\hbar \boldsymbol{k}}{2})\exp(i\boldsymbol{k}\cdot\boldsymbol{r}_i) \\
&+ \quad \frac{e^2}{2V\epsilon_0 m_e}\sum_{\substack{0<k,k'<k_c \\ \boldsymbol{k}\neq-\boldsymbol{k}'}} Q_k Q_{k'}\hat{k}\cdot\hat{k}'\sum_i \exp[i(\boldsymbol{k}+\boldsymbol{k}')\cdot\boldsymbol{r}_i]. \qquad (2.27)
\end{aligned}$$

We have here used that $Q_{-k} = -Q_k^*$ and $P_k = -P_{-k}^*$, as follows by choosing a real-valued vector potential $\boldsymbol{A}$. The first three terms in (2.27) are written solely in "electron variables" $(\boldsymbol{p}_i, \boldsymbol{r}_i)$, and hence describe electrons interacting via a short-range (screened) two-body potential. The fourth term describes collective modes, with variables (Q_k, P_k). The fifth term is an interaction between the screened electrons and the collective modes, while the sixth term, finally, corresponds to interactions among the collective modes themselves. The latter are called (or *plasmons* ["quantized plasma oscillations"]). Their origin can be visualized by imagining that we insert a test electron into the electron liquid. This electron pushes on the surrounding electrons, which overshoot and relax, and as the effect propagates through the liquid, a collective charge oscillation (of frequency ω_p) is created.

To continue the analysis, one writes H_{el} in (2.27) on second-quantized form, performs a canonical transformation, and then analyzes the resulting Hamiltonian perturbatively. We shall not carry out this program, which – although straightforward – is calculationally somewhat cumbersome. The outcome of the analysis is that the electron-plasmon interaction in the fifth term of (2.27) can be largely eliminated by a renormalization of the electron mass, $m \to m^*$. Similarly, most of the plasmon- plasmon interaction (sixth term in (2.27)) is absorbed by a renormalization of the plasmon frequency ω_p. The picture that emerges is that the strongly interacting particles in the electron liquid can be described as a collection of weakly interacting *quasiparticles* (electrons with renormalized masses and interacting only over short distances via a screened potential) and *plasmons* (quantized collective charge oscillations). As it turns out, the picture can be made sharper and its origin shown to be deeper and more ubiquitous than what our perturbative sketch may suggest. To see how, we turn to Landau Fermi liquid theory.

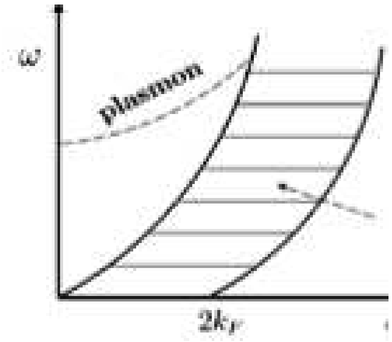

Figure 2.11: Plasmon mode and particle-hole excitation continuum (hatched area) of an electron liquid ($d = 3$).

2.2.2 Fermi liquid theory

In the previous section we saw how the electron liquid may be described as a system of weakly interacting elementary excitations. These came in two brands: *Quasiparticles* (screened electrons with renormalized masses) and *collective excitations* (plasmons with renormalized frequencies). This structure of excitations is not unique to the electron liquid, but applies to all known *normal Fermi liquids*. Roughly speaking, a "normal Fermi liquid" is a quantum liquid of fermions with no broken symmetries (that is, the groundstate of the liquid respects all the symmetries of the Hamiltonian), and with the fermions interacting via a repulsive potential that does not cause singular scattering (i.e. all scattering amplitudes remain finite).[14] At temperatures below the degeneracy temperature T_F the available phase space for scattering of the fermions becomes highly restricted, making possible a description of the quantum liquid in terms of a dilute collection of weakly interacting elementary excitations. This was the great insight of Lev Landau in 1956. Landau developed the theory for explaining the properties of liquid ^{3}He (the simplest known Fermi liquid), but soon thereafter people realized that the theory is more general, and can be applied also, for example, to the electron liquid.

Landau's original approach was phenomenological. He started with some general assumptions, considering only macrosocopic phenomena where momenta and frequencies of the experimental probes (particles, fields,...) are much smaller than the Fermi momentum and energy of the system under study. Landau then showed that the response to these probes can be obtained by considering a collection of effectively noninteracting fermionic quasiparticles and collective excitations (oscillatory modes of the Fermi surface) with properties encoded by a set of parameters (effective mass, effective frequencies, and *Landau parameters)* that can be determined from a few experiments. Given this, the theory then gains predictive power. The results of Landau's theory (and also some of his assumptions!) were subsequently derived "microscopically" (that is, starting from a microscopic Hamiltonian) using perturbative quantum field theory (Pitaevskii 1959, Luttinger and Nozieres 1962,...).

In what follows we shall basically follow Landau. As his formulation has a certain "airy" quality, it is easy to get misled and believe that the theory is maybe not all that useful. This is wrong! Landau's theory is deep and subtle, it has made unexpected predictions, and, most importantly, it serves as a conceptual pinnacle for thinking about a variety of condensed matter problems. As the reader will realize when starting applying it (even at the mundane level of homework problems), there are quite a few surprises in store! A modern and elegant perspective on the theory from the vantage point of the renormalization group has been given by Shankar. Quoting Shankar: "For many readers of Landau's work there was an element of mystery surrounding some of the manipulations. This had to be so, since Landau substituted forty years of subsequent developments with his remarkable intuition." This comment will be appreciated as we continue.

Basics

As a starter, we recall that the ground state of a system of non-interacting fermions is characterized by the fact that all single particle energy levels below the Fermi energy are occupied and all states above the Fermi level are empty. Hence, the occupation probability in ground

[14] This is a textbook statement that is somewhat oversimplified. There are other mechanisms besides broken symmetries and singular scattering that may invalidate the use of Fermi liquid theory. Some of them are fairly well understood today, like the presence of certain types of fermion-impurity interactions. Others remain to be clarified. We shall return to this issue in section 2.2.3.

state is $n_0(\mathbf{p}) = \Theta(\mu - \epsilon(\mathbf{p}))$. All excited states correspond to different occupation probabilities $n(\mathbf{p}) = n_0(\mathbf{p}) + \delta n(\mathbf{p})$, where the deviation from the ground state is given to an eigenstate of the Hamiltonian, then the distribution function changes with time, and after some decay time τ the deviation from ground state differs significantly from the initial one $\delta n(\mathbf{p}, t = 0)$.

The simplest candidate for an excitation is a single particle excitation with either $\delta n(\mathbf{p}) = +1$ for some $\mathbf{p}$ such that $\epsilon(\mathbf{p}) > \mu$ or $\delta n(\mathbf{p}) = -1$ for some $\mathbf{p}$ such that $\epsilon(\mathbf{p}) < \mu$. This type of excitation is an exact eigenstate of the Hamiltonian if the system is non-interacting and, following Landau, we propose that it may even be a good approximate eigenstate of an interacting Hamiltonian. A heuristic argument proceeds via a "Gedanken Experiment": Starting with the non-interacting system, let us imagine that we "turn on" the interaction very slowly, so slow that the system has time to adjust itself, with the original noninteracting eigenstates smoothly turning into new states of the interacting system. With a 1-to-1 correspondence between non-interacting fermion states and states of the interacting system, we can use the *same* quantum numbers to label the new states and also the *same* Fermi distribution. This latter feature is supported by a rigorous theorem proven by Luttinger (1961): "The Fermi surface of an interacting system encloses the same volume as a corresponding non-interacting system with the same particle density." In the case of an interacting system the ground state may differ from that of a non-interacting system and the dispersion relation $\epsilon(\mathbf{p})$ may be modified by interactions, so it is not quite appropriate to call these excitations simply particles and holes, instead, the terms *quasiparticle* and *quasihole* are used. A quasiparticle state can be pictured "perturbatively" as in Fig. (2.12) where a fermion added to some single-particle state $|\boldsymbol{k}\rangle$ causes other fermions to be excited out of the Fermi sphere (because of the interaction between particles). If the resulting state $-$ which is a superposition of many different single-particle states $-$ can be labelled by the momentum $\boldsymbol{k}$ of the "original" single-particle state (as ascertained by Landau!), we may treat it as being effectively occupied by some particle, and this is what we call the *quasiparticle*.

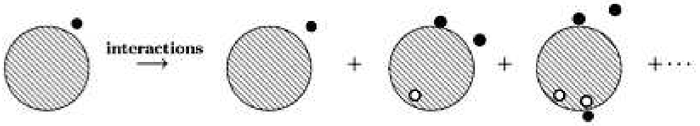

Figure 2.12: Illustration of a perturbative expansion of the change of a single-particle state of an added fermion due to interactions with fermions in the Fermi sphere.

This type of excitation is a good approximation to the eigenstate of the Hamiltonian provided that the state does not decay too fast (in case the Gedanken Experiment above would not make any sense: the new states would decay before we have had time to switch on the full interaction!). The decay is due to interactions between quasiparticles — hence, we can determine the plausibility of Landau's quasiparticle picture by calculating the scattering rate between elementary excitations.

Let us calculate the decay rate for a quasiparticle with momentum $\mathbf{p}_1$ in a d-dimensional system. The state of the system can change if a quasiparticle with momentum $\mathbf{p}_1$ scatters of another quasiparticle with momentum $\mathbf{p}_2$ and after the scattering event the two quasiparticles move with momenta $\mathbf{p}_3$ and $\mathbf{p}_4$. If the number of excitations is small, the state $\mathbf{p}_2$ is only occupied if $p_2 < p_F$, and the states $\mathbf{p}_3$ and $\mathbf{p}_4$ are available to be scattered into if $p_3 > p_F$ and

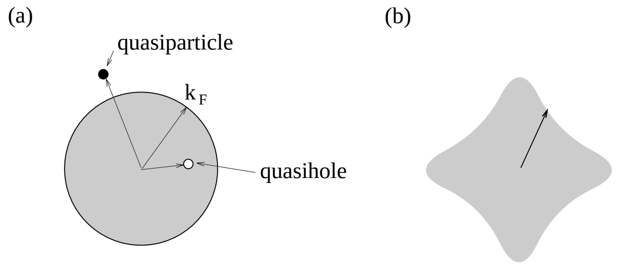

Figure 2.13: (a) Fermi surface, a quasiparticle, and a quasihole. (b) Collective excitation.

$p_4 > p_F$ as shown in Figure 2.14. Hence, taking into both momentum and energy conservation as well as an interaction strength which depends on the momentum exchange we obtain

$$\begin{aligned}\Gamma(\mathbf{p}_1) \propto \; & \int (dp_2) \int (dp_3) \int (dp_4) \Theta(p_F - p_2)\Theta(p_3 - p_F)\Theta(p_4 - p_F) \\ & |V(|\mathbf{p}_1 - \mathbf{p}_3|)|^2 \delta(\mathbf{p}_1 + \mathbf{p}_2 - \mathbf{p}_3 - \mathbf{p}_4)\delta(\epsilon(p_1) + \epsilon(p_2) - \epsilon(p_3) - \epsilon(p_4))\end{aligned}$$

where $(dp_i) = \frac{d^d p_i}{(2\pi)^d}$. Assuming a parabolic dispersion $\epsilon(p) = \frac{p^2}{2m}$ and using the two conservation laws we can write $\Theta(p_4 - p_F) = \Theta(p_1^2 + p_2^2 - p_3^2 - p_F^2)$ and simplify the argument of the δ-function so that we get

$$\begin{aligned}\Gamma(\mathbf{p}_1) \propto \; & \int_{\sqrt{2p_F^2 - p_1^2}}^{p_F} dp_2 \, p_2^{d-1} \int_{p_F}^{\sqrt{p_1^2 + p_2^2 - p_F^2}} dp_3 \, p_3^{d-1} \\ & \int d\Omega_2 \int d\Omega_3 |V(|\mathbf{p}_1 - \mathbf{p}_3|)|^2 \delta(2\mathbf{p}_1 \cdot \mathbf{p}_2 - 2\mathbf{p}_3 \cdot \mathbf{p}_4)\end{aligned}$$

where Ω_i denotes the direction angles of $\mathbf{p}_i$. Hence, if $p_1 \gtrsim p_F$, both p_2 and p_3 lie in the vicinity of the Fermi surface, and we can take $(p_2 p_3)^{d-1} \approx p_F^{2(d-1)}$ outside as a constant factor to obtain

$$\begin{aligned}\Gamma(\mathbf{p}_1) \propto \; & \int_{\sqrt{2p_F^2 - p_1^2}}^{p_F} dp_2 \int_{p_F}^{\sqrt{p_1^2 + p_2^2 - p_F^2}} dp_3 \\ & \int d\Omega_2 \int d\Omega_3 |V(2p_F \sin(\theta_{13}))|^2 \delta(2\mathbf{p}_1 \cdot \mathbf{p}_2 - 2\mathbf{p}_3 \cdot \mathbf{p}_4).\end{aligned}$$

Since the lengths of all the four momenta are nearly equal, the δ-function forces the angle between $\mathbf{p}_3$ and $\mathbf{p}_4$ to be nearly the same as the angle between $\mathbf{p}_1$ and $\mathbf{p}_2$. In general, this angular correlation is uninteresting, and we get

$$\Gamma(\mathbf{p}_1) \propto (p_1 - p_F)^2 \propto |\epsilon(p_1) - \mu|^2.$$

Hence, the scattering phase space is greatly reduced near the Fermi surface, and quasiparticles are indeed good approximations for low-energy excited states. At a positive temperature the distribution functions are smoothened and we obtain $\Gamma(\mathbf{p}_1) \propto \max[(k_B T)^2, |\epsilon(p_1) - \mu|^2]$. This means that near the Fermi level quasiparticles behave essentially as free particles, and a perturbative treatment of interactions may well succeed. Whether or not perturbation theory actually does yield reliable results must be investigated more carefully using renormalization

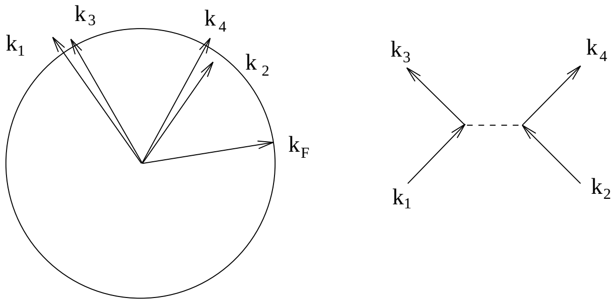

Figure 2.14: Scattering of quasiparticles near the Fermi level.

group (RG) techniques, but at least we can be initially somewhat optimistic. However, after a second thought one realizes that there is an obvious loophole in the preceding analysis: If $\boldsymbol{p}_1 = -\boldsymbol{p}_2$ then $\boldsymbol{p}_3 = -\boldsymbol{p}_4$ can take *any* value, and we are no longer assured that the quasiparticles essentially behave as free particles near the Fermi surface. Do the quasiparticles take advantage of this possibility so as to invalidate the Fermi liquid picture? An RG analysis shows that they don't as long as the interaction is repulsive. However, for a weak attractive interaction they do, and this leads to a superfluid state (superconductivity, in the case of electrons).

The above analysis also breaks down in one dimension. In one dimension, it is not possible for the angle $\angle(\mathbf{p}_3, \mathbf{p}_4)$ to be "nearly the same as" as the angle $\angle(\mathbf{p}_1, \mathbf{p}_2)$: all angles are either 0 or π. In general, for a nonlinear dispersion relation $\epsilon(p)$ the only possible scattering events in one dimension are $(\mathbf{p}_1, \mathbf{p}_2) \to (\mathbf{p}_1, \mathbf{p}_2)$ or $(\mathbf{p}_1, \mathbf{p}_2) \to (\mathbf{p}_2, \mathbf{p}_1)$. Therefore, we cannot argue that the interaction strength is effectively reduced near Fermi level, and the validity of a perturbative approach in one dimension cannot be ascertained with phase space arguments. We shall look closer at the one-dimensional case in Sec. 2.2.4.

Ordinary Fermi liquids have also other types of excitations in addition to the quasiparticles discussed above. These collective excitations correspond to different deformations $\delta n(\mathbf{p})$, and can be visualized as breathing modes of $n_0(\mathbf{p})$. The simplest of these modes, the isotropic breathing $\delta n(\mathbf{p}) = f(p)$, corresponds to changing the particle density, and can therefore only occur if it has a spatial dependence $\delta n(\mathbf{r}, \mathbf{p}) = f(\mathbf{r}, p)$ such that the total number of particles in the system remains constant (*i.e.* $\int\int d^d r d^d p\, \delta n(\mathbf{r}, \mathbf{p}) = 0$). In charged systems (like in the electron liquid) this density oscillation corresponds to a plasmon (c.f. the previous section), and has a dispersion relation $\omega(q) \sim q^{(3-d)/2}$ as can be seen by dimensional analysis ($\omega^2 \sim (1/m)(e^2/4\pi\epsilon)\rho q^\alpha$). For a neutral Fermi liquid such as liquid ^{3}He the isotropic breathing mode is known as zero sound and has a linear dispersion relation $\omega = v_0 q$. Note that in one dimension even the plasmon dispersion relation is linear, $\omega(q) = v_p q$. These types of collective excitations exist even in one-dimensional conductors as we will see in the next sections.

Applications

In this section we shall look at a few applications, mostly so as to get aquainted with the spirit of Landau's theory. To keep things simple we assume that we are dealing with liquid ^{3}He: an isotropic translationally invariant neutral Fermi liquid wih short-range repulsive (”hard-core”) interactions between the particles. At the end of the section we will discuss

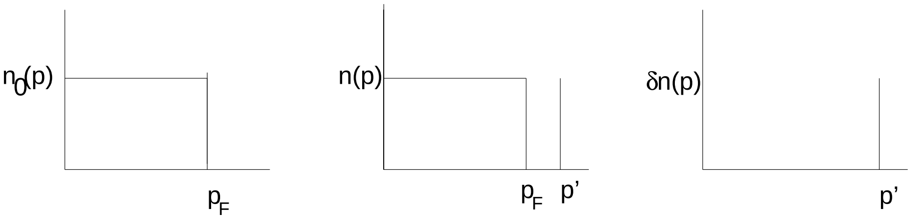

Figure 2.15: Distribution functions for the groundstate ($n_0(\boldsymbol{p})$), an excited state with an added quasiparticle at $\boldsymbol{p} = \boldsymbol{p}'$ ($n(\boldsymbol{p})$), and the corresponding single quasi-particle excitation $\delta n(\boldsymbol{p}) = \delta(\boldsymbol{p} - \boldsymbol{p}')$.

what complications arise when applying the theory to conduction electrons in a metal or a semiconductor.

Let us start by considering the simplest type of elementary excitation in liquid ^{3}He; a single-particle excitation obtained by adding an extra atom to the liquid. This process is depicted in Fig. 2.15, with $n_0(\boldsymbol{p})\,[n(\boldsymbol{p})]$ the Fermi-Dirac distribution in the groundstate [excited state].

Note that this innocent-looking figure rests on the strong assumption that the states of the *interacting* fermion system are distributed in the same way as those of a non-interacting Fermi gas (for which, by definition, a Fermi-Dirac distribution applies). Introducing the notation $E\,[E_0]$ for the energy of the groundstate [excited state], and $F = E - \mu N\,[F_0 = E_0 - \mu N_0]$ the corresponding free energies at zero temperature (where $\mu \equiv \epsilon_F = \partial E_0/\partial N$ is the chemical potential of the groundstate), Landau proposed the following expansion of $F - F_0$:

$$F - F_0 = \sum_{\boldsymbol{p}} (\epsilon(\boldsymbol{p}) - \mu)\delta n(\boldsymbol{p}) + \frac{1}{2}\sum_{\boldsymbol{p}\boldsymbol{p}'} f(\boldsymbol{p},\boldsymbol{p}')\delta n(\boldsymbol{p})\delta n(\boldsymbol{p}') + \mathcal{O}(\delta n^3(\boldsymbol{p})). \tag{2.28}$$

The first term in (2.28) describes propagating free quasiparticles with kinetic energy $\epsilon(\boldsymbol{p}) \equiv \delta E/\delta n(\boldsymbol{p})$, where $\delta E = E - E_0$ and $\delta n(\boldsymbol{p}) = n(\boldsymbol{p}) - n_0(\boldsymbol{p})$. The second term in (2.28) encodes the leading contribution to the interaction between quasiparticles, with interaction energy $f(\boldsymbol{p},\boldsymbol{p}') = \delta^2 E/\delta n(\boldsymbol{p})\delta(\boldsymbol{p}')$. It is important to realize that on the order of the free energy F both $\epsilon_p - \mu$ and δn_p are infinitesimal, $\sim \delta$, and the first and second term are of the same order of magnitude (assuming that $f(\boldsymbol{p},\boldsymbol{p}') \sim \mathcal{O}(1)$). Next, it is useful to define a *renormalized single quasi-particle energy* $\bar{\epsilon}(\boldsymbol{p})$ that includes the lowest-order effect of the interaction,

$$\bar{\epsilon}(\boldsymbol{p}) = \epsilon(\boldsymbol{p}) + \frac{1}{2}\sum_{\boldsymbol{p}'} f(\boldsymbol{p},\boldsymbol{p}')\delta n(\boldsymbol{p}'), \tag{2.29}$$

and a *quasiparticle effective mass* m^*,

$$\frac{1}{m^*} \equiv \frac{v_F}{p_F} = \frac{|\nabla_{\boldsymbol{p}}\epsilon(\boldsymbol{p})|}{p_F}|_{p=p_F} \tag{2.30}$$

with $\boldsymbol{v}_F = \nabla_{\boldsymbol{p}}\epsilon(\boldsymbol{p}_F)$ the group velocity at the Fermi surface. To include spin in the formalism one endows the interaction energy with spin indices: $f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}')$, where for liquid ^{3}He, $\sigma, \sigma' = \pm 1/2$ are the projections of the nuclear spins on some chosen quantization axis. In the absence of a magnetic field the theory is invariant under time-reversal invariance which implies that

$$f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}') = f_{-\sigma,-\sigma'}(-\boldsymbol{p},-\boldsymbol{p}'). \tag{2.31}$$

Reflection invariance of the spherical Fermi surface in turn implies that

$$f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}') = f_{\sigma,\sigma'}(-\boldsymbol{p},-\boldsymbol{p}') = f_{-\sigma,-\sigma'}(\boldsymbol{p},\boldsymbol{p}'), \tag{2.32}$$

and it follows that $f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}')$ can only depend on the relative spin orientation, $\sigma\sigma' = \pm 1/4$. We can thus write

$$f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}') = f(\boldsymbol{p},\boldsymbol{p}') + 4\sigma\sigma'\varphi(\boldsymbol{p},\boldsymbol{p}'). \tag{2.33}$$

At the low energies at which we are working $|\boldsymbol{p}| \approx |\boldsymbol{p}'| \approx p_F$, and the pair of momenta $(\boldsymbol{p},\boldsymbol{p}')$ can be specified by their relative polar angle ϑ (taking advantage of the spherical symmetry of the Fermi surface which makes the value of the azimuthal angle immaterial). It follows that $f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}')$ in (2.33) can be expressed as

$$\begin{aligned} f_{\sigma\sigma'}(\boldsymbol{p},\boldsymbol{p}') &= f(\vartheta) + 4\sigma\sigma'\varphi(\vartheta) \\ &= \sum_{L=0}^{\infty}(f_L + 4\sigma\sigma'\varphi_L)P_L(\cos\vartheta) \end{aligned} \tag{2.34}$$

where in the second line we have expanded the functions $f(\vartheta)$ and $\varphi(\vartheta)$ in Legendre polynomials $P_L(\cos\vartheta)$. In simple applications, as for liquid ^{3}He, $f_0 > f_1 \gg f_2 \gg ...$ and $\varphi_0 > \varphi_1 \gg \varphi_2 \gg ...$, and one may truncate the theory to contain only a small number of parameters. This yields a "good" phenomenological theory that is well prescribed by fitting these *Landau parameters* to known experimental data. Given this, the theory can then be used to make predictions of new experiments.

To illustrate the theory, let us sketch how to calculate the specific heat and the compressibility of a Fermi liquid. These are both observables that measure the response of the system to a weak perturbation at equilibrium, and can be obtained by studying properties of the quasiparticles only. Other observables, like the intriguing "zero sound" of liquid ^{3}He, or spin density waves, are instead determined by the collective modes, which, roughly speaking, correspond to oscillations of the Fermi surface. However, we shall not make an endeavor on this here.

Specific heat ($T \to 0$ limit)

The specific heat per unit volume is defined by $c_V \equiv T(\partial s/\partial T)_V$ where

$$s = \frac{-k_B}{V}\sum_{\boldsymbol{p}\sigma}[n_\sigma(\boldsymbol{p})\ln(n_\sigma(\boldsymbol{p})) + (1 - n_\sigma(\boldsymbol{p})\ln(1 - n_\sigma(\boldsymbol{p}))] \tag{2.35}$$

is the entropy per unit volume. Note that the expression in (2.35) is identical to that for an ideal Fermi gas: the entropy only depends on the combinatorics of states, and since – by assumption – the states in the Fermi liquid are in one-to-one correspondence with those in the Fermi gas, the combinatorics is the same. Inserting

$$n_\sigma(\boldsymbol{p}) = \frac{1}{1 + \exp[(\bar{\epsilon}(\boldsymbol{p}) - \mu)/k_B T]} \tag{2.36}$$

into (2.35), with $\bar{\epsilon}(\boldsymbol{p})$ the renormalized quasiparticle energy from Eq. (2.29), and replacing the sum over $n(\boldsymbol{p})$ by an integral, one obtains

$$s = \frac{\pi^2}{3}g(\mu)k_B^2 T + ... \tag{2.37}$$

Here "..." indicates higher-order terms which are irrelevant in the $T \to 0$ limit, and $g(\mu)$ is the density of states at the Fermi surface,

$$g(\mu) = \sum_{\boldsymbol{p}} \delta(\epsilon(\boldsymbol{p}) - \mu) = \frac{m^* p_F}{\pi^2 \hbar^3} \tag{2.38}$$

(which is the same as for a Fermi gas, but with $m \to m^*$). Differentiating (2.37) with respect to T one obtains

$$c_V \approx \frac{m^* p_F}{3\hbar} k_B^2 T. \tag{2.39}$$

The linear temperature dependence of the specific heat is one of the hallmarks of a Fermi liquid. Note in particular that the quasiparticle effective mass can be directly obtained by measuring the specific heat of the Fermi liquid. A simple physical explanation of the linearity is that only those fermions that are within $k_B T$ of the Fermi energy (hence, a fraction of $k_B T/\epsilon_F$ of all fermions) are affected by temperature, and their energies change roughly by $k_B T$ due to thermal excitations, so that the energy change is proportional to T^2 and the specific heat to the derivative of this, *i.e.* to T. The fermionicity of this argument lies of course in the first part.

Compressibility

As a further illustration, let us calculate the compressibility κ of a Fermi liquid, related to the speed of sound c by

$$c^2 = \frac{1}{\kappa m \rho}, \tag{2.40}$$

where ρ is the particle density. To set the stage, we first recall the definition of a pressure. At zero temperature,

$$P \equiv -\frac{\partial E_0}{\partial V} = -f(\rho) + \rho \partial f(\rho) \partial \rho, \tag{2.41}$$

with $f(\rho) = E_0/V$ the energy density. The inverse compressibility $1/\kappa$ is then given by

$$\frac{1}{\kappa} \equiv -V \frac{\partial P}{\partial V} = \rho^2 \frac{\partial^2 f}{\partial \rho^2}, \tag{2.42}$$

where the second identity follows from Eq. (2.41). The chemical potential $\mu = \partial E_0/\partial N$ can be written as

$$\mu = \frac{\partial f}{\partial \rho}, \tag{2.43}$$

and combined with Eq. (2.42) this yields that

$$\frac{1}{\kappa} = N\rho (\frac{\partial \mu}{\partial N})_V. \tag{2.44}$$

Thus, to obtain the compressibility we must find out how the number of particles N changes as we change the chemical potential μ. Let us first ask a different question: "How does the renormalized quasiparticle energy change if we change the chemical potential by an amount $d\mu$?" *On* the Fermi surface its change is clearly equal to $d\mu$, but it is still instructive to track down its two distinct contributions: One comes from the change of momenta of the quasiparticles,

$$d\epsilon_{(1)} = \nabla_{\boldsymbol{p}} \epsilon(\boldsymbol{p}) \cdot d\boldsymbol{p}_F = v_p dp_F, \tag{2.45}$$

the other from a change $\delta n(\boldsymbol{p})$ in the number of other quasiparticles:

$$d\epsilon_{(2)} = \sum_{\boldsymbol{p}'} f(\boldsymbol{p}, \boldsymbol{p}')\delta n(\boldsymbol{p}'). \tag{2.46}$$

Adding Eqs. (2.45) and (2.46), dividing by $d\mu = d\epsilon_{(1)} + d\epsilon_{(2)}$, and using that

$$\delta n(\boldsymbol{p}) = -\frac{\partial n_0(\boldsymbol{p})}{\partial \epsilon(\boldsymbol{p})}\frac{\partial \epsilon(\boldsymbol{p})}{\partial p}dp_F = \delta(\epsilon(\boldsymbol{p}) - \mu)v_p dp_F, \tag{2.47}$$

we obtain an equation for $dp_F/d\mu$ telling us how the Fermi surface changes as the chemical potential changes:

$$v_p\frac{dp_F}{d\mu} + \sum_{\boldsymbol{p}'} f(\boldsymbol{p}, \boldsymbol{p}')\delta(\epsilon_{\boldsymbol{p}'} - \mu)v_{p'}\frac{dp'_F}{d\mu} = 1. \tag{2.48}$$

How does this help us to obtain $dN/d\mu$? Well, from (2.47) we learn that

$$dN = \sum_{\boldsymbol{p}} \delta n(\boldsymbol{p}) = \sum_{\boldsymbol{p}} \delta(\epsilon(\boldsymbol{p}) - \mu)v_p dp_F, \tag{2.49}$$

implying that

$$\frac{dN}{d\mu} = \sum_{\boldsymbol{p}} \delta(\epsilon(\boldsymbol{p}) - \mu)[v_p\frac{dp_F}{d\mu}]. \tag{2.50}$$

The expression within the square bracket in (2.50) is most easily obtained from (2.48) by replacing the sum by an integral, $\sum_{\boldsymbol{p}} ... \to 2\pi\int p^2 \sin\vartheta dp d\vartheta$. Inserting the spin index on $f(\boldsymbol{p}, \boldsymbol{p}')$ and summing over it, and using the expansion in Eq. (2.34) for $f_{\sigma\sigma'}(\boldsymbol{p}, \boldsymbol{p}')$, we find that

$$v_p\frac{dp_F}{d\mu} = \frac{1}{1 + F_0}, \tag{2.51}$$

where

$$\begin{aligned} F_0 &= \frac{\hbar^3}{4}g(\mu)\sum_{\sigma'}\sum_{L=0}^{\infty}\int_0^{\pi}(f_L + 4\sigma\sigma'\varphi_L)P_L(\cos\vartheta)\sin\vartheta d\vartheta \\ &= \frac{\hbar^3}{4}g(\mu)f_0. \end{aligned} \tag{2.52}$$

Combining Eqs. (2.40), (2.42), and (2.50) - (2.52), we can finally write a closed expression for the sound velocity c:

$$c^2 = \frac{N}{mg(\mu)}(1 + F_0). \tag{2.53}$$

We have gone to some length in deriving the result in Eq. (2.53), for two reasons. First, it well illustrates how the interaction among quasiparticles gets encoded by a *Landau parameter*, F_0, on top of the mass renormalization $m \to m^*$. Also, the result in Eq. (2.53) is instructive as it reveals a potential "instability" of the Fermi liquid: A value of $F_0 < -1$ results in an unphysical imaginary sound velocity, implying that $F_0 \to -1$ signals a transition to a new phase of matter that cannot be described as a Fermi liquid. Whereas this particular instability never happens, other do as we shall see below.

The approach used to derive the sound velocity in Eq. (2.53) can easily be adapted to obtain other thermodynamic (equilibrium) properties, such as the magnetic susceptibility χ of a Fermi liquid,

$$\chi = g(\mu)\frac{\mu_B^2}{1+Z_0}, \tag{2.54}$$

where $Z_0 = g(\mu)\varphi_0$. As $Z_0 \to -1$, the susceptibility diverges, signaling a transition to a ferromagnetic state. This instability also does not happen for liquid ^{3}He, but it *may* occur in the electron liquid – or to be more precise – for electrons in metals. More generally one can show that whenever

$$F_L \leq -(2L+1) \; and/or \; Z_L \leq -(2L+1), \tag{2.55}$$

Landau Fermi liquid theory breaks down. The only experimentally identified case of such a *Landau-Pomeranchuk instability* is the transition to a ferromagnetic state, signalled by a divergence of χ in Eq. (2.54), with $Z_0 \to -1$.

At this point, let us briefly comment upon the description of the *non-equilibrium* properties of a Fermi liquid. Restricting ourselves to the case where the system is close to equilibrium, the dynamics gets governed by a Boltzmann equation for the *local* quasiparticle distribution $n_{\boldsymbol{p}}(\boldsymbol{r},t)$. One finds that in addition to excited single quasiparticles, also collective modes – corresponding to oscillations of the Fermi surface – may now be excited. A case in point are the *plasmons* of the electron liquid (cf. Sec 2.2.1). Another famous example is the *zero sound mode* of liquid ^{3}He. At small frequencies ω and quasiparticle collision times τ, $\omega\tau \ll 1$, sound in a Fermi liquid behaves as ordinary hydrodynamic ("first") sound, with the velocity parameterized as in Eq. (2.53). As the temperature is lowered, however, we know that the average quasiparticle scattering time increases as T^{-2}, and this eventually leads to a situation where $\omega\tau \gg 1$. The quasiparticles then no longer have time to relax during one period of the sound wave, and one would have guessed that the excess quasiparticles in a region of space caused by a compression would simply diffuse away, and no sound would propagate. However, Landau noted that if there are quasiparticle interactions, a local change of the density may still drive a change in the neighboring density, thus setting up a sound-like longitudinal collective mode of the liquid, known as *zero sound*.[15]

What about electrons in metals? Like ^{3}He atoms in the liquid phase these also form a Fermi liquid, but there are some complications. First, the electrons are charged and experience a long-ranged Coulomb interaction. This can be taken care of in exact analogy with our perturbative approach in the previous section where we split the quasiparticle interaction in a long-range Coulombic part plus a short-range part. The fact that the electrons are charged, however, implies that spin and spatial degrees of freedom get coupled via a spin-orbit interaction $\sim g\mu_B\boldsymbol{\sigma}\cdot(\boldsymbol{v}\times\boldsymbol{E})$, with $\boldsymbol{E}$ the electric field from surrounding electrons (and ions), and where $g\mu_B\boldsymbol{\sigma}$ is the effective electron magnetic moment. As a consequence, the spin- and spatial rotational symmetries are no longer independent, in contrast to ^{3}He. For most metals the spin-orbit interaction is weak and can be included as a perturbation, with the Fermi liquid serving as an unperturbed "reference system". However, for a strong spin-orbit interaction the applicability of Fermi liquid theory becomes doubtful.

The lattice of ions in which the electrons move not only contributes to the spin-orbit interaction, but also breaks the translational and rotational space symmetries by introducing a *band structure*. Moreover, the lattice produces additional interactions with phonons, and also

[15] The first measurements of zero sound in pure liquid ^{3}He were carried out by Abel, Anderson, and Wheatley in 1966, showing excellent agreement with Landau Fermi liquid theory

with whatever defects and impurities that are embedded in the lattice. Starting with the band structure, this can easily be taken care of when the Fermi surface remains almost spherical, as is the case for *e.g.* the alkali metals. Roughly speaking, the role of the Landau quasiparticle is now taken by a ”quasi-Bloch electron”, with an effective mass that encodes both the presence of the lattice and the electron-electron interaction. The case of non-spherical Fermi surfaces is more tricky, in particular when the bands are narrow, with a multiband structure. This typically leads to a strong renormalization of the effective mass and/or the Landau parameters, and in addition may require that other types of parameters are introduced. The details of the problem then become intractable within a Landau Fermi liquid description, although the theory may still be used to conceptualize the basic physics. The strong renormalization can be understood qualitatively as coming from a change of the density of states induced by the bands tructure, leading to an enhanced quasiparticle interaction. In some cases this may cause an instability at a critical temperature or pressure, resulting in a transition to a new phase (such as an *itinerant ferromagnet* or a *Mott insulator*). As to the effect of interactions with phonons, these may be absorbed by a renormalization of the effective mass if the electron frequencies are smaller than the phonon Debye frequency. (In the opposite limit the electron ”shakes off” the cloud of phonons and no mass renormalization is required.) However, at low temperatures the phonons may mediate an *attractive* interaction between quasiparticles, leading to a BCS (or Cooper) instability where the metal turns into a superconductor. Again, this phase cannot be described by Fermi liquid theory. Considering finally the interactions with impurities and defects, their effect have to be carefully analyzed on a case-to-case basis. Dilute concentrations of local potential scatterers are usually harmless for the applicability of Fermi liquid theory. In contrast, a dense array of magnetic moments that interact dynamically with the electrons (a *Kondo lattice)* may lead to a breakdown of Landau's theory. Let us here stress that all of the above applies to conduction electrons moving in a three- (or two-dimensional) lattice. As we shall see in Sec. 2.2.4, in one dimension Fermi liquid theory breaks down for *any* fermionic system, be it conduction electrons in a semiconducting quantum wire, electrons in a metallic carbon nanotube, or a cold fermionic gas trapped in a one-dimensional optical lattice.

Home problem 3: Compressibility of a Fermi liquid
Try to fill out the gaps and holes in the somewhat sketchy derivation of the compressibility of a Fermi liquid as given above. In particular, give a careful argument for Eq. (2.47), and tighten the link to the final result in Eq. (2.53).

2.2.3 Non-Fermi liquids: Broken symmetries, quantum criticality, disorder, and all that...

See A. J. Schoefield, *Non-Fermi Liquids*, Contemporary Physics **40**, 95 (1999).
http://fy.chalmers.se/ tfkhj/QuantumMatter/Schoefield.pdf

2.2.4 The Importance of Dimensionality: Luttinger liquid

In Section 2.2.2 we sketched an argument suggesting that Fermi liquid theory breaks down in one spatial (1D) dimension. It is instructive to look a bit closer at what goes wrong. For that purpose, let us consider a gas of free fermions under the influence of a weak potential $\phi(\boldsymbol{r})$. The rearrangement of the fermion density due to the potential can be described by an induced fermion density $\rho(\boldsymbol{r})$. Going over to Fourier transforms we have the relation

$$\rho(\boldsymbol{k}) = \chi(\boldsymbol{k})\phi(\boldsymbol{k}), \tag{2.56}$$

where $\chi(\boldsymbol{k})$ is the free-particle response function. The Danish physicist Jens Lindhard derived an expression for this function, which is therefore called the *Lindhard function*:

$$\chi(\boldsymbol{k}) = \int \frac{d\boldsymbol{k}'}{(2\pi)^d} \frac{f(\boldsymbol{k}') - f(\boldsymbol{k}' + \boldsymbol{k})}{\epsilon(\boldsymbol{k}') - \epsilon(\boldsymbol{k}' + \boldsymbol{k})}. \tag{2.57}$$

Here $f(\boldsymbol{k})$ is the Fermi distribution, $\epsilon(\boldsymbol{k})$ is the energy of a fermion, and d denotes the dimension. We shall not attempt to rederive this function but only note that the integrand in Eq. (2.57) tells us that (at $T = 0$) the dominant response comes from fermions weakly excited out of the Fermi sea, giving rise to low-energy particle-hole pairs, in agreement with intuition. The response function can be calculated and plotted in one, two, and three dimensions. One finds that the 1D curve diverges at $k = 2k_F$, while in 2D (3D) the curves have a cusp (a logarithmic divergent derivative) at this wave vector. This has implications for the response of a free fermion gas to a potential. For example, if we consider the potential coming from ions on a lattice, phonons will couple to the electrons leading to a lattice instability in 1D *(the Peierls instability)*, while in 2D and 3D the electron-phonon coupling results in a *Kohn anomaly* in the phonon dispersion curve.

The potential $\phi(\boldsymbol{k})$ does not necessarily originate from an applied field or the ions on a lattice. Much more important for our discussion is that it can be effectively induced by the other fermions in the gas *if these are interacting*. Let us assume that we move one of the fermions a small distance. If the system consists of non-interacting fermions nothing will happen, but if we consider a system with *interacting* fermions this displacement will change the potential felt by the other fermions and they will adjust according to Eq. (2.56). These adjustments will in turn lead to a new induced potential, leading to new rearrangements. In higher dimensions these adjustments will be damped since the response functions are finite and the system will return to its initial state. This is equivalent to the statement that quasiparticles have finite lifetime. In 1D, in contrast, the original disturbance will not be damped but turn into something new that cannot be described by Fermi liquid theory.

The hypothesis of the one-to-one correspondence between eigenstates in a degenerate Fermi gas and the quasiparticle states of a Fermi liquid is represented by the requirement of a nonsingular "adiabatic switching on" of the interaction. The instability in the 1D Fermi gas described above will thus ruin Landau's fundamental hypothesis: Turning on interactions among fermions in 1D will *not* create quasiparticles, at least not of the Fermi liquid brand. The origin of the singular response function is the topology of the 1D Fermi surface, as we shall see now. As we have already noted, the main contribution to the integral in Eq. (2.57) comes from particle- hole excitations where the particle and the hole have almost the same energy, causing the denominator of the integrand to almost vanish. Given that the Fermi surface in 1D consists of two distinct points $\pm\boldsymbol{k}_F$ one immediately realizes that such particle-hole pairs are connected by a momentum difference $\Delta\boldsymbol{k} = 2\boldsymbol{k}_F$. In two and three dimensions the measure keeps the integral finite while this is not the case in 1D; thus a singularity appears.

The non-existence of Landau quasiparticles in 1D is thus entirely due to the structure of the particle-hole pair spectrum, dictated by the topology of the 1D Fermi surface. The particle-hole spectrum is easily constructed from that of the free Fermi gas single-particle spectrum and is depicted in Fig. (2.16). The characteristic feature of the particle-hole spectrum prohibiting the creation of quasiparticles is the empty region between $\boldsymbol{k} = 0$ and $\boldsymbol{k} = \pm 2\boldsymbol{k}_F$. As we have discussed, a quasiparticle can be visualized as an electron with a cloud of low-energy excitations around it, and, since these low-energy particle-hole excitations are very rare in 1D (occurring only on lines through $\boldsymbol{k} = 0, \pm 2\boldsymbol{k}$), quasiparticles are not easily formed. This is to be contrasted to the situation in three dimensions, with a broad band of low-energy particle-hole excitations (see Fig. (2.11)).

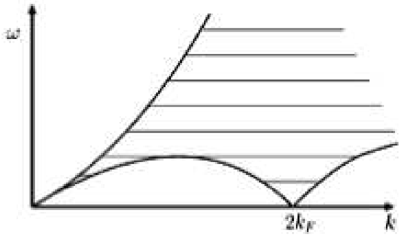

Figure 2.16: Particle-hole pair spectrum in 1D.

A final word before we go on. Interacting fermions, of course, exist also in one dimension, but they cannot form quasiparticles, and these interacting fermions must therefore be described using another theory than Landau Fermi liquid theory. The basis of such a theory, replacing that of a Fermi liquid, is the *Luttinger liquid* where, as we shall see, interactions are included from the very beginning. To set up this theory we shall use the second quantized formalism that is reviewed in the Appendix C.

Let us start by considering a non-interacting system of fermions on a 1D lattice with lattice constant a. The dispersion relation is given by $\epsilon(k) = -2t\cos(ka)$ where t is the hopping matrix element between adjacent lattice sites. This dispersion relation is rather cumbersome to deal with, but as long as we are interested in low energy properties only, we can linearize $\epsilon(k)$ near the Fermi energy E_F as shown in Figure 2.17. Since there are two Fermi points $\pm k_F$, we get two linear branches so that one has $\frac{\partial \epsilon}{\partial k} > 0$ corresponding to right-moving particles and the other has $\frac{\partial \epsilon}{\partial k} < 0$ corresponding to left-moving particles. Although the linearization is a good approximation of the dispersion relation only in the vicinity of the Fermi points, we use the linear dispersion relations for *all* k — as long as we are only interested in the low-energy excitations this extension of the approximation has no physical consequences, but mathematically it turns out to be very convenient. The resulting Hamiltonian with two linear branches is given by

$$H_0 = v_F \sum_{k,s} \Big[(k - k_F) c^\dagger_{+,k,s} c_{+,k,s} + (-k - k_F) c^\dagger_{-,k,s} c_{-,k,s} \Big] \tag{2.58}$$

where the operator $c_{r,k,s}$ annihilates a left- or right-moving fermion (corresponding to $r = \pm$) with momentum k and spin s. The Hamiltonian H_0 is known as the *Luttinger model* for non-interacting fermions; later we will see how we can include interactions in the model as well.

The analysis of H_0 is greatly simplified by introducing the so-called partial densities $\rho_{\pm,s}(q)$ for left-movers and right-movers,

$$\rho_{+,s}(q) \;=\; \sum_k c^\dagger_{+,k+q,s} c_{+,k,s}$$

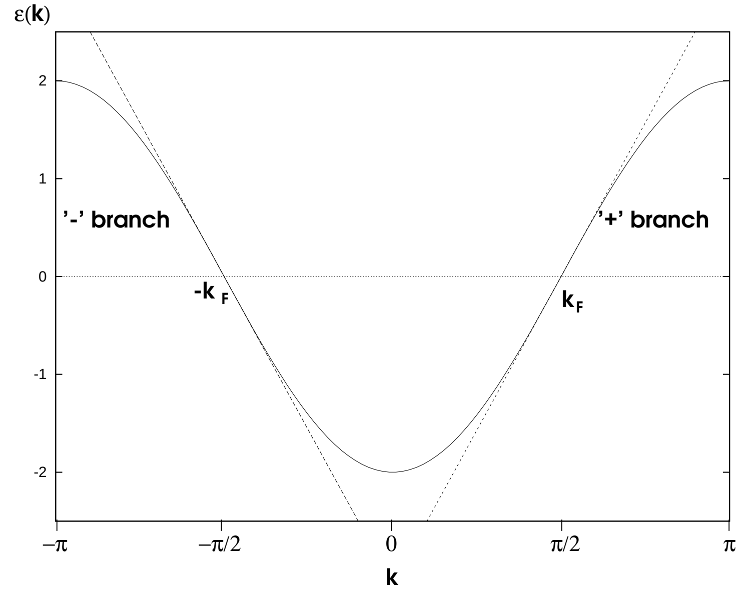

Figure 2.17: Dispersion relation for non-interacting fermions on a lattice, and the linearized approximation for $\epsilon(k) \approx E_F$. In the plot $E_F = 0$ corresponding to a half-filled band.

$$\rho_{-,s}(q) \quad = \quad \sum_k c^\dagger_{-,k+q,s} c_{-,k,s}.$$

In the following we will need the commutation relations between $\rho_{r,s}(q)$, and it turns out that most of them vanish; the only potentially complicated ones are

$$\begin{aligned}
&[\rho_{r,s}(-q), \rho_{r,s}(q)] \\
&= \sum_{k,k'} (c^\dagger_{r,k-q,s} c_{r,k,s} c^\dagger_{r,k'+q,s} c_{r,k',s} - c^\dagger_{r,k'+q,s} c_{r,k',s} c^\dagger_{r,k-q,s} c_{r,k,s}) \\
&= \sum_{k,k'} (\; -c^\dagger_{r,k-q,s} c^\dagger_{r,k'+q,s} c_{r,k,s} c_{r,k',s} + c^\dagger_{r,k-q,s} c_{r,k',s} \delta_{k,k'+q} \\
&\qquad + c^\dagger_{r,k'+q,s} c^\dagger_{r,k-q,s} c_{r,k',s} c_{r,k,s} - c^\dagger_{r,k'+q,s} c_{r,k,s}) \delta_{k-q,k'}) \\
&= \sum_k (c^\dagger_{r,k-q,s} c_{r,k-q,s} - c^\dagger_{r,k,s} c_{r,k,s}) \\
&= \sum_k [n_{r,s}(k-q) - n_{r,s}(k)]
\end{aligned}$$

where we used the fermionic anticommutation relations $\{c^\dagger_{r,k,s}, c_{r',k',s'}\} = \delta_{r,r'} \delta_{k,k'} \delta_{s,s'}$ If the sum over k extended only over a finite range, the last expression would yield zero. Now, however, we have extended the branches '+' and '−' for all k, and we have to be more careful. Let us therefore assume that all states with $rk < k_0$ are occupied (*i.e.* $n_{+,s}(k) = 1$ for $k < k_0$ and $n_{-,s}(k) = 1$ for $k > -k_0$), and carefully cancel out the (infinite) contributions from states that are assumed to be fully occupied. Then the last expression is

$$\begin{aligned}
&\sum_k (n_{r,s}(k-q) - n_{r,s}(k)) \\
= &\left(\sum_{r(k-q)>k_0} n_{r,s}(k-q) + \sum_{rk<k_0+rq} 1\right) - \left(\sum_{rk>k_0} n_{r,s}(k) + \sum_{rk<k_0} 1\right) \\
= &\left(\sum_{rk>k_0} n_{r,s}(k) - \sum_{rk>k_0} n_{r,s}(k)\right) + \left(\sum_{rk<k_0+rq} 1 - \sum_{rk<k_0} 1\right) \\
= & rq\frac{L}{2\pi}.
\end{aligned}$$

Note that this result is independent of k_0, and we can safely take $k_0 \to -\infty$ which amounts to assuming that states which are infinitely far below the Fermi level are fully occupied.[16] In conclusion, we have

$$[\rho_{r,s}(-q), \rho_{r',s'}(q')] = \delta_{rr'}\delta_{ss'}\delta_{qq'} r\frac{qL}{2\pi} \tag{2.59}$$

Comparing with the usual commutation relations for bosons, $[b, b^\dagger] = 1$, we see that[17] for $q > 0$ the partial densities $\rho_{+,s}(-q)$ and $\rho_{-,s}(q)$ correspond to boson destruction operators and $\rho_{+,s}(q)$ and $\rho_{-,s}(-q)$ to boson creation operators. This is not all that surprising since $\rho_{r,s}(q)$ are constructed from products of two fermion operators.

The usefulness of the partial densities becomes apparent once we evaluate the commutators between them and the Hamiltonian:

$$\begin{aligned}
&[H_0, \rho_{r,s}(q)] \\
= & v_F \sum_{k,k'} (rk - k_F)[c^\dagger_{r,k,s}c_{r,k,s}c^\dagger_{r,k'+q,s}c_{r,k',s} - c^\dagger_{r,k'+q,s}c_{r,k',s}c^\dagger_{r,k,s}c_{r,k,s}] \\
= & v_F \sum_{k,k'} (rk - k_F)[\ -c^\dagger_{r,k,s}c^\dagger_{r,k'+q,s}c_{r,k,s}c_{r,k',s} + c^\dagger_{r,k,s}c_{r,k',s}\delta_{k,k'+q} \\
& \qquad\qquad +c^\dagger_{r,k'+q,s}c^\dagger_{r,k,s}c_{r,k',s}c_{r,k,s} - c^\dagger_{r,k'+q,s}c_{r,k,s}\delta_{k,k'}] \\
= & v_F \sum_k (rk - k_F)[c^\dagger_{r,(k-q)+q,s}c_{r,k-q,s} - c^\dagger_{r,k+q,s}c_{r,k,s}] \\
= & v_F rq \sum_k c^\dagger_{r,k+q,s}c_{r,k,s} \\
= & v_F rq \rho_{r,s}(q)
\end{aligned}$$

Hence, the commutator is proportional to the partial density operator — note that this only works because the spectrum is linear, which allowed us to change variables on the third line to cancel the factors $(rk - k_F)$.

Another way to see why linear dispersion is crucial is to note that the operator $c^\dagger_{r,k+q,s}c_{r,k,s}$ can be thought of as creating a particle-hole pair with momentum q. Since the dispersion law is linear, the energy of this pair is independent of k, and particle-hole excitations with different k (but same q) get mixed by energy-conserving scattering processes. The partial densities $\rho_{r,s}(q)$ represent such linear combinations.

[16] We are tacitly assuming that the operators $n(k)$ only act on states which satisfy this condition; if the system is subjected to strong enough perturbation so that deep states are not fully occupied, the commutation relations must be modified. An example of this is the fractional quantum Hall effect, where the occupation probability of single particle states far below the Fermi level is a fraction p/q (with q odd) rather than one.

[17] Apart from a scale factor.

Similar commutation relations can be obtained between $\rho_{r,s}(q)$ and the Hamiltonian $\widetilde{H}_0 = v_F \frac{2\pi}{L} \sum_{s,k>0} [\rho_{+,s}(k)\rho_{+,s}(-k) + \rho_{-,s}(-k)\rho_{-,s}(k)]$ which is quadratic in the operators $\rho_{r,s}(q)$,

$$\begin{aligned} & [\widetilde{H}_0, \rho_{r,s}(q)] \\ = \; & v_F \frac{2\pi}{L} \sum_{k>0} [\rho_{r,s}(k)\rho_{r,s}(-k), \rho_{r,s}(q)] \\ = \; & v_F \frac{2\pi}{L} r\rho_{r,s}(-k) \frac{qL}{2\pi} \delta_{-k,q}\Theta(k) + v_F r \rho_{r,s}(k) \frac{qL}{2\pi} \delta_{k,q}\Theta(k) \\ = \; & v_F r q \rho_{r,s}(q). \end{aligned}$$

This suggests that H_0 and $\widetilde{H}_0$ are in some sense equivalent: most physical quantities can be expressed in terms of commutators between different operators, and since H_0 and $\widetilde{H}_0$ give the same commutators, they can be expected to yield similar physical results.

Let us at this point not worry too much about to what extent H_0 and $\widetilde{H}_0$ are indeed equivalent, but rather examine why the description in terms of $\widetilde{H}_0$ might be useful. This will be clear if we consider what happens when we introduce an interaction term to the Hamiltonian. In the language of electron operators $c_{r,k,s}$ an electron-electron interaction term involves a product of four operators, and the Hamiltonian becomes very difficult to analyze. In terms of the density operators $\rho_{r,s}(q)$, however, the electron-electron interaction can be expressed as a product of only two operators, and the Hamiltonian remains quadratic. The simplest type of an interaction does not scatter right-moving and left-moving electrons into each other. Since the directions of motion of all electrons are conserved, this interaction is known as forward scattering. It is described by[18]

$$\begin{aligned} H_{FW} = \frac{1}{L} \sum_{s,s',q>0} & \Big\{ 2g_2(q)\rho_{+,s}(q)\rho_{-,s'}(-q) + \\ & g_4(q)[\rho_{+,s}(q)\rho_{+,s'}(-q) + \rho_{-,s'}(-q)\rho_{-,s}(q)] \Big\} \end{aligned}$$

In most physical cases the interactions are equally strong between the two branches as they are within a branch meaning that $g_2(q) = g_4(q)$ but for generality it is useful to keep the two processes formally separate. The precise form of the interaction is not important for the following discussion, and we set both $g_2(q)$ and $g_4(q)$ to q-independent constants g_2 and g_4. These can be viewed as the long wavelength limits of the original interaction constants. Since the pure Coulomb interaction is long range and has a diverging Fourier transform for small q, we implicitly assume that there are some mobile charges in the vicinity of the one dimensional system that act to screen out the long range part of the interaction and result in finite values for g_2 and g_4.

Before we diagonalize the Hamiltonian $\widetilde{H}_0 + H_{FW}$, it is instructive to consider what happens if only one spin state is present. The corresponding Hamiltonian, the so-called spinless Luttinger model, reads

$$\begin{aligned} H' = \frac{1}{L} \sum_{q>0} & \{(2\pi v_F + g_4)[\rho_+(q)\rho_+(-q) + \rho_-(-q)\rho_-(q)] \\ & +2g_2\rho_+(q)\rho_-(-q)]\} \end{aligned} \tag{2.60}$$

[18] The names g_2 and g_4 for the two coupling constants are historical, and originate from the so-called g-ology — see *e.g.* Solyom.

This Hamiltonian can be diagonalized with a Bogolubov transformation $\rho_j(q) = \alpha_j \rho_+(q) + \alpha'_j \rho_-(q)$ where $j \in \{1, 2\}$. Requiring that ρ_j obey similar commutation relations as ρ_+ and ρ_- for $j = 1$ and $j = 2$, respectively, we get $\alpha_1 = \cosh(\varphi_1)$, $\alpha'_1 = \sinh(\varphi_1)$, $\alpha_2 = \sinh(\varphi_2)$, and $\alpha'_2 = \cosh(\varphi_2)$. Requiring furthermore that ρ_1 and ρ_2 commute gives $\varphi_1 = \varphi_2 = \varphi$. The transformation angle φ is determined by requiring that the Hamiltonian can be written as

$$H' = \frac{2\pi}{L} \sum_{q>0} [v_1 \rho_1(q)\rho_1(-q) + v_2 \rho_2(-q)\rho_2(q)],$$

which yields

$$\begin{aligned} v_1 \cosh^2(\varphi) + v_2 \sinh^2(\varphi) &= v_F + g_4/(2\pi) \\ v_1 \sinh^2(\varphi) + v_2 \cosh^2(\varphi) &= v_F + g_4/(2\pi) \\ (v_1 + v_2)\cosh(\varphi)\sinh(\varphi) &= g_2/(2\pi) \end{aligned}$$

or $\tanh(2\varphi) = g_2/(2\pi v_F + g_4)$ and $v_2 = v_1 = v = [g_2/(2\pi)]/\sinh(2\varphi)$. Using functional relations between hyperbolic functions this can be simplified to

$$v = \sqrt{[v_F + g_4/(2\pi)]^2 - [g_2/(2\pi)]^2}. \tag{2.61}$$

The diagonalized Hamiltonian is now given by

$$H' = \frac{2\pi}{L} v \sum_{q>0} [\rho_1(q)\rho_1(-q) + \rho_2(-q)\rho_2(q)]. \tag{2.62}$$

It is customary to define the interaction parameter $g = e^{-2\varphi} = \sqrt{\frac{2\pi v_F + g_4 - g_2}{2\pi v_F + g_4 + g_2}}$ so that $g < 1$ corresponds to the physical case of repulsion between charge carriers ($g_2 > 0$), $g = 1$ is the limiting case of a non-interacting system, and $g > 1$ corresponds to an attractive interaction.

In the case of fermions with spin it is useful to first transform the partial densities into charge (c) and spin (σ) degrees of freedom by defining ($r = \pm$)

$$\rho_{rc}(q) = \frac{1}{\sqrt{2}} (\rho_{r\uparrow}(q) + \rho_{r\downarrow}(q)) \tag{2.63}$$

$$\rho_{r\sigma}(q) = \frac{1}{\sqrt{2}} (\rho_{r\uparrow}(q) - \rho_{r\downarrow}(q)) \tag{2.64}$$

so that we have $\widetilde{H}_0 = v_F \frac{2\pi}{L} \sum_{k>0, \alpha=c,\sigma} [\rho_{+\alpha}(k)\rho_{+\alpha}(-k) + \rho_{-\alpha}(-k)\rho_{-\alpha}(k)]$ and

$$H_{FW} = \frac{2}{L} \sum_{k>0} [4g_2 \rho_{+c}(-k)\rho_{-c}(k) + 2g_4 (\rho_{+c}(k)\rho_{+c}(-k) + \rho_{-c}(-k)\rho_{-c}(k))].$$

Hence, the interaction only affects the charge branches $\rho_{\pm c}$, and the Hamiltonian can be diagonalized as before to obtain

$$H = \frac{2\pi}{L} \sum_{q>0,\alpha} v_\alpha [\rho_{1\alpha}(q)\rho_{1\alpha}(-q) + \rho_{2\alpha}(-q)\rho_{2\alpha}(q)] \tag{2.65}$$

where $v_c = \sqrt{[v_F + g_4/\pi]^2 - [g_2/\pi]^2}$ is the charge velocity and $v_s = v_F$ is the spin velocity. The charge operators ρ_{1c} and ρ_{2c} are related to $\rho_{\pm c}$ through a similar Bogolubov transformation as in the spinless case, and the spin operators are given by $\rho_{1s} = \rho_{+s}$ and $\rho_{2s} = \rho_{-s}$.

We have now exactly diagonalized the bosonic Hamiltonian H, and we must next address the question to what extend it is equivalent with the original fermion Hamiltonian. The operators $\rho_{\pm c}(q)$ and $\rho_{\pm\sigma}(q)$ create charge and spin waves that travel either to the right or to the left. As is often the case with collective excitations, there is no excitation of this type at zero wave vector — this can be seen for instance from the commutation relations (2.59). An excitation with zero wave vector would correspond a uniform change in either the charge (or spin) density, or in the (electric or spin) current density; while such excitation are clearly possible in the fermionic system, they are not present in the bosonized Hamiltonian, and must be added by hand.[19] These so-called zero modes or topological excitations contribute an extra term in the Hamiltonian that is given by

$$H_{\text{zero modes}} = \frac{\pi}{2L}\left(\frac{v_c}{g_c}N_c^2 + g_c v_c J_c^2 + \frac{v_\sigma}{g_\sigma}N_\sigma^2 + g_\sigma J_\sigma^2\right) \tag{2.66}$$

where $g_c = e^{-2\varphi_c}$ and $g_\sigma = 1$, N_c is the particle number, J_c the particle current, N_σ the total spin, and J_σ the spin current.[20] Note that the zero mode contribution is inversely proportional to the system size L, and we can therefore associate for instance the term $\frac{\pi}{2L}\frac{v_c}{g_c}N_c^2$ with the usual charging term $Q^2/(2C)$.[21] It can be shown quite rigorously[22] that the bosonized Hamiltonian with the zero mode terms corresponds exactly to the original fermion Hamiltonian, and the two can be used interchangeably depending on which form is more convenient. Usually, of course, the bosonic form is preferred since it is only second order in the boson operators whereas the fermionic Hamiltonian is quadratic in the fermion operators.

The bosonic Hamiltonian (2.65) shows that the spin and charge waves travel with different velocities (except in the non-interacting limit), which is known as spin-charge separation. This means that the excitations in a one-dimensional metal cannot be mapped continuously to the quasiparticles of a non-interacting fermion system — quasiparticles carry both charge and spin so that the charge and spin degrees of freedom are tied together. Consequently, one dimensional metals are qualitatively different from their higher dimensional counterparts, and do not belong to the category of Fermi liquids. In one dimension the Luttinger model possesses a similar status as the non-interacting Fermi gas in higher dimensions: most one-dimensional systems can be described starting from the Luttinger model, perhaps with some renormalized parameters. Therefore, one-dimensional metals are often called *Luttinger liquids* — they may not be exactly described by the Luttinger model, but they are qualitatively similar in the same way as higher dimensional metals are not exactly described by non-interacting Fermi gas but are qualitatively similar to it. However, sometimes strange things happen even in one dimension, and the behavior of a system may deviate qualitatively from a Luttinger model. Usually this type of behavior results from the opening of a gap at the Fermi level (*cf.* superconductivity in 3D), which may arise, *e.g.*, due to a scattering process that transfers right-moving fermions into left-moving fermions and *vice versa*. This backscattering process, when translated into the bosonic language, results in a complicated additional term in the boson Hamiltonian that renders the problem no longer exactly diagonalizable. Fortunately, it can be shown using renormalization group arguments that such backscattering terms are

[19] They can, however, also be implicitly included by meticulously defining the $q \to 0$ limit for the bosonic operators.

[20] By 'current' we mean the difference in the numbers of left moving and right moving charges or spins.

[21] The analogy is not perfect since the term $\frac{\pi}{2L}\frac{v_c}{g_c}N_c^2$ is non-zero even for non-interacting particles ($g_c = 1$) and hence contains a kinetic energy contribution as well.

[22] F.D.M. Haldane, J. Phys. C **14**, 2585 (1981).

only important (relevant) if the Fermi level is exactly at $E_F = 0$ corresponding to a half-filled band.

Bosonization

Although we have now obtained the bosonized form of the Luttinger Hamiltonian, we are not quite done. Most physical quantities can be related to the expectation values of products of *electron* operators and are readily expressible in the fermion language, but it is not clear how one can write them in term of the bosonic operators. Hence, we need a fermion-boson dictionary that allows us to translate operators in one language to operators in another language.

What is an electron? Fundamentally, it is a particle that carries charge 1 and spin 1 (in units of $-e$ and $\frac{\hbar}{2}$, respectively), and obeys fermionic anticommutation relations. Hence, if we can (uniquely) construct an operator (in the bosonic picture) that satisfies these conditions, we have found a bosonic expression for a fermion operator.

Let us start by laying out the strategy, and worry about mathematical details only later. Consider charge 1 particles in general (spin is treated similarly). An operator $\tilde{\psi}^\dagger(x)$ creates a charge 1 at position x if, as a result of an application of $\tilde{\psi}^\dagger(x)$, the charge density $\rho_c(x')$ increases by $\delta(x'-x)$, that is, if we have $[\rho_c(x'), \tilde{\psi}^\dagger(x)] = \delta(x-x')\tilde{\psi}^\dagger(x)$. This looks like a doable task since we have a bosonic representation for the density operators, and we may be able construct a bosonic version of $\tilde{\psi}^\dagger(x)$. The operator $\tilde{\psi}^\dagger(x)$ is not quite what we need since it does not satisfy fermionic commutation relations. To do this final step we employ the Jordan-Wigner transformation and write $\psi(x) = e^{i\pi N_L(x)}\tilde{\psi}(x)$ where $N_L(x)$ is the total number of charges to the left of the position x. Also $N_L(x)$ is likely to be expressible in terms of boson operators since it is just $\int_{-L/2}^{x} dx'\rho(x')$. To show that the Jordan-Wigner transformation does the trick and results in fermionic anticommutation relations, let us evaluate the anticommutator $\{\psi(x), \psi(x')\}$ in a state that contains N_1 electrons to the left of x, N_2 electrons to the left of x', and assume $x' > x$. Neglecting the commuting parts we have $\{\psi(x), \psi(x')\} \propto (-1)^{N_1}(-1)^{N_2} + (-1)^{N_2-1}(-1)^{N_1} = 0$ since acting with $\psi(x)$ reduces the number of electrons to the left of x' by one while acting with $\psi(x')$ has no effect on the number of electrons to the left of x. Formally, we need to be quite careful with the implementation of this idea since we usually deal with infinite systems so the both N_1 and N_2 are infinite, and therefore we may easily get results like $N_2 - 1 = N_2$.

Now that we have a plan we can proceed with the mathematical construction of $\psi^\dagger(x)$. First we need a density operator $\rho_\Sigma(q) = \rho_+(q) + \rho_-(q)$ (we ignore spin for the moment). It is useful to introduce even the other linear combination $\rho_\Delta(q) = \rho_+(q) - \rho_-(q)$ which is related to particle current. Using the identity $[e^A, B] = e^A[A, B]$ (valid if $[A, [A, B]] = 0$) we see that the commutation relation $[\rho_\Sigma(x'), \tilde{\psi}^\dagger(x)] = \delta(x-x')\tilde{\psi}^\dagger(x)$ is satisfied by the operator $\tilde{\psi}^\dagger(x) = e^{A(x)}$ if $[\rho_\Sigma(x'), A(x)] = \delta(x-x')$. Fourier transforming we obtain $[\rho_\Sigma(-q'), A(q)] = L\delta_{q,q'}$ which is solved by $A(q) = \frac{\pi}{q}\rho_\Delta(q)$ or $A(x) = \sum_q e^{-iqx}\frac{\pi}{Lq}\rho_\Delta(q)$. For the Jordan-Wigner transformation we need $N_L(x) = \int_{-L/2}^{x} dx'\rho_\Sigma(x')$ and we arrive at our first guess for the electron operator,

$$\psi^\dagger(x) \overset{?}{=} e^{\sum_q \left[i\pi \int_{-L/2}^{x} dx' \left(\frac{1}{L}e^{-iqx'}\right)\rho_\Sigma(q) + e^{-iqx}\frac{\pi}{Lq}\rho_\Delta(q)\right]}$$

Two questions remain: is this a possible electron creation operator, and if so, is this construction unique? Unfortunately, the answer to both these questions is no. Since we started by dividing the electron spectrum into right-moving and left-moving branches, we must have *two* electron operators, one for each branch. Furthermore, the bosonic operators $\rho_\pm(q)$ do not exist for $q = 0$, and therefore our present form for $\psi^\dagger(x)$ cannot create a net charge, it can only redistribute existing charges in the system. Hence, we must introduce four more operators $U_\pm^\dagger$ and $U_\pm$ which add and remove, uniformly, charge in the two branches thereby changing the zero-mode quantum numbers $N_\pm$ by one. These additional ladder operators for different branches anticommute, $\{U_+, U_-^\dagger\} = 0$ and $\{U_+, U_-\} = 0$, while they commute for the same branch $U_+U_+^\dagger = U_+^\dagger U_+$. There is even another ambiguity in our first guess: we could have chosen the Jordan-Wigner factor to be $e^{-i\pi N_L(x)}$ just as well as $e^{i\pi N_L(x)}$, and we could even have chosen the reference point to be anything and not just $-L/2$. Finally, the units of $\psi^\dagger(x)$ are not correct, and the sum over wave vectors does not converge (large q behavior $\sim q^{-1}$), and we must fix both these problems. In the end we arrive at a new guess ($r = \pm$)

$$\psi_r^\dagger(x) = \frac{1}{\sqrt{2\pi\alpha}} e^{-irk_F x - r\sum_q e^{-\frac{1}{2}\alpha|q|}\left[-\frac{2\pi}{Lq}e^{iqx}\rho_r(q)\right] - irN_r\frac{2\pi x}{L}} U_r^\dagger \tag{2.67}$$

where α is a convergence factor that will be set to zero ($\alpha \to 0^+$) at the end of the calculations. The proof that the definition (2.67) indeed yields the proper fermionic commutation relations is given in Appendix D.

Power laws

Having obtained the bosonic form of the electron operator, we are ready to determine some physical properties of the system. To demonstrate one of the differences between a Luttinger liquid and Fermi liquid we calculate the occupation probability $n_+(k) = \langle\psi_+^\dagger(k)\psi_+(k)\rangle$ at zero temperature. We carry out the calculation for a spinless Luttinger model but the result is the same for model with spin since the interaction does not mix spin states. For free fermions (in any dimension) we have $n_+(k) = \Theta(k_F - k) \sim |k - k_F|^0$ where the second form is for future convenience. For a Luttinger model it is easiest to calculate $n_+(k)$ through its Fourier transform as $n_+(k) = L^{-1}\int_{-L/2}^{L/2} dx \int_{-L/2}^{L/2} dx' e^{ik(x-x')}\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0}$. Inserting the bosonized forms for the fermion operators we get

$$\begin{aligned}\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0} = \;& \frac{1}{2\pi\alpha}\left\langle e^{-i(k_F+\frac{2\pi N_+}{L})(x-x')}\right\rangle_{T=0} e^{-\frac{1}{2}\sum_q \frac{2\pi}{Lq}e^{-\alpha|q|+iq(x-x')}} \\ & \left\langle e^{\sum_q \frac{2\pi}{Lq}e^{-\frac{1}{2}\alpha|q|}[e^{iqx}\rho_+(q)+e^{-iqx'}\rho_+(-q)]}\right\rangle_{T=0}\end{aligned}$$

where the first average only depends on the zero modes and the second average only on the bosonic forms. We obtained this form by using $e^A e^B = e^{A+B}e^{\frac{1}{2}[A,B]}$ to combine the two exponentials into one, which resulted in the commutator term that appears as the second exponential factor. The zero mode average is easy at $T = 0$ when only $N_+ = 0$ contributes, but the average over the bosonic modes is more complicated since ρ_+ does not appear in the bosonized Hamiltonian (2.62). It is convenient to write $\rho_+(q)$ in terms of the eigenmodes as $\rho_+(q) = \cosh(\varphi)\rho_1(q) - \sinh(\varphi)\rho_2(q)$ and separate the resulting exponential as $e^{C+D} = e^C e^D e^{-\frac{1}{2}[C,D]}$ where C only contains creation operators and D only annihilation operators. Since the zero temperature average involves only the ground state and since annihilation operators acting on a ground state yield zero, the operator factors $e^C e^D$ collapse into unity, and only the factor containing the commutator remains. This gives, after some algebra,

$$\begin{aligned}\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0} = e^{-ik_F(x-x')}\frac{1}{2\pi\alpha}e^{-\frac{1}{2}\sum_{q>0}\frac{2\pi}{Lq}[e^{-q(\alpha-i(x-x'))}-e^{-q(\alpha+i(x-x'))}]}\\ e^{-\cosh(2\varphi)\frac{1}{2}\sum_{q>0}\frac{2\pi}{Lq}[2e^{-\alpha q}-e^{-q(\alpha-i(x-x'))}-e^{-q(\alpha+i(x-x'))}]}.\end{aligned}$$

The sums we can do using the Taylor expansion of logarithm $\sum_{n=1}^\infty \frac{1}{n}z^n = -\ln(1-z)$, and we find

$$\begin{aligned}\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0} = \; & e^{-ik_F(x-x')}\frac{1-e^{-\frac{2\pi\alpha}{L}}}{2\pi\alpha}\frac{1}{1-e^{-\frac{2\pi}{L}(\alpha+i(x-x'))}}\\ & \left[\frac{1-e^{-\frac{2\pi\alpha}{L}}}{\left(1-e^{-\frac{2\pi}{L}(\alpha+i(x-x'))}\right)\left(1-e^{-\frac{2\pi}{L}(\alpha-i(x-x'))}\right)}\right]^{2\sinh^2(\varphi)}.\end{aligned}$$

Now we let $L \to \infty$ and get the simple expression

$$\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0} = \frac{1}{2\pi i}e^{-ik_F(x-x')}\frac{1}{x-x'-i\alpha}\left[\frac{\alpha^2}{(x-x')^2+\alpha^2}\right]^{\sinh^2(\varphi)}$$

so that the occupation probability is given by

$$n_+(k) = \frac{1}{2\pi i}\int_{-\infty}^{\infty}dx\frac{e^{i(k-k_F)x}}{x-i\alpha}\left[\frac{\alpha^2}{x^2+\alpha^2}\right]^{\sinh^2(\varphi)}.$$

By changing variables to $\zeta = x/\alpha$ we find that the integral can be written in terms of the function $F_a(q) = \int_0^\infty d\zeta\,\frac{\cos(q\zeta)}{(1+\zeta^2)^{a+1}}$ (with $a = \sinh^2(\varphi)$); this integral satisfies a differential equation that allows it to be connected to the modified Bessel functions, and in the end we get

$$n_+(k) \sim |k-k_F|^{2\sinh^2(\varphi)} \tag{2.68}$$

which shows that for $\varphi \neq 0$ there is no discontinuity of occupation at k_F, showing that a Luttinger model is fundamentally different from a Fermi liquid for all $g \neq 1$. Recalling the connection between g and φ we see that the occupation probability near Fermi surface obeys a power law with exponent $\nu - 1 = \frac{1}{2}\left(g+\frac{1}{g}\right) - 1 \geq 0$. This result is typical for Luttinger liquids: most physical dependences take the form of power laws with exponents depending on the interaction parameter g.

Home problem 4: Density of states of a Luttinger liquid

The "tunneling density of states" $D(\omega)$, which measures how easy or difficult it is for an extra electron to enter the system by a local tunneling process, can be obtained as a Fourier transform of $F(t) = \langle 0|e^{iHt}\Psi^\dagger(0)e^{-iHt}\Psi(0)|0\rangle$ where H is the Hamiltonian, $\Psi(x) = \psi_+(x) + \psi_-(x)$ is the physical electron operator taking into account that electrons can travel either to the left or to the right, and $|0\rangle$ the ground state. For simplicity, set $T = 0$, ignore spin, and neglect the zero-mode contribution to the Hamiltonian. Note that the cross terms involving products like $\psi_-^\dagger(x')\psi_+(x)$ vanish, and note further that if you know the result for the $+$-branch, the result for the $-$-branch can be obtained be replacing $v \to -v$. Hence, it is sufficient to only consider the $+$-branch.

1. Following the same procedure as in the calculation of $\langle\psi_+^\dagger(x)\psi_+(x')\rangle_{T=0}$, obtain $F(t)$

Hints: (a) show that $\mathcal{U}^\dagger \exp(A)\mathcal{U} = \exp(\mathcal{U}^\dagger A\mathcal{U})$
where $\mathcal{U}$ is an arbitrary unitary operator
(expand $\exp(A)$ as a power series).
Hence, since e^{-iHt} is unitary, it is sufficient to know
$e^{iHt}\rho_\pm e^{-iHt}$ to obtain $e^{iHt}\Psi^\dagger(0)e^{-iHt}$

(b) write $\rho_\pm$ in terms of the diagonalizing fields $\rho_{1,2}$, and
obtain $\rho_{1,2}(q,t) \equiv e^{iHt}\rho_{1,2}(q)e^{-iHt}$

(c) use the same trick as in the calculation of $n_+(k)$
to combine all factors to the form a single
exponential e^{A+B}

(d) separate the exponential to factors $e^C e^D F(t)$
where C only contains bosonic creation operators,
D only contains bosonic annihilation operators, and
$F(t)$ is a function of time that contains no operators.

(e) note that an annihilation operator acting on the
ground state yields zero, so $e^D|0\rangle = |0\rangle$,
and the final result is given by the function $F(t)$.

2. Fourier transform this to show that $D(\omega) \equiv \mathcal{F}[F(t)]$ behaves like a power law near the Fermi energy, $D(\omega) \sim |\omega|^{\nu-1}$, where the exponent ν depends on the interaction parameter g.

3. Show that $\nu - 1 > 0$ for all g, and find the interaction strength g such that $\nu - 1 = 0$, corresponding to a constant tunneling density of states near the Fermi level.

In the home problem you will find out that the density of states in a Luttinger liquid vanishes for small energies, *i.e.*, near the Fermi surface. Again the functional form is given by a power law $D(\omega) \sim |\omega|^{\nu-1}$. This behavior means that it is very difficult to add particles to a Luttinger liquid without supplying some energy in excess of ϵ_F. This can be physically understood as an orthogonality catastrophe: the ground state of a system with $(N+1)$ particles cannot be obtained from the ground state of an N particle system simply by occupying one more single particle state — a state so obtained would be nearly orthogonal to the true ground state regardless of how the new single particle state was chosen. Instead, due to interactions between the particles, the original N particles must adjust their quantum state to accommodate the extra one roughly the same way as pearls in a necklace must move aside when a new pearl is added. This rearrangement is achieved more easily if there is some energy available to carry it through.[23] The difference between one dimension and higher dimensions is that in 1D such rearrangements affect the entire system, while in higher dimensions they are restricted to some vicinity of the perturbation.

The orthogonality catastrophe has far-reaching consequences for a number of physical quantities. The additional energy that facilitates change in the relatively rigid one-dimensional systems usually comes from either temperature or an external voltage source, and consequently most observables obey power laws as a function of T or V. Since the naive power law exponent is determined by dimensional analysis, the anomalous power laws can appear only as $(T/T_0)^\gamma$ or $(V/V_0)^\gamma$ where $k_B T_0$ and eV_0 are some energy scales that are determined by "high energy" phenomena where our approximations begin to fail. Such energy scales may come from *e.g.* length dependence of interactions ($g(k) \approx g(k=0)$ break down), curvature of the kinetic energy dispersion ($\epsilon_{\rm kin} \approx \pm v_F(k \mp k_F)$ breaks down), some other effects due to the underlying lattice, or effects connected with the finite length of the one-dimensional wire (this energy scale is typically $\approx v/L$).

[23] An alternative, more rigorous way is to say that while the naive construction of occupying one additional single particle state results in an $(N+1)$ electron state that is orthogonal to the true ground state, it is not orthogonal to excited states so that if there is some excitation energy available, it is possible (albeit still quite difficult) to add particles by single particle tunneling.

Chapter 3

Examples

3.1 Quantum wires

3.1.1 Carbon nanotubes

Carbon is one of the most versatile elements. Since historical times it is known to possess two allotropes, the stable graphite and the quasistable diamond (the title of the old James Bond film *Diamonds Are Forever* is misleading). In 1985 a new form of carbon was discovered, a class of structures known as fullerenes. The most common fullerene is C_{60} which is like a miniature soccer ball consisting of 60 carbon atoms in a geometrical structure that combines hexagons and pentagons, but similar carbon structures exist for atom number ranging from 20 to the hundreds. The structures that are usually regarded as the most recently discovered form of carbon, the carbon nanotubes, were actually seen by Morinobu Endo already in 1976, but they were largely ignored until their rediscovery by Sumio Iijima in 1991.

To understand the structure of carbon nanotubes it is useful to consider graphite first. Graphite is made up of two-dimensional sheets — graphene layers — that are stacked above each other. Interactions that keep the layers together are quite weak van der Waals forces which makes graphite one of the best dry lubricants. Each graphene layer is a honeycomb lattice of carbon atoms that are covalently bonded through sp^2 hybridization (the three-dimensional structure of diamond arises as a result of sp^3 hybridization). In terms of its electric properties graphene is a semimetal where the valence and conductance bands touch each other at six points of the Brillouin zone as shown in Fig. 3.1.

A carbon nanotube can be thought of as a wrapped up graphene sheet. There are many ways of wrapping the sheet to form a cylinder, and different wrappings result in nanotubes with different structures. One consequence of the wrapping is that the **k**-vector in the circumferential direction φ can only assume such values that $k_\varphi 2\pi R$, where R is the nanotube radius, equals an integer times 2π — this is required for the wave function to be single valued. Hence, only discrete values of the circumferential wave vector are allowed. This implies that it is not at all obvious that the K-points of the graphene Brillouin zone correspond to allowed values of the nanotube **k**-vector; if they are not among the allowed values, the valence and conduction bands of the nanotube do not touch, and the nanotube is an insulator or, at best, a semiconductor.[1]

[1] The difference between these two being the size of the gap compared to the temperature. Since the gaps in nanotubes are quite small, it is often better to regard the non-metallic nanotubes as semiconductors rather than as insulators.

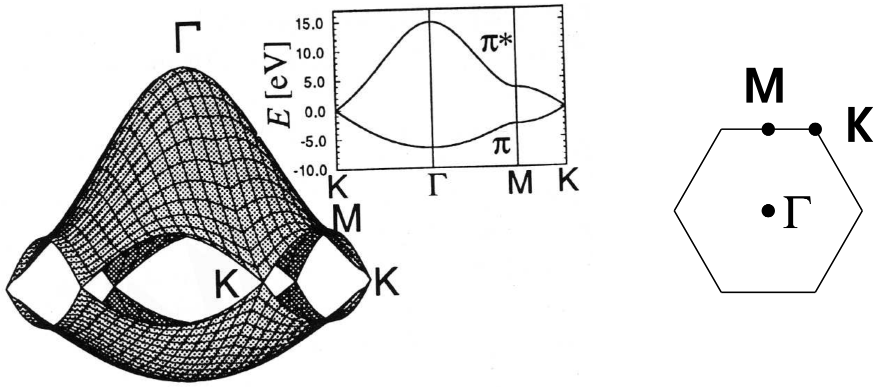

Figure 3.1: Graphene band structure showing that the occupied ($E < 0$) and unoccupied ($E > 0$) bands touch only at the six corners (K-points) of the Brillouin zone.

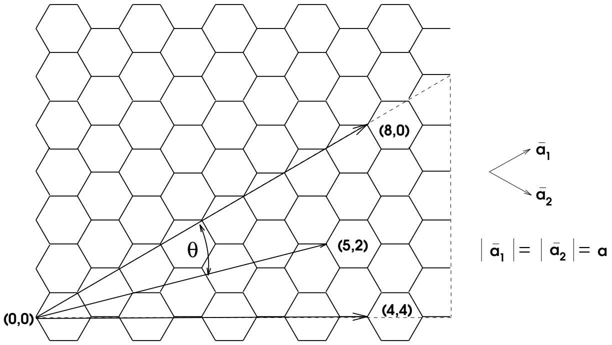

Figure 3.2: Graphene sheet showing the basis vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ and some wrapping vectors.

Whether the graphene K-points are allowed wave vectors for nanotubes depends on how the graphene sheet is wrapped, which means that we need a description of different wrapping procedures in order to be able to classify nanotubes. This is usually done by introducing two basis vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ for the graphene sheet as shown in Fig. 3.2. Wrapping means that the carbon atom in the origin will be folded on top of another carbon atom in the sheet, and the separation between these two defines the wrapping. If the two atoms are separated by vector $n\mathbf{a}_1 + m\mathbf{a}_2$, the resulting nanotube is known as a (n, m) tube. The nanotubes that are of the type $(n, 0)$ are known as zig-zag tubes, those of type (n, n) are known as armchair tubes, and the rest are known as chiral tubes. It turns out that only tubes for which $n - m$ is an integer multiple of three are metallic, and the rest are semiconducting (some deviations to this rule exist for the very smallest tubes).

The mechanical properties of carbon nanotubes are determined by the strong carbon-carbon bonds. It turns out that carbon nanotubes are the stiffest material known to man, with Young's modulus of approximately 1 TPa, compared to for instance 200 GPa for steel.[2] They are also the strongest material, and can stretch by about 30% before breaking. Nanotubes are also tend to deform elastically so that when the external forces are removed, they return to their original shapes. These unique mechanical properties of nanotubes make them very interesting for a wide range of applications ranging from badminton rackets and golf clubs to reinforcements in concrete or mechanical support in cell phone batteries. Many other applications utilize the electrical properties of nanotubes, and both nanotube displays and nanotube transistors are being developed. A particularly interesting category of applications relies on combining electrical and mechanical functionalities in a field known as nanoelectromechanics, which is an important research direction at Chalmers and Göteborg University.

In reality carbon nanotubes are of course not made by peeling sheets of graphene of pencils are rolling them into small cylinders. There are by now many methods of producing nanotubes, all of which involve heating some carboneous substance to a high temperature and then letting the carbon gas cool down and form new structures. If the circumstances are right, a larger or smaller fraction of the structures are nanotubes. There are many ways of providing the heat — initially this was done either by an electric discharge or by a laser — but the method that is now most popular is chemical vapor deposition (CVD). CVD growth requires a catalyst (for instance iron), and by controlling the placement of catalyst particles one can grow nanotubes at selected positions. The direction of growth can be controlled by applying an AC electric field, and a combination of these two techniques has been developed to quite a versatile method. There are still, however, some difficulties: first and foremost, one cannot, yet, selectively grow nanotubes with specific (n, m) values, the thickness of the grown tubes is not perfectly controlled, and electric contacts between the tubes and external circuits are difficult. Also, often the nanotubes consist of several concentric cylinders (multi-walled nanotubes, MWNTs) and coupling between the different shells (walls) is not perfectly understood yet. In many applications, however, multi-walled tubes are preferred since they are predominantly metallic (any metallic shell is sufficient to make the whole tube metallic), they are larger than the single-walled tubes (SWNTs) which makes them sturdier mechanically and easier to contact, and their growth conditions are not quite as restricted as those of SWNTs.

From the theoretical point of view SWNTs are a very interesting system. They have very

[2] Young's modulus E gives the relationship between the strain $\frac{\delta L}{L}$ and the stress σ. Larger E implies that the material stretches less when subjected to the same force per area of cross section.

small radii, of the order of one nanometer, they have few impurities, and they can be metallic: they are very close to being ideal one-dimensional wires. The armchair metallic tubes have a band structure where two one-dimensional bands cross the Fermi level at $k = +k_F$ and two at $k = -k_F$, implying that they are effectively like quantum wires with two transverse modes. Since each mode can carry electrons with spin up or spin down, the maximal conductance of a single armchair SWNT is predicted to be $4\frac{e^2}{h}$ corresponding to a resistance of about 6.5 kΩ. The bands of SWNTs are linear over quite a large energy range near $k = \pm k_F$, which suggests that the Luttinger approximation should be quite good.

Fact or Fiction?

Experiments carried by Paul McEuen's group at UC Berkeley and Cees Dekker's group at TU Delft measured the conductance of a SWNT connected to external electrodes by tunnel junctions that were placed either near the end of the nanotube, or far from the ends. The conductance was measured as a function of both the voltage and the temperature, and compared to predictions based on the Luttinger model. The free parameters in the model were the interaction constant g which could be roughly estimated, and should be the same for the two experiments, and the proportion of the applied voltage that actually drops across the tunnel junction rather than elsewhere in the circuit. The theoretical prediction is ($\alpha = \frac{1}{2}\left(\frac{1}{g} - 1\right)$)

$$\begin{aligned}\frac{dI}{dV} \sim\ & T^\alpha \sinh\left(\frac{eV}{2k_BT}\right) \left|\Gamma\left(1 + \frac{\alpha}{2} + i\frac{eV}{2\pi k_BT}\right)\right|^2 \\ & \left\{\frac{1}{2}\coth\left(\frac{eV}{2k_BT}\right) - \frac{1}{\pi}\mathrm{Im}\left[\psi\left(1 + \frac{\alpha}{2} + i\frac{eV}{2\pi k_BT}\right)\right]\right\}\end{aligned}$$

where $\psi(z)$ is the logarithmic derivative of the gamma function.
The agreement between theory and experiment appears quite good except at the lowest temperatures where the Coulomb blockade of the SWNT becomes important. The interaction parameter g that gives the best fit ir roughly $g \approx 0.24$ which compares quite favorably with the theoretical estimate $g \approx 0.28$. Hence, the interaction is very strong, $g \ll 1$, and in particular $g < 0.5$ which is the smallest g-value that is compatible with a short range interaction, implying that the long range of the Coulomb interaction is important. However, the curve does not exhibit many features (essentially just the slope and the position of the kink at $k_BT \approx eV$), and other explanations have been suggested, such as the role of the electromagnetic environment which also leads to power laws similar to those predicted by the Luttinger model. Consequently, not everybody is convinced that nanotubes behave as Luttinger liquids.
What would it take to convince those in doubt? Detecting the separation between spin and charge would probably constitute smoking gun evidence that would be hard to reject. This could be done, for instance, by measuring the so-called spectral function $A(k, \omega)$ that exhibits peaks at such combinations of k and ω that are connected by a dispersion relation $\omega = \omega_\alpha(k)$: in Luttinger liquids there are the charge $\alpha = \rho$ and spin $\alpha = \sigma$ excitations have different velocities and hence different dispersions ($\omega = vk$), so a spin-charge separation should result in the splitting of peaks in the spectral function. Unfortunately, the spectral function is not directly measurable, and its value must be inferred from experiments that are usually sensitive to some integral of $A(k, \omega)$, and an integral can assume the same value due to two small peaks or one large one.

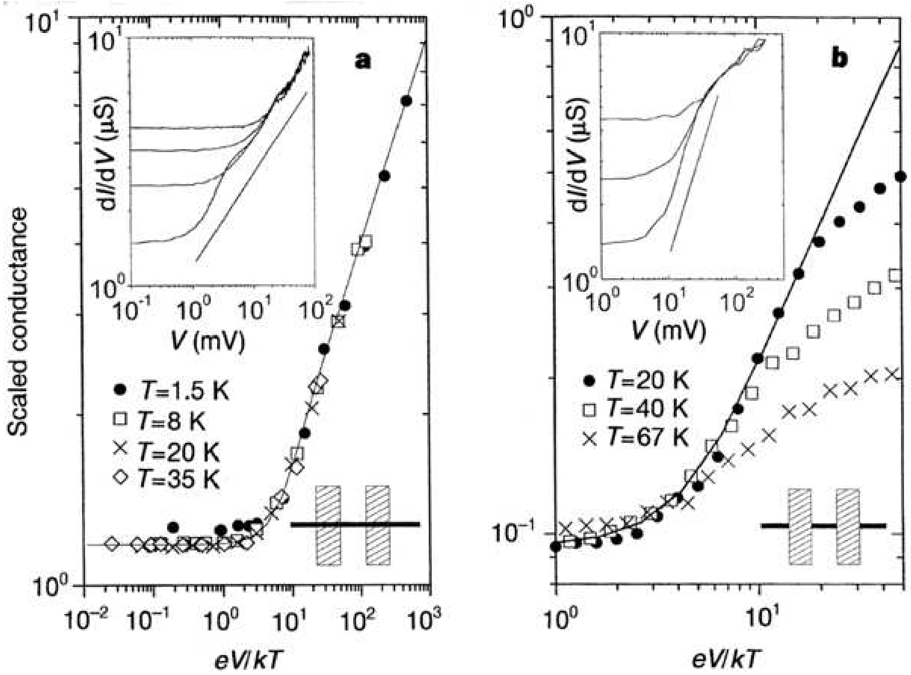

Figure 3.3: Current through a SWNT, multiplied by a power of temperature, as a function of eV/k_BT. The solid line is the prediction of the Luttinger model. Data by Marc Bockrath *et al.*, Nature **397, 598, 1999.**

3.1.2 Other quantum wires

Carbon nanotubes can be described as bottom-up fabricated quantum wires: one starts from simpler structures, carbon atoms, and creates suitable circumstances so that the ingredients coalesce to form the desired end results. Bottom-up fabrication is what one finds in Nature, but the conventional engineering approach has been to take a large piece of material, and work on it to create the desired end structure. This method has been applied all the way from the Stone Age, when it was used to make arrow heads, to the Silicon Age of microchips, and it has served countless generations of engineers quite well — we should at least investigate if the method can be applied to create quantum wires, too.

To create a quantum wire we need confinement in two directions, call them z and y, in order to create a wire in the third, x, direction. Confinement in z direction is straightforward: one can either (i) coat a semiconducting material such as silicon with an insulating oxide, build a metal gate on top of the oxide, and apply a voltage to the gate to attract charge carriers to the interface between the semiconductor and the insulator, or (ii) bring two different semiconductors together which generically results in carrier flow (diffusion) from one material to another until an internal electric field builds up to stop the flow, with the end result that there are electrons on one side of the interface. The first method is the standard engineering recipe to create MOSFETs, and was used *e.g.* in the study of the integer quantum Hall effect. The second method typically employs GaAs and AlGaAs as the two semiconductors and results in the creation of thin electron layer on the GaAs side of the interface. In both cases the potential well that confines carriers to the interface is so narrow that only the lowest mode in the z-direction is occupied, rendering the carrier systems effectively two dimensional (2DES). The advantage of the latter method is that it allows for higher mobilities, *i.e.* lower resistivities, which makes it easier to study many of the more exotic phenomena.

Confinement in the lateral (y) direction is usually achieved lithographically. In order for the quantum wire of width W to be one-dimensional, the Fermi level must not be too high, in particular,

$$\frac{\hbar^2 k_F^2}{2m} + \frac{\hbar^2}{2m}\left(\frac{\pi}{W}\right)^2 < \frac{\hbar^2}{2m}\left(\frac{2\pi}{W}\right)^2$$

where the left hand side is the kinetic energy of an electron on the Fermi level and in the lowest transverse mode, and the right hand side is the energy of an electron in the next lowest transverse mode and with no longitudinal momentum. Hence, the Fermi wave vector must be $k_F < \frac{\pi\sqrt{3}}{W}$. This Fermi wave vector corresponds to a two-dimensional carrier density $n = k_F^2/(2\pi) = \frac{3\pi}{2W^2}$. It is hard to maintain good mobility in a 2DES if the carrier density falls below 10^{10} cm^{-2} (the Fermi sea is so shallow that the impurities can completely block carrier flow), which implies that the channel width must be smaller than about 200 nm. In practice the carrier density must be higher since quantum wires are more sensitive to impurities, and the maximal quantum wire width falls well below 100 nm. In the above analysis we assumed that the confining potential was like a square well, which is a reasonable model for etched quantum wires. Another way to create quantum wires using a top-down method is to pattern closely spaced metallic gates on top of the 2DES, and expel electrons underneath the gates to create a narrow channel in the 2DES between the gates. This method results in a considerably softer confinement, roughly similar to a parabolic potential, and must be modeled accordingly.

Lithographically defined quantum wires have been seen to exhibit quantized conductance in accordance with the Landauer model, but they are quite prone to disorder effects — impurities or surface roughness — which has hampered effects to detect Luttinger-type interaction effects in them. This is in part due to the asymmetry of the fabrication technique: control in the z-direction (growth direction) is on the level of individual atomic layers, about 0.4 nm, so interfaces are very sharp, whereas control in the lateral direction is accurate only to about 10 nm. An obvious way to try to improve the situation is to use the growth control even in the lateral direction. This results in the so-called cleaved edge overgrowth technique developed by Amir Yacoby and coworkers in the late 1990s. In this technique one grows a conventional AlGaAs-GaAs sandwich, stops the growth, cleaves the sample along a suitable crystal plane, rotates the sample, and starts to grown on what used to be the edge of the sample. The wires produced by this technique are the cleanest semiconductor wires to date, and Luttinger-type effects may have been seen in them; however, just like in the case of carbon nanotubes, the evidence is not entirely convincing.

Apart from carbon nanotubes, there are other bottom-up -type quantum wires. In the category of linear organic molecules probably the most famous is DNA, but the conductivity of DNA is a matter of some debate at present (in the words of one researcher in the field, the conductivity of DNA is comparable to that of silicon dioxide), and DNA may not be the best example of a one-dimensional metal. Other molecules such as polyacetylene are promising in some aspects but prohibitively difficult to work with (polyacetylene combusts spontaneously). The category that is perhaps most promising is a group of organic salts known as Bechgaard salts, which have exhibited power law dependences characteristic of Luttinger liquids but also some signs of spin-charge separation in both optical and transport experiments.

3.2 Fractional Quantum Hall Effect

3.2.1 Bulk properties

In the Chapter on coherence effects we discussed the Integer Quantum Hall Effect, IQHE, which could be explained entirely without reference to interactions between electrons. The Fractional Quantum Hall Effect (FQHE), on the other hand, owes its existence entirely to electron-electron interactions.

Before embarking on the analysis of FQHE in particular let us first consider the relative importances of interactions in electron fluids. The kinetic energy of an electron in the absence of a magnetic field is $\frac{\hbar^2 k^2}{2m}$ and at $T = 0$ so many **k**-states are occupied that all electrons can be accommodated. The wave vector of the highest occupied state is hence given by $N = 2V\int_0^{k_F} S_{d-1}\frac{dk}{(2\pi)^d}k^{d-1} = 2VS_{d-1}\frac{1}{(2\pi)^d d}k_F^d$ or $k_F = 2\pi\left(\frac{dn}{S_{d-1}}\right)^{1/d}$ where N is the number of electrons, V is the volume of the sample, $n = N/V$ is the electron density, and S_{d-1} is the area of the $(d-1)$-dimensional surface of a d-dimensional unit sphere, *i.e.*, $S_2 = 4\pi$, $S_1 = 2\pi$ and $S_0 = 2$. Consequently, the kinetic energy of electrons on the Fermi level scales as $n^{2/d}$. The Coulomb interaction between electrons scales as their inverse separation, *i.e.* as $n^{1/d}$. This means that the relative importance of interactions *vs.* kinetic energy scales as $n^{-1/d}$, which means that dense electron systems are effectively weakly interacting (their energetics is dominated by the kinetic energy) while sparse electron systems are strongly interacting. Also, we see that the density dependence is stronger in low-dimensional systems, implying that strong interaction effects are more likely to be seen in systems with reduced dimensionality.

The above argument relies on the specific form of the kinetic energy that is valid at zero magnetic field. In the discussion of the IQHE we saw that in large magnetic fields the kinetic energy falls into Landau levels with energies $\epsilon_n = (n + \frac{1}{2})\hbar\omega_c$ where $\omega_c = eB/m$ and $n = 0, 1, 2, \ldots$. Each Landau level is hugely degenerate, and in the absence of interactions it makes no difference, from an energetics point of view, in which order states in a Landau level are occupied. If a Landau level is completely full, all states in it must be occupied, but for a partially filled Landau level there are many ways to choose the occupied states: if the level has N_1 degenerate states that must accommodate $N < N_1$ electrons, there are $\begin{pmatrix} N_1 \\ N \end{pmatrix}$ ways of choosing the occupied states. Consider now a 1/3-filled Landau level of a 1 cm^2 sample at a magnetic field of one tesla: the number of ways of choosing the occupied states is $\begin{pmatrix} 2.5\times 10^{10} \\ (2.5/3)\times 10^{10} \end{pmatrix} \approx 10^{7\times 10^9}$. In the presence of interactions, this degeneracy is broken, and among all the $10^{7\times 10^9}$ states there is one with the lowest energy, the ground state, and a few that have energies slightly above that of the ground state. We will now set out to find these states. An exhaustive search does not seem feasible; instead, at some point along the way, we must think.

In the discussion of the IQHE we employed the transverse gauge but for the FQHE discussion it is simpler to use the symmetric gauge that we encountered in the analysis of persistent currents. Not all too surprisingly, the Schrödinger equation in this gauge also becomes a harmonic oscillator (this is almost obvious from the beginning — physical features such as energy spectrum are gauge independent, and in the transverse gauge we obtained a harmonic oscillator). The appropriate quantum numbers are now the principal quantum number n, which is the Landau level index and determines the energy $\epsilon_n = (n + \frac{1}{2})\hbar\omega_c$, and the angular momentum quantum number ℓ. In the symmetric gauge the (unnormalized) states can be written as

$$\psi_{n,\ell}(r,\theta) = r^{\ell} e^{-i\ell\theta} e^{-\frac{r^2}{4\ell_c^2}} L_n^{\ell}\left(\frac{r^2}{2\ell_c^2}\right)$$

where $\ell_c = \sqrt{\hbar/(eB)}$ is the magnetic length and $L_n^{\ell}(x)$ an associated Laguerre polynomial. In the following we only consider the lowest Landau level when the Laguerre polynomial is identically equal to unity. It is convenient to introduce the complex position variable $z = re^{i\theta}$ so that the lowest Landau level wave functions can be written as $z^{\ell} e^{-|z|^2/4}$ where all lengths are measured in units of ℓ_c.

The fractional quantum Hall state will turn out to be a true many-body state that cannot be described in terms of single-electron wave functions. Therefore, we will need the many-particle wave functions that depend on all electrons' coordinates simultaneously. To this end we begin by considering the many-particle state describing the full lowest Landau level. Since electrons are fermions, the many-particle state $\Psi(z_1, \ldots, z_N)$ must be odd under particle exchange, $\Psi(z_1, \ldots, z_i, \ldots, z_j, \ldots, z_N) = -\Psi(z_1, \ldots, z_j, \ldots, z_i, \ldots, z_N)$, which can be accomplished by a Slater determinant. The Slater determinant is the determinant of a matrix whose entries are $\psi_i(z_j) = z_j^i e^{-|z_j|^2/4}$. When we expand the determinant, we see that each term contains the same Gaussian factor, and the many-particle state for a full lowest Landau level can be written as

$$\Psi_1(z_1, \ldots, z_N) = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_N \\ \ldots & & & \\ z_1^{N-1} & z_2^{N-1} & \ldots & z_N^{N-1} \end{vmatrix} e^{-\frac{1}{4}\sum_{j=1}^{N} |z_j|^2}$$

This determinant can be evaluated in a closed form by noticing that (i) the polynomial prefactor is of order $0 + 1 + \ldots + (N-1) = N(N-1)/2$ since each term in the determinant contains one factor from each row, and (ii) the polynomial is odd under exchange $z_i \leftrightarrow z_j$, which implies that it must be proportional to (an odd power of) $(z_i - z_j)$ for any choice of indices $i \neq j$. The second observation implies that the polynomial prefactor is proportional to $\prod_{i<j}(z_i - z_j)$, possibly multiplied by a symmetric polynomial of the coordinates, but this polynomial is already of the order determined by argument (i), so the symmetric polynomial must be just a constant. Inspection reveals that the constant is one, and hence the many-particle state describing a full Landau level is uniquely given by $\Psi_1(z_1, \ldots, z_N) = \prod_{i<j}(z_i - z_j) e^{-\frac{1}{4}\sum_{j=1}^{N} |z_j|^2}$.

A possible many-electron state Ψ describing a partially filled lowest Landau level shares many features with Ψ_1. Since it is made up from the same single particle states, it must contain the same Gaussian factor, multiplied by an analytic function of the coordinates z_i — analytic since it is a polynomial of z_i and does not involve any complex conjugates z_i^*. Since the wave function must be odd with respect to particle exchange, it must also be proportional to $\prod_{i<j}(z_i - z_j)$, but unlike Ψ_1, it may (and must) be multiplied by some symmetric polynomial of the coordinates. The choice of symmetric multipliers may be reduced by noticing that since the system is cylindrically symmetric (this is where the gauge choice is useful), the angular momentum L_z is a good quantum number. The angular momentum of a single particle state is simply ℓ, and the angular momentum of a many-particle state is given by the sum of the ℓ-quantum numbers. Hence, if Ψ is an eigenstate of L_z, each term in the polynomial multiplying the Gaussian factor must have the same order. Now, following Robert Laughlin, we postulate

that Ψ has the form[3]

$$\Psi(z_1,\dots,z_N)=\prod_{i<j} f(z_i-z_j)e^{-\frac{1}{4}\sum_{j=1}^{N}|z_j|^2}$$

where $f(z)$ is some odd polynomial of a uniform degree. "Polynomial of a uniform degree" is simply a power, $f(z)=z^m$, where m must be odd. Hence, provided that Ψ is of the form above, the only possibilities are

$$\Psi_m(z_1,\dots,z_N)=\prod_{i<j}(z_i-z_j)^m e^{-\frac{1}{4}\sum_{j=1}^{N}|z_j|^2}, \quad m\ odd.$$

Now we have established that Ψ_m is a possible many-electron state describing electrons on the lowest Landau level and consistent with the conservation of angular momentum. What kind of a state is Ψ_m, in particular, what is the density of the electron system it describes? To address this point, we again follow Laughlin's neat trick that relies on experience from two seemingly disjoint fields of physics. In quantum mechanics the probability density of finding particles at positions $\{z_i\}$ is given by $|\Psi_m(\{z_i\})|^2$. In classical statistical mechanics, the probability density of finding particles at positions $\{z_i\}$ is proportional to $e^{-\beta H_{\rm cl}(\{z_i\})}$ where $\beta=(k_BT)^{-1}$ and $H_{\rm cl}(\{z_i\})$ is the classical Hamiltonian. Laughlin asked "What is the classical Hamiltonian that yields the probability density $|\Psi_m(\{z_i\})|^2$ in classical statistical mechanics?" — if we know that, we can apply our knowledge of the classical system to infer the nature of the state $\Psi_m(\{z_i\})$. This is easy, simply a matter of taking a square and a logarithm, and yields

$$H_{\rm cl}(\{z_i\})=\beta\frac{1}{2}\sum_{j=1}^{N}|z_j|^2-2\beta m\sum_{i<j}\ln|z_i-z_j|$$

where β is arbitrary. The second term looks like what you get for a collection of particles at positions $\{z_i\}$ that interact with each other through a logarithmic potential. Choosing $\beta=m/2$ yields a form where we can identify m as some kind of a charge of the particles and $-\ln r$ as the interaction potential.

Consider now, for no obvious reason, a purely two-dimensional universe. In that Flatland (as it was called in a famous novel by Edwin A. Abbott) the Coulomb potential still obeys the Poisson equation $\nabla^2\varphi(r)=-2\pi\rho(r)$ whose solution for a point charge is $\varphi(r)=-\ln(r)$. Hence, the second term of $H_{\rm cl}$ describes point charges in Flatland. The first term can be identified by considering the Flatland potential that is connected to a constant charge distribution, which is obtained by solving $\nabla^2V(r)=-2\pi\rho$. Integrating twice yields $V(r)=-\frac{1}{2}\pi\rho r^2$, which shows that the first term of $H_{\rm cl}$ describes Flatland point charges of magnitude m interacting with a uniform background with charge density $\rho=\frac{1}{2\pi}$. Thus, we have established that the classical Hamiltonian describes a system of Flatland point charges interacting with each other and a uniform background charge. In the limit of large N, the energy of this system blows up unless the density of the point charges precisely neutralizes the background charge, which implies that the classical probability density is maximized for a uniform distribution of point charges, and that the uniform density is given by $\frac{1}{2\pi m}$. Since the Flatland point charges are in a one-to-one correspondence with the electrons of the FQHE system, we have now

[3] The choice of this form was motivated in part by analysis of interaction nucleon systems in nuclei, where this Ansatz is known as the Jastrow form.

established that the probability density associated with $\Psi_m(\{z_i\})$ is maximized for a uniform distribution of electrons with density $\rho_m = \frac{1}{2\pi m \ell_c^2}$, corresponding to filling factor $\nu = 1/m$. Thus, the Laughlin wave function can only be constructed a particular densities.

A more complete study — *e.g.* exact diagonalization of a small system — shows that the Laughlin states are extraordinarily good approximations to the true ground states at the corresponding densities. Further evidence to their particular role can be obtained by considering excitations above these ground states. We will not go into details here, but merely point out that in the Flatland analogy a possible excitation is to create some inhomogeneities in the charge distribution. Since the corresponding wave function must still be of the form of an antisymmetric polynomial times a Gaussian (lest we wish to pay the kinetic energy cost of involving higher Landau levels), the simplest excitation to create is a charge deficiency: simply multiply the ground state by factor $\prod_i (z_i - z_0)$, which results in a reduced electron density near the point z_0. The Flatland analogy can be used to establish that proportion $1/m$ of an electron charge is missing near the point z_0 (compared to the uniform charge distribution). This charge deficiency may move around, and forms an elementary excitation (known as a quasihole) in the FQHE system. A corresponding charge bump, or quasielectron, can also be constructed. If the system is subjected to electric fields or other perturbations, the quasielectrons and quasiholes move almost as independent particles, except for the sum rule that the total charge in the system must be an integer times the electron charge. These fractionally charged excitations have been seen experimentally in shot noise measurements, and they constitute a serious deviation from the Landau Fermi liquid theory in which all excitations are in one-to-one correspondence with excitations of a non-interacting system.

The Hall conductance can be calculated as in the Integer Quantum Hall Effect, and is again found to have the classical form $\nu\frac{e^2}{h}$. Indeed, experiments reveal that at high enough magnetic fields two-dimensional electron systems exhibit quantized Hall conductances at values $\frac{1}{m}\frac{e^2}{h}$ where m is an odd integer. However, the experimental system exhibits quantized Hall conductance at many other fractional multiples p/q of the quantum conductance as shown in Figure 3.4, but only at fractions where the denominator q is odd. This hierarchy of levels can be explained in many ways (they may be equivalent in the end) where the basic idea is that the higher hierarchy levels can be interpreted as quantum Hall effects of the excitations of the lower states in the hierarchy — for instance, the excitations of the 1/3 state can give rise to the 2/5 state *etc.*.

We will not pursue this analysis any longer in this course, but conclude that a particularly simple hierarchy construction has been provided by Jainandra Jain who considered fictitious particles that carry not only an electric charge but also an even number of magnetic flux quanta $\Phi_0 = h/e$. In the Jain construction, for instance, the 1/3 state is understood as a full lowest Landau level of of fictitious particles carrying two flux quanta: the total flux per particle is Φ_0 since it is a full lowest Landau level plus $2\Phi_0$ since each particles carries two flux quanta, so that the total flux per particle is $3\Phi_0$, which implies filling factor 1/3. Similarly, if these fictitious particles fill p Landau levels, the total flux per particle is $2 + 1/p$ and the filling factor is $(2+1/p)^{-1} = p/(2p+1)$, *i.e.* 1/3, 2/5, 3/7... Binding four flux quanta to each particle results in filling factors 1/5, 2/9, 3/13... The Jain construction can even be used to write down wave functions for the different states and the results agree with the Laughlin construction for the $1/m$ states.

The Jain construction can even be employed at filling factor $\nu = 1/2$, which does not allow for a fractional quantum Hall state, and experimentally no quantized Hall conductance

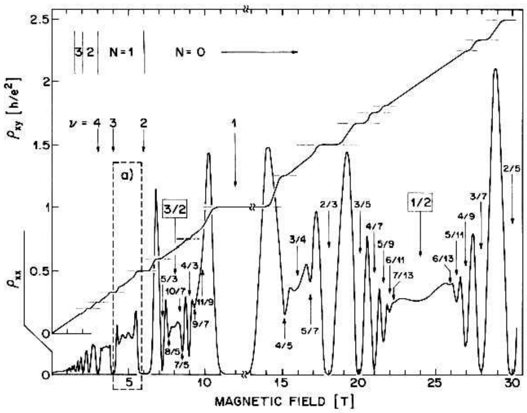

FIG. 1. Overview of diagonal resistivity ρ_{xx} and Hall resistance ρ_{xy} of sample described in text. The use of a hybrid magnet with fixed base field required composition of this figure from four different traces (breaks at ≃12 T). Temperatures were ≈150 mK except for the high-field Hall trace at $T=85$ mK. The high-field ρ_{xx} trace is reduced in amplitude by a factor 2.5 for clarity. Filling factor ν and Landau levels N are indicated.

Figure 3.4: Hall resistance and longitudinal resistance of a very pure two-dimensional electron system as a function of the magnetic field. Note the quantized values of ρ_{xy} at $\frac{q}{p}\frac{h}{e^2}$ where q is an odd integer. The quantization plateaux are accompanied by minima of the longitudinal resistance indicating a gap in the excitation spectrum.

is observed at this filling factor. However, experiments reveal that even at this filling factor the electron fluid forms a special kind of a state with unusual properties. This special state corresponds to Jain's fictitious particles in zero magnetic field (as all of the magnetic flux is accounted for by the flux quanta bound to particles), and turns out to be a Fermi liquid, albeit an unusual one — the special nature of these even denominator states has prompted some researchers to introduce a classification of the quantized Hall effects as the Integer Quantum Hall Effect, the Fractional Quantum Hall Effect, and the Unquantized Quantum Hall Effect (the last term is not yet in universal use). This $\nu = \frac{1}{2}$ state was first studied theoretically in a joint publication by Bertrand Halperin, Patrick Lee, and Nicholas Read in the early 1990s.

3.2.2 Edges

In the Integer Quantum Hall Effect the edges of the sample played an important role, which was essentially due to the fact that there was a gap to excitations in the bulk — such excitations would require promoting a particle to the next Landau level at the cost of $\hbar\omega_c$ — while excitations near the edge were gapless and could respond to small perturbing fields. In the Fractional Quantum Hall Effect the excitations in the bulk are quasielectrons and quasiholes whose creation also requires a finite amount of energy (of the order of $0.03e^2/(4\pi\epsilon_0\ell_c) \approx 5$ meV $= 60$ K at a field of 10 teslas. Thus, FQHE experiments must be carried out at higher magnetic fields and at low temperatures so that thermally excited quasiparticles do not smear out the effect). Hence, even in the FQHE, the edges can be expected to pay a large role in the response of the system.

The role of the FQHE edge states has been studied in particular by Xiao-Gang Wen and coworkers, who have concluded that the FQHE edges differ from the IQHE edges in one fundamental aspect. Both edge states are chiral, that is, they only support excitations that move in one direction along the edge; this we saw in the IQHE case. The difference is that while the IQHE edges behave like non-interacting uni-directional (chiral) one-dimensional wires, the FQHE edges behave as interacting uni-directional one-dimensional wires. In view of the crucial role that interactions play in the FQHE, it is hardly surprising that they are important at the edge, too. The precise arguments leading to a theoretical description of the FQHE edges are too involved to be considered here, but the end result is that the FQHE edges can be described as chiral Luttinger liquids. For the simple filling factors $\nu = 1/m$ the interaction constant of the edge Luttinger liquid is given by $g = \nu$, that is, its value is uniquely determined by the bulk properties.[4] Therefore, the FQHE system provides an interesting opportunity to test the Luttinger liquid model as the predicted power law exponents are well known, and are predicted to change in a well-defined fashion as the filling factor is changed. A difficulty in this testing method is that when an FQHE sample is connected to a measurement apparatus, the intricate correlations inside the sample must somehow evolve into the ordinary Fermi liquid inside the apparatus. That evolution is difficult to describe, and it is quite likely that the Luttinger effects do not survive the invasive measurement; instead, contactless measurements or measurements that are sensitive to internal correlations are needed. One such measurement focused on scattering across a thin neck in an FQHE system, and showed that the scattering was consisted with transport of fractionally charged quasiparticles or the chiral edge Luttinger liquid model.

3.3 Quantum magnets

HJ

3.4 Kondo effect

HJ

3.5 Mott insulators

HJ

3.6 Bose-Einstein condensation

HJ

[4] At more complicated filling factors the edge is also more complicated, and may consist of several chiral Luttinger liquids propagating in different directions and only weakly interacting with one another.

Chapter 4

Toolbox

4.1 Renormalization group

4.1.1 Position space renormalization group

Renormalization group (RG) technique was developed by Kenneth Wilson in the early 1970s as a formal implementation of scaling ideas of Leo Kadanoff. It has become the standard tool for analyzing the properties of a wide range of models in theoretical physics, and the ideas have been applied to many other fields as well (e.g. turbulence). The basic idea behind renormalization group is to relate the system's behavior on a macroscopic, large length scale to its description on a microscopic scale by systematically inspected how small scale phenomena manifest themselves on longer length scales. This is done by carefully eliminating those degrees of freedom that describe small scale variations — in real space, short range fluctuations, or in fourier space, modes with larger wave vectors — and determining how the small scale properties lead to couplings between the degrees of freedom in a larger scale. We have already seen results that have the flavor of an RG analysis when we discussed localization: the scaling function $\beta(G)$ tells how conductance on one length scale relates to conductance on a larger length scale. This scaling behavior allowed us to understand why in three dimensions some systems behave as metals while others behave as insulators; such an identification of macroscopic phases is one common outcome of an RG analysis as we will see in the following.

The method is easiest to illustrate in real space, and particularly easy if there is only one spatial dimension. Real applications of the technique are usually carried out in momentum space, so for the purposes of this course we start with position space renormalization group, move then on to a more general discussion of the method, and conclude with a momentum space application.

Decimation

We will start our study of the renormalization group technique by considering a special case, decimation, which can often be carried out exactly for one-dimensional models. In higher dimensions exact decimation is usually not possible and some approximations are needed.

Let us consider the infinite one-dimensional Ising model

$$H = -J \sum_{i=-\infty}^{\infty} s_i s_{i+1} - h \sum_{i=-\infty}^{\infty} s_i \tag{4.1}$$

where $s_i = \pm 1$ are dimensionless spin variables on lattice sites i and J and h are parameters that describe the interactions between adjacent spins and the effect of external magnetic field, respectively. Both J and h have units of energy. The partition function is given by

$$Z(T,h) = \sum_{\{s_i\}_{i=-\infty}^{\infty}} \exp\left[K \sum_{i=-\infty}^{\infty} s_i s_{i+1} + \beta h \sum_{i=-\infty}^{\infty} s_i\right] \tag{4.2}$$

where $K = \beta J$. This would be easy to evaluate if the different spins were not coupled — in that case we could do sum over each spin independently — but the term $s_i s_{i+1}$ couples two neighboring spins, which makes the evaluation of Z more complicated. However, we can divide all spins into two subsets in such a way that spins in one subset interact only with spins in the other subset[1]: the neighbors of an even site are odd, and the neighbors of an odd site are even. To take advantage of this separation into mutually non-interacting lattices we write the partition function as

$$Z(T,h) = \sum_{\{s_{2n-1}\}_{n=-\infty}^{\infty}} \sum_{\{s_{2l}\}_{l=-\infty}^{\infty}} \exp\left[K \sum_{k=-\infty}^{\infty} (s_{2k-1}s_{2k} + s_{2k}s_{2k+1}) + \beta h \sum_{k=-\infty}^{\infty} s_{2k} + \beta h \sum_{k=-\infty}^{\infty} s_{2k+1}\right], \tag{4.3}$$

where we separated the contributions from even and odd sites. Let us now consider the sum over the spin at a particular even site $2l$. There are only three terms in the exponent that depend on s_{2l}, so the sum over s_{2l} gives

$$\sum_{s_{2l}=\pm 1} \exp\left[K s_{2l}(s_{2l-1} + s_{2l+1}) + \beta h s_{2l}\right] = 2\cosh\left[K(s_{2l-1} + s_{2l+1}) + \beta h\right].$$

We can do the same for all other even terms, and find that the partition function can be written as

$$Z(T,h) = \sum_{\{s_{2n-1}\}_{n=-\infty}^{\infty}} \prod_{n=-\infty}^{\infty} 2\cosh\left[K(s_{2n-1} + s_{2n+1}) + \beta h\right] e^{\frac{1}{2}\beta h(s_{2n-1}+s_{2n+1})}, \tag{4.4}$$

where I wrote the last term in a symmetric way for convenience.

Now we have expressed the partition function of the original system as a sum of only half of the degrees of freedom — only the odd spins. We can imagine that the partition function (4.4) could arise as a partition function for a system consisting of only the odd spins, but since it is not in the usual form $\mathrm{Tr}e^{-\beta H'}$, it is not immediately clear what Hamiltonian H' of the odd-spin-system would result in the partition function (4.4). Let us proceed by making the *Ansatz* that the Hamiltonian H' is also of the Ising type, i.e.

$$H' = -J' \sum_{n=-\infty}^{\infty} s_{2n-1}s_{2n+1} - \sum_{n=-\infty}^{\infty} (h' s_{2n-1} + g'). \tag{4.5}$$

[1] Lattices for which such a division is possible are called bipartite, and they are frequently much simpler to deal with than more general lattices. The difference is particularly important in the case of an antiferromagnet: a bipartite lattice has a vanishing entropy as $T \to 0$ but a non-bipartite lattice has a large number of degenerate ground states and therefore does not obey the third law of thermodynamics, i.e. the entropy does not vanish as temperature goes to zero.

s_{2l-1}	s_{2l+1}	LHS	RHS
1	1	$2e^{\beta h}\cosh[\beta h + 2K]$	$e^{K'+\beta h'+\beta g'}$
-1	1	$2\cosh[\beta h]$	$e^{-K'+\beta g'}$
1	-1	$2\cosh[\beta h]$	$e^{-K'+\beta g'}$
-1	-1	$2e^{-\beta h}\cosh[\beta h - 2K]$	$e^{K'-\beta h'+\beta g'}$

Table 4.1: Decimation equations (4.7). LHS = "left hand side", RHS = "right hand side".

Here J' describes the interaction between the odd spins, h' describes their coupling to the external magnetic field, and g' is included to allow for a shift in the zero of energy. The parameter g' does not affect the behavior of the model.

It is by no means clear that this *Ansatz* is correct, that is, that we can find constants J', h', and g' so that the Hamiltonian H' gives rise to the partition function (4.4). To proceed we write down the partition function for the Hamiltonian H',

$$Z'(T,h') = \sum_{\{s_{2n-1}\}_{n=-\infty}^{\infty}} \exp\left[K' \sum_{n=-\infty}^{\infty} s_{2n-1}s_{2n+1} + \beta \sum_{n=-\infty}^{\infty} (\frac{1}{2}h'(s_{2n-1}+s_{2n+1}) + g')\right]. \tag{4.6}$$

We isolate the terms in (4.4) and in Z' that depend on spins s_{2l-1} and s_{2l+1}, and set them equal to each other — that is necessary if the partition functions are to agree. That gives the equations

$$\begin{aligned} &2\cosh\left[K(s_{2l-1}+s_{2l+1}) + \beta h\right] e^{\frac{1}{2}\beta h(s_{2l-1}+s_{2l+1})} \\ &= \exp\left[K' s_{2l-1}s_{2l-1} + \tfrac{1}{2}\beta h'(s_{2l-1}+s_{2l+1}) + \beta g'\right] \end{aligned} \tag{4.7}$$

where $s_{2l-1} = \pm 1$ and $s_{2l+1} = \pm 1$. Thus, we have four equations and only three unknowns K', h' and g', and the correctness of our *Ansatz* appears to be in doubt. It is convenient to write out all the equations, which I have done in Table 4.1. We see that the second and third equations collapse into one, so that there are only three independent equations and therefore we may expect to find a solution. Dividing the first equation by the fourth yields

$$h' = h + \frac{1}{2\beta}\log\left[\frac{\cosh[2K+\beta h]}{\cosh[-2K+\beta h]}\right]; \tag{4.8}$$

dividing the product of the first and fourth equations by the product of the second and third equations gives

$$K' = \frac{1}{4}\log\left[\frac{\cosh[2K+\beta h]\cosh[-2K+\beta h]}{\cosh^2[\beta h]}\right]. \tag{4.9}$$

A few remarks are now in order. We have obtained the result that if we perform a partial sum over half of the degrees of freedom, i.e. the even spins, the resulting partition function describes an Ising-type interaction between the remaining degrees of freedom. The form of the Hamiltonian is exactly the same as before but the coupling constant has changed, $K \to K'$ as has the effective magnetic field, $h \to h'$. Nothing prevents us from doing another partial sum over half of the remaining degrees of freedom (say, sites $4l+1$). If we were to do that, we would find that the partition function after the partial summation would describe an Ising-interaction between the remaining degrees of freedom (s_{4l+3}). The coupling constant of the resulting model is K'' and the effective magnetic field is h'', which describe the behavior of the model at a length scale that is four times the original length scale (the separation between the remaining degrees of freedom has increased by a factor of four). Thus, we can coarse-grain the original Ising chain by successively removing every second spin, and the result at each stage is another Ising model with a new coupling constant $K^{(n)}$ and a new effective magnetic field $h^{(n)}$. The equations (4.8) and (4.9) that describe how the parameters of the Hamiltonian change when we remove some degrees of freedom are called renormalization group equations.

Let us analyze the renormalization group equations a little more carefully. First of all, we notice that if the original magnetic field h vanishes, so does the renormalized field h': no magnetic field is spontaneously generated in the rescaling process. This has to be the case since a zero magnetic field implies that there is no preferred direction for the spins (either up or down); if there is no preferred direction in the small length scale description, there cannot be one in the large length scale description either. Secondly, if the external magnetic field vanishes, we have $K' = \frac{1}{2}\log\cosh[2K]$. If $K' < K$, the interaction between the coarse-grained spins is weaker than between the adjacent spins. This means that if two spins are very far apart, interaction between them is very weak, and their directions are not correlated. If on the other hand $K' > K$, the interactions between coarse-grained spins is stronger than between adjacent spins and their directions are strongly correlated. In the intermediate case $K' = K$ the interaction strength does not depend on the separation between spins. The equation $K' = \frac{1}{2}\log\cosh[2K] = K$ has two solutions, $K = 0$ and $K = \infty$ which correspond to the intermediate case when the effective coupling is independent of the distance between the spins. Since $K = \beta J$, this implies that the interaction strength is scale independent if $T = 0$ or $T = \infty$. In the first case interactions are strong, and the system is ordered, whereas in the second (infinite temperature) case interactions are very weak and the system is disordered. In this particular model we have $K' < K$ for all $0 < K < \infty$, so unless the temperature is exactly zero, the interactions get weaker with coarse-graining and the system is disordered at sufficiently large length scales. Hence, the one-dimensional Ising model is disordered at all non-zero temperatures, meaning that the knowledge of spin orientation at one point does not allow one to determing the spin orientation far away from the initial point.[2] This result was first obtained by Ernst Ising in his doctoral thesis in 1920.

It is useful to describe the renormalization group equations graphically as I have done in Fig. 4.1. The starting points of the arrows indicate the original parameters (K, h), and the end points indicate the renormalized parameters (K', h') after one decimation procedure, i.e. on a length scale that is twice the original length scale. We see that upon successive decimations, the parameters flow towards one of four different points, depending on the original parameters (K, h): if $h > 0$, the flow is towards $(K^* = 0, h^* = +\infty)$, if $h < 0$ the parameters flow towards $(K^* = 0, h^* = -\infty)$, and if $h = 0$, they flow towards $(K^* = 0, h^* = 0)$ unless $K = \infty$ in which case the parameters remain $(K = \infty, h^* = 0)$. The flow diagram tells us that even a small magnetic field changes the system's large scale behavior qualitatively — the magnetic field becomes more and more important upon successive decimations as the remaining degrees of freedom represent more and more spins — and therefore the magnetic field is called a relevant variable. We can imagine that some other perturbations get weaker at longer length scales, in which case they are called irrelevant. The concepts of renormalization group flow and relevant and irrelevant variables will be discussed in more detail in later sections.

[2] Technically, the correlation function $\langle s_i s_j \rangle$ decays exponentially as $e^{-|i-j|a/\xi}$ where a is the lattice constant and ξ the correlation length.

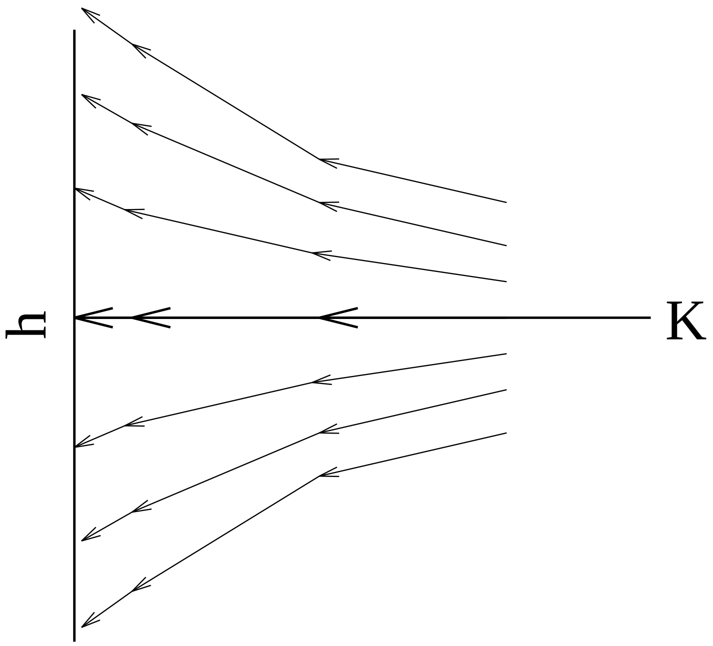

Figure 4.1: Renormalization group flows for the one-dimensional Ising chain under decimation.

General properties of a renormalization group

Let us consider a Hamiltonian $H(\{s_j\})$ which describes a system of spins $\{s_j\}$ on a d-dimensional lattice with lattice constant a. The Hamiltonian has some parameters like the external magnetic field, the coupling constant etc., which we will call $\{K_n\}$ — thus, we can have for instance

$$\mathcal{H} = -\beta H = -\beta(-J\sum_{\langle i,j\rangle} s_i s_j - h\sum_i s_i) = K_1 \sum_{\langle i,j\rangle} s_i s_j + K_2 \sum_i s_i, \tag{4.10}$$

where we multiplied by $-\beta$ for convenience. The space $\{K_n\}$ is called the parameter space of the Hamiltonian. The parameters $\vec{K} = \{K_n\}$ determine all the properties of the system. In particular, they determine the probability that the spins $\{s_i\}$ assume particular values $\{\sigma_j\}$ — this is given by $P(\{s_j = \sigma_j\}) = \frac{1}{Z}\exp[-\beta H(\{\sigma_j\})]$. They also determine the correlation length $\xi = af(\vec{K})$ where $f(\vec{K})$ is some function of the parameters. Note that the correlation length is proportional to the lattice spacing a.

An RG transformation of the Hamiltonian consists of several steps. First we form block spins, in the way that was introduced by Kadanoff, by grouping Λ^d spins together to create block spins $\{s_j^{(b)}\}$. Block spins are effective, macroscopic degrees of freedom that describe the blocks. They are chosen to be same type of objects as the microscopic degrees of freedom (spins $s_i = \pm 1$ on the microscopic lattice), and their values are defined by the microscopic degrees of freedom within the blocks they represent; typically in position space renormalization the block spin values are determined by a majority rule. The distance between adjacent block spins is Λ times larger than the distance between the original spins so that the new lattice constant is $a' = \Lambda a$. Since the block spins are functions of the original spins, the probability distribution of the block spins must be related to the probability distribution of the original spins. Writing the probability distribution for the block spins as $P(\{s_j^{(b)} = \sigma'_j\}) = \frac{1}{Z'}\exp[-\beta H'(\{\sigma'_j\})]$ we can define the block spin Hamiltonian $H'(\{s_j^{(b)}\})$. We assert that the block spin Hamiltonian H' is of the same form as the original Hamiltonian H, and only the parameters $\{K_n\}$ may change.[3] The parameters $\vec{K}' = \{K'_n\}$ of the block spin Hamiltonian are connected to the original parameters $\vec{K} = \{K_n\}$ through a transformation R_Λ,

$$\vec{K}' = R_\Lambda \vec{K}. \tag{4.12}$$

The transformation R_Λ is called a renormalization group transformation. Note that R_Λ is a mapping from one vector in the parameter space to another — if R_Λ were a linear mapping it would be a matrix, but in general it is nonlinear; equation (4.12) is nothing but a shorthand notation for $K'_j = F_j(\Lambda; K_1, K_2, \ldots, K_n)$. It is usually not possible to find R_Λ exactly, instead, we will have to resort to various approximate techniques some of which we will study in the subsequent sections. Furthermore, there is no unique way to rescale a particular problem so the transformations R_Λ are not uniquely determined by the microscopic Hamiltonian.

Equation (4.12) tells how the parameters describing the system depend on the length scale that we are considering. Each RG transformation corresponds to an additional level of coarse graining through the construction of block spins. The new, coarse-grained description ignores some short scale structure of the original microscopic description, but preserves the long wavelength properties of the model. If we successively increase the scale by a factor Λ, the parameters $\{K_n\}$ move along some trajectories in the parameter space. This motion is called the renormalization group flow.

If we perform two scale transformations R_{Λ_1} and R_{Λ_2} in succession, their combined effect is to change the scale by the factor $\Lambda = \Lambda_1\Lambda_2$; thus, the transformations R_Λ form a semigroup satisfying

$$R_{\Lambda_1\Lambda_2} = R_{\Lambda_1} R_{\Lambda_2}. \tag{4.13}$$

However, in general there is no way to reconstruct the small scale Hamiltonian if we know the system's large scale behavior and therefore the inverse transformation $(R_\Lambda)^{-1}$ does not exist — consequently, the operations R_Λ do not form a group and the term renormalization group is something of a misnomer.

Fixed points and renormalization group flow. There are usually a number of points in the parameter space that satisfy $R_\Lambda \vec{K}^* = \vec{K}^*$, *i.e.* points that are invariant under the renormalization group transformation R_Λ. These *fixed points* are particularly important: for instance, in the localization analysis in the beginning of the course, the fixed point $\beta(G_c) = 0$ could be identified as the point separating metallic and insulating behavior. To see the importance of fixed points, recall that the correlation length is given by $\xi = af(\vec{K})$ where a is the lattice constant and $f(\vec{K})$ is some function of the parameters $\{K_n\}$. Since the correlation function is a measurable quantity, it cannot depend on whether we describe the system in terms of the microscopic spins $\{s_i\}$ or block spins $\{s_j^{(b)}\}$. When we perform the RG transformation that takes us over from the microscopic description $\mathcal{H}$ to the coarse-grained

[3] This is actually no restriction since the original Hamiltonian can include all kinds of terms, for instance we could write

$$\mathcal{H} = K_1 \sum_{\langle i,j\rangle} s_i s_j + K_2 \sum_i s_i + K_3 \sum_{i,j \text{ n.n.n}} s_i s_j + K_4 \sum_{\langle i,j,k,l\rangle} s_i s_j s_k s_l \tag{4.11}$$

where the last two terms run over pairs of next-nearest neighbors and and groups of four neighboring spins, respectively. If the last two terms are not present on the microscopic level we just set $K_3 = K_4 = 0$.

description $\mathcal{H}'$, the parameters change from $\vec{K}$ to $R_\Lambda \vec{K}$, and the lattice constant changes from a to Λa. Thus, we have the result

$$\xi = af(\vec{K}) = \Lambda a f(R_\Lambda \vec{K}). \tag{4.14}$$

Now, if $\vec{K}$ is a fixed point $\vec{K}^*$ then $R_\Lambda \vec{K}^* = \vec{K}^*$ and we have $\xi^* = af(\vec{K}^*) = \Lambda a f(\vec{K}^*)$. This equation has only two solutions: either $\xi^* = 0$, or $\xi^* = \infty$. The first case corresponds to a completely uncorrelated phase whereas the second case is a critical point: the correlation length is divergent. Thus, all points $\vec{K}^*$ in the parameter space that are invariant under RG transformations correspond either to an uncorrelated phase, or to a critical point. The former are called trivial fixed points and the latter are called critical fixed points.

Each critical fixed point has its basin of attraction which consists of those points in the parameter space that flow towards the fixed point, that is, of points $\vec{K}$ such that $\lim_{\Lambda\to\infty} R_\Lambda \vec{K} = \vec{K}^*$.[4] This basin of attraction is called the critical manifold. Let us consider a point $\vec{K}$ on the critical manifold, and perform successive RG transformations R_Λ so that $\vec{K}^{(n)} = (R_\Lambda)^n \vec{K}$. Since ξ is independent of n we have $\xi = af(\vec{K}) = \Lambda a f(\vec{K}^{(1)}) = \ldots = \Lambda^n a f(\vec{K}^{(n)}) = \ldots$. Since $\Lambda > 1$ and $f(\vec{K}^*) = \infty$ the right hand side diverges as $n \to \infty$, and therefore $\xi = \infty$. Thus, the correlation length is divergent for all parameter values $\vec{K}$ that lie in the basin of attraction of a critical fixed point, which justifies the term critical manifold.

Thus, we have determined that the RG analysis allows us, by determining the fixed points of the RG transformation, to determine the critical points and critical manifolds. We are often interested not only in the fixed points that descrbe the different macroscopic behaviors of the system but also how the properties of actual systems approach the fixed points as the system size is increased. This is described by the so-called critical, or scaling exponents. The fixed point and scaling exponents are the main outcomes of an RG analysis. At the fixed points the system possesses a new type of symmetry — scale invariance — that is exact at the fixed point, but near a fixed point it is only approximate. We knew from experience that symmetries often lead to simplifications in physical problems [5]. In the example of quantum mechanics, identifying rotational symmetry allows us to organize and understand atomic spectra in great detail. So far RG falls short of such great expectations. This suggests that we have not yet fully utilized the potential of the symmetry in the present problem. In particular, we have not found any "quantum numbers". Recalling that angular momentum quantum numbers appear in quantum mechanics as eigenvalues of rotation operators suggests that if we wish to find some counterpart of quantum numbers in the present problem, we should analyze the eigenvalue problem for the symmetry operations.

Since the symmetry in the present case is only approximate, and valid only near criticality when the correlation length is large, we proceed by analyzing the RG equation near the fixed points. Writing $\vec{K} = \vec{K}^* + \delta\vec{K}$ and applying the transformation R_Λ we get

$$\vec{K}' = \vec{K}^* + \delta\vec{K}' = R_\Lambda(\vec{K}^* + \delta\vec{K}) \approx \vec{K}^* + \left(\frac{\partial R_\Lambda}{\partial \vec{K}}\right)_{\vec{K}^*} \delta\vec{K} \tag{4.15}$$

[4] Sometimes the basin of attraction of a fixed point only contains the point itself, which is then known as a repulsive fixed point.

[5] Scale invariance implies that the equations describing critical phenomena are invariant under global scale transformations (stretching). It is believed nowadays that the invariance extends to *local* scale transformations as well (spatially varying stretching). This more general *conformal invariance* leads to great simplifications in particular for two-dimensional models.

where we linearized the RG transformation near the fixed point. In component notation we now have linearized RG equations

$$\delta K'_m = \sum_n \left(\frac{\partial K'_m}{\partial K_n}\right)_{K^*} \delta K_n \tag{4.16}$$

and we can define the corresponding eigenvalue problem. However, since we do not know anything about the matrix $\hat{M}^{\Lambda}_{mn} = \left(\frac{\partial K'_m}{\partial K_n}\right)_{K^*}$ (except that it is real), it is not clear that the matrix is diagonalizable or that its eigenvalues are real.

For simplicity we assume that the matrix is diagonalizable. We can then write down the eigenvalue problem

$$\sum_n \hat{M}^{\Lambda}_{mn} e^n_\sigma = \lambda_{\Lambda,\sigma} e^m_\sigma \tag{4.17}$$

where e^n_σ is the n^{th} component of the σ^{th} eigenvector and $\lambda_{\Lambda,\sigma}$ is the corresponding eigenvalue. Since two successive RG transformations are equivalent to one RG transformation with a larger rescaling factor, $R_{\Lambda_1} R_{\Lambda_2} = R_{\Lambda_1 \Lambda_2}$, the matrices $\hat{M}$ satisfy $\hat{M}^{\Lambda_1} \hat{M}^{\Lambda_2} = \hat{M}^{\Lambda_1 \Lambda_2}$ and therefore

$$\lambda_{\Lambda_1,\sigma} \lambda_{\Lambda_2,\sigma} = \lambda_{\Lambda_1 \Lambda_2,\sigma}. \tag{4.18}$$

This implies $\lambda_{\Lambda,\sigma} = \Lambda^{y_\sigma}$ where y_σ is a number that depends on σ but not on Λ. The usefulness of the eigenvalue analysis becomes apparent if we write in (4.16) $\delta\vec{K} = \sum_\sigma a_\sigma \vec{e}_\sigma$ so that $\delta\vec{K}' = \sum_\sigma a_\sigma \lambda_{\Lambda,\sigma} \vec{e}_\sigma$. We see now that if $|\lambda_{\Lambda,\sigma}| < 1$ and $a_\sigma \neq 0$ the renormalized parameters K' are closer to the fixed point that the original parameters. Therefore the critical manifold is spanned by those eigenvectors whose eigenvalue has a modulus less than one.

We can now be more specific about the definition of relevant, irrelevant, and marginal directions in the parameter space. The direction $\vec{e}_\sigma$ is called

- relevant, if $|\lambda_{\Lambda,\sigma}| > 1$ or $y_\sigma > 0$
- irrelevant, if $|\lambda_{\Lambda,\sigma}| < 1$ or $y_\sigma < 0$
- marginal, if $|\lambda_{\Lambda,\sigma}| = 1$ or $y_\sigma = 0$

Thus, the critical manifold is spanned by the directions that are irrelevant near the critical fixed point (hence the terminology: it does not matter if the system is displaced from a critical point to an irrelevant direction, upon RG transformations it will flow towards the critical point and reach it in the infinitely large scale limit). Relevant directions, in contrast, are directions that take the system away from the critical fixed point. Marginal directions require more careful analysis and often lead to logarithmic corrections in various quantities.

Position-Space Renormalization

In the previous discussion of the renormalization group there was one crucial step that we did not address in any detail: once we have $\mathcal{H} = -\beta H$, how do we get $\mathcal{H}' = -\beta H'$? If we can obtain $\mathcal{H}'$, it is (at least in principle) straightforward to find the fixed points, identify stable phases, linearize the transformations near critical fixed points, and obtain critical exponents. In this section we will discuss one approximate scheme to obtain the renormalized Hamiltonian for a set of lattice models. The method is called position-space renormalization. It is best

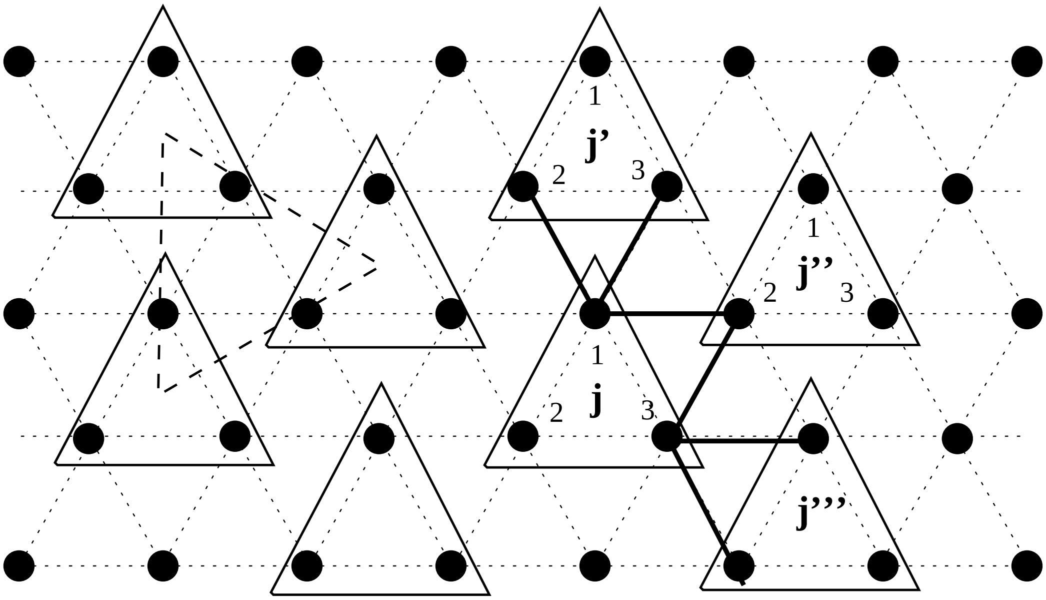

Figure 4.2: Triangular lattice and the block spin construction. The blocks j, j', j'', and j''' have been named and interactions between them are indicated by solid lines.

illustrated with an example, and the easiest example to deal with is once again the Ising model. This time we will study the model in two dimensions on a triangular lattice; the advantage of concentrating on this model is that it was solved exactly by the Norwegian physicist Lars Onsager in 1944, which allows us to determine the accuracy of the approximations we will make.

The triangular lattice is shown in Figure 4.2. We start by constructing block spins $s_j^{(b)}$ from three spins s_{j1}, s_{j2} and s_{j3} that lie in the corners of a triangle j. The natural way to assign a value to the block spin is to use *majority rule*: if the majority of the spins s_{j1}, s_{j2} and s_{j3} are $+1$, the block spin $s_j^{(b)} = +1$, otherwise $s_j^{(b)} = -1$. The majority rule can be generalized to all position-space renormalization problems, although if the number of spins in the block is even we must come up with a special rule for the case when equally many spins in the block are $+1$ and -1. Note that the block spins also form a triangular lattice but the lattice constant has increased from a to $\sqrt{3}a$, hence, the RG parameter is $\Lambda = \sqrt{3}$ in this case.

The second step is to rewrite the original Hamiltonian

$$\mathcal{H} = -\beta H = K \sum_{\langle ij \rangle} s_i s_j + \beta h \sum_i s_i \tag{4.19}$$

as a sum of two terms $\mathcal{H}_0 + \mathcal{H}_1$ so that the term $\mathcal{H}_0$ does not couple different blocks. This gives

$$\begin{aligned} \mathcal{H}_0 &= \sum_j [K(s_{j1}s_{j2} + s_{j2}s_{j3} + s_{j3}s_{j1}) + \beta h(s_{j1} + s_{j2} + s_{j3})] && (4.20) \\ \mathcal{H}_1 &= \sum_j K(s_{j1}s_{j'2} + s_{j1}s_{j'3} + s_{j1}s_{j''2} + s_{j3}s_{j''2} + s_{j3}s_{j'''1} + s_{j3}s_{j'''2}) && (4.21) \end{aligned}$$

where j' is the block right above block j and j'' is the block that is slightly above and to the right of block j, see Fig. 4.2. Since j runs over all blocks, including the interactions between the block j and its neighbors to the left or below would result in double counting in the expression for $\mathcal{H}_1$. The probability for a particular block spin configuration is given by the sum of the probabilities of those microscopic spin configurations that give rise to the right block spins. The coarse-grained Hamiltonian is thus given by

$$e^{\mathcal{H}'(\{s_j^{(b)}\})} = \sum_{\{s_i\}}{}' e^{\mathcal{H}_0(\{s_i\})+\mathcal{H}_1(\{s_i\})} \tag{4.22}$$

where the sum runs over configurations $\{s_i\}$ such that $s_j^{(b)} = \text{sign}(s_{j1}+s_{j2}+s_{j3})$ for all blocks j.

The difficulty in evaluating this expression comes from the term $e^{\mathcal{H}_1(\{s_i\})}$ which couples neighboring blocks. We will treat the difficult term perturbatively. To do that, we first define the average of the observable A with respect to $\mathcal{H}_0$ as

$$\langle A(\{s_j^{(b)}\})\rangle_0 = \frac{\sum'_{\{s_i\}} e^{\mathcal{H}_0(\{s_i\})} A(\{s_i\})}{\sum'_{\{s_i\}} e^{\mathcal{H}_0(\{s_i\})}}. \tag{4.23}$$

The average is a function of the block spins $\{s_j^{(b)}\}$ because the sums are restricted to correspond to a particular block spin configuration. With this definition the equation for $\mathcal{H}'$ can be written as

$$e^{\mathcal{H}'(\{s_j^{(b)}\})} = \sum_{\{s_i\}}{}' e^{\mathcal{H}_0(\{s_i\})} \langle e^{\mathcal{H}_1(\{s_j^{(b)}\})}\rangle_0. \tag{4.24}$$

The first factor in this expression is easy to evaluate since it contains terms only within one block. Therefore,

$$e^{\mathcal{H}_0(\{s_i\})} = \prod_j e^{[K(s_{j1}s_{j2}+s_{j2}s_{j3}+s_{j3}s_{j1})+\beta h(s_{j1}+s_{j2}+s_{j3})]} \tag{4.25}$$

and

$$Z_0(\{s_j^{(b)}\}) = \sum_{\{s_i\}}{}' e^{\mathcal{H}_0(\{s_i\})} =$$
$$\prod_j \sum_{\substack{s_{j1}=\pm 1\\ s_{j2}=\pm 1\\ s_{j3}=\pm 1}} \delta[\text{sign}(s_{j1}+s_{j2}+s_{j3}), s_j^{(b)}] e^{[K(s_{j1}s_{j2}+s_{j2}s_{j3}+s_{j3}s_{j1})+\beta h(s_{j1}+s_{j2}+s_{j3})]} \tag{4.26}$$

where $\delta(i,j)$ is the Kronecker delta. Here $Z_0(\{s_j^{(b)}\}) = \prod_j z_0(s_j^{(b)})$ is defined as an analog to the partition function. The sum is easiest to evaluate if we tabulate all the possible combinations of s_{j1}, s_{j2} and s_{j3} and the corresponding $s_j^{(b)}$. This is done in Table 4.2. From the table we can read

$$z_0(s_j^{(b)}) = \exp[3K + 3s_j^{(b)}\beta h] + 3\exp[-K + s_j^{(b)}\beta h]. \tag{4.27}$$

For simplicity we will now concentrate on the case of no external magnetic field so that $z_0(s_j^{(b)}) = \exp[3K] + 3\exp[-K]$ is independent of $s_j^{(b)}$ and $Z_0 = [\exp(3K) + 3\exp(-K)]^{N/3}$ where $N/3$ is the number of blocks (N is the number of spins).

s_{j1}	s_{j2}	s_{j3}	$s_j^{(b)}$	$\exp[\mathcal{H}_0]$
+1	+1	+1	+1	$\exp(3K+3\beta h)$
+1	+1	-1	+1	$\exp(-K+\beta h)$
+1	-1	+1	+1	$\exp(-K+\beta h)$
-1	+1	+1	+1	$\exp(-K+\beta h)$
+1	-1	-1	-1	$\exp(-K-\beta h)$
-1	+1	-1	-1	$\exp(-K-\beta h)$
-1	-1	+1	-1	$\exp(-K-\beta h)$
-1	-1	-1	-1	$\exp(3K-3\beta h)$

Table 4.2: Evaluation of $z_0(s_j^{(b)})$.

It remains to evaluate the average $\langle e^{\mathcal{H}_1}\rangle_0$. We do that using the *cumulant expansion*: we have

$$\begin{aligned}\langle e^{\mathcal{H}_1}\rangle_0 &= 1+\langle\mathcal{H}_1\rangle_0+\tfrac{1}{2}\langle\mathcal{H}_1^2\rangle_0+\dots\\ &= \exp\left[\langle\mathcal{H}_1\rangle_0+\tfrac{1}{2}\left(\langle\mathcal{H}_1^2\rangle_0-\langle\mathcal{H}_1\rangle_0^2\right)+\dots\right].\end{aligned}\tag{4.28}$$

The last line is called a cumulant expansion, and the first term in the exponent is called the first cumulant, the second one is the second cumulant *etc.*. The cumulant expansion is frequently much more accurate than the simple Taylor expansion. We will be satisfied with the first order cumulant expansion so we only need to evaluate $\langle\mathcal{H}_1\rangle_0$. The only operators that appear in $\mathcal{H}_1$ are products of two spins, and therefore we must evaluate terms like $\langle s_{j1}s_{j'2}\rangle_0$. Since the average is performed relative to $\mathcal{H}_0$ which does not couple different blocks, we have simply $\langle s_{j1}s_{j'2}\rangle_0=\langle s_{j1}\rangle_0\langle s_{j'2}\rangle_0$. The averages $\langle s_{j1}\rangle_0$ can be read from Table 4.2, which gives (for $h=0$)

$$\langle s_{j1}\rangle_0=\frac{1}{z_0}s_j^{(b)}\left[1\times\exp(3K)+(1+1-1)\times\exp(-K)\right]=s_j^{(b)}\frac{e^{3K}+e^{-K}}{e^{3K}+3e^{-K}}.\tag{4.29}$$

The averages of s_{j2} and s_{j3} are obtained similarly. Inserting this into the expression for $\mathcal{H}_1$ gives

$$\langle\mathcal{H}_1(\{s_j^{(b)}\})\rangle_0=2K\sum_{\langle j,j'\rangle}s_j^{(b)}s_{j'}^{(b)}\left(\frac{e^{3K}+e^{-K}}{e^{3K}+3e^{-K}}\right)^2.\tag{4.30}$$

Here the factor two arises since in our expression for $\mathcal{H}_1$ there are six products of two spins which describe the interactions between the block j and its nearest neighbors. Thus, each pair of nearest neighbors is coupled by two interaction lines. Hence, we have

$$e^{\mathcal{H}'}=[e^{3K}+3e^{-K}]^{N/3}e^{2K\sum_{\langle j,j'\rangle}s_j^{(b)}s_{j'}^{(b)}\left(\frac{e^{3K}+e^{-K}}{e^{3K}+3e^{-K}}\right)^2}.\tag{4.31}$$

Taking the logarithms gives

$$\mathcal{H}'=\frac{N}{3}\log[e^{3K}+3e^{-K}]+2K\left(\frac{e^{3K}+e^{-K}}{e^{3K}+3e^{-K}}\right)^2\sum_{\langle j,j'\rangle}s_j^{(b)}s_{j'}^{(b)}.\tag{4.32}$$

The first term is an additive constant which we are not interested in, and the second term looks like an Ising coupling between the block spins. This is exactly what we hoped to get, and now we just read off the renormalized coupling constant

$$K' = 2K\left(\frac{e^{3K}+e^{-K}}{e^{3K}+3e^{-K}}\right)^2 \tag{4.33}$$

which is our RG equation.

The fixed point K^* is given by

$$\frac{1}{\sqrt{2}} = \frac{e^{3K^*}+e^{-K^*}}{e^{3K^*}+3e^{-K^*}} \tag{4.34}$$

which gives $e^{4K^*} = 2\sqrt{2}+1$ or $K^* = \frac{1}{4}\log[2\sqrt{2}+1] \approx 0.336$ or $T_c \approx 2.98J/k_B$. The exact result is $K_c = \frac{1}{4}\log[3] \approx 0.275$ so we are off by about 20%. To calculate the critical exponents we must differentiate the expression (4.33) with respect to K. This is most conveniently done with Mathematica or some other symbolic manipulation software, which gives

$$\frac{dK'}{dK} = 2\frac{(1+e^{4K})(3+4e^{4K}+e^{8K}+16e^{4K}K)}{(3+e^{4K})^3} \tag{4.35}$$

which yields at $K = K^*$

$$\lambda_t = \frac{dK'}{dK} = 2(1+\sqrt{2})\frac{4+3\sqrt{2}+\log(1+2\sqrt{2})}{(2+\sqrt{2})^3} \approx 1.62352. \tag{4.36}$$

Recalling $\Lambda = \sqrt{3}$ and $\lambda_t = \Lambda^{y_t}$ gives $y_t \approx 0.882203$; the exact result is $y_t = 1$ so that the exponent is off by 12%.

An obvious way to improve the analysis is to include the next term in the cumulant expansion. That, however, gets rather complicated since the second order cumulant introduces interactions between next-nearest neighbor blocks and third-nearest neighbor blocks. The analysis can nevertheless be carried through and the result is $K_c \approx 0.2575$ and $y_t \approx 1.042$, which is a significant improvement over the first order result. However, conceptually the possibility of improving the calculation systematically is a major improvement over the previous theories: we now have a tool, position space renormalization, that allows us to analyze large scale manifestations of small scale interactions. On the level that we have discussed the tool is still rather rough and actual calculations can get rather involved. Most practical applications of the renormalization group method are carried out in wavevector space rather than in position space, but the general idea is the same: careful elimination of those degrees of freedom that correspond to small scale structure (*i.e.* short length scales or large wavevectors).

Home problem 2: Position-space renormalization: 2D Ising model — Consider an Ising model on a two-dimensional square lattice (zero external magnetic field). Form block spins by grouping together four spins of the original lattice — this will lead to a difficulty in treating the case when two of the original spins are up and two down (the majority voting rule ends up in a tie). Distribute the Boltzmann weights of these contributions equally among the $s^{(b)} = +1$ and $s^{(b)} = -1$ cases (i.e. put a factor of $\frac{1}{2}$ in front of the Boltzmann weights of the tied configurations and include them in both block spin states). Apply the position-space renormalization group techniques to determine the critical temperature and the critical exponents α and ν. Compare the PSRG results with the exact ones.

4.1.2 Momentum space renormalization group

HJ

4.2 Approximate analytic techniques

4.2.1 Mean field theory

HJ

4.2.2 Dynamic mean field theory

HJ

4.3 Exact analytic techniques

4.3.1 Bethe Ansatz

HJ

4.3.2 Bosonization

HJ

4.4 Numerical techniques

4.4.1 Density functional theory

In materials science, and increasingly in chemistry, one of the most successful computational strategies in describing systems comprising a large number of interacting particles is known as density functional theory, DFT. The origins of DFT are quite old, in the 1920s, but it was developed into a complete tool by Walter Kohn and co-workers only in the 1960s.

The basic starting point of density functional theory is the discovery by Kohn and Pierre Hohenberg that the ground state energy of an interacting many-particle system is in a 1-to-1 correspondence with the density of the system. Hence, if we somehow manage to find out the bijection $n(\mathbf{r}) \leftrightarrow E_0$, and can determine the ground state density, we can obtain the ground state energy.

Some parts of the ground state energy can be easily related to the density — for instance an external potential $V_{\text{ext}}(\mathbf{r})$ results in the energy contribution $\int d^3r\, V_{\text{ext}}(\mathbf{r})n(\mathbf{r})$ — while for others the connection is more complicated. Kohn, together with Lu Sham, devised a scheme that works out quite well. The idea is that the ground state energy is written as

$$E_0 = T_0[n] + \int d^3r\, V_{\text{ext}}(\mathbf{r})n(\mathbf{r}) + \frac{1}{2}\int\int d^3r d^3r'\, n(\mathbf{r})U(\mathbf{r}-\mathbf{r}')n(\mathbf{r}') + E_{xc}[n]$$

where the first term is the kinetic energy of a non-interacting system with density $n(\mathbf{r})$, the second term is the energy associated with the external potential, the third term is the classical (Hartree, or direct) interaction energy arising from an interaction potential $U(\mathbf{r})$, and the final term is whatever is needed to make the equation valid. Obviously, all problems have been placed in evaluating the final term, known as the exchange-correlation energy. The usefulness of this approach arises from the fact the last term is usually quite small. Let us postpone the discussion of the exchange-correlation energy for a little while and consider the other terms first.

Another term that leads to difficulties is the kinetic energy which is not readily expressible in terms of the density. To address this problem, Kohn and Sham wrote the density in the form $n(\mathbf{r}) = \sum_\alpha |\psi_\alpha(\mathbf{r})|^2$ using single-particle wave functions $\psi_\alpha(\mathbf{r})$. By requiring that the energy is at its minimum, one can derive a set of (Euler-Lagrange) equations for the wave functions ψ_α, which typically assume the form

$$-\frac{\hbar^2}{2m}\psi_\alpha(\mathbf{r}) + V_{\rm ext}(\mathbf{r})\psi_\alpha(\mathbf{r}) + \int d^3r' U(\mathbf{r}-\mathbf{r}')n(\mathbf{r}')\psi_\alpha(\mathbf{r}) + \frac{\delta E_{xc}}{\delta n(\mathbf{r})}\psi_\alpha(\mathbf{r}) = \epsilon_\alpha\psi_\alpha(\mathbf{r}) \tag{4.37}$$

which resembles the Schrödinger equation for a single particle. Here ϵ_α is technically a Lagrange multiplier and, technically, $\psi_\alpha(\mathbf{r})$ is just a calculational tool to construct a density. In practice, however, one often regards both ϵ_α and ψ_α as physical quantities — the energy and wave function of a meaningful single-particle state. In an interacting many-particle system this interpretation cannot be justified, but experience has shown that it has more value than can be rigorously proven.[6]

Now the many-particle problem has been reduced to a set of single-particle equations for the functions $\psi_\alpha(\mathbf{r})$. However, since the left hand side of the Kohn-Sham equation 4.37 depends on the density $n(\mathbf{r})$ which depends on the functions ψ_α, the equations must be solved iteratively, and at each cycle the density $n(\mathbf{r})$ must be constructed from the functions $\psi_\alpha(\mathbf{r})$ whose Lagrange multipliers ϵ_α are lowest. For spinless fermions, the number of functions used equals the number of particles, while for fermions with spin, the number of functions used equals $[(N-1)/2]+1$ with $[N/2]$ functions with smallest ϵ_α contributing to the density by $2|\psi_\alpha(\mathbf{r})|^2$.

There are many numerical implementations of how to solve the Kohn-Sham equations, and we shall not discuss them here; several of the implementations have been commercialized. Typically, the solution time of a set of Kohn-Sham equations increases as N^p with $p \approx 2-3$ but there is some hope to improve the scaling to $p = 1$ by exploiting the fact that, typically, the particles are near-sighted: their interactions with particles far away can be described in an average fashion, and only the nearby particles are treated by pairwise interactions. Presently, the number of electrons that can be reasonably described using DFT ranges from a few hundred on a PC to a few thousand on a more powerful computer.

The Achilles' heel of DFT is the fact that we do not know how the exchange-correlation energy depends on the density. There are many approaches to determine this dependence. Historically the first was the Thomas-Fermi approximation, which is based on treating a uniform electron system in the Hartree-Fock approximation: this results, in three dimensions, in an exchange (Fock) contribution to the total energy that can be written as $-V\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\frac{e^2}{4\pi\epsilon_0}n^{4/3}$. Assuming that the exchange-correlation energy of a non-uniform electron system can be written as an integral over space with an integrand that only depends on the local density results in the approximation

$$E_{xc} = -\int d^3r \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\frac{e^2}{4\pi\epsilon_0}n(\mathbf{r})^{4/3}.$$

[6] It turns out that the Lagrange multiplier for the highest occupied level can be rigorously interpreted as an ionization energy (Koopman's theorem).

This is quite a rough approximation since it only includes exchange, and assumes that the local density is the only relevant degree of freedom. The first aspect can be resolved by doing higher order perturbation theory on the uniform electron system — typically, in a more accurate formulation, one sums certain perturbative contributions to infinite order in the perturbation theory (*e.g.* random phase approximation), and thereby obtains a more accurate description of a uniform system. This *local density approximation*, LDA, form of the exchange-correlation energy is usually written as

$$E_{xc}^{LDA} = \int d^3 r n(\mathbf{r}) \epsilon_{xc}^{LDA}(\mathbf{r})$$

where $\epsilon_{xc}^{LDA}(\mathbf{r})$ is the exchange-correlation energy of an electron in a uniform electron gas with density $n(\mathbf{r})$; in the Kohn-Sham equations this results in a potential term $v_{xc}^{LDA} = \frac{\delta E_{xc}^{LDA}}{\delta n(\mathbf{r})} = \frac{\partial n(\mathbf{r})\epsilon(\mathbf{r})}{\partial n(\mathbf{r})}$. Extending the discussion to non-uniform charge densities is harder, and the first attempts in this direction actually gave results that were inferior to those obtained in the local density approximation. The reason of the failure of these so-called gradient corrections was identified as having to do with certain exact constraints being violated — simple gradient corrections describe a system that is fundamentally non-physical — and improved scheme known generalized gradient approximation (GGA) were introduced. The GGA has proven quite successful.

Apart from perturbative methods, one can find approximations for the exchange-correlation energy by using other numerical techniques in connection with analytic constraints. The analytic constraints typically require specific limits (form at low or high density, positive compressibility or pressure *etc.*), and the numerical methods are usually some techniques that allow high accuracy for small systems such as exact diagonalization or Quantum Monte Carlo. Even available experimental results may be used to constrain the form for the exchange-correlation energy.

The density functional theory was developed for the purposes of materials science where it has been used successfully to obtain electronic structures of large varieties of materials and individual molecules. It forms, together with psudopotential methods that describe the most energetic (valence) electrons of an atom, the backbone of computational materials science. The technique has also been used to discuss the properties of interacting electron systems such as quantum dots where it has revealed intricate correlated states.

A limitation of DFT is that in its original form it only addresses the ground state properties, and in most systems in condensed matter, the ground state is inert and completely uninteresting. The interesting physics is associated with the excitations above the ground state: they determine the system's dynamics and response to external stimuli. Also, in its original form DFT has difficulty dealing with systems with large degeneracy such as the quantum Hall effect, or systems where spin plays a crucial role, or where time dependence is of fundamental importance (*e.g.* most non-equilibrium systems), or many other cases.

However, DFT has been extended to cover many of these applications: in the presence of magnetic field one has to use current density functional theory (CDFT) with $E_{xc}[n, \mathbf{j}]$; in the case of spin, spin-DFT (SDFT, $E_{xc}[n_\uparrow, n_\downarrow]$) is used; to describe dynamics time-dependent DFT is used (TDDFT, $E_{xc}[n]$ used with a time-dependent Kohn-Sham equations); if degeneracy is important, one can use ensemble DFT *etc.* — in the worst case, you may need time-dependent ensemble spin current DFT, or its extension to transport in non-equilibrium problems. All these extensions come at their cost in terms of complications and, sometimes, in terms of having to relax the rigor of the Kohn-Hohenberg *theorem* to some weaker form of assumed truth. They also come with limitations: for instance, the time-dependent DFT runs out of steam after a few femtoseconds, which makes it purely suited for most applications that involve macroscopic timescales. Lately much effort has been invested in combining the power of DFT with description of transport in different systems, and while it is generally believed that this is a reasonable direction to aim at, there is no real consensus of the merit of the implementations used thus far.

The extensions result even in practical difficulties in addition to the formal ones. In the case of the spin-DFT, for instance, the dependence of the exchange-correlation energy on the two spin densities cannot be deduced from the spin-polarized version $E_{xc}[n(\mathbf{r})]$: the exchange term acts only between electrons with the same spin, $E_x[n_\uparrow, n_\downarrow] = E_x[n_\uparrow] + E_x[n_\downarrow]$ but the correlation part does not separate in the same way. Typically one resolves this problem by making the Ansatz that the exchange-correlation energy depends on the total density $n = n_\uparrow + n_\downarrow$ and the polarization $\xi = (n_\uparrow - n_\downarrow)/n$ and invokes the interpolation formula $\epsilon_{xc}(n,\xi) = \epsilon_{xc}^{\mathrm{pol}}(n) + f(\xi)[\epsilon_{xc}^{\mathrm{unpol}}(n) - \epsilon_{xc}^{\mathrm{pol}}(n)]$ where $\epsilon_{xc}^{\mathrm{pol}}$ and $\epsilon_{xc}^{\mathrm{unpol}}$) are the exchange-correlation energy densities for fully polarized and fully unpolarized systems, respectively, and $f(\xi)$ is an interpolation form that yields the correct polarization dependence for the dominant exchange part. The exchange-correlation energy densities for $\xi = 0$ and $\xi = 1$ must be extracted using any of the standard methods. This method has been applied quite successfully to, *e.g.*, the determination of ground state electron structures of semiconductor quantum dots. However, from a fundamental point of view it is not as firmly established as the stadard DFT, and one can actually construct cases when there is no one-to-one correspondence between the spin densities and the ground state energy (K. Capelle and G. Vignale, Phys. Rev. Lett. **86**, 5546 (2001)) thereby violating the fundamental assumption behind SDFT; fortunately, these examples appear to be of little practical consequence.

For the case of the current-DFT the current density must be introduced as an extra parameter; for systems with time reversal invariance the current density must vanish in the ground state but in cases when the time reversal symmetry is broken by an external magnetic field the ground state may well carry a current as we saw in the persistent current discussion. In the same way as the density dependence of E_{xc} results in an extra potential v_{xc}^{LDA} in the Kohn-Sham equations, the current dependence results in an extra vector potential A_{xc}^{LCDA}. The current density dependence of E_{xc} is restricted by gauge invariance, and it turns out that the exchange-correlation energy can only depend on current densities through the vorticity $\nabla \times \mathbf{j}(\mathbf{r})/n(\mathbf{r})$. While approximations for the exchange-correlation vector potential can be derived in the same as those for the exchange-correlation potential, one usually employs a mapping that allows A_{xc}^{LCDA} to be related to v_{xc}^{LDA} of a uniform system in a fictitious magnetic field whose strength is related to the current density of the original system. The CDFT method has also been applied to study the electronic structure of variety of nanoscale systems in external magnetic fields, with results that are in good agreement with those obtained by other methods. Yet, similar non-uniqueness concerns apply to CDFT as to SDFT.

Despite the complications and shortcomings listed above, even rudimentary applications of DFT often yield surprisingly good results for interacting electron systems. As an example we consider analysis of Coulomb blockade in a two-dimensional semiconducting quantum dot in the presence of a perpendicular magnetic field. This system has been studied extensively both experimentally and theoretically since the late 1980s. The experimental structure typically comprises a two-dimensional electron system (*e.g.* at the interface between GaAs and AlGaAs)

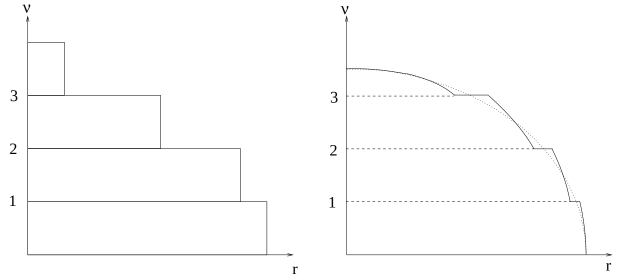

Figure 4.3: Electron densities for a parabolic quantum dot as predicted by the simplest capacitive model ("cake model) and the simplest DFT model.

and some way to deplete the electron system except in a small region that forms the quantum dot. The depletion can be described by an external potential which typically assumes a parabolic form $V(\mathbf{r}) = \frac{1}{2}m\omega_0^2 r^2$ where the potentials curvature ω_0 depends on the design; typical values are $\hbar\omega_0 \approx$ 1-2 meV. The system's leading behavior is similar to what we saw in our analysis of Coulomb blockade in metallic quantum dots: the conductance through the dot is in general small, except at a series of gate voltage values at which two charge states are degenerate, and conductance has a maximum. These conductance maxima are equally spaced on the gate voltage axis, and the peak spacing is related to the gate capacitance by $\Delta V_g = e/C_g$. This simple picture suggests that magnetic field plays no role — the capacitances are simple geometric quantities, unaffected by a magnetic field.

Experimentally, however, it is seen that the positions of conductance peaks change with B. The first explanation that was suggested for this behavior was based on the Hamiltonian

$$H = \sum_\alpha \epsilon_\alpha c_\alpha^\dagger c_\alpha + \frac{Q^2}{2C}$$

where the first term is the single particle energy of the states $|\alpha\rangle$ obtained by solving the Schrödinger equation for electrons in a strong magnetic field and subject to a parabolic confining potential, and the second term is the interaction energy in the capacitance approximation. Although the capacitance is independent of B, the single-particle energies depend on the magnetic field, and consequently the gate voltage values at which charge degeneracy occurs is B-dependent.[7] The resulting electron distribution resembles a multi-layered wedding cake: the states on different Landau levels and with different spins are occupied up to some maximum angular momentum value, *i.e.* up to some radius, and within this disk they have a nearly constant density corresponding to $n_1 = (2\pi\ell_c^2)^{-1}$. While this explanation seems to be able to explain some experimental features, it requires certain unphysical assumptions such as that the electron density in the dot must be higher than that in the leads; this is profoundly unreasonable as quantum dots are created by expelling electrons from a 2DES so that only a small puddle of the electron liquid remains.

The next level of theoretical sophistication was to implement a Thomas-Fermi-type density functional description. The kinetic energy is now quite simple if the system is in a strong

[7] In metallic structures the density of levels is so high that the magnetic field has little impact — as the field is changed, the set of occupied levels changes but the total energy is rather insensitive to B. In semiconducting dots the level separation is large so that the set of occupied levels is not easily affected, and therefore the magnetic field dependence of individual energy levels is more important.

magnetic field as is the case in the experiments: the kinetic energy is quantized to Landau levels which can accommodate electrons up to a maximum density that depends on the magnetic field. In this first DFT description the exchange-correlation energy was simply set to zero, and the electron density was written as a sum of contributions from occupied single particle states as $n(\mathbf{r} = \sum_{n,\sigma} \psi^{\dagger}_{n,\sigma}(\mathbf{r})\psi_{n,\sigma}(\mathbf{r})$ where the states $|n,\sigma\rangle$ satisfy Kohn-Sham equations. Compared to the "cake model, the electron density resulting from this model is considerably smoother. It can be understood from the classical limit: the classical electron density for a two-dimensional charge distribution in a parabolic external potential is $n(\mathbf{r}) = n_0\sqrt{r^2 - R_0^2}$ where R_0 is the radius of the charge distribution.[8] Quantum effects manifest themselves in the fact that as long as $n(\mathbf{r})$ is below $n_1 = (2\pi\ell_c^2)^{-1}$, all electrons have equal kinetic and Zeeman energies, for the range $n_1 < n < 2n_1$ two spin states (with different Zeeman energies) are needed, for $2n_1 < n$ at least two Landau levels (with different kinetic energies) are needed *etc.*. Since the electrostatic energy scale dominates in the problem, deviations from the classical distribution are rather small; typically, the electron density follows the classical form except near the points when $n(\mathbf{r})$ is approximately an integer times n_1 when it is preferable from an energy point of view to transfer some charge from a higher spin or Landau state to a lower one, thereby losing some Coulomb energy but gaining Zeeman or kinetic energy. This results in the formation of (narrow) density plateaux $n(\mathbf{r}) = pn_1$, $p \in \mathbb{N}$, which are known as incompressible strips. This level of description results in an excellent agreement between theoretical and experimental results as shown in Fig. 4.4.

The next level of improvement is to include a non-zero exchange-correlation term. This modifies the results of the Thomas-Fermi theory particularly at the edges of the charge distribution where the system is compressible. A fair amount of effort has been invested in the regime where $n(\mathbf{r}) = n_1$ for a large range of radii r, which is known as the maximum density droplet as it corresponds to the maximum density consistent with the lowest spin and Landau state. When the magnetic field is increased, the lowest level could occupy even more electrons but that would cost more Coulomb energy so the system faces a dilemma: on one hand the confining potential favors compact, small electron densities, but on the other hand the electron-electron interaction favors more spread out charge densities. The energetically favorable distribution turns out to be one where the edge of the maximum density droplet undergoes a reconstruction and the compact central droplet is surrounded by circular charge density rings — the electron density no longer decreases monotonically as a function of the distance from the center of the quantum dot as shown in Fig. 4.5. The exchange-correlation effects are needed to describe these edge reconstruction effects accurately. Further refinement of the theoretical description in terms of current density functional theory predicts detailed internal structures in the charge and spin densities inside the quantum dots as shown in Fig. 4.6 (see S.M. Reimann and M. Manninen, Rev. Mod. Phys. **74**, 1283 (2002)). However, these higher correlation get increasingly more difficult to detect experimentally.

4.4.2 Quantum Monte Carlo

Monte Carlo methods are a stochastic techniques for evaluating definite integrals.[9] A simple argument shows that they are likely to be more powerful than deterministic techniques

[8] Deriving this result is referred to as *an amusing exercise* by the mathematical physicist Elliot Lieb at Princeton University. Feel welcome to check if you agree with him (it's not that hard).

[9] This section follows closely the lecture notes Matthias Troyer presented at the Boulder summer school in Colorado, USa, in 2004.

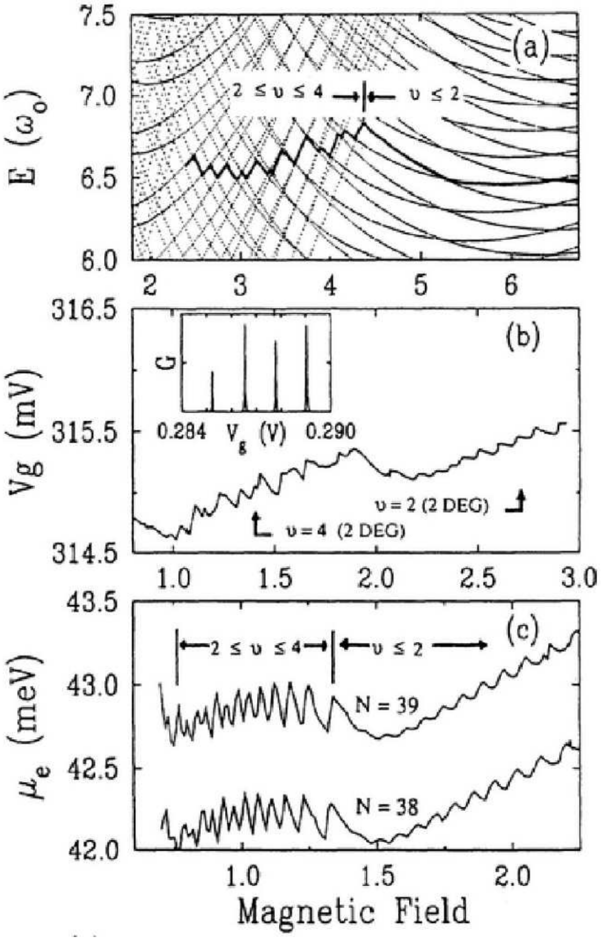

Figure 4.4: Comparison between the predictions of the "cake model" (top), experiment (middle), and Thomas-Fermi model (bottom). The positions of conductance peaks are seen to move to higher energies with an increasing magnetic field. The simple model predicts that the oscillations in the peak position disappear when only the lowest Landau level is occupied while the experiment and the more sophisticated model show a different behavior.

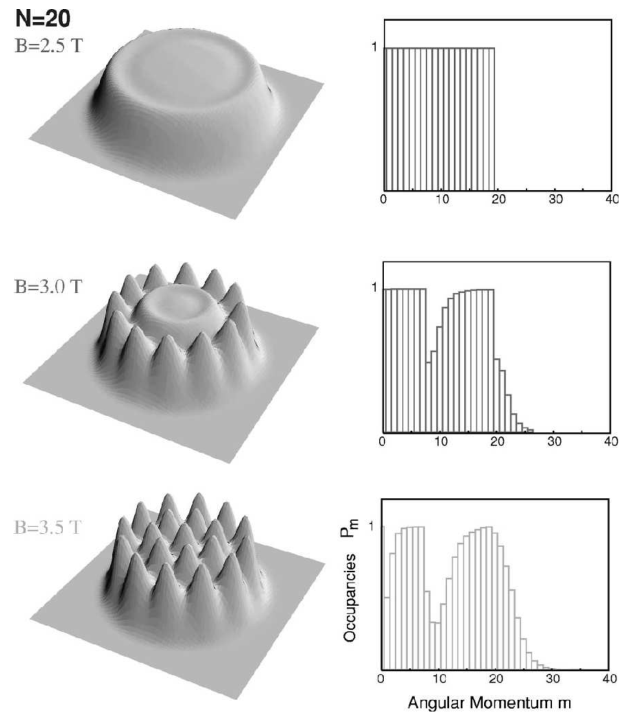

Figure 4.5: Reconstruction of the maximum density droplet edge as the magnetic field is increased, calculated by current spin density functional theory for a semiconducting quantum dot with 20 electrons — note the non-monotonic occupation of single-particle states (Kohn-Sham orbitals) in terms of their angular momentum, and the intricate internal structure at even higher magnetic fields. From S.M. Reimann and M. Manninen, Rev. Mod. Phys. **74**, 1282 (2002).

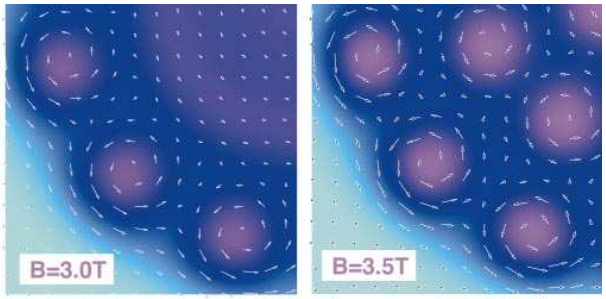

Figure 4.6: Ground states currents calculated using CSDFT for a quantum dot with 20 electrons. From S.M. Reimann and M. Manninen, *op.cit.*.

provided that the domain of integration is at least two-dimensional: the error of a stochastic method decreases typically as $N^{-1/2}$ where N is the number of sampling points, while the error of a deterministic method decreases typically as the lattice spacing a of the discretization lattice; if the N points are equally distributed in all dimensions, the number of point in any given direction is $N^{1/d}$ and the lattice spacing, and integration error, is given by $N^{-1/d}$ so that for $d = 1$ a deterministic method converges faster, for $d \geq 3$ a stochastic method is faster, and for $d = 2$ the two have comparable scaling properties. Since problems that involve many interacting degrees of freedom typically lead to multi-dimensional integrals, Monte Carlo methods seem well suited for them.[10]

In classical physics Monte Carlo methods are often used to evaluate thermal expectation values

$$\langle A \rangle = \frac{\int d^d r \, p(\mathbf{r}) A(\mathbf{r})}{\int d^d r \, P(\mathbf{r})}$$

where $P(\mathbf{r})$ is a probability distribution that in equilibrium is given by $e^{-\beta E(\mathbf{r})}$. Typically, one creates an ensemble which has the property that a point $\mathbf{r}$ is sampled with probability $P(\mathbf{r})$, which can be done by *e.g.* the well-known Metropolis algortithm where one starts from a position $\mathbf{r}_i$, suggests a move to a position $\mathbf{r}_{i+1}$, and accepts or rejects the move with a probability that corresponds to a detailed balance criterion determined by the energies of the two positions involved. This method works well on paper, and even in reality provided that the sampled points are statistically independent, that is, that they are separated by a distance that is large compared to the correlation length of the system. If the two points are not independent, the number of statistically independent samples is less than the actual number of samples, and convergence is poorer than expected. This often happens near phase transitions where the correlation length grows dramatically, and results in a phenomenon known as critical slowing down.

More generally the integral in the Monte Carlo method is a more general sum (or integral) over the degrees of freedom of the system at hand — for a model of magnetic material one typically sums over all possible orientations of the magnetics moments (spins) at all lattice sites. In this case critical slowing down is encountered near the Curie (for ferromagnets) or Neel (for antiferromagnets) temperature at which the materials becomes magnetically ordered. The reason for critical slowing down is easy to identify: near the ferromagnetic transition, the material forms domains such that magnetization is constant within a domain but varies from one domain to another. It is energetically very costly to flip one spin in a domain but flipping all spins in a domain would carry much lower a cost. This also suggests a way to fight the effects of critical slowing down: instead of updating one spin at a time, one tries to identify clusters of spins and update them all at once. Many implementations of cluster algorithms have been suggested, all of them consisting of two parts that first identify suitable clusters and then decide how the cluster spins should be updated to create the next, statistically independent sampling point. A key requirement for these algorithms is that they must obey ergodicity, that is, each configuration must be reachable with a finite number of steps from any initial configurations, and that the configurations the algorithm generates must follow the appropriate distribution function. Two of the best known cluster Monte Carlo algorithms

[10] It turns out that there is an intermediate category of integration techniques known as quasi-random algorithms based on so-called minimum discrepancy sequences that are, simple stated, irregular but fixed multi-dimensional lattices. For these minimum discrepancy methods the number of lattice points needed to achieve accuracy ϵ varies as $N \sim \epsilon^{-1}[\ln(\epsilon^{-1}]^{(d-1)/2}$, *cf.* Monte Carlo $N \sim \epsilon^{-2}$ and regular lattice $N \sim \epsilon^{-d}$. For further information see H. Woźniakowski, Bull. Am. Math. Soc. **24**, 185 (1991).

are the Wolff algorithm (U. Wolff, Phys. Rev. Lett. **62**, 361 (1989)) and the Swendsen-Wang algorithm (R.H. Swendsen and J.-S. Wang, Phys. Rev. Lett. **58**, 86 (1987)).

In quantum mechanics we have a further complication that specifying the degrees of freedom of a quantum system does not usually specify its energy. Typical example is a quantum particle in an external potential: specifying the particle's momentum allows its kinetic energy to be calculated, but makes it impossible to determine the potential energy, while specifying the particle's position results in opposite problems; simultaneous specification of both momentum and position is not allowed. This would seem to render Monte Carlo methods unusable in quantum mechanics.

Fortunately, the problem can be overcome. We will illustrate this with a spin model, the quantum mechanical Heisenberg model, with the Hamiltonian

$$H = -J \sum_{\langle i,j \rangle} \left(S_i^x S_j^x + S_i^y S_j^y + S_i^z S_j^z \right)$$

where J is a coupling constant with units of energy, $S_i^{x,y,z}$ are the (dimensionless) components of the spin at site i, and $\langle i, j\rangle$ is a set of nearest neighbors. Hence, the model describes a lattice of atoms where each atom carries a spin and spins between nearest neighbor atoms interact. The quantum nature of this model is reflected by the fact that for any atom i we can only specify one component of the spin due to the commutation relation $[S^x, S^y] = iS^z$ and its cyclic permutations — the different components of spin are like the position and momentum of a quantum particle.

The thermal average of an operator A is now given by

$$\langle A \rangle = \frac{\mathrm{Tr}\left(Ae^{-\beta H}\right)}{\mathrm{Tr}\left(e^{-\beta H}\right)}$$

where $\mathrm{TrA} = \sum_{\{S_i^z\}} \langle \{S_i^z\}|A|\{S_i^z\}\rangle$ is the trace of the operator A, *i.e.* sum over all diagonal matrix elements in some complete basis (here I chose the z-component of the spin at each site). The problem is that the matrix elements are not straightforward to calculate. The way around is to introduce a large integer M such that $\beta = M\Delta\tau$ and $e^{-\beta H} = e^{-M\Delta\tau H} = \left(e^{-\Delta\tau H}\right)^M$. For sufficiently large M the quantity in the paranthesis can be approximated as $1 - \Delta\tau H$ so that the denominator of the above expression can be written as

$$\begin{aligned} &Z = \mathrm{Tr}\left[(1 - \Delta\tau H)^M\right] = \textstyle\sum_{\{S_{0,i}^z\}} \sum_{\{S_{1,i}^z\}} \sum_{\{S_{2,i}^z\}} \cdots \sum_{\{S_{M-1,i}^z\}} \\ &\langle\{S_{0,i}^z\}|(1 - \Delta\tau H)|\{S_{1,i}^z\}\rangle\langle\{S_{1,i}^z\}|(1 - \Delta\tau H)|\{S_{1,i}^z\}\rangle\langle\{S_{2,i}^z\}|\dots(1 - \Delta\tau H)|\{S_{0,i}^z\}\rangle \end{aligned}$$

where I inserted a complete set of states ($1 = \sum_i |i\rangle\langle i|$) between each of the factors $(1-\Delta\tau H)$. For future refererence, we can write Z as $Z = \sum_{\{S_{k,i}^z\}} P(\{S_{k,i}^z\})$ where P can, hopefullly, be interpreted as the unnormalized probability of a configuration $\{S_{k,i}^z\}$. A similar expansion can be written for the numerator.

The price of this has been that the number of dimensions has effectively increased by one: the configurations $\{S_{k,i}^z\}$ have acquired a new label, k, that determines between which factors this complete set of states has been inserted. This extra dimension is known as imaginary time, for reasons we will not go into, and is a general feature of equilibrium problems: a quantum problem has effectively one more degree of freedom than the corresponding classical problem. Thus, we have the expression

$$\langle A \rangle = \frac{\sum P(\{S_{k,i}^z\})A(\{S_{k,i}^z\})}{\sum P(\{S_{k,i}^z\})}$$

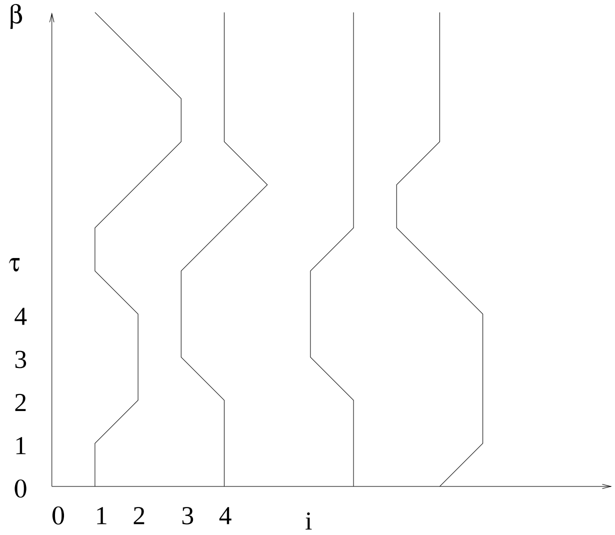

Figure 4.7: Worldlines showing the trajectories of up-spins in imaginary time. The vertical direction is given by $k\hbar\Delta\tau$ with $0 \leq k \leq M-1$ so that it extends from 0 to (nearly) β; the y-axuis is often multiplied by $\hbar$ so that the direction has units of time.

for the thermal expectation value. Here $A(\{S^z_{k,i}\})$ represents a product of the relevant matrix elements; in the ordinary case the operator A acts only at one instance of the artificially introduced imaginary time argument and $A(\{S^z_{k,i}\}) = \langle\{S^z_{0,i}\}|A|\{S^z_{1,i}\}\rangle$.

Now that the Hamiltonian no longer appears in the exponent, evaluating the matrix elements is much easier. Introducing new operators $S^+ = S^x + iS^y$ and $S^- = S_x - iS^y$ so that $[S^z, S^+] = iS^y + S^x = S^+$ which means that S^+ increases the z-component of spin by one. With this notation we can write

$$H = -J\sum_{\langle i,j\rangle}\left(S^z_i S^z_j + \frac{1}{2}(S^+_i S^-_j S^-_i S^+_j)\right)$$

which shows that the non-zero factors in the product expression for Z are either diagonal, $S^z_{k,i} = S^z_{k+1,i}$, or involve two opposite spin flips, $\left(S^z_{k,i}, S^z_{k,j}\right) \rightarrow \left(S^z_{k+1,i} \pm 1, S^z_{k+1,j} \mp 1\right)$. Hence, if one draws lines in the k-direction connecting up-spins, the lines are either vertical, or move diagonally between adjacent lattice points. These so-called world lines, shown in Fig. 4.7, form a basis for quantum cluster algorithms, usually known as loop algorithms, that introduce larger moves than single spin updates, thereby reducing the effects of critical slowing down.

The Quantum Monte Carlo method as described above works quite well for boson problems. For fermion problems, however, there is a serious difficulty: the quantities P are not positive definite, and cannot therefore be interpreted as probabilities, which makes Metropolis-type algorithms unfeasible. This is the infamous *fermion sign problem.* Formally, however, there is no problem: we just define a new (unnormalized) probability density through $|P|$ and write the expectation value as

$$\langle A\rangle = \frac{\frac{\sum |P(\{n^z_{k,i}\})|\mathrm{sign}(P(\{n^z_{k,i}\}))A(\{n^z_{k,i}\})}{\sum |P(\{n^z_{k,i}\})|}}{\frac{\sum |P(\{n^z_{k,i}\})|\mathrm{sign}(P(\{n^z_{k,i}\}))}{\sum |P(\{n^z_{k,i}\})|}}.$$

where n_i is the probability of finding a fermion at site i — we simply interpret the up-spins as full sites and the down-spins as empty sites on some discrete lattice.[11] In practice, however, the average of the sign of P can be very small — typically the expectation value of the sign varies as $e^{-\beta N}$ — which leads to numerical difficulties. Great deal of effort has been invested in overcoming the sign problem, and every now and then researchers announce that they have found the solution to the fermion sign problem, although thus far the announcers have failed to convince their colleagues. One authority in the field has promised the Nobel prize (without any authority to do so!) to whoever solves the problem, so you are welcome to try your luck.

[11] More complicated fermions problems can also be treated using QMC, this simple example of a single fermion species on a lattice serves only as an example.

Appendices

Appendix A

Coulomb Blockade at a Single Barrier

So far we have ignored the effects that the appearance or disappearance of an electron might have in the two leads. If the leads are large metallic bodies, such effects are vanishingly small — the charge disturbance is screened within a few plasma oscillations — but if the leads are narrow wires, the effects may be more significant. We now turn our attention to charge redistribution in the leads and examine how it affects tunneling across a tunnel junction.

We can model a lead as a lossless transmission line which is terminated by a capacitance C_0. The capacitance C_0 we interpret as the junction capacitance so that a tunneling event corresponds to charging the end capacitor by charge $\pm e$. For simplicity we use a discrete model for the transmission line as indicated in Figure A.1. The continuum limit can be obtained by letting the inductance L, capacitance C, and cell size a approach zero such that the specific inductance $\ell = L/a$ and specific capacitance $c = C/a$ remain constants.

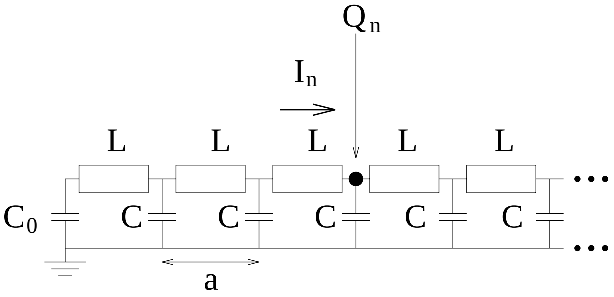

Figure A.1: Coupling of the capacitor C_0 to a (lossless) transmission line

We define variables Q_n and I_n so that $\dot{Q}_n = -I_{n+1} + I_n$. The inductive and capacitive energies are then $\frac{1}{2}L\sum_{n=1}^{\infty} I_n^2$ and $\frac{1}{2}C_0^{-1}Q_0^2 + \frac{1}{2}C^{-1}\sum_{n=1}^{\infty} Q_n^2$. The inductive energy can be identified as a kinetic term and the capacitive energy as a potential term so that the

Lagrangian is[1]

$$\mathbb{L} = \frac{1}{2}L\sum_{n=1}^{\infty} I_n^2 - \frac{1}{2}C^{-1}\sum_{n=1}^{\infty} Q_n^2 - \frac{1}{2}C_0^{-1}Q_0^2$$

For the quantum mechanical treatment we would like to have a Hamiltonian, and therefore we need the momenta conjugate to the charges Q_n. We have (note $I_n = -\sum_{k=0}^{n-1}\dot{Q}_k$)

$$P_n = \frac{\partial \mathbb{L}}{\partial \dot{Q}_n} = L\sum_{k=1}^{\infty} I_k \frac{\partial I_k}{\partial \dot{Q}_n} = -L\sum_{k=n+1}^{\infty} I_k$$

so that $I_n = L^{-1}(P_n - P_{n-1})$, $\dot{Q}_0 = -L^{-1}(P_1 - P_0)$, and $\dot{Q}_{n\neq 0} = -L^{-1}(P_{n-1} - 2P_n - P_{n+1})$. The Hamiltonian is then

$$\mathbb{H} = \sum_{n=1}^{\infty} P_n \dot{Q}_n - \mathbb{L} = \frac{1}{2L}\sum_{n=1}^{\infty}(P_n - P_{n-1})^2 + \frac{1}{2C}\sum_{n=1}^{\infty} Q_n^2 + \frac{1}{2C_0}Q_0^2$$

The task we now set for ourselves is to investigate what happens if the charge Q_0 suddenly changes due to a tunneling process — how does the transmission line react to this disturbance? We will find that the transmission line's inability to instantaneously carry away the extra charge results in a modification of the tunneling rates Γ, and consequently also changes the current-voltage characteristics of the structure.[2]

Although $\mathbb{H}$ is nothing but a collection of harmonic oscillators, exact diagonalization is cumbersome due to lack of translational invariance. Let us therefore make the approximation that the dynamics of the transmission line is not greatly affected by the difference of the terminal capacitance C_0 and the other capacitances C. The Hamiltonian with $C_0 = C$ is diagonalized by the transformation

$$\begin{aligned} Q_n &= \sqrt{\frac{2}{\pi}}\int_0^{\pi} ds\, Q(s)\cos[s(n+\frac{1}{2})] \\ P_n &= \sqrt{\frac{2}{\pi}}\int_0^{\pi} ds\, P(s)\cos[s(n+\frac{1}{2})] \end{aligned}$$

which yields

$$\mathbb{H} \approx \frac{1}{2}\int_0^{\pi} ds \left[\frac{4\sin^2(s/2)}{L}P^2(s) + \frac{1}{C}Q^2(s)\right]$$

This is analogous to a collection of elastic string Hamiltonians with masses $m(s) = L/[4\sin^2(s/2)]$, spring constants $K(s) = 1/C$, frequencies $\omega(s) = \frac{2}{\sqrt{LC}}\sin(s/2)$, and length scales $\alpha(s) = (L/C\hbar^2)^{1/4}[2\sin(s/2)]^{-1/2}$. The classical ground state is $Q_j = 0 = P_j$, but quantum mechanically we cannot simultaneously specify the values of Q_j and P_j. Instead, the quantum mechanical ground state of the transmission line is given by the wave functions

$$\psi_s^{(0)}(q) = A_s^{(0)} e^{-\frac{1}{2}\alpha^2(s)q^2}$$

[1] In order to avoid confusion between the inductance L and the Lagrangian I have chosen to denote the latter by $\mathbb{L}$, and for consistency the Hamiltonian is denoted by $\mathbb{H}$.

[2] The analysis we are about to carry out is quite a standard one in condensed matter physics: we consider a harmonic system (*e.g.* a crystal with phonons), perturb the system suddenly and locally, and see how the perturbation is smeared out by vibrations. Usually the analysis is carried out using the second quantized formalism (see *e.g.* G.D. Mahan, *Many-Particle Physics*, Plenum, New York, 1991, chapter 4.3.), but we will instead use the more elementary first quantized approach.

where $A_s^{(0)}$ is a normalization constant. Hence, the probability density that the charge in mode s deviates from the classical ground state charge by amount q is given by $|\psi_s^{(0)}(q)|^2$, and a similar distribution is obtain for the deviation between momentum in mode s and the classical ground state momentum. From now on we will concentrate on the $T = 0$ case so that initially the transmission line is in its ground state.

Let us now consider what happens if the charge Q_0 changes instantaneously by amount δQ due to a tunneling event. Since Q_0 does not have a well-specified value before the tunneling event (a distribution of Q_0-values is possible), it is at first sight not clear how to describe the state of the system after tunneling. Actually, it is very easy! If the initial distribution of charge is given by $\psi_s(q)$, a shifted distribution is $\psi_s(q - \delta Q(s))$, and we only need to determine the proper $\delta Q(s)$ for each mode using the diagonalizing transformation. An equivalent way to look at the problem is to say that the tunneling event does not instantaneously change the quantum state of the transmission line, but instead it shifts the classical ground state of the harmonic oscillators by $\delta Q(s) = \delta Q \sqrt{2/\pi} \cos(s/2)$, and the quantum mechanical wave functions after a tunneling event are centered around these new minima. The new eigenfunctions are given by

$$\phi_s^j(q) = A_s^{(j)} e^{-\frac{1}{2}\alpha^2(s)[q-\delta Q(s)]^2} H_j[\alpha(s)(q - \delta Q(s))]$$

where $H_j(z)$ is a Hermite polynomial. Since the transmission line is described as a collection of harmonic oscillators, the energy of $\phi_s^j(q)$ is $(j + 1/2)\hbar\omega(s)$.

The instantaneous charging of C_0 shifts the ground states of the harmonic oscillators but it does not change the quantum state $\{\psi_s^{(0)}(q)\}$ of the transmission line. Therefore, after the tunneling event has taken place, the transmission line is no longer in an eigenstate of $\mathbb{H}$, and in particular it is not in its ground state. Instead, after tunneling each mode s is in linear combination

$$\psi_s^{(0)}(q) = \sum_{j=0}^{\infty} \gamma_{0j}(s) \phi_s^j(q)$$

of eigenstates. The coefficients $\gamma_{0j}(s)$ can be obtained using the properties of Hermite polynomials, and after some algebra we get $\gamma_{0j}(s) = (j!)^{-1/2} e^{-\alpha^2(s)\delta Q^2(s)/4} [-\alpha(s)\delta Q(s)/\sqrt{2}]^j$. (Note that $\sum_{j=0}^{\infty} |\gamma_{0j}(s)|^2 = 1$ as must be the case.)

This has some rather remarkable consequences. Since we assume that energy is conserved in each tunneling process, the initial state (one extra electron to the left of C_0, transmission line in its ground state) and the final state (one extra electron to the right of C_0, transmission line in some excited state) must have the same energies. The probability of finding extra energy $j\hbar\omega(s)$ in the s^{th} mode of the transmission line is given by $|\gamma_{0j}(s)|^2$ and therefore the probability density of having excess energy ϵ in mode s is

$$A(s, \epsilon) = \sum_{j=0}^{\infty} |\gamma_{0j}(s)|^2 \delta(\epsilon - j\hbar\omega(s)).$$

The energy in the initial state can be distributed in many ways among the different modes s, and it is easier to consider first what happens if there are S discrete modes. The probability density $A(E)$ of finding the transmission line with total excess energy E is given by a product of the probabilities associated with different modes, subject to the condition that the total excitation energy is E, *i.e.*

$$A(E) = \int_{-\infty}^{\infty} \prod_s d\epsilon_s \left[\prod_s A(s, \epsilon_s) \right] \delta(E - \sum_s \epsilon_s).$$

By writing the delta function in terms of its Fourier transform and by defining the Fourier transform of $A(s,\epsilon)$ to be $\exp[B(s,\theta)]$, we get

$$A(E) = \int_{-\infty}^{\infty} \frac{d\theta}{2\pi} e^{-i\theta E} e^{\sum_s B(s,\theta)} = \int_{-\infty}^{\infty} \frac{d\theta}{2\pi} e^{-i\theta E} e^{\int_0^\pi ds B(s,\theta)}$$

where we took the continuum limit again.

To obtain the tunneling rates we divide the energy of the system after tunneling has taken place into two parts: some of the energy is taken up by a single-particle energy of the extra electron (*e.g.* its kinetic energy), while some is used to excite the transmission line. In our approximation that all energy-conserving tunneling events are equivalent, the tunneling rate to charge the capacitor C_0 is given by

$$\Gamma(V) = \frac{1}{R_t e^2} \int_{-\infty}^{\infty} dE \int_{-\infty}^{\infty} dE' f(E)[1 - f(E')] A(E + eV - E')$$

where $f(E)$ counts the number of occupied states to the left of C_0, $[1 - f(E')]$ counts the number of available states to the right of C_0, and $A(E + eV - E')$ accounts for the possibility of exciting the transmission line with the available energy (note that the Fermi levels on the two sides of C_0 differ by eV). At zero temperature — which is the only case we can consider since we have assumed the transmission line to be initially in its ground state — the tunneling rate reduces to

$$\Gamma(V) = \frac{1}{R_t e^2} \int_{-\infty}^{0} dE \int_{0}^{\infty} dE' A(E + eV - E') = \frac{1}{R_t e^2} \int_{0}^{eV} dE \int_{0}^{E} dE' A(E')$$

where we used that $A(E) = 0$ for $E < 0$. The current through the junction is given by

$$I(V) = e\Gamma(V) = \frac{1}{R_t} \left[V \int_0^{eV} dE A(E) - \frac{1}{e} \int_0^{eV} dE E A(E) \right]$$

since at zero temperature the tunneling rate in the reverse direction vanishes.

Hence, we need to find $A(E)$ and therefore the Fourier transform of $A(s,\epsilon)$. Some algebra yields $A(s,\theta) = \exp\left[\left(e^{i\theta\hbar\omega(s)} - 1\right) \frac{1}{2}\alpha^2(s)\delta Q^2(s)\right]$ so that

$$A(E) = \int_{-\infty}^{\infty} \frac{d\theta}{2\pi} e^{-i\theta E} e^{\int_0^\pi ds \frac{1}{2}\alpha^2(s)\delta Q^2(s)\left(e^{i\theta\hbar\omega(s)} - 1\right)}$$

and the task of obtaining $A(E)$ has been reduced to a Fourier transform, albeit a complicated one. We carry out the transform in Appendix B, and find $A(E) \propto E^{\frac{e^2}{h}\sqrt{\frac{L}{C}}-1} \equiv E^{g-1}$. Consequently, the current at small voltage is $I(V) \propto V^{g+1}$. For a more general transmission line with resistive elements one finds $g = Z(0)/(h/e^2)$ where $Z(\omega)$ is the (frequency dependent) impedance of the transmission line. For an LC-line we have $Z(\omega) = \sqrt{L/C}$.

We have now obtained the main result of this appendix: the impedance of the environment manifests itself in the power law current-voltage characteristics. A simple way to understand this result qualitatively is to think of the environmental impedance as a quantity that determines the charge relaxation rate, the rate at which the junction capacitance can be charged or discharged. If the relaxation rate is fast, a tunneling charge disappears in the leads before the next tunneling event, and we cannot obtain Coulomb blockade. If the relaxation is slow, corresponding to a high impedance, a charge that has tunneled through the capacitor provides an extra barrier for the next tunneling event, and current flow is impeded. Figuratively, you can think of difference of lifting water or sand over a barrier using a bucket: the effort to lift one bucket remains the same as the previously poured water flows away, while the effort for lifting sand gets harder and harder as the sand piles up.

At this point we must examine the validity of our earlier approximation that the transmission line dynamics is unaffected by the difference between C and C_0. Mathematically, having a capacitance $C_0 \gg C$ at the end of a transmission line is equivalent to having a large weight at the end of an elastic string. Since the disturbance is localized at one end of the string, it has little effect on the low-energy modes, and therefore our relatively simple analysis can be expected to be quite accurate in the low-voltage limit. The high-energy modes, however, are affected by the difference between C_0 and C. This we can see *e.g.* by calculating the how much the total energy of the transmission line changes as a result of a tunneling event. The energy expectation value in mode s after tunneling is given by $E(s) = \sum_{j=0}^{\infty} |\gamma_{0j}(s)|^2 \hbar\omega(s)(j+1/2) = \hbar\omega(s)(\frac{1}{2}\alpha^2(s)\delta Q^2(s)+\frac{1}{2})$. The last term is nothing but the zero point energy, which is present even in the ground state, so the energy change in mode s due to the tunneling event is in our approximation given by $\delta E(s) = \frac{1}{2}\hbar\omega(s)\alpha^2(s)\delta Q^2(s)$, and the total energy change due to tunneling is $\delta E = \int_0^{\pi} ds \delta E(s) = \delta Q^2/(2C)$. This result is not quite correct: since the relevant capacitance is C_0 rather than C, the energy should also be $\delta Q^2/(2C_0)$ rather than $\delta Q^2/(2C)$. Classically, the energy of the transmission line immediately after a tunneling event is $\delta Q^2/(2C_0)$ (the classical state is $Q_n = \delta Q \delta_{n,0}$, $P_n = 0$) and later this energy is only redistributed between different inductors and capacitors but not lost. Quantum mechanically the energy of the transmission line after tunneling is not fully determined — the line is not in an energy eigenstate — but the energy expectation value $\int_0^{\infty} dEEA(E)$ must nevertheless be given by its classical value $\delta Q^2/(2C_0)$. This error is a consequence of our earlier substitution $C_0 \rightarrow C$, which was necessary to explicitly find out the eigenmodes. The total excitation energy can also be obtained as $\int_0^{\infty} dEEA(E)$ which in our approximation yields $e^2/(2C)$ rather than the exact value $E_C = e^2/(2C_0)$ showing that our result for $A(E)$ is too large at larger energies. Using this exact sum rule we find that at large voltages ($eV \gg E_C$) the current voltage characteristics are approximately given by $I(V) = \frac{1}{R_t}[V - e/(2C_0)]$ so that the IV-curve is shifted to slightly larger voltages.[3]

[3] Actually, it is possible to carry out the analysis without explicitly diagonalizing the transmission line thereby avoiding the replacement $C_0 \rightarrow C$. The result agrees with our simpler approach.

Appendix B

The Minnhagen method

Let us consider the function $F(x) = e^{if(x)}$ and obtain its Fourier transform $F(k)$. Often direct integration is quite difficult unless $f(x)$ is particularly simple. Another way of finding $F(k)$ was introduced by Petter Minnhagen of Umeå University. We start by differentiating $F(x)$ and finding the Fourier transform of $F'(x)$. On one hand it is $ikF(k)$, but on the other hand we have, by direct differentiation,

$$\begin{aligned}\mathcal{F}[F'(x)] = & \textstyle\int_{-\infty}^{\infty} dx e^{ikx} i f'(x) F(x) \\ = & \textstyle\int_{-\infty}^{\infty} \frac{dk'}{2\pi} F(k') \int_{-\infty}^{\infty} dx e^{i(k-k')x} i f'(x) \\ = & \textstyle\int_{-\infty}^{\infty} \frac{dk'}{2\pi} F(k') i g(k-k')\end{aligned}$$

where $g(k)$ is the Fourier transform of $f'(x)$. Hence, we have the integral equation

$$kF(k) = \int_{-\infty}^{\infty} dk' g(k') F(k-k').$$

While this integral equation is in general quite difficult to solve, we can easily find a solution if we have some additional information about $F(k)$ and $g(k)$. Often in the problems of physical interest $F(k-k')$ vanishes if $k-k' < 0$ — this is for instance the case if k measures excitation energy and $F(k)$ is a spectral function. Furthermore, if the small-k value of $g(k)$ approaches a constant $g(k \approx 0) \approx g$, we get

$$kF(k) = g \int_0^k dk' F(k-k')$$

which has the solution $F(k) = k^{g-1}$.

In the specific case of tunneling into a lossless transmission line we must obtain the Fourier transform of $A(\theta)$, which is exactly of the type we discussed above if we identify

$$f(\theta) = -i\frac{1}{2} \int_0^{\pi} ds \alpha^2(s) \delta Q^2(s) \left[e^{i\theta\hbar\omega(s)} - 1 \right].$$

Taking the derivative, Fourier transforming, and setting $E \to 0^+$ yields

$$g(0) = \frac{\delta Q^2}{C} \frac{2}{\pi} \int_0^{\pi/2} du \cos^2 u \int_{-\infty}^{\infty} \frac{d\theta}{2\pi} \exp\left[i \frac{2\theta\hbar}{\sqrt{LC}} \sin u - iE\theta \right]_{E\to 0} = \frac{1}{h} \delta Q^2 \sqrt{\frac{L}{C}}$$

and hence $A(E) \approx E^{\frac{1}{h}\delta Q^2 \sqrt{\frac{L}{C}} - 1}$.

Appendix C

Second quantization

The formalism that goes under the somewhat mysterious name of **second quantization**[1] underpins all theories of quantum many-body systems. A basic feature of these is that one is dealing with identical particles: For bosons [fermions] one thus has to work with states that are symmetric [anti-symmetric] under the exchange of two particles. Although conceptually straightforward, this would become a daunting task if trying to use the ordinary ("first-quantized") formalism of quantum mechanics. Just imagine explicitly calculating matrix elements in a basis of the symmetrized [anti-symmetrized] many-particle states

$$|\,\lambda_1, \lambda_2, ..., \lambda_N\rangle \equiv \frac{1}{\sqrt{N! \prod_{i=1}^{m} n_{\lambda_i}!}} \sum_{\mathcal{P}} \zeta^{\mathrm{sgn}\mathcal{P}} \,|\,\lambda_{\mathcal{P}1}\rangle \otimes |\,\lambda_{\mathcal{P}2}\rangle \otimes ... \otimes |\,\lambda_{\mathcal{P}N}\rangle \tag{C.1}$$

when the particle number N is large! Here $\lambda_1, \lambda_2, ..., \lambda_N$ are quantum numbers labeling the single-particle states $|\,\lambda_1\rangle$,$|\,\lambda_2\rangle$, ...,$|\,\lambda_N\rangle$, n_{λ_i} is the number of particles in the single-particle state$|\,\lambda_i\rangle$, with m the number of distinct quantum numbers.[2] $\zeta = 1[-1]$ for bosons [fermions], and sgn$\mathcal{P} \equiv 0(1)$ when the number of transpositions that take $(1, 2, ..., N)$ to the permutation $(\mathcal{P}1, \mathcal{P}2, ...\mathcal{P}N)$ is even (odd). (Note that the product of single-particle states on the right-hand side of Eq. (C.1) is ordered: the first state$|\,\lambda_{\mathcal{P}1}\rangle$ is that of the "first" particle, the second state $|\,\lambda_{\mathcal{P}2}\rangle$ is that of the "second" particle, and so on...; "first", "second", etc. referring to some fixed book keeping of the (identical!) particles.)

Another difficulty with the "first-quantized" description is that the particle number N is assumed to be fixed, with a Hilbert space

$$\mathcal{H}^N = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes ... \otimes \mathcal{H}_N, \tag{C.2}$$

[1] The notion of "second quantization" first appeared in work by Paul Jordan in 1927, where he tried to extend the canonical quantization scheme of ordinary quantum mechanics to the quantization of fields. (A definite breakthrough came a year later with a paper that Jordan co-authored with Oskar Klein and Eugene Wigner.) Conceptually, however, there is only *one* step to be taken when quantizing a classical theory, with "second quantization" being a formalism adapted for quantizing a classical *field theory.* In the present context of many-particle physics (where, depending on the particular application, a classical field may, or may not be present), "second quantization" simply offers a convenient language in which to describe the statistics of many identical particles.

[2] Note that the piece $1/\sqrt{\prod_{i=1}^{m} n_{\lambda_i}!}$ of the normalization factor only comes into play for bosons, since for fermions there is only one particle in each single-particle state. The expression $n_{\lambda_i}!$ counts the number of permutations of particles in the *same* single-particle state$|\,\lambda_i\rangle$. To account for all states, one then forms the product $\prod_{i=1}^{m} n_{\lambda_i}!$.

where $\mathcal{H}_j$ is the Hilbert space of the j:th particle. Now, we know from statistical physics that it is often convenient to allow the particle number to fluctuate *(grand canonical ensemble)*. The "first-quantized" formalism, however, is ill-adapted to handle this situation. An additional, maybe less obvious reason why one would like to develop a more powerful formalism is that quantum many-body systems generically exhibit *collective excitations* which are hard to come to grips with in the "first-quantized" formalism. One reason for this – among others – is that these excitations (the quantized plasmon excitations of the electron liquid being an example that we discussed in the the chapter on Fermi liquids) also fluctuate in number.

Second quantization removes all these difficulties in one stroke. The idea behind the formalism is as simple as beautiful. One first introduces a convenient way of denoting the many-particle states in Eq. (C.1), the so-called **occupation number representation**: Instead of enumerating all the single particles, specifying in which states they are (cf. the left-hand side of Eq. (C.1)), one simply specifies how many particles there are in each available single-particle state. As an example, consider a state in Eq. (C.1) with, say, $N=7$, and with $\lambda_1=\lambda_2=\lambda_5\equiv\lambda$, $\lambda_3=\lambda_6\equiv\mu$, and $\lambda_4=\lambda_7\equiv\nu$. In the notation introduced in Eq. (C.1), this state would be written as $|\,\lambda,\lambda,\mu,\nu,\lambda,\mu,\nu\rangle$. In the occupation number representation one writes this state more economically as $|\,3,2,2\rangle$, saying that there are three particles in the single-particle state $|\,\lambda\rangle$, and two particles each in $|\,\mu\rangle$ and $|\,\nu\rangle$, (having agreed on the ordering λ,μ,ν of the quantum numbers). Generalizing, $|\,n_{\lambda_1},n_{\lambda_2},...,n_{\lambda_m}\rangle$ is the symmetrized [anti-symmetrized] state with n_{λ_i} particles in the single-particle state $|\,\lambda_i\rangle, i=1,2,...,m$, with $\sum_{i=1}^m n_{\lambda_i}=N$, N being the number of particles. The parameters n_{λ_i} are the *occupation numbers* which give the name to this representation. Note that, by the Pauli principle, fermionic occupation numbers can only take the values 0 or 1. Also note that when specifying a state, we write out only those occupation numbers that take non-zero values. The states in Eq. (C.1), compactly denoted[3] by $|\,n_{\lambda_1},n_{\lambda_2},...,n_{\lambda_m}\rangle$ in the occupation number representation, form an orthonormal basis of the *physical Hilbert space* $\mathcal{F}^N\subset\mathcal{H}^N$, "physical" meaning that $\mathcal{F}^N$ contains only states that are properly symmetrized [anti-symmetrized]. Thus, *any* symmetrized [anti-symmetrized] state $|\,\psi\rangle$ in $\mathcal{F}^N$ can be written as

$$|\,\psi\rangle=\sum_{n_{\lambda_1},n_{\lambda_2},...}c_{n_{\lambda_1},n_{\lambda_1},...}\,|\,n_{\lambda_1},n_{\lambda_2}...\rangle, \tag{C.3}$$

with $\sum_i n_{\lambda_i}=N$, and where

$$\langle n_{\lambda_1},n_{\lambda_2}...\mid n'_{\lambda_1},n'_{\lambda_2}...\rangle=\delta_{n_{\lambda_1}n'_{\lambda_1}}\delta_{n_{\lambda_2}n'_{\lambda_2}}... \tag{C.4}$$

To allow for fluctuating particle numbers one forms the so-called **Fock space**[4]

$$\mathcal{F}=\mathcal{F}^0\oplus\mathcal{F}^1\oplus\mathcal{F}^2...\oplus\mathcal{F}^N, \tag{C.5}$$

[3] Let us here comment that one often simplifies the notation by replacing n_{λ_j} by n_j, assuming that we have agreed upon an ordered set of quantum numbers $\{\lambda_j\}$ to label the single-particle states. We can then do the replacement $|\,n_{\lambda_1},n_{\lambda_2},...,n_{\lambda_m}\rangle\rightarrow$
$|\,n_1,n_2,...,n_m\rangle$. In this tool box, however, we shall choose to stick to the more pedantic notation where the quantum numbers are explicitly written out.

[4] Vladimir Fock, 1898 - 1974, was one of the leading theoretical physicists in the former Soviet Union. He did important work in the foundations of quantum mechanics, general relativity, and elasticity theory. He has given his name not only to the concept of a *Fock space*, but to a number of other contributions, maybe the best known being *Hartree-Fock theory*, an extension of the original Hartree theory that Fock developed in 1930.

where $\mathcal{F}^N$ is the physical Hilbert space of N particles *(Fock N-particle subspace)*. Note that the peculiar 0-particle space $\mathcal{F}^0$ contains only one state, the "empty" state $|\,0\rangle$ *(vacuum state)*. A basis of $\mathcal{F}$ is obtained by taking all of our previous basis states, dropping the condition $\sum_{i=1}^m n_{\lambda_i} = N$ on the occupation numbers. Given this, one may now form linear superpositions of states $|\,n_{\lambda_1}, n_{\lambda_2}, ...\rangle$ containing different numbers of particles. This is not only convenient, but sometimes crucial for properly describing the physics at hand, a famous example being the BCS ground state.

Although the occupation number representation is nice and handy, by itself it does not offer a big improvement over the "first-quantized" formalism. All that we have done so far is to replace the "first-quantized" way of labeling a symmetrized [anti-symmetrized] many-particle state (cf. the left-hand side of Eq. (C.1)) by a more compact notation, using occupation numbers, then removing the condition that the particle number is fixed. To make some real progress we have to introduce operators that take us from one Fock subspace to another:

$$a_\lambda^\dagger : \mathcal{F}^N \longrightarrow \mathcal{F}^{N+1}, \qquad a_\lambda : \mathcal{F}^N \longrightarrow \mathcal{F}^{N-1}. \tag{C.6}$$

Here $a_\lambda^\dagger$ is a *creation operator* that adds one particle in the single-particle state $|\,\lambda\rangle$ to the system. Similarly, a_λ is a *destruction operator* that removes one particle in the state $|\,\lambda\rangle$ from the system. (The "dagger" on $a_\lambda^\dagger$ signifies that $a_\lambda^\dagger$ is the Hermitian adjoint of a_λ.) This step may seem innocent, but a second thought makes us realize that it is far from obvious how to consistently construct these operators. Recall that $\mathcal{F}^N$ and $\mathcal{F}^{N\pm1}$ are *physical* Hilbert spaces, that is, $a_\lambda^\dagger$ must have the property of mapping a symmetrized [anti-symmetrized] state with N particles onto a symmetrized [anti-symmetrized] state with $N+1$ particles (with a_λ effecting the inverse mapping). One may worry that this non-trivial property will lead to complicated relations between operators labeled by different quantum numbers, making the whole approach intractable.

The magic of "second quantization" is that this is not so! All that we have to require is that the creation- and destruction operators satisfy the algebras

$$\text{BOSONS:} \quad [a_\lambda^\dagger, a_\mu] = \delta_{\lambda\mu}, \quad [a_\lambda^\dagger, a_\mu^\dagger] = [a_\lambda, a_\mu] = 0 \tag{C.7}$$

$$\text{FERMIONS:} \quad \{a_\lambda^\dagger, a_\mu\} = \delta_{\lambda\mu}, \quad \{a_\lambda^\dagger, a_\mu^\dagger\} = \{a_\lambda, a_\mu\} = 0 \tag{C.8}$$

and that their actions on an arbitrary Fock basis state is such that

$$a_{\lambda_j}^\dagger\,|n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}, ...\rangle \;\equiv\; \sqrt{n_{\lambda_j}+1}\,\zeta^{s_j}\,|n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}\!+\!1, ...\rangle \tag{C.9}$$

$$a_{\lambda_j}\,|n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}, ...\rangle \;=\; \sqrt{n_{\lambda_j}}\,\zeta^{s_j}\,|n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}\!-\!1, ...\rangle \tag{C.10}$$

As before, $\zeta = 1[-1]$ for bosons [fermions], with $s_j = \sum_{i=1}^{j-1} n_i$. It may be worth pointing out that the fact that a_{λ_j} and $a_{\lambda_j}^\dagger$ are Hermitian adjoints of each other implies that Eq. (C.10) is a consequence of the definition in Eq. (C.9). This is the reason why in Eq. (C.9) the symbol "$\equiv$" appears, whereas in Eq. (C.10) we have an equals sign. More importantly, note that the non-trivial sign-factor $(-1)^{s_j}$ for fermions follows from the algebra in (C.8). To see how, let us look at the simple case of two fermions. Eq. (C.9) here implies that

$$\begin{aligned} a_{\lambda_2}^\dagger a_{\lambda_1}^\dagger\,|0\rangle &= -\,|n_{\lambda_1}\!=\!1, n_{\lambda_2}\!=\!1\rangle \\ a_{\lambda_1}^\dagger a_{\lambda_2}^\dagger\,|0\rangle &= |n_{\lambda_1}\!=\!1, n_{\lambda_2}\!=\!1\rangle \end{aligned} \tag{C.11}$$

as is indeed enforced by the anti-commutator in (C.8). This example makes it clear and simple that the antisymmetry of a fermionic many-particle state is built into the second quantization formalism by the algebra in (C.8): When exchanging two particles[5] the anti-commutator produces the required minus sign! The analogous observation holds for bosons: The symmetry of a bosonic many-particle state is ensured by the operator algebra in (C.7).

It is convenient to introduce a *number operator*

$$\hat{n}_\lambda \equiv a^\dagger_\lambda a_\lambda \tag{C.12}$$

that counts the number of particles in the single-particle state $|\,\lambda\rangle$. That is,

$$\hat{n}_{\lambda_j}\,|\,n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}, ...\rangle = n_{\lambda_j}\,|\,n_{\lambda_1}, n_{\lambda_2}, ..., n_{\lambda_j}, ...\rangle. \tag{C.13}$$

To develop some intuition, the reader is encouraged to convince herself how Eq. (C.13) follows from Eqs. (C.9) and (C.10).

The formalism encoded by Eqs. (C.7), (C.8), (C.9), and (C.10), with the states in (C.9) and (C.10) being basis states of the Fock space $\mathcal{F}$ in Eq. (C.5), is what we call **second quantization**. Its power and economy becomes immediate when noting that by iteration of Eq. (C.9) we can generate *any* basis state $|\,n_{\lambda_1}, n_{\lambda_2}, ...\rangle$ in $\mathcal{F}$ by simply hitting the vacuum state $|\,0\rangle$ by the appropriate product of creation operators:

$$|\,n_{\lambda_1}, n_{\lambda_2}, ...\rangle = \prod_i \frac{1}{\sqrt{n_{\lambda_i}!}} (a^\dagger_{\lambda_i})^{n_{\lambda_i}}\,|\,0\rangle, \tag{C.14}$$

with $1/\sqrt{n_{\lambda_i}!}$ being a normalization factor. Compare the simplicity and elegance of Eq. (C.14) with the horror lurking behind Eq. (C.1)! Instead of $N!$ permutations that yield the unwieldy structure of the state in (C.1), the Fock states $|\,n_{\lambda_1}, n_{\lambda_2}, ...\rangle$ are simply generated by applying a product of creation operators to a single reference state (the vacuum state $|\,0\rangle$). The symmetry properties of the states automatically come out right via the operator algebras in Eqs. (C.7) and (C.8). There is no need to symmetrize [anti-symmetrize] "by hand" as in the first-quantized formalism. The symmetry [anti-symmetry] of a many-particle state is piece and parcel of the second-quantized language! We note in passing that the anti-commutator $\{a^\dagger_\lambda, a^\dagger_\mu\} = 0$ in (C.8) implies that $(a^\dagger_\lambda)^2 = 0$, and hence (reassuringly!) we can never obtain a state where two fermions are in the same single-particle state.

As in any application of quantum mechanics, a *change of basis* is sometimes what makes the day! This is easily carried out in second quantization. Writing $a^\dagger_\lambda\,|\,0\rangle = |\,n_\lambda = 1\rangle \equiv |\,\lambda\rangle$ and $a^\dagger_{\lambda'}\,|\,0\rangle = |\,n_{\lambda'} = 1\rangle \equiv |\,\lambda'\rangle$, and using the resolution of the identity $\mathbf{1} = \sum_\lambda |\,\lambda\rangle\langle\lambda\,|$ (assuming, as before, that the states $\{|\,\lambda\rangle\}$ form a complete single-particle basis) we have that

$$\begin{aligned} a^\dagger_{\lambda'}\,|\,0\rangle &= \sum_\lambda |\,\lambda\rangle\langle\lambda\,|\,a^\dagger_{\lambda'}\,|\,0\rangle && \text{(C.15)}\\ &= \sum_\lambda |\,\lambda\rangle\langle\lambda\mid\lambda'\rangle && \text{(C.16)}\\ &= \sum_\lambda \langle\lambda\mid\lambda'\rangle a^\dagger_\lambda\,|\,0\rangle. && \text{(C.17)} \end{aligned}$$

[5] The exchange of two particles is here emulated by an exchange of two creation operator, $a^\dagger_{\lambda_1}$ and $a^\dagger_{\lambda_2}$. Since these operators do not carry particle indices, we "keep track" on the particles by the order in which they are "created": In the first (second) line of Eq. (C.11) the "first" ("second") particle is that in the single-particle state $|\,\lambda_1\rangle$, while the "second" ("first") particle is that in the single-particle state $|\,\lambda_2\rangle$.

Since any state can be generated from the vacuum state $|\,0\rangle$ (cf. Eq. (C.14)), it follows that Eq. (C.15) can be elevated to an operator identity:

$$a^\dagger_{\lambda'} = \sum_\lambda \langle \lambda \mid \lambda' \rangle a^\dagger_\lambda. \tag{C.18}$$

The Hermitian adjoint of (C.18) reads:

$$a_{\lambda'} = \sum_\lambda \langle \lambda' \mid \lambda \rangle a_\lambda. \tag{C.19}$$

In the case of a continuous set of quantum numbers (like the coordinates $\lambda \equiv x$ in configuration space) one usually uses the bracket notation $a^\dagger(x), a(x)$ instead of $a^\dagger_x, a_x$. It goes without saying that the sum in Eq. (C.18) now gets replaced by an integral. As an example, consider the transformation between a coordinate $(\lambda = x)$ and a momentum $(\lambda' = k)$ basis in a one-dimensional system of length ℓ. An adaption of Eq. (C.18) yields that

$$a^\dagger_k = \int_0^\ell dx\, \langle x \mid k \rangle a^\dagger(x) \tag{C.20}$$

with the inverse transformation

$$a^\dagger(x) = \sum_k \langle k \mid x \rangle a^\dagger_k. \tag{C.21}$$

For a translationally invariant system (periodic boundary conditions), we have that $\langle x \mid k \rangle = \langle k \mid x \rangle^* = \exp(ikx)/\sqrt{\ell}$, allowing us to write Eqs. (C.20) and (C.21) as the Fourier transforms

$$a^\dagger_k = \frac{1}{\sqrt{\ell}} \int_0^\ell dx\, \exp(ikx) a^\dagger(x), \qquad a^\dagger(x) = \frac{1}{\sqrt{\ell}} \sum_k \exp(-ikx) a^\dagger_k. \tag{C.22}$$

To complete the machinery of second quantization, we must specify how an arbitrary operator gets represented in terms of the creation- and annihilation operators. In the case of a *one-body operator* $\hat{\mathcal{O}}_1 = \sum_n \hat{o}_n$, with $\hat{o}_n$ an operator that acts on the n:th particle only (reverting to the language of first quantization where we put labels on the particles!), this is readily achieved. Consider a basis in which the operators $\hat{o}_n$ are diagonal:

$$\hat{o}_n = \sum_i c_i \,|\, \lambda_i \rangle_{nn} \langle \lambda_i \,|, \tag{C.23}$$

with $c_i = {}_n\langle \lambda_i \mid \hat{o}_n \mid \lambda_i \rangle_n$, and where $\{|\lambda_i\rangle_n\}$ is a complete set of single-particle states in the Hilbert space of the n:th particle.[6] Given (C.23) we have that

$$\langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid \hat{\mathcal{O}}_1 \mid n_{\lambda_1}, n_{\lambda_2}, ... \rangle = \sum_n \sum_i c_i \langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid \lambda_i \rangle_{nn} \langle \lambda_i | n_{\lambda_1}, n_{\lambda_2}, ... \rangle. \tag{C.24}$$

[6] For book keeping purposes, having regressed to first-quantized language, we here label the single particle states not only by the quantum numbers λ_i but also by the "particle index" n, thus keeping track on which particular single-particle Hilbert space a state belongs to. Note that the coefficients $c_i = {}_n\langle \lambda_i \mid \hat{o}_n \mid \lambda_i \rangle_n$ are independent of n, since the expectation values ${}_n\langle \lambda_i \mid \hat{o}_n \mid \lambda_i \rangle_n$ do not care about in which particular single-particle Hilbert space $\mathcal{H}_n$ they have been evaluated, all states $|\, \lambda_i \rangle_1, |\, \lambda_i \rangle_2, ..., |\, \lambda_i \rangle_n, ...$ being identical copies of each other!

Recall that $|\, n_{\lambda_1}, n_{\lambda_2}, ...\rangle$ represents a symmetrized [anti-symmetrized] state, with each term containing factors $...|\,\lambda_i\rangle_1 \otimes |\,\lambda_i\rangle_2 \otimes ... \otimes |\,\lambda_i\rangle_{n_{\lambda_i}}....$ As in Eq. (C.23) we have here labeled the single-particles states by the quantum numbers λ_i *and* the particle indices $n = 1, 2, ..., n_{\lambda_i}$, so as to keep track on the single-particle space to which a state belongs (in contrast to the right-hand side of Eq. (C.1) where we relied on a fixed ordering of the single-particle states). When acting with $c_i\,|\,\lambda_i\rangle_{nn}\langle\lambda_i\,|$ on the state $|\, n_{\lambda_1}, n_{\lambda_2}, ...\rangle$ it is then easy see that when $1 \le n \le n_{\lambda_i}$ we get back this state, multiplied by c_i. On the other hand, when the condition $1 \le n \le n_{\lambda_i}$ is not satisfied the state gets "killed", by the orthogonality of the single-particle states, giving an outcome $= 0$. It follows from Eqs. (C.23) and (C.24) that

$$\langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid \hat{\mathcal{O}}_1 \mid n_{\lambda_1}, n_{\lambda_2}, ...\rangle = \sum_i c_i n_{\lambda_i} \langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid n_{\lambda_1}, n_{\lambda_2}, ...\rangle \qquad \text{(C.25)}$$

Using Eq. (C.12) we can mold Eq. (C.25) on the form

$$\langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid \hat{\mathcal{O}}_1 \mid n_{\lambda_1}, n_{\lambda_2}, ...\rangle = \langle n'_{\lambda_1}, n'_{\lambda_2}, ... \mid \sum_i c_i \hat{n}_{\lambda_i} \mid n_{\lambda_1}, n_{\lambda_2}, ...\rangle. \qquad \text{(C.26)}$$

Since this equality is valid for any set of states, we infer that

$$\hat{\mathcal{O}}_1 = \sum_i c_i \hat{n}_{\lambda_i} = \sum_i a^\dagger_{\lambda_i} \langle \lambda_i \mid \hat{o} \mid \lambda_i\rangle a_{\lambda_i}, \qquad \text{(C.27)}$$

where we have used that $\hat{n}_{\lambda_i} = a^\dagger_{\lambda_i} a_{\lambda_i}$ and $c_i = \langle \lambda_i \mid \hat{o} \mid \lambda_i\rangle$ to write the expression in the last term. (The matrix element $\langle \lambda_i \mid \hat{o} \mid \lambda_i\rangle$ is a c-number and can of course be placed anywhere in the product. Our choice to sandwich it between the two operators $a^\dagger_{\lambda_i}$ and a_{λ_i} is the conventional one.)

Going to an arbitrary basis, using Eqs. (C.18) and (C.19), we finally obtain from (C.27):

$$\hat{\mathcal{O}}_1 = \sum_{i,j} a^\dagger_{\lambda_i} \langle \lambda_i \mid \hat{o} \mid \lambda_j\rangle a_{\lambda_j}. \qquad \text{(C.28)}$$

This is our desired representation of a single-body operator in second quantization.[7]

To put some flesh on the bones, let us look at two ubiquitous examples of one-body operators: **(i)** the *spin operator* $\hat{\boldsymbol{S}}$ for a system of spin-1/2 fermions, and **(ii)** the *Hamiltonian* $\hat{H}$ for non-interacting particles (fermions or bosons).

(i) Starting with the single-particle spin-1/2 operator $\hbar\boldsymbol{\sigma}/2 = (\hbar/2)(\sigma_x, \sigma_y, \sigma_z)$, with σ_x, σ_y, and σ_z the Pauli matrices, we identify $|\uparrow\rangle$ and $|\downarrow\rangle$ as the eigenstates of σ_z which form the single-particle basis. Reading off from Eq. (C.28), setting $\lambda_i \equiv \lambda$, $\lambda_j \equiv \mu$, we then find the second-quantized representation of the spin operator:

$$\hat{\boldsymbol{S}} = \sum_{\lambda,\mu=\uparrow,\downarrow} a^\dagger_\lambda \boldsymbol{\sigma}_{\lambda\mu} a_\mu, \qquad \text{(C.29)}$$

where $\boldsymbol{\sigma}_{\lambda\mu} = \langle \lambda \mid \hat{\boldsymbol{\sigma}} \mid \mu\rangle$. Quite frequently there are additional quantum numbers hanging around, labeling e.g. the sites α of the lattice on which the particles live. In such cases one

[7] Note that we have suppressed the particle indices n in Eqs. (C.27) and (C.28), since these have now become immaterial: All single-particle Hilbert spaces $\mathcal{H}_j$ are spanned by the same complete set of states $\{|\,\lambda_i\rangle\}$, and there is no need to single out one of them as a particular representative. The particle indices n *only* make sense as a book-keeping device in first-quantized language!

may write a generalized version of Eq. (C.29):

$$\hat{\boldsymbol{S}} = \sum_{\alpha} a^{\dagger}_{\alpha\lambda} \boldsymbol{\sigma}_{\lambda\mu} a_{\alpha\mu}, \tag{C.30}$$

using the "Einstein convention" to sum over the repeated spin indices $\lambda =\uparrow, \downarrow$ and $\mu =\uparrow, \downarrow$ (without writing out the summation explicitly).

(ii) Taking off from a basis of momentum states where the (one-dimensional) kinetic energy operator $K = -p^2/2m$ is diagonal, and then changing to a coordinate basis where $K = -(\hbar^2/2m)\partial_x^2$, Eq. (C.27) implies that

$$\hat{H} = \int dx\, a^{\dagger}(x) \left(-\frac{\hbar^2}{2m} \partial_x^2 + V(x) \right) a(x), \tag{C.31}$$

replacing the sum in (C.27) by an integral. The extension to higher dimensions is immediate:

$$\hat{H} = \int d^d r\, a^{\dagger}(\boldsymbol{r}) \left(-\frac{\hbar^2}{2m} \nabla^2 + V(\boldsymbol{r}) \right) a(\boldsymbol{r}). \tag{C.32}$$

Turning to the case of two-body operators $\hat{\mathcal{O}}_2 = \sum_{n,n'} \hat{o}_{nn'}$, as needed for describing the pairwise interaction among particles (labeled by n and n'), we may use an analysis that is a blue-copy of that for a single-body operator. Although straightforward, its explicit execution is a bit cumbersome, though. For this reason, we here only give the result, which is simple and transparent. In fact, it is precisely what one would have guessed given the result for the one-body operator in Eq. (C.28):

$$\hat{\mathcal{O}}_2 = \sum_{\lambda\lambda'\mu\mu'} a^{\dagger}_{\mu} a^{\dagger}_{\mu'} \langle \mu, \mu' \mid \hat{o} \mid \lambda, \lambda' \rangle a_{\lambda'} a_{\lambda}. \tag{C.33}$$

Here $\hat{o}$ is a two-particle operator acting on the two-particle state $|\, \lambda, \lambda' \rangle = |\, \lambda \rangle \otimes |\, \lambda' \rangle$, in exact analogy with the case of the one-particle operator above (which we also called $\hat{o}$ after having erased the particle index!). Given the result in Eq. (C.33) we easily read off the second-quantized expression for a two-body potential $V(\boldsymbol{r}, \boldsymbol{r}')$ that acts between two particles with coordinates $\boldsymbol{r}$ and $\boldsymbol{r}'$:

$$\hat{V} = \frac{1}{2} \int d^d r \int d^d r'\, a^{\dagger}(\boldsymbol{r}') a^{\dagger}(\boldsymbol{r}) V(\boldsymbol{r}, \boldsymbol{r}') a(\boldsymbol{r}) a(\boldsymbol{r}'). \tag{C.34}$$

We have here used that the two-particle potential operator V is diagonal in the two-particle basis $|\, \boldsymbol{r}, \boldsymbol{r}' \rangle$, with matrix elements $V(\boldsymbol{r}, \boldsymbol{r}')$.

The extension to n-body operators with $n > 2$ is more or less automatic, but not particularly relevant for applications. We may thus stop here, having exposed the reader to the basic facts about second quantization.

Appendix D

Bosonic form for $\psi_r(x)$

The bosonic form for the fermion operators is given by Eq. (2.67)

$$\psi_r^\dagger(x) = \frac{1}{\sqrt{2\pi\alpha}} e^{-irk_F x - r\sum_q e^{-\frac{1}{2}\alpha|q|}\left[-\frac{2\pi}{Lq}e^{iqx}\rho_r(q)\right] - irN_r\frac{2\pi x}{L}} U_r^\dagger. \tag{D.1}$$

We will prove this assertion by evaluating the anticommutation relations $\{\psi_r(x), \psi_r^\dagger(x')\}$ and showing that they yield $\delta_{rr'}\delta(x-x')$ as required.

The anticommutation relations between different branches are satisfied automatically due to the fact that the partial density operators between different branches commute and the ladder operators anticommute. The anticommutation relations within a branch must, in contrast, be checked explicitly, and we obtain

$$\begin{aligned}
&\{\psi_+(x), \psi_+^\dagger(x')\} = \tfrac{1}{2\pi\alpha} e^{ik_F(x-x')} \\
&\times \Bigg[U_+ e^{-\sum_q \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha|q| - iqx}\rho_+(q) + iN_+\frac{2\pi x}{L}} e^{-\sum_q \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha|q| + iqx'}\rho_+(-q) - iN_+\frac{2\pi x'}{L}} U_+^\dagger \\
&\quad + e^{-\sum_q \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha|q| + iqx'}\rho_+(-q) - iN_+\frac{2\pi x'}{L}} U_+^\dagger U_+ e^{-\sum_q \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha|q| - iqx}\rho_+(q) + iN_+\frac{2\pi x}{L}} \Bigg].
\end{aligned}$$

The ladder operators can be commuted to the right by noticing that $U_+^\dagger$ increases N_+ by one, and $U_+^\dagger U_+ = 1 = U_+ U_+^\dagger$. It is useful to divide the sums over q into $q > 0$ and $q < 0$ parts, which results in an expression like $e^{i\frac{2\pi(x-x')}{L}} e^{A-B} e^{C-D} + e^{C-D} e^{A-B}$ where $A = -\sum_{q>0} \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q - iqx}\rho_+(q)$, $B = -\sum_{q>0} \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q + iqx}\rho_+(-q)$, $C = -\sum_{q>0} \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q + iqx'}\rho_+(-q)$, and $D = -\sum_{q>0} \frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q - iqx'}\rho_+(q)$. Applying the relation $e^A e^B = e^{A+B+\frac{1}{2}[A,B]}$ repeatedly yields

$$\begin{aligned}
&\{\psi_+(x), \psi_+^\dagger(x')\} \\
= \;& \frac{e^{ik_F(x-x') + iN_+\frac{2\pi(x-x')}{L}}}{2\pi\alpha} \\
& e^{-\sum_{q>0}\frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q}\left(e^{-iqx} - e^{-iqx'}\right)\rho_+(q)} e^{\sum_{q>0}\frac{2\pi}{qL} e^{-\frac{1}{2}\alpha q}\left(e^{iqx} - e^{iqx'}\right)\rho_+(-q)} \\
& \left[e^{i\frac{2\pi(x-x')}{L}} e^{\sum_{q>0}\frac{2\pi}{qL} e^{-\alpha q + iq(x-x')}} + e^{\sum_{q>0}\frac{2\pi}{qL} e^{-\alpha q - iq(x-x')}} \right] e^{-\sum_{q>0}\frac{2\pi}{qL} e^{-\alpha q}}.
\end{aligned}$$

The sums over q can be evaluated using $q = n\frac{2\pi}{L}$ and $\sum \frac{1}{n} z^n = -\ln(1-z)$. This yields

$e^{-\sum_{q>0}\frac{2\pi}{qL}e^{-\alpha q}} = 1 - e^{-\frac{2\pi\alpha}{L}}$ and

$$\begin{aligned}
& \{\psi_+(x), \psi_+^\dagger(x')\} \\
= \quad & \frac{1-e^{-\frac{2\pi\alpha}{L}}}{2\pi\alpha} e^{ik_F(x-x')+iN_+\frac{2\pi(x-x')}{L}} \\
& e^{-\sum_{q>0}\frac{2\pi}{qL}e^{-\frac{1}{2}\alpha q}\left(e^{-iqx}-e^{-iqx'}\right)\rho_+(q)} e^{\sum_{q>0}\frac{2\pi}{qL}e^{-\frac{1}{2}\alpha q}\left(e^{iqx}-e^{iqx'}\right)\rho_+(-q)} \\
& e^{i\frac{\pi(x-x')}{L}} \left[\frac{1}{e^{-i\frac{\pi(x-x')}{L}} - e^{-\frac{2\pi\alpha}{L}} e^{i\frac{\pi(x-x')}{L}}} + \frac{1}{e^{i\frac{\pi(x-x')}{L}} - e^{-\frac{2\pi\alpha}{L}} e^{-i\frac{\pi(x-x')}{L}}}\right]
\end{aligned}$$

and taking the $\alpha \to 0^+$ limit gives

$$\begin{aligned}
& \{\psi_+(x), \psi_+^\dagger(x')\} \\
= \quad & \frac{1}{L} e^{ik_F(x-x')+iN_+\frac{2\pi(x-x')}{L}} \\
& e^{-\sum_{q>0}\frac{2\pi}{qL}\left(e^{-iqx}-e^{-iqx'}\right)\rho_+(q)} e^{\sum_{q>0}\frac{2\pi}{qL}\left(e^{iqx}-e^{iqx'}\right)\rho_+(-q)} \\
& e^{i\frac{\pi(x-x')}{L}} \left[\frac{iL}{2\pi}\frac{1}{x-x'+i0^+} - \frac{iL}{2\pi}\frac{1}{x-x'-i0^+}\right]
\end{aligned}$$

or $\{\psi_+(x), \psi_+^\dagger(x')\} = \delta(x-x')$ as required. The anticommutator $\{\psi_+(x), \psi_+(x')\}$ is evaluated similarly.

Appendix E

Alternate forms for the Luttinger H

For future reference it is useful to write the Hamiltonian in terms of the operators ρ_Σ and ρ_Δ. Expressing $\rho_\pm$ in terms of the sum and difference fields yields immediately (apart from constants)

$$H' = \frac{\pi}{L} \sum_{q>0} \left[(v_F + g_4 + g_2) \rho_\Sigma(q) \rho_\Sigma(-q) + (v_F + g_4 - g_2) \rho_\Delta(q) \rho_\Delta(-q) \right].$$

In terms of the velocity v defined in Eq. (2.61) this can be written as

$$H' = \frac{\pi}{L} \sum_{q>0} \left[\frac{v}{g} \rho_\Sigma(q) \rho_\Sigma(-q) + gv \rho_\Delta(q) \rho_\Delta(-q) \right]$$

where g is the interaction parameter. Note that although this expression appears to be diagonal, the commutation relation $[\rho_\Sigma(-q'), \rho_\Delta(q)] = \frac{Lq}{\pi}$ makes this form less convenient than the form (2.62). Using the Heisenberg equation of motion $\partial_t \rho_\Sigma = -i[\rho_\Sigma, H']$ shows that the time derivative of the particle number density is $\partial_t \rho_\Sigma(q) = (iq) gv \rho_\Delta(q)$, and therefore if we define the field $(2\pi)^{-1/2}\Phi(x)$ that counts the amount of charge left of x, $\Phi(x) = \sqrt{2\pi} \int^x dx' \rho_\Sigma(x')$, we see that $\partial_t \Phi(x) = gv\sqrt{2\pi}\rho_\Delta(x)$. It is customary to rename $\sqrt{2\pi}\rho_\Delta(x) = \partial_x \phi(x)$so that the Hamiltonian is given by

$$H' = \frac{v}{2} \int_{-L/2}^{L/2} dx \left[g^{-1} (\partial_x \Phi(x))^2 + g (\partial_x \phi(x))^2 \right].$$

Alternative notation is to denote $\partial_x \phi(x) = \Pi(x)$ to obtain

$$H' = \frac{v}{2} \int_{-L/2}^{L/2} dx \left[g^{-1} (\partial_x \Phi(x))^2 + g \Pi^2(x) \right].$$

Both these forms are often used in the literature.