Manual for DFT+GW+BSE and the related matters

Jinyuan Wu

February 27, 2026

Contents

Chapter 1

Preliminaries

1.1 Second quantization

Here are some handy formulae for building second-quantized Hamiltonian from wave functions or stuff like that that are generated by first-principle softwares.

1.1.1 For electrons

In second quantization, for fermions we have

$$\{c_i, c_j^\dagger\} = \delta_{ij}, \tag{1.1}$$

and Fock states are given by

$$|n_1, n_2, \ldots\rangle = (c_1^\dagger)^{n_1} (c_2^\dagger)^{n_2} \cdots |0\rangle . \tag{1.2}$$

This gives us the expected (-1) sign change when two fermions are switched. A single particle Hamiltonian

$$H_0 = \sum_{i=1}^{N} h_i \tag{1.3}$$

is to be rewritten as

$$H_0 = \sum_{\alpha, \beta} \langle \alpha | h | \beta \rangle \, c_\alpha^\dagger c_\beta, \tag{1.4}$$

and a two-particle Hamiltonian

$$H_{\mathrm{I}} = \frac{1}{2} \sum_{i \neq j=1}^{N} V_{ij} = \sum_{\text{pair } i,\, j} V_{ij} \tag{1.5}$$

is to be rewritten as

$$H_{\mathrm{I}} = \frac{1}{2} \sum_{\alpha, \beta, \delta, \gamma} \langle \alpha\beta | V | \delta\gamma \rangle \, c_\alpha^\dagger c_\beta^\dagger c_\delta c_\gamma. \tag{1.6}$$

Note that in the first-quantized Hamiltonian, i and j label particles, while in the second-quantized Hamiltonian, $\alpha, \beta, \delta, \gamma$ label single-particle wave functions. The $1/2$ factor is there to counter the double counting of ij and ji.

The representation switching formulae are

$$c_\alpha^\dagger = \sum_{\tilde\alpha} \langle \tilde\alpha | \alpha \rangle \, c_{\tilde\alpha}^\dagger \tag{1.7}$$

and

$$c_\alpha = \sum_{\tilde\alpha} \langle \alpha | \tilde\alpha \rangle \, c_{\tilde\alpha}. \tag{1.8}$$

Specifically, suppose $\{\alpha\}$ is a single-particle basis, we say

$$\psi^\dagger(\boldsymbol{x}) = \sum_\alpha \langle \boldsymbol{x} | \alpha \rangle \, c_\alpha^\dagger \tag{1.9}$$

and $\psi(\boldsymbol{x})$ are **field operators**. Most frequently, α is the momentum (in the Bloch representation) or the index of primitive unit cells or atoms (in the Wannier representation). When α is the momentum, the relation between the field operator and the creation operator is just Fourier transform. Usually we just choose the normalization of the field operators to satisfy

$$\{\psi(\boldsymbol{x}), \psi^\dagger(\boldsymbol{x}')\} = \delta(\boldsymbol{x} - \boldsymbol{x}'). \tag{1.10}$$

Under this convention, if we define

$$\psi_{\boldsymbol{k}} = \frac{1}{\sqrt{V}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}}, \tag{1.11}$$

we find

$$\{c_{\boldsymbol{k}_1}, c_{\boldsymbol{k}_2}^\dagger\} = \delta_{\boldsymbol{k}_1 \boldsymbol{k}_2}. \tag{1.12}$$

This convention should be kept in mind in the rest of this note. A large benefit of this convention is it lifts the burden to worry about normalization concerning δ-functions if we work in the momentum space: no δ-function appears in the rules now! To carry out Feynman diagrammatic calculations in an actual computer, a cutoff on the density of $\boldsymbol{k}$ points is needed (essentially, a cutoff on the size of the system V). This means we should replace $\int \mathrm{d}^3\boldsymbol{k}\,/2\pi$ with properly normalized $\sum_{\boldsymbol{k}}$, where $\boldsymbol{k}$ goes over the discretely infinite $\boldsymbol{k}$-grid. Also, this convention means the inner product is to be defined as

$$\langle \psi | \varphi \rangle = \int \mathrm{d}^3\boldsymbol{r} \, \psi^*(\boldsymbol{r}) \varphi(\boldsymbol{r}) \tag{1.13}$$

without any prefactors like $1/V$, where $\boldsymbol{r}$ goes over every point in the sample, and

$$\int \mathrm{d}^3\boldsymbol{r} \cdot 1 = V. \tag{1.14}$$

The above conventions have following consequences:

$$\langle \boldsymbol{r} | \boldsymbol{r}' \rangle = \delta(\boldsymbol{r} - \boldsymbol{r}'), \tag{1.15}$$

and

$$\frac{1}{V} \sum_{\boldsymbol{k}} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \Rightarrow \langle \psi | \varphi \rangle = \sum_{\boldsymbol{k}} \langle \psi | \boldsymbol{k} \rangle \, \langle \boldsymbol{k} | \varphi \rangle. \tag{1.16}$$

8

Note that the $\langle \boldsymbol{r}'|\cdot|\boldsymbol{r}\rangle$ matrix element of a local potential operator contains a $\delta(\boldsymbol{r} - \boldsymbol{r}')$ factor: we have

$$\langle \boldsymbol{r}'|\hat{\boldsymbol{r}}|\boldsymbol{r}\rangle = \boldsymbol{r}\,\langle \boldsymbol{r}'|\boldsymbol{r}\rangle = \boldsymbol{r}\delta(\boldsymbol{r} - \boldsymbol{r}').$$

This, however, doesn't cause any problem in ordinary calculations. For example we can verify that we have

$$\langle \psi|V(\boldsymbol{r})|\varphi\rangle = \int \mathrm{d}^3\boldsymbol{r}\,\psi(\boldsymbol{r})^*V(\boldsymbol{r})\varphi(\boldsymbol{r}),$$

without any normalization factors, and similarly

$$\langle \boldsymbol{r}|V|\psi\rangle = V(\boldsymbol{r})\psi(\boldsymbol{r}).$$

1.2   Details in diagrammatics

This section briefly goes through some tricky aspects of Feynman diagram techniques that may seem puzzling when we do concrete calculations.

1.2.1   Infinitesimals

There are some infinitesimals in Feynman rules that are often ignored. The first is about the illdefinedness of $\mathcal{T}\,\langle c(t)c^\dagger(0)\rangle$ when $t = 0$. We want to make the propagator to be the particle number (so that if we evaluate the tadpole diagram, we get the Hartree term). Therefore, the contribution of an electron line is

$$\xrightarrow{}_{k} \quad := \mathcal{T}\,\langle c_{\boldsymbol{k}}(t - 0^+)c_{\boldsymbol{k}}^\dagger(0)\rangle$$

$$= \int \frac{\mathrm{d}\omega}{2\pi}\mathrm{e}^{-\mathrm{i}\omega(t-0^+)}\underbrace{\frac{\mathrm{i}}{\omega - \xi_{\boldsymbol{k}} + \mathrm{i}0^+\,\mathrm{sgn}(\xi_{\boldsymbol{k}})}}_{\mathrm{i}G_0(\omega,\boldsymbol{k})} = \int \frac{\mathrm{d}\omega}{2\pi}\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{e}^{\mathrm{i}\omega 0^+}\mathrm{i}G_0(\omega, \boldsymbol{k}).$$

(1.17)

In the notation used here, $G(t, 0)$ is proportional to $\mathcal{T}\,\langle c(t)c^\dagger(0)\rangle$ and therefore is not well-defined when $t = 0$; the infinitesimal therefore is introduced explicitly by adding the $\mathrm{e}^{\mathrm{i}\omega 0^+}$ factor; some may *define* that $\mathrm{i}G$ is the correct propagator, and the $\mathrm{e}^{\mathrm{i}\omega 0^+}$ factor or the small time displacement is embedded into the definition of G.

The necessity of this $\mathrm{e}^{\mathrm{i}\omega 0^+}$ factor can also be seen by explicitly doing the integration: when $t = 0$, if we ignore the $\mathrm{e}^{\mathrm{i}\omega 0^+}$ factor, we get

$$\int \frac{\mathrm{d}\omega}{2\pi}\frac{\mathrm{i}}{\omega - \xi_{\boldsymbol{k}} + \mathrm{i}0^+\,\mathrm{sgn}(\xi_{\boldsymbol{k}})}.$$

This integral is not zero, but we want it to be zero when $\xi_{\boldsymbol{k}} > 0$, so we have to add a $\mathrm{e}^{\mathrm{i}\omega 0^+}$ factor to make the integrand approaches zero quickly enough in the upper plane, so we can construct an integration contour in the upper plane, in which there is no pole, and

$$\int_{|\omega|=R\gg 1}\mathrm{e}^{\mathrm{i}\omega 0^+}\frac{\mathrm{d}\omega}{2\pi}\frac{\mathrm{i}}{\omega - \xi_{\boldsymbol{k}} + \mathrm{i}0^+\,\mathrm{sgn}(\xi_{\boldsymbol{k}})} = 0.$$

Note that an interaction line should also receive the same mini-regularization outlined for Green functions, because it's obtained by integrating out some intermediate states. For bare Coulomb interaction this is not needed, because we don't have ω dependence in the potential, and it makes no sense to discuss the poles when we change ω. It does make sense to talk about retardation in the relativistic origin of Coulomb interaction: the Coulomb interaction is mediated by virtual photons, and is therefore proportional to the off-shell (i.e. $\omega \to 0$) limit of the photon propagator, which has $\omega^2 - \boldsymbol{q}^2 + \mathrm{i}0^+$ as the denominator, and we get

$$V(q) = \frac{4\pi e^2}{\boldsymbol{q}^2 - \omega^2 - \mathrm{i}0^+}. \tag{1.18}$$
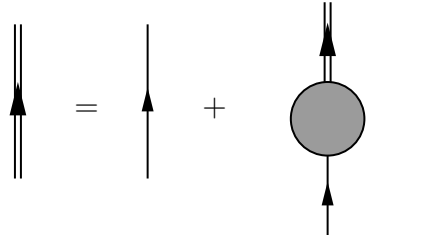
Also, for screened Coulomb interaction, the correct retardation is important, because now something looking like (1.18) appears again.

The 1^+ in $\Sigma(1,2) = \mathrm{i}W(1^+,2)G(1,2)$ seems to be from the infinitesimal in the electronic Green function instead of the infinitesimal in the Coulomb interaction line (§ 4.1.2).

1.2.2 The expression of Σ and W

In this section I only consider how many imaginary units there are in front of Green functions, self energies, etc. Normalization factors like 2π or V involved in summation of $\boldsymbol{r}$ or $\boldsymbol{k}$ are not considered.

The self-energy correction is visualized as the follows:

$$\tag{1.19}$$

and from it we have

$$\mathrm{i}G = \mathrm{i}G_0 + \mathrm{i}G_0\mathrm{i}G \times \quad .$$

It's then a good idea to define

$$-\mathrm{i}\Sigma = \quad , \tag{1.20}$$

because in this case, we have

$$G = G_0 + GG_0\Sigma, \tag{1.21}$$

and therefore

$$\underbrace{\omega - \xi_{\boldsymbol{k}}^0}_{1/G_0} = \underbrace{\omega - \xi_{\boldsymbol{k}}}_{1/G} + \Sigma, \tag{1.22}$$

which agrees with the definition of the self energy as the shift of single-particle energy from the free dispersion.

It should be noted that the derivations above are mainly about how to set the position of i correctly: I left the $e^{i\omega 0^+}$ factor (which can be found in (1.17)) out. After taking these factors into account, we find (1.21) becomes

$$e^{i\omega 0^+} iG = e^{i\omega 0^+} iG_0 + e^{i\omega 0^+} iG_0 \times \bigcirc \times e^{i\omega 0^+} iG. \tag{1.23}$$

We can eliminate the common $ie^{i\omega 0^+}$ factor (note that the magnitude of 0^+ doesn't matter, so a finite product of $e^{i\omega 0^+}$ collapses into a single $e^{i\omega 0^+}$), and thus we get (1.20) again; but note that for the same "collapse of finite product of $e^{i\omega 0^+}$" reason outlined above, the $e^{i\omega 0^+}$ factor in the RHS of (1.20), if any, can also be removed.

Similarly, we define the corrected interaction line as

$$-iW = \quad \lessgtr \quad , \tag{1.24}$$

because in this way, when there is no interaction corrections, we have

$$W = \frac{e^2}{r} =: v. \tag{1.25}$$

1.2.3 About "antiparticles"

The directions of momentum lines indicate whether the particles are created or annihilated. When the momentum arrow goes against the arrow on a line, we say this line is an antiparticle line. But here is a puzzle: we know there is no such thing as positrons in condensed matter physics, so what does "antiparticle" mean?

Here the problem lies on what it means to be the antiparticle of a kind of particle. In particle physics, when we do the particle-antiparticle transformation to an electron state whose polarization is one of the Dirac basis, a ψ^+ particle is flipped into a ψ^- particle and vice versa. In condensed matter physics, the label of ψ^+ and ψ^- (or ϕ and χ as people often call them: $\psi = (\phi, \chi)$) is no longer there: the χ modes in the Dirac field have already been integrated out. So electrons in condensed matter physics don't really have antiparticles in the context of high energy physics.

Indeed, if we are still dealing with scattering problems in the non-relativistic limit, the antiparticle lines don't appear at all! And similar to the case in QED (which can be checked in Peskin (A.6)), no separate momentum arrows parallel to the internal lines are needed: When calculating the propagator, the processes of both "an electron traveling forward" and "a hole traveling backward" are automatically covered together.

The antiparticle lines only appear when there are electrons in the ground state, which usually indicates there is a $-\mu N$ term in the Hamiltonian so having some preexisting electrons lower the energy further, and this differs with the scattering case only in the rules pertaining to the external lines. For external lines, we now have diagrams like the following:

,
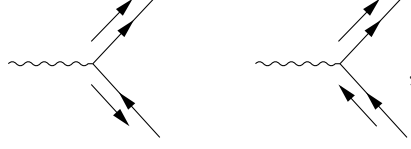
because now it's possible to annihilate a preexisting electron in the ground state, but for internal lines, momentum labels can still be directly attached to the internal lines. This can be also seen by reckoning how Feynman rules are derived: the series we obtain by expanding $\mathrm{e}^{-\mathrm{i}Ht}$ contains field operators, not single creation or annihilation operators, and after Wick expansion, the correlation functions we get are all like $\langle \bar{\psi}\psi \rangle$, and of course an annihilation operator appearing in the expression of the out state in terms of the ground state can be contracted with a creative operator in $\mathrm{e}^{-\mathrm{i}Ht}$, and this is visualized as an "antiparticle" external line with an outward momentum line. So here, the "particle-antiparticle transformation" is just swapping $c_{\boldsymbol{k}}$ and $c_{\boldsymbol{k}}^{\dagger}$ – this operation is still legit in condensed matter physics, because it doesn't involve the χ field; of course, the operation doesn't create that kind of antiparticle in high energy physics.

No real modification happens to the propagator when there are electrons in the ground state. We have

$$\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t\, \mathcal{T}\, \langle c_{\boldsymbol{k}}(t)c^{\dagger}(0)\rangle = \frac{\mathrm{i}}{\omega - \epsilon_{\boldsymbol{k}} + \mu}, \tag{1.26}$$

which can be straightforwardly obtained by looking at

$$H = \sum_{\boldsymbol{k}} \epsilon_{\boldsymbol{k}} c_{\boldsymbol{k}}^{\dagger} c_{\boldsymbol{k}} - \mu N \tag{1.27}$$

without doing any calculation.

Now we have to face the tough question: if antiparticle lines are there when there is a Fermi ball in the ground state, then why poles corresponding to antiparticles (whatever they are) are absent in the propagator? The answer is, for a $\boldsymbol{k}$ on an antiparticle line appearing in diagrams, the corresponding pole can indeed be understood as a pole of an antiparticle: for an antiparticle line with momentum $\boldsymbol{k}$, $\boldsymbol{k}$ has to be under the Fermi surface in the ground state, so $\omega_{\boldsymbol{k}} = \epsilon_{\boldsymbol{k}} - \mu < 0$, and the point $\omega = \omega_{\boldsymbol{k}}$ thus may be understood as an antiparticle pole. But here is a rather strong antisymmetry between particles and antiparticles: in external lines, when particles appear ($\boldsymbol{k}$ over Fermi surface), antiparticles never appear; when antiparticles appear ($\boldsymbol{k}$ below Fermi surface), particles never appear. The spectrum of electrons is split into two halves: for the part over the Fermi surface, only particles are visible, while for the part below the Fermi surface, only antiparticles are visible.

This means we can do away with antiparticle lines. By defining

$$b_{\boldsymbol{k}} = \begin{cases} c_{\boldsymbol{k}}, & \epsilon_{\boldsymbol{k}} > \mu, \\ c_{\boldsymbol{k}}^{\dagger}, & \epsilon_{\boldsymbol{k}} < \mu, \end{cases} \tag{1.28}$$

for $\epsilon_{\boldsymbol{k}} < \mu$, we have

$$\int_{-\infty}^{\infty} e^{i\omega t} \, dt \, \mathcal{T} \, \langle b_{\boldsymbol{k}}(t) b^{\dagger}(0) \rangle = \frac{i}{\omega - \mu + \epsilon_{\boldsymbol{k}}}, \tag{1.29}$$

and now all poles have positive energies. It's also easy to replace c operators in all interaction vertices with b operators, so now, in the theory in terms of b operators, there is no antiparticle poles or Feynman diagrammatic antiparticle lines. Indeed, b operators give the true free excitation spectrum in a system with a non-zero chemical potential.

For $\epsilon_{\boldsymbol{k}} < \mu$, $b_{\boldsymbol{k}}^{\dagger}$ is said to *create* a **hole**. The energy of a hole is still positive: the energy of a state with a hole with momentum $\boldsymbol{k}$ is

$$\sum_{\boldsymbol{k}' \neq \boldsymbol{k}, \epsilon_{\boldsymbol{k}'} < \mu} \epsilon_{\boldsymbol{k}'} - \mu(N-1),$$

and compared with the ground state, the energy of the hole is

$$\begin{aligned} E &= \sum_{\boldsymbol{k}' \neq \boldsymbol{k}, \epsilon_{\boldsymbol{k}'} < \mu} \epsilon_{\boldsymbol{k}'} - \mu(N-1) - \left(\sum_{\epsilon_{\boldsymbol{k}'} < \mu} \epsilon_{\boldsymbol{k}'} - \mu N \right) \\ &= \mu - \epsilon_{\boldsymbol{k}} > 0. \end{aligned} \tag{1.30}$$

So, we may say a hole is the antiparticle of an electron, but when we talk about the former, the latter is just a part of the background. Unlike the case in particle physics, where the electron and the positron are definitely two things, the hole and the electron are basically two *representations* of the *same* thing. (But this doesn't make talking about "annihilation between an electron and a hole" nonsense, because in an annihilation-between-electron-and-hole process, the n and $\boldsymbol{k}$ numbers of the electron and the hole are different, so there is no problem of double counting the same mode in two representations, etc.) Once we choose the picture in which there are holes, the electron modes corresponding to the holes should be be considered.

1.2.4 The direction of momentum on the interaction line

Box 1.1: Convention when defining Fourier transformation

In this note, when we define the Fourier transformation for a function, we always do the integral over the *variable name*. That's to say, since we have

$$W(\omega) = \int dt \, e^{i\omega t} W(t),$$

the Fourier integral of $W(-t)$ is

$$\int dt \, e^{i\omega t} W(-t).$$

Thus, we can say "changing t to $-t$ means changing ω to $-\omega$". This convention makes doing actual calculation easier.

What is demonstrated in the last section, essentially, is "changing the direction of momentum line attached to a propagator in a diagram doesn't create a new diagram". The same works for the bare interaction line: changing the 4-momentum from p to $-p$ doesn't create a new diagram. Thus, the two diagrams below are actually *one* diagram, and only one of them should be calculated, or otherwise we have double counting:
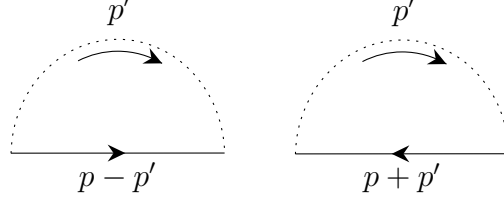
If we write down the expressions of these diagrams, we get

$$\int \frac{\mathrm{d}^4 p'}{(2\pi)^4} G(p + p') W(p'),$$

and

$$\int \frac{\mathrm{d}^4 p'}{(2\pi)^4} G(p - p') W(p') = \int \frac{\mathrm{d}^4 p'}{(2\pi)^4} G(p + p') W(-p').$$

To make the two diagrams equal to each other, what we need is $W(p) = W(-p)$ ...or do we? It should be noted that if we replace $W(\boldsymbol{p}, \omega)$ with $W(-\boldsymbol{p}, -\omega)$, the physical rules of the system *always* remains the same, as long as we have translational symmetries for time and space (if $\boldsymbol{p}$ is a lattice momentum, then we only need translational symmetry for the lattice). This is *not* due to the time reversal symmetry! Instead, this is due to the fact that $W(\boldsymbol{p}, \omega)$ only makes sense in

$$S_{\mathrm{int}} = \sum_{p_1, p_2, q} c^\dagger_{p_1+q} c^\dagger_{p_2-q} W(q) c_{p_2} c_{p_1}.$$

By manipulating the variables, we easily find

$$S_{\mathrm{int}} = \sum_{p_1, p_2, q} c^\dagger_{p_1-q} c^\dagger_{p_2+q} W(-q) c_{p_2} c_{p_1}$$

$$= \sum_{p_1, p_2, q} c^\dagger_{p_2+q} c^\dagger_{p_1-q} W(-q) c_{p_1} c_{p_2}$$

$$= \sum_{p_1, p_2, q} c^\dagger_{p_1+q} c^\dagger_{p_2-q} W(-q) c_{p_2} c_{p_1},$$

where in the first line, we change q into $-q$, and in the second line, we swap the order of the operators, and get a $(-1)^2 = 1$ factor, and in the third line, we swap the dummy variables p_1 and p_2. Thus we can see $W(q)$, although not necessarily *equal* to $W(-q)$, is always *equivalent* to $W(-q)$. $W(q) = W(-q)$ is of course true for the unscreened Coulomb interaction, and this directionless feature eventually originates from the directionless feature of the photon propagator; but it doesn't have to be true if we want to replace $W(q)$ with $W(-q)$: indeed, this means if $W(q) \neq W(-q)$, maybe it's a good idea to *symmetrize* $W(q)$.

Some additional comments are needed when we put the time variable in the frequency space but keep the spatial coordinates in the real space. The $W(q)$-equivalent-to-$W(-q)$ conclusion now is that $W(\boldsymbol{r},\boldsymbol{r}',\omega)$ equals to $W(\boldsymbol{r}',\boldsymbol{r},-\omega)$. We don't need to add a minus sign to $\boldsymbol{r}$ or $\boldsymbol{r}'$, but we need to swap them, so that

$$\int \mathrm{d}^3\boldsymbol{r}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot(\boldsymbol{r}-\boldsymbol{r}')}W(\boldsymbol{r}',\boldsymbol{r},-\omega) = W(-\boldsymbol{p},-\omega),$$

while

$$\int \mathrm{d}^3\boldsymbol{r}\, \mathrm{e}^{-\mathrm{i}\boldsymbol{p}\cdot(\boldsymbol{r}-\boldsymbol{r}')}W(\boldsymbol{r},\boldsymbol{r}',\omega) = W(\boldsymbol{p},\omega).$$

The question, then, is what corresponds to the time reversal symmetric. Although the time reversal operation doesn't change $\boldsymbol{r}$ and t, it changes the *order* of the coordinates: thus, $W(\boldsymbol{r},\boldsymbol{r}',\omega)$ equals $W(\boldsymbol{r}',\boldsymbol{r},\omega)$. Note that time reversal symmetry doesn't change ω, if we swaps $\boldsymbol{r}$ and $\boldsymbol{r}'$: this can be explicitly verified for the Green function. We have

$$\psi(\boldsymbol{x},t) \xrightarrow{T} \mathrm{i}\psi^\dagger(\boldsymbol{x},-t), \tag{1.31}$$

and therefore

$$\begin{aligned}
\mathcal{T}\,\langle \psi(\boldsymbol{x},t)\psi^\dagger(\boldsymbol{x}',0)\rangle &\xrightarrow{T} \mathcal{T}\,\langle \mathrm{i}\psi^\dagger(\boldsymbol{x},-t)\mathrm{i}\psi(\boldsymbol{x}',0)\rangle \\
&= -\mathcal{T}\,\langle \psi^\dagger(\boldsymbol{x},-t)\psi(\boldsymbol{x}',0)\rangle \\
&= \mathcal{T}\,\langle \psi(\boldsymbol{x}',0)\psi^\dagger(\boldsymbol{x},-t)\rangle.
\end{aligned} \tag{1.32}$$

When the time translational symmetry is present, the last line becomes $\mathcal{T}\,\langle \psi(\boldsymbol{x}',t)\psi^\dagger(\boldsymbol{x},0)\rangle$. This can be explained quite intuitively: after time reversal operation, a process in which the electron moves from $\boldsymbol{x}'$ to $\boldsymbol{x}$ becomes a process in which the electron from $\boldsymbol{x}$ to $\boldsymbol{x}'$, costing exactly the same amount of time. Thus after time reversal operation, we need to replace $G(\boldsymbol{x},t;\boldsymbol{x}',0)$ with $G(\boldsymbol{x}',t;\boldsymbol{x},0)$, and therefore we need to replace $G(\boldsymbol{r},\boldsymbol{r}',\omega)$ with $G(\boldsymbol{r}',\boldsymbol{r},\omega)$.

You may ask: so why can't we just replace t with $-t$? We can, actually: but now we *shouldn't* change the order of $\boldsymbol{r}$ and $\boldsymbol{r}'$! Let's have a look at what it means to change t to $-t$:

$$\begin{aligned}
W(\boldsymbol{r},t;\boldsymbol{r}',0) &\xrightarrow{t\to -t} W(\boldsymbol{r},-t;\boldsymbol{r}',0) \\
&\simeq W(\boldsymbol{r}',0;\boldsymbol{r},-t) \\
&= W(\boldsymbol{r}',t;\boldsymbol{r},0).
\end{aligned}$$

The second line comes from the aforementioned fact that the initial and end labels in W can be swapped, which may change the value but doesn't influence the final result; the third line comes from time translational symmetry. So we see keeping $\boldsymbol{r}$ and $\boldsymbol{r}'$ untouched and changing t to $-t$ is equivalent to keeping t untouched and swap $\boldsymbol{r}$ and $\boldsymbol{r}'$. Both representations of the time reversal operation are good: but remember only to do one of them – don't do both of them!

In the $(\boldsymbol{r},\boldsymbol{r}',\omega)$ representation, if we take the first representation of the time reversal operation, we have

$$W(\boldsymbol{r},\boldsymbol{r}',\omega) \xrightarrow{T} W(\boldsymbol{r}',\boldsymbol{r},\omega), \tag{1.33}$$

while if we take the second representation, we get

$$W(\boldsymbol{r}, \boldsymbol{r}', \omega) \xrightarrow{T} W(\boldsymbol{r}, \boldsymbol{r}', -\omega). \tag{1.34}$$

Note that $W(\boldsymbol{r}, \boldsymbol{r}', \omega)$ is always equivalent to $W(\boldsymbol{r}', \boldsymbol{r}, -\omega)$, but the former is only equivalent to $W(\boldsymbol{r}', \boldsymbol{r}, \omega)$ when we have time reversal symmetry.

The above discussion is about a time reversal symmetry in a single term in the interaction Hamiltonian. It's possible to have a term that doesn't have time reversal symmetry in H_{int}, but then the complex conjugate of the aforementioned complex term appears in another diagram, and changing the direction of the momentum line merely swaps two diagrams.

1.3 "Quantitative" Feynman rules

The above discussions are all pretty loose; in this section I will deal with Feynman diagrams more "quantitatively" and rigorously.

1.3.1 In the free space

The Coulomb interaction Hamiltonian, in the momentum space, is

$$H_{\text{e-e}} = \frac{1}{2} \sum_{\{\boldsymbol{k}_i\},\{\sigma_i\}} \int \mathrm{d}^3\boldsymbol{r} \int \mathrm{d}^3\boldsymbol{r}' \left\langle \boldsymbol{k}_1\sigma_1, \boldsymbol{k}_2\sigma_2 \left| \frac{e^2}{|\boldsymbol{r}-\boldsymbol{r}'|} \right| \boldsymbol{k}_3\sigma_2, \boldsymbol{k}_4\sigma_4 \right\rangle c^\dagger_{\boldsymbol{k}_1\sigma_1} c^\dagger_{\boldsymbol{k}_2\sigma_2} c_{\boldsymbol{k}_3\sigma_3} c_{\boldsymbol{k}_4\sigma_4}. \tag{1.35}$$

By definition (if you are unsure about how $\boldsymbol{k}_i$'s correspond to $\boldsymbol{r}$ and $\boldsymbol{r}'$, rewrite everything in terms of creation and annihilation operators) we have

$$\left\langle \boldsymbol{k}_1\sigma_1, \boldsymbol{k}_2\sigma_2 \left| \frac{e^2}{|\boldsymbol{r}-\boldsymbol{r}'|} \right| \boldsymbol{k}_3\sigma_3, \boldsymbol{k}_4\sigma_4 \right\rangle$$
$$= \frac{\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_1\cdot\boldsymbol{r}}}{\sqrt{V}} \frac{\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_2\cdot\boldsymbol{r}'}}{\sqrt{V}} \frac{e^2}{|\boldsymbol{r}-\boldsymbol{r}'|} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}_3\cdot\boldsymbol{r}'}}{\sqrt{V}} \frac{\mathrm{e}^{\mathrm{i}\boldsymbol{k}_4\cdot\boldsymbol{r}}}{\sqrt{V}} \delta_{\sigma_2\sigma_3}\delta_{\sigma_1\sigma_4},$$

and then we need to integrate over $\boldsymbol{r}$ and $\boldsymbol{r}'$. By redefining the variables, we have

$$\int \mathrm{d}^3\boldsymbol{r} \int \mathrm{d}^3\boldsymbol{r}' \left\langle \boldsymbol{k}_1\sigma_1, \boldsymbol{k}_2\sigma_2 \left| \frac{e^2}{|\boldsymbol{r}-\boldsymbol{r}'|} \right| \boldsymbol{k}_3\sigma_2, \boldsymbol{k}_4\sigma_4 \right\rangle$$
$$= \frac{1}{V^2} \delta_{\sigma_2\sigma_3}\delta_{\sigma_1\sigma_4} \int \mathrm{d}^3\boldsymbol{r} \int \mathrm{d}^3\boldsymbol{r}' \, \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}_1+\boldsymbol{k}_2-\boldsymbol{k}_3-\boldsymbol{k}_4)\cdot\boldsymbol{r}'} \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}_1-\boldsymbol{k}_4)\cdot(\boldsymbol{r}-\boldsymbol{r}')} \frac{e^2}{|\boldsymbol{r}-\boldsymbol{r}'|}$$
$$= \frac{1}{V} \delta_{\sigma_2\sigma_3}\delta_{\sigma_1\sigma_4} \cdot \delta_{\boldsymbol{k}_1+\boldsymbol{k}_2,\boldsymbol{k}_3+\boldsymbol{k}_4} \int \mathrm{d}^3\boldsymbol{r} \, \mathrm{e}^{-\mathrm{i}(\boldsymbol{k}_1-\boldsymbol{k}_4)\cdot\boldsymbol{r}} \frac{e^2}{|\boldsymbol{r}|}.$$

We can define

$$\boldsymbol{q} = \boldsymbol{k}_1 - \boldsymbol{k}_4 = \boldsymbol{k}_3 - \boldsymbol{k}_2, \tag{1.36}$$

and we find

$$H = \frac{1}{2V} \sum_{\boldsymbol{k},\boldsymbol{k}',\boldsymbol{q},\sigma,\sigma'} c^\dagger_{\boldsymbol{k}+\boldsymbol{q},\sigma} c^\dagger_{\boldsymbol{k}'-\boldsymbol{q},\sigma'} V(\boldsymbol{q}) c_{\boldsymbol{k}'\sigma'} c_{\boldsymbol{k}\sigma}, \tag{1.37}$$

where

$$V(\boldsymbol{q}) = \int \mathrm{d}^3\boldsymbol{r}\,\mathrm{e}^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}}\,\underbrace{\frac{e^2}{|\boldsymbol{r}|}}_{V(\boldsymbol{r})} = \frac{4\pi e^2}{\boldsymbol{q}^2}. \tag{1.38}$$

Note that the above formula can be generalized: we can change $V(\boldsymbol{r})$, and still get the momentum space Hamiltonian with the same shape of (1.37).

What's the Feynman rules corresponding to the TODO: n interaction line $\Rightarrow$ n independent momenta, not all of which are on Coulomb interaction lines

TODO: the $V \to \infty$ limit.

1.3.2 In a crystal

There are several things that need attention concerning Feynman rules in condensed physics:

- The propagator is no longer $1/(\omega - \boldsymbol{k}^2/2m)$, but $\propto \sum_{\boldsymbol{k}} \psi_{n\boldsymbol{k}}(\boldsymbol{r})\psi^*_{n\boldsymbol{k}}(\boldsymbol{r}')/(\omega - \xi_{n\boldsymbol{k}})$;
- Usually we limit all momenta to the **first Brillouin zone (1BZ)**, and therefore a sum over an unbounded momentum variable $\boldsymbol{k}$ or $\boldsymbol{q}$ has to be recast into $\sum_{\boldsymbol{G}}\sum_{\boldsymbol{k}} f(\boldsymbol{k} + \boldsymbol{G})$.

Below I describe a set of Feynman rules that are compatible with the notation in [20]. Note that there is no controversy over normalization in the *real space* Feynman diagrams: although the inner structure of $G(1, 2)$ has changed in the presence of the crystal potential, its relation with other diagrammatic components as abstract entities is never changed. Thus, starting from the space-in-real-space-and-time-in-frequency-space formalism may be a good idea: the free-space momentum formalism however is an important reference for normalization constants. In the space-in-real-space-and-time-in-frequency-space formalism, in a diagram containing n Coulomb interaction lines, the variables summed over are n frequency variables, and $2n$ space variables, and we just need to do

$$\int \frac{\mathrm{d}\omega_1}{2\pi}\cdots\int \frac{\mathrm{d}\omega_n}{2\pi}\int \mathrm{d}^3\boldsymbol{r}_1\cdots\int \mathrm{d}^3\boldsymbol{r}_{2n}. \tag{1.39}$$

We choose a normalization scheme for $\psi_{n\boldsymbol{k}}$ such that

$$c^\dagger_{n\boldsymbol{k}} = \int \mathrm{d}^3\boldsymbol{r}\,\underbrace{\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}u_{n\boldsymbol{k}}(\boldsymbol{r})}_{\psi_{n\boldsymbol{k}}}\psi^\dagger(\boldsymbol{r}), \quad \{c_{n\boldsymbol{k}}, c^\dagger_{n'\boldsymbol{k}'}\} = \delta_{nn'}\delta_{\boldsymbol{k}\boldsymbol{k}'}, \tag{1.40}$$

where $\psi^\dagger(\boldsymbol{r})$ is the complex conjugate of the non-relativistic electron field operator. This leads to the normalization condition

$$\int \mathrm{d}^3\boldsymbol{r}\,\psi^*_{n\boldsymbol{k}}(\boldsymbol{r})\psi_{n'\boldsymbol{k}'}(\boldsymbol{r}) = \delta_{nn'}\delta_{\boldsymbol{k}\boldsymbol{k}'}, \tag{1.41}$$

and

$$G^0(\boldsymbol{r}, \boldsymbol{r}', \omega) = \sum_{n,\boldsymbol{k}} \frac{\psi_{n\boldsymbol{k}}(\boldsymbol{r})\psi^*_{n\boldsymbol{k}}(\boldsymbol{r}')}{\omega - \xi_{n\boldsymbol{k}} + \mathrm{i}0^+\,\mathrm{sgn}(\omega)}. \tag{1.42}$$

In the free-electron case, we expect

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}) = \frac{1}{\sqrt{V}} e^{i(\boldsymbol{k}+\boldsymbol{G}_n)\cdot\boldsymbol{r}}. \tag{1.43}$$

This normalization scheme is used in this note, because it makes the anti-commutation algebra of creation and annihilation operators the simplest. Therefore (note that the only possibility for $\boldsymbol{G}_n + \boldsymbol{k}$ to be equal to $\boldsymbol{G}_{n'} + \boldsymbol{k}'$ is for $\boldsymbol{G}_n$ to be $\boldsymbol{G}_{n'}$ and $\boldsymbol{k}$ to be $\boldsymbol{k}'$)

$$\text{LHS of (1.41)} = \frac{1}{V} \int d^3\boldsymbol{r} \, e^{i\boldsymbol{r}(-\boldsymbol{G}_n-\boldsymbol{k}+\boldsymbol{G}_{n'}+\boldsymbol{k}')} = \frac{1}{V} V \delta_{\boldsymbol{G}_n+\boldsymbol{k},\boldsymbol{G}_{n'}+\boldsymbol{k}'} = \delta_{nn'}\delta_{\boldsymbol{k}\boldsymbol{k}'},$$

and

$$\text{RHS of (1.42)} = \frac{1}{V} \sum_{n,\boldsymbol{k}} e^{i(\boldsymbol{k}+\boldsymbol{G}_n)\cdot(\boldsymbol{r}-\boldsymbol{r}')} = \frac{1}{V} \sum_{\text{unbounded }\boldsymbol{k}} e^{i\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')}$$

$$= \int \frac{d^3\boldsymbol{k}}{(2\pi)^3} e^{i\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')} =: \text{LHS of (1.42)},$$

so (1.41) and (1.42) go back to the free-space case.

Now we consider what happens when we encounter a Coulomb interaction line. The structure in which a Coulomb interaction line is connected to four electron lines gives rise to the following factor in the interpretation of the diagram:

$$M = \int d^3\boldsymbol{r} \int d^3\boldsymbol{r}' \, iG(\boldsymbol{r}_4,\boldsymbol{r},\omega_1) iG(\boldsymbol{r}_3,\boldsymbol{r}',\omega_2) iG(\boldsymbol{r}',\boldsymbol{r}_2,\omega_3) iG(\boldsymbol{r},\boldsymbol{r}_1,\omega_4)(-i)v(\boldsymbol{r}-\boldsymbol{r}'). \tag{1.44}$$

Note that here ω_4 is just a shorthand of $\omega_1 + \omega_2 - \omega_3$ and this condition is *not* imposed by a δ-function factor according to the rules listed above for the free-space case (if we want to make this expression really symmetric, we should add a $2\pi\delta(\sum\omega)$ factor, and let, say, $\int d\omega_4/2\pi$ explicitly impose the energy conservation condition for us). Inserting (1.42) into the above expression, and using the definition

$$v(\boldsymbol{r}-\boldsymbol{r}') = \int \frac{d^3\boldsymbol{p}}{(2\pi)^3} e^{i\boldsymbol{p}\cdot(\boldsymbol{r}-\boldsymbol{r}')} v(\boldsymbol{p}), \tag{1.45}$$

we have

$$\begin{aligned}
M = \,& \psi_4(\boldsymbol{r}_4)\psi_3(\boldsymbol{r}_3)\psi_2^*(\boldsymbol{r}_2)\psi_1^*(\boldsymbol{r}_1) \\
& \times \int \frac{d^3\boldsymbol{q}}{(2\pi)^3} \sum_{1,2,3,4} \frac{i}{\omega_1 - \xi_4 + i0^+ \operatorname{sgn}(\omega_1)} \frac{i}{\omega_1 - \xi_3 + i0^+ \operatorname{sgn}(\omega_2)} \\
& \times \frac{i}{\omega_1 - \xi_2 + i0^+ \operatorname{sgn}(\omega_3)} \frac{i}{\omega_1 - \xi_1 + i0^+ \operatorname{sgn}(\omega_4)} \\
& \times \int d^3\boldsymbol{r} \, \psi_4^*(\boldsymbol{r})\psi_1(\boldsymbol{r}) e^{i\boldsymbol{q}\cdot\boldsymbol{r}} \int d^3\boldsymbol{r}' \, \psi_3^*(\boldsymbol{r}')\psi_2(\boldsymbol{r}') e^{-i\boldsymbol{q}\cdot\boldsymbol{r}'} v(\boldsymbol{q}),
\end{aligned} \tag{1.46}$$

where the label 1, 2, etc. mean $(n_1,\boldsymbol{k}_1), (n_2,\boldsymbol{k}_2)$, etc., which *doesn't* contain ω variables. Since V is always large enough, we have

$$\int \frac{d^3\boldsymbol{q}}{(2\pi)^3} = \frac{1}{V} \sum_{\boldsymbol{G}} \sum_{\boldsymbol{q}},$$

and eventually we get

$$M = \psi_4(\boldsymbol{r}_4)\psi_3(\boldsymbol{r}_3)\psi_2^*(\boldsymbol{r}_2)\psi_1^*(\boldsymbol{r}_1)$$
$$\times \frac{1}{V}\sum_{\boldsymbol{G}}\sum_{\boldsymbol{q}}\sum_{1,2,3,4} \frac{i}{\omega_1 - \xi_4 + i0^+\,\mathrm{sgn}(\omega_1)} \frac{i}{\omega_1 - \xi_3 + i0^+\,\mathrm{sgn}(\omega_2)}$$
$$\times \frac{i}{\omega_1 - \xi_2 + i0^+\,\mathrm{sgn}(\omega_3)} \frac{i}{\omega_1 - \xi_1 + i0^+\,\mathrm{sgn}(\omega_4)}$$
$$\times \langle 4|e^{i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|1\rangle \, \langle 3|e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|2\rangle \, v(\boldsymbol{q}+\boldsymbol{G}). \tag{1.47}$$

The structure of this expression reveals the Feynman rules for a crystal:

- For each external electron line, according to its direction, write down $\psi(\boldsymbol{r})$ (outgoing) or $\psi^*(\boldsymbol{r})$ (incoming).

- For each inner electron line, i.e. propagator, write down

$$\xrightarrow{\;\;n,k\;\;} \quad = \frac{i}{\omega - \xi_{n\boldsymbol{k}} + i0^+\,\mathrm{sgn}(\omega)} =: iG^0_{n\boldsymbol{k}}(\omega), \tag{1.48}$$

  where $k = (\omega, \boldsymbol{k})$.

- For each Coulomb interaction line, write

$$\cdots\cdots^{\,q,\,\boldsymbol{G}}\cdots\cdots \quad = -i\frac{1}{V}v(\boldsymbol{q}+\boldsymbol{G}) \tag{1.49}$$

  When an interaction line appears, enforce the momentum and energy conservation laws by hand. That's so say, when you see
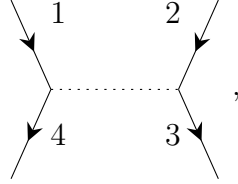
  we need to replace k_4 by $k_1 + k_2 - k_3$ manually: we just do a find-and-replace operation in the interpretation of this diagram. We *don't* integrate over k_4: k_4 has already been replaced in every part of the interpretation of this diagram and is not an integration variable.

- For each vertex – here I mean the two-electron-line-one-dotted-line structure, not the four-electron-line structure (i.e. not the diagrammatic component corresponding to H_{int}), we write $\langle \mathrm{out}|e^{\pm i\boldsymbol{r}\cdot(\boldsymbol{q}+\boldsymbol{G})}|\mathrm{in}\rangle$:

$$= \langle n, \boldsymbol{k}-\boldsymbol{q}|e^{-i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|n'\boldsymbol{k}\rangle =: M_{n'n}(\boldsymbol{k}-\boldsymbol{q},\boldsymbol{q},\boldsymbol{G})^*, \tag{1.50}$$

and

$$\langle n, \boldsymbol{k} + \boldsymbol{q}|e^{i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|n'\boldsymbol{k}\rangle =: M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}). \qquad (1.51)$$

The sign is decided by the direction of $\boldsymbol{q}$. The direction of $\boldsymbol{q}$ only matters when deciding the signs of momenta in momentum conservation: we *don't* sum over the two directions; thus, the above two Feynman rules are actually one rule, and this fact can be easily checked explicitly: (1.50), for example, can also be written as $M_{nn'}(\boldsymbol{k}, -\boldsymbol{q}, -\boldsymbol{G})$. The value of $\boldsymbol{q}$ is determined by the values of free momentum variables; if $\boldsymbol{q}$ is not selected as a free momentum variable, it shouldn't be integrated over. Here the normalization of the wave functions should be (1.41).

We can go further and rewrite $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$ in the $\boldsymbol{G}$-space, instead of the real space. It can be verified that under the following normalization

$$\psi_{n\boldsymbol{k}}(\boldsymbol{r}, \sigma) = e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \cdot \underbrace{\frac{1}{\sqrt{V}} \sum_{\boldsymbol{G}} c_{n\boldsymbol{k}}(\boldsymbol{G}, \sigma)e^{i\boldsymbol{G}\cdot\boldsymbol{r}}}_{u_{n\boldsymbol{k}}(\boldsymbol{r},\sigma)}, \qquad (1.52)$$

we have

$$M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) = \sum_{\boldsymbol{G}',\sigma} c^*_{n,\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{G}', \sigma)c_{n'\boldsymbol{k}}(\boldsymbol{G}' - \boldsymbol{G}, \sigma). \qquad (1.53)$$

The $\langle \text{out}|e^{\pm i(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|\text{in}\rangle$ are there to replace the real space integral

$$\int d^3\boldsymbol{r} \int d^3\boldsymbol{r}' \, \psi^*_m(\boldsymbol{r})\psi^*_n(\boldsymbol{r}')\frac{e^2}{|\boldsymbol{r} - \boldsymbol{r}'|}\psi_r(\boldsymbol{r}')\psi_s(\boldsymbol{r}),$$

which incorporates the influence of the crystal potential field.

Now we compare the above Feynman rules with the momentum space Feynman rules in § 1.3.1. Note that here we distribute the space and time parts of $(2\pi)^4$ differently: the space part, $(2\pi)^3$, becomes $1/V$, because we replace $\int d^3\boldsymbol{q}$ by $\sum_{\boldsymbol{G}}\sum_{\boldsymbol{q}}$.

Since there are n free momentum variables and n Coulomb interaction lines, we can attribute the $1/V$ factor to the interaction line, and we can also attribute it to the sum over $\boldsymbol{q}$. In [20], the $1/V$ factor comes together with $v(\boldsymbol{q} + \boldsymbol{G})$. Indeed, they define

$$v(\boldsymbol{q} + \boldsymbol{G}) = \frac{4\pi e^2}{V|\boldsymbol{q} + \boldsymbol{G}|^2}. \qquad (1.54)$$

The time part, 2π, now comes with integrations of ω's. Thus, a ring diagram should be interpreted as

$$\int \frac{d\omega}{2\pi} \sum_{m,n} iG^0_n(\boldsymbol{k}, \omega)iG^0_m(\boldsymbol{q} - \boldsymbol{k}, \omega_0 - \omega), \qquad (1.55)$$

where $q = (\omega_0, \boldsymbol{q})$.

Another way to make sense of the Feynman rules in this section is to write down the Coulomb interaction Hamiltonian in the crystal momentum representation before deriving Feynman rules. The interaction Hamiltonian – the Coulomb repulsion Hamiltonian – is

$$H = \sum_{1,2,3,4} c_4^\dagger c_3^\dagger H_{4321} c_2 c_1, \tag{1.56}$$

$$H_{4321} = \frac{1}{2} \int \mathrm{d}^3 \boldsymbol{r} \int \mathrm{d}^3 \boldsymbol{r}' \, \langle 4|\boldsymbol{r}\rangle \, \langle 3|\boldsymbol{r}'\rangle \, \langle \boldsymbol{r}'|2\rangle \, \langle \boldsymbol{r}|1\rangle \, v(\boldsymbol{r} - \boldsymbol{r}'), \tag{1.57}$$

Now again invoking the Fourier transformation from $v(\boldsymbol{r} - \boldsymbol{r}')$ to $v(\boldsymbol{q})$, and discretize the integration over $\boldsymbol{q}$, the $1/V$ factor appears again. The $1/2$ prefactor is canceled in the same way it's canceled in the free-space case.

1.3.3 Diagrammatic components as tensors

Both the Coulomb interaction line and the band electron propagator line contains one discrete index. And there are actually *two* indices: for an interaction-corrected Coulomb line, the two edges of a W line can have different $\boldsymbol{G}$ vectors, as is clearly shown below:
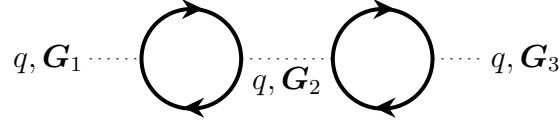
For the renormalized electron Green function, the same applies.

Note that for a free-space electron, the band index label n is essentially $\boldsymbol{G}_n$, but once the crystal potential and the self-energy is added, electrons can be scattered from one $\boldsymbol{G}$ to another. One may want to diagonalize the effective Hamiltonian, but now the band index after diagonalization no longer corresponds to the index of $\boldsymbol{G}$ vectors.
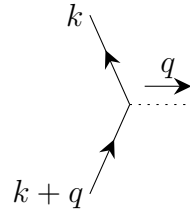
Thus, the $\boldsymbol{G}$- and n-summations mentioned in § 1.3.2 can be recast into matrix forms:

 $= \mathrm{i}G_{\boldsymbol{k}}^0(\omega) := \mathrm{i}[\delta_{nn'}G_{n\boldsymbol{k}}^0(\omega)]_{nn'}, \tag{1.58}$

 $= -\mathrm{i}v(q) := -\mathrm{i}[\delta_{\boldsymbol{GG}'}v(\boldsymbol{q} + \boldsymbol{G})]_{\boldsymbol{GG}'}, \tag{1.59}$

and

 $= M := [\, \langle n, \boldsymbol{k} + \boldsymbol{q}|\mathrm{e}^{\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}}|n'\boldsymbol{k}\rangle]_{nn'\boldsymbol{G}}, \tag{1.60}$

$$= M^* = [\, \langle n'\boldsymbol{k}| \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}} |n, \boldsymbol{k}+\boldsymbol{q}\rangle]_{nn'\boldsymbol{G}}. \tag{1.61}$$

How to contract the indices can be done by following the standard quantum mechanic convention: the index of what happens later appears on the left side.

1.4 Quasiparticles

1.4.1 Overview

In the context of *ab initio* calculations, the term quasiparticle usually means renormalized band electrons and holes (which may also include things like polarons). Bosonic modes are just called "excitations". Non-electron-like fermionic modes like spinons are simply absent here because up to now, most *ab initio* techniques only work for weakly correlated systems, so this terminology creates no confusion.

In most *ab initio* tasks, we expect the single-electron Green function to look like a "broadened" version of the free electron Green function. In this case, since $G \sim 1/E - H$, we say H is the single quasiparticle Hamiltonian, and the poles of the single-electron Green function are said to reflect the dispersion curve $\varepsilon_{n\boldsymbol{k}}$ of quasiparticles. The correction to the $\varepsilon_{n\boldsymbol{k}}$ is the most salient influence of Coulomb repulsion.

This however doesn't mean the spectrum of the system *only* contains quasiparticles: for example, as is demonstrated in RPA of the empty-lattice model, we always have the plasmon mode besides quasiparticles; in insulators, the bound state of an electron and a hole is called an **exciton**, again something not a quasiparticle but a legit component of the system's spectrum.

Even for quasiparticles, the effective single-electron energy is still dependent to the particle number distribution, and quasiparticle states have finite lifetime – they are not eigenstates of the system, after all. It should be noted in the full quantum treatment of the system, even when $T = 0$, we can still talk about damping: many-body wave functions containing a single quasiparticle are not strictly eigenstates of the system, so if we start with a single quasiparticle state, we will find that because of unitarity, probabilistic weight is "dissipated" into states we are not interested in or even can't describe. A physical picture is that quantum fluctuation makes scattering possible when $T = 0$.

> **Box 1.2: Quantum fluctuation**
>
> A more rigorous formulation of the idea is the follows: "quantum fluctuation" is a fancy word for "non-commuting terms in Hamiltonian". Each term has to be, of course, meaningful in a basis that has direct physical meaning: in the basis of the eigenstates of the total Hamiltonian, nothing is non-commutative; but in the basis of, say, Fock states, we have two terms in the total Hamiltonian, the single electron term and the Coulomb repulsion term, each of which has direct meaning but doesn't commute with the other. This

non-commutation means an eigenstate of the system can never be a Fock state, but a mix of Fock states. Thus $\langle$quasiparticle $n, \boldsymbol{k}|$ground state$\rangle \neq 1$, and an initial state containing one quasiparticle always evolves away, and we say "quantum fluctuation enables scattering".

The **Fermi liquid theory** describes the behavior of these (semi-)well-defined, bare electron or hole-like quasiparticles, if any. TODO: derive the follows

$$E = \varepsilon_{n\boldsymbol{k}} n_{n\boldsymbol{k}} + fnn. \tag{1.62}$$

The relation between diagrammatic components and the ε, f functions are given in [18].

1.4.2   The single-electron Green function

When Coulomb interaction is turned off, the single-electron Green function

$$G(\boldsymbol{r}, t, \boldsymbol{r}', t') = \langle 0 | \psi(\boldsymbol{r}, t) \psi^\dagger(\boldsymbol{r}', t') | 0 \rangle \tag{1.63}$$

is just the Green function of the single-electron Schrodinger equation (i.e. the EOM of the electron field operator), and because of the equation (1.63), we can use the Green function as a time-evolution operator:

$$\psi(\boldsymbol{r}, t) = \int \mathrm{d}^3 \boldsymbol{r}' \, G(\boldsymbol{r}, \boldsymbol{r}', t, t') \psi(\boldsymbol{r}', t'). \tag{1.64}$$

(1.63) – commonly referred to as i$G^>$ – is not the usual Green function used in § 1.2 and § 1.3, however, because it allows "evolution to the past". What is usually used is the time-ordered Green function

$$G_{\alpha\beta}(\boldsymbol{r}, t, \boldsymbol{r}', t') = -\mathrm{i} \langle \mathcal{T} \, \psi_\alpha(\boldsymbol{r}, t) \psi_\beta^\dagger(\boldsymbol{r}', t') \rangle. \tag{1.65}$$

The presence of the time-ordering operator makes the Wick theorem true and thus the time ordered Green function appears in Feynman diagrams as propagators. This Green function makes it easy to manipulate the relative order of the two field operators by adding an infinitesimal part to t and t'. For example we are able to find the electron density expectation

$$\langle n(\boldsymbol{x}) \rangle = -\mathrm{i} G(\boldsymbol{x}, t; \boldsymbol{x}, t^+). \tag{1.66}$$

Note that since the time-order operator is defined differently for bosons and fermions, when using infinitesimal part to control the relative order of the two field operators, we need to occasionally add a minus sign. For example, for bosons we have

$$\langle n(\boldsymbol{x}) \rangle = \mathrm{i} G(\boldsymbol{x}, t; \boldsymbol{x}, t^+). \tag{1.67}$$

We consider the operator version of the Wick theorem:

$$A_1 A_2 \cdots A_n =: A_1 A_2 \cdots A_n : + \sum_{i,j} \overline{A_i A_j} : A_1 \cdots A_{i-1} A_{i+1} \cdots A_{j-1} A_{j+1} \cdots A_n : + \cdots , \tag{1.68}$$

where

$$\overline{A_i A_j} = A_i A_j - :A_i A_j: \,. \tag{1.69}$$

If all the A_i operators are creation and annihilation operators, and we are working with a non-interactive system, then at the ground state, the expectation values of normal ordered operators all vanish, and contraction (1.69) is a number and not an operator, and thus we find

$$\langle A_1 A_2 \cdots A_n \rangle = \begin{cases} 0, & \text{with odd } n, \\ \sum_P (-1)^P \overline{A_{i_1} A_{i_2}} \cdots \overline{A_{i_{n-1}} A_{i_n}}, & \text{with even } n, \end{cases} \tag{1.70}$$

where P is the permutation used to map $1, 2, \ldots, n$ to $i_1, \ldots, i_n$. We can now apply the Wick theorem to a time ordered operator, and

The free electron Green function can be derived as follows: The ground state is the vacuum state and we have

$$\begin{aligned} G^{(0)}_{\boldsymbol{k}\sigma}(t - t') &= -\mathrm{i}\,\langle 0|\mathcal{T}\, c_{\boldsymbol{k}\sigma}(t) c^\dagger_{\boldsymbol{k}\sigma}(t')|0\rangle \\ &= -\mathrm{i}\theta(t - t')\mathrm{e}^{-\mathrm{i}\varepsilon_{\boldsymbol{k}}(t - t')}, \end{aligned} \tag{1.71}$$

because when $t < t'$, the time ordered operator in the bracket is $c^\dagger c$, and therefore vanishes after we take the expectation. Now we can do Fourier transform and get

$$G^{(0)}_{\boldsymbol{k}\sigma}(\omega) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}(\omega + \mathrm{i}\eta)(t - t')} \cdot G^{(0)}_{\boldsymbol{k}\sigma}(t - t')\, \mathrm{d}t = \frac{1}{\omega - \varepsilon_{\boldsymbol{k}} + \mathrm{i}\eta}, \tag{1.72}$$

where η is an infinitesimal quantity used to make the integral converge – here it should be 0^+.

We can also work with a ground state containing a Fermi sea: in this case we have

$$G^{(0)}_{\sigma}(\boldsymbol{k}, t - t') = -\mathrm{i}\theta(t - t')\theta(k - k_{\mathrm{F}})\mathrm{e}^{-\mathrm{i}\varepsilon_{\boldsymbol{k}}(t - t')} + \mathrm{i}\theta(t' - t)\theta(k_{\mathrm{F}} - k)\mathrm{e}^{-\mathrm{i}\varepsilon_{\boldsymbol{k}}(t - t')}. \tag{1.73}$$

The second term, when applied Fourier transform, also requires a convergence factor but not it should be 0^- because we are evolving backwards to $-\infty$. So in the end we get (note that the notation is changed for a form comparable to the notation in particle physics)

$$G^{(0)}_{\sigma}(\boldsymbol{k}, \omega) = \frac{\theta(k - k_{\mathrm{F}})}{\omega - \varepsilon_{\boldsymbol{k}} + \mathrm{i}0^+} + \frac{\theta(k_{\mathrm{F}} - k)}{\omega - \varepsilon_{\boldsymbol{k}} - \mathrm{i}0^+}. \tag{1.74}$$

1.4.3 Spectral density

When the interaction is turned off, the electron spectrum can be obtained by

$$\rho(\boldsymbol{k}, \omega) = -\frac{1}{\pi}\mathrm{sgn}(k - k_{\mathrm{F}})\,\mathrm{Im}\,G_{\sigma}(\boldsymbol{k}, \omega) = \delta(\omega - \xi_{\boldsymbol{k}}). \tag{1.75}$$

This means we can define

$$A_{\sigma}(\boldsymbol{k}, \omega) = -\frac{1}{\pi}\mathrm{sgn}(k - k_{\mathrm{F}})\,\mathrm{Im}\,G_{\sigma}(\boldsymbol{k}, \omega) \tag{1.76}$$

even when the interaction is turned on; if what Coulomb interaction does is merely self-energy correction, we can capture the self-energy corrected dispersion relation; if we end up in a strongly corrected system, the fact that no clear peak can be identified in $A_{\sigma}(\boldsymbol{k}, \sigma)$ gives us numerical evidence for that.

24

1.5 System of units

In the $T = 0$ case the physics is determined by the dimensionless time evolution operator $\mathrm{e}^{-\mathrm{i}Ht}$, and therefore the only important dimensionless quantities in condensed matter physics systems that determine the time evolution are

$$\frac{\hbar^2 \nabla^2}{2m_e} \frac{t}{\hbar}, \quad \frac{\hbar^2 \nabla^2}{2m_\mathrm{ion}} \frac{t}{\hbar}, \quad \frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \frac{t}{\hbar}, \quad \frac{1}{4\pi\epsilon_0} \frac{eQ}{r} \frac{t}{\hbar}, \quad \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{r} \frac{t}{\hbar}. \tag{1.77}$$

We expect the five dimensionless quantities to be the same *regardless of what system of units we use*. In principle, a simulation software only needs to care about the five quantities; the system of units is merely something that's needed when the program reports its output to a human readable form. Almost everyone will write the ion charges $Q_{1,2,...}$ as multiples of e, and m_ion can be given as its ratio with m_e, so what really matter in defining the unit are just $\frac{\hbar^2 \nabla^2}{2m_e} \frac{t}{\hbar}$ and $\frac{1}{4\pi\epsilon_0} \frac{e^2}{r} \frac{t}{\hbar}$, or, since we are only interested in the system of units, $\frac{\hbar^2}{2m_\mathrm{e}} \frac{1}{r} \frac{4\pi\epsilon_0}{e^2}$ and $H_\mathrm{Hydrogen\ atom} t/\hbar$.

We can define a system of units by assigning certain values to factors appearing in the two quantities mentioned right above; the definition of the new system of units therefore is given by the following statements: "in the the new system of units, this quantity has this fixed value, and that quantity has that fixed value, and the values of $\frac{\hbar^2}{2m_\mathrm{e}} \frac{1}{r} \frac{4\pi\epsilon_0}{e^2}$ and $H_\mathrm{Hydrogen\ atom} t/\hbar$ are the same as their values in SI". Step by step, we can assign values to the following quantities without causing a conflict and without leaving undefined units:

1. The unit of r i.e. all lengths.

2. $\hbar$, which appears in $\mathrm{e}^{-\mathrm{i}Ht/\hbar}$. When the units of quantities in H are all defined, $\hbar$ define the unit of time; alternatively for convenience we may want to have $\hbar = 1$, and this then imposes an additional constraint to the units appearing in $\frac{\hbar^2}{2m_\mathrm{e}} \frac{1}{r} \frac{4\pi\epsilon_0}{e^2}$.

3. $e^2/4\pi\epsilon_0$, which is the coupling strength.

4. e, not really important in the calculation but is used to report the charge density once the calculation is finished.

5. m_e, also not really important in the calculation because m_ion can always be given as multiples of m_e.

The most natural length unit seems to be based on the Bohr radius. Even if we know nothing about the exact form of the Bohr radius, by analyzing the structure of the hydrogen atom Hamiltonian, we have (below m means m_e)

$$\begin{aligned} H &= \frac{\hbar^2 \nabla^2}{2m} - \frac{e^2}{4\pi\epsilon_0} \frac{1}{r} \\ &= \frac{\hbar^2}{m} \left(\frac{1}{2} \frac{\partial^2}{\partial \boldsymbol{r}^2} - \frac{me^2}{4\pi\epsilon_0 \hbar^2} \frac{1}{r} \right) \\ &= \frac{me^4}{16\pi^2\epsilon_0^2\hbar^2} \left(\frac{1}{2} \frac{\partial^2}{\partial \left(\frac{me^2}{4\pi\epsilon_0 \hbar^2} \boldsymbol{r} \right)^2} - \frac{1}{\frac{me^2}{4\pi\epsilon_0 \hbar^2} r} \right), \end{aligned}$$

and naturally to have the beautiful form of $\nabla^2/2 - 1/r$ of the Hamiltonian (we still want the $1/2$ factor to remind us that the kinetic energy is $\boldsymbol{p}^2/2m$), we define the following quantity, which happens to be the Bohr radius, to be the atomic length unit:

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{me^2}. \tag{1.78}$$

That's to say, $\left.\frac{4\pi\epsilon_0\hbar^2}{me^2}\right|_{\text{our system of units}}$ should be one. This definition of the length unit in terms of $\epsilon_0, \hbar, m, e$ implicitly satisfies the constraint from $\frac{\hbar^2}{2m_e}\frac{1}{r}\frac{4\pi\epsilon_0}{e^2}$ because now

$$\frac{\hbar^2}{2m_e}\frac{1}{r}\frac{4\pi\epsilon_0}{e^2} = \frac{a_0}{r}$$

indeed is a dimensionless quantity without any dependence on the system of units. Next we require that under our system of units, $\hbar = 1$, so whatever the system of units for ϵ_0, m, e is, the unit of time is automatically given. So we have the constraint $\left.\frac{4\pi\epsilon_0}{me^2}\right|_{\text{our system of units}} = 1$. The next move to take is to decide what is $4\pi\epsilon_0/e^2$; a common practice is to use Gaussian system of units to remove all appearances of ϵ_0; the definitions of magnetic quantities in Gaussian units are different from those in SI, but for the electric part, to switch to Gaussian units we simply dictate that the arrangement of units has to make $4\pi\epsilon_0 = 1$.

Now we are left the unit of time, energy, mass and charge to define, the constraint that the unit of time is to be defined by setting $\hbar = 1$, and the constraint $\left.me^2\right|_{\text{our system of units}} = 1$. It seems we still have one additional choice to make. We take another move similar to the definition of the length unit and note that the Hydrogen ground-state energy

$$E = -\frac{e^2}{2a_0} \tag{1.79}$$

is given by the definition of e in our units. Hence, by choosing the value of e, or the unit of energy, or the value of m, we can finalize our derivation and arrive at one specific atomic system of units. The **Hartree atomic units** takes the following:

$$\hbar = m_e = e^2 = 4\pi\epsilon_0 = 1. \tag{1.80}$$

This means $E_{\text{H},n=1} = 0.5$ Hartree unit, and thus we find one Hartree atomic unit of energy is $2E_{\text{H},n=1} \approx 27.2\,\text{eV}$. On the other hand, we can also dictate that $E_{\text{H},n=1}$ is exactly one atomic unit of energy, and this means

$$\hbar = 2m_e = \frac{e^2}{2} = 4\pi\epsilon_0 = 1. \tag{1.81}$$

This is known as the **Rydberg atomic units**.

Chapter 2

Experimental characterization methods

2.1 Light-matter interaction

2.1.1 Interaction Hamiltonian

The non-relativistic minimal coupling is given by replacing $\boldsymbol{p}$ with $\boldsymbol{p} - q\boldsymbol{A}$. For electrons in a local potential field $V(\boldsymbol{r})$, the Hamiltonian

$$H = \frac{(\boldsymbol{p} - q\boldsymbol{A})^2}{2m} + V(\boldsymbol{r}) = \frac{(\boldsymbol{p} + e\boldsymbol{A})^2}{2m} + V(\boldsymbol{r}), \qquad (2.1)$$

and therefore the full light-matter interaction Hamiltonian is

$$H_{\text{light-matter}} = \frac{e\boldsymbol{p} \cdot \boldsymbol{A}}{m} + \frac{e^2 \boldsymbol{A}^2}{2m}. \qquad (2.2)$$

Usually, the double-photon process is ignored, and we get

$$H_{\text{light-matter}} = \frac{e\boldsymbol{p} \cdot \boldsymbol{A}}{m}. \qquad (2.3)$$

Note that here $\boldsymbol{p}$ is the original momentum operator, i.e. $-\mathrm{i}\boldsymbol{\nabla}$, and not the lattice momentum.

For pseudopotential calculations, as well as some self-energy methods, however, the potential field is no longer local. That's to say $V(\boldsymbol{r}, \boldsymbol{r}')$ is not diagonal in the coordinate representation. If we stick to writing the Hamiltonian in $(\boldsymbol{x}, \boldsymbol{p})$, then essentially this means we have $\boldsymbol{p}$ in V: the minimal coupling therefore changes V (and this is quite reasonable, since V includes the influence of inner electronic orbitals, which of course are also influenced by the external field), and now the light-matter interaction Hamiltonian contains infinite $\boldsymbol{A}^n$ terms, and even the lowest term is not as simple as (2.3).

2.1.2 Dipole approximation

Fortunately, usually we only work with light fields that are weak enough, and even when we talk about nonlinear response, usually it comes from *multiple appearance* of the external field line in one Feynman diagram, instead of *multiple-photon*

vertices. Moreover, usually the external field is smooth enough compared with the characteristic length of the system, i.e. lattice constants. This leads to the **dipole approximation**. The Hamiltonian is

$$H_{\text{light-matter}} = -\boldsymbol{d} \cdot \boldsymbol{E}. \tag{2.4}$$

It can be derived when the effect of the outside electromagnetic field is predominantly electrostatic, and the system in question is restricted to a relatively small region. With the above two approximations, we can attribute the light-matter interaction to

$$H_{\text{light-matter}} = q\varphi \approx \text{const} + q\boldsymbol{r} \cdot \boldsymbol{\nabla}\varphi \simeq -\underbrace{q\boldsymbol{r}}_{\boldsymbol{d}} \cdot \boldsymbol{E} = e\boldsymbol{r} \cdot \boldsymbol{E}. \tag{2.5}$$

Of course, when magnetic response of the system is important, we also need to add a magnetic dipole interaction term, etc.

The dipole approximation has another form

$$H_{\text{light-matter}} = e\boldsymbol{v} \cdot \boldsymbol{A}, \quad \boldsymbol{v} = \frac{1}{\mathrm{i}\hbar}[\boldsymbol{r}, H] \tag{2.6}$$

in radiation gauge $\boldsymbol{\nabla} \cdot \boldsymbol{A} = 0$. This can be derived as follows. In (2.5), we attribute the electric field to φ, and assume $\boldsymbol{A}$ is small (and thus the magnetic coupling can be ignored). But of course we can take the *radiational gauge*

$$\varphi = 0, \quad \boldsymbol{\nabla} \cdot \boldsymbol{A} = 0, \tag{2.7}$$

and this means

$$\boldsymbol{E} = -\frac{\partial \boldsymbol{A}}{\partial t}. \tag{2.8}$$

Note that this is valid only when the charges creating the electromagnetic field is not in the spatial region we are dealing with, or otherwise the two conditions conflict with each other: the $\varphi = 0$ condition leads to (2.8), but (2.8) immediately leads to $\frac{\partial}{\partial t}\boldsymbol{\nabla} \cdot \boldsymbol{E} = 0$, which generally doesn't hold if the source of the electric field is within the system in question; in condensed matter physics however this is definitely true, because the light source is always outside the sample. When the potential term is non-local, we can always rewrite $V(\boldsymbol{r}, \boldsymbol{r}')$ into $V(\boldsymbol{r}, \boldsymbol{p})$, and with appropriate normal ordering, we have

$$\frac{\partial}{\partial \boldsymbol{p}} V(\boldsymbol{r}, \boldsymbol{p}) = \frac{1}{\mathrm{i}\hbar}[\boldsymbol{r}, V].$$

Therefore the single-body Hamiltonian after electromagnetic coupling is

$$\begin{aligned}
H &= \frac{(\boldsymbol{p} + e\boldsymbol{A})^2}{2m} + V(\boldsymbol{r}, \boldsymbol{p} + e\boldsymbol{A}) \\
&\approx H|_{\boldsymbol{A}=0} + e\boldsymbol{A} \cdot \frac{\partial H}{\partial \boldsymbol{p}} \\
&= \frac{\boldsymbol{p}^2}{2m} + V(\boldsymbol{r}, \boldsymbol{p}) + e\boldsymbol{A} \cdot \dot{\boldsymbol{r}},
\end{aligned}$$

where

$$\dot{\boldsymbol{r}} = \boldsymbol{v} = \frac{1}{\mathrm{i}\hbar}[\boldsymbol{r}, H] = \frac{\boldsymbol{p}}{m} + \frac{1}{\mathrm{i}\hbar}\underbrace{[\boldsymbol{r}, V]}_{\neq 0}. \tag{2.9}$$

This is an illustration of why (2.3) itself doesn't suffice as the whole light-matter interaction Hamiltonian, even when the field is weak.

Still, in many cases the contribution of the non-local terms in the pseudopotential is small, and (2.3) is still a good approximation. In BerkeleyGW this is known as `use_momentum`; the velocity-based, theoretically correct coupling Hamiltonian is known as `use_velocity`.

2.2 The dielectric constant

2.2.1 Local field correction

Local field correction is often demonstrated by the Clausius-Mossotti relation or equivalently the Lorentz-Lorenz equation equation; this correction is absent in any component of the screening of the Coulomb interaction line; instead, it is to be found by integrating out microscopic *photon* modes, not any intrinsic degree of freedom in the condensed matter system.

Let's start with the most generalized theory. Suppose a rather smooth external electric field $V(\boldsymbol{r})$ is applied to the material; we want it to be smooth so that we can do dipole approximation, etc. Since the electric field is very smooth, we know that after doing Fourier transform on it and confining the momentum variable into the first Brillouin zone, we have

$$V(\boldsymbol{q}, \boldsymbol{G}) = V(\boldsymbol{q})\delta_{\boldsymbol{G}=0}, \tag{2.10}$$

as the wave vector $\boldsymbol{q}$ is never out of the first Brillouin zone. In practice, this condition is satisfied even for soft X-ray. The screened potential, then, is

$$V^{\text{screened}}(\boldsymbol{q}, \boldsymbol{G}, \omega) = \sum_{\boldsymbol{G}'} \epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}(\boldsymbol{q}, \omega) V(\boldsymbol{q}, \boldsymbol{G}') = \epsilon_{\boldsymbol{G}\boldsymbol{G}'=0}^{-1}(\boldsymbol{q}, \omega) V(\boldsymbol{q}). \tag{2.11}$$

Here in the symbol $\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}$ means to first take the inverse of ϵ and then find its $(\boldsymbol{G}, \boldsymbol{G}')$ matrix element. So now V^{screened} plays the role of $\boldsymbol{E}$ in the material, while V plays the role of $\boldsymbol{D}$, since the latter is only about "free charges" i.e. charges from outside of the material. But in real experimental measurement what is measured is not $V^{\text{screened}}(\boldsymbol{q}, \boldsymbol{G}, \omega)$, but its smoothed version: since we are working with crystal systems, a natural cutoff is the first Brillouin zone, and thus only components with $\boldsymbol{G} = 0$ contributes to the macroscopic value of V^{screened}. Thus we find that the macroscopic potential is

$$\bar{V}^{\text{screened}}(\boldsymbol{r}, \omega) = \frac{1}{V} \sum_{\boldsymbol{q}} e^{i\boldsymbol{q}\cdot\boldsymbol{r}} V^{\text{screened}}(\boldsymbol{q}, \boldsymbol{G}=0, \omega),$$

$$\bar{V}^{\text{screened}}(\boldsymbol{q}, \omega) = \int d^3\boldsymbol{r}\, e^{-i\boldsymbol{q}\cdot\boldsymbol{r}} \bar{V}^{\text{screened}}(\boldsymbol{r}, \omega) = (\epsilon^{-1}(\boldsymbol{q}, \omega))_{\boldsymbol{G}=0, \boldsymbol{G}'=0} V(\boldsymbol{q}). \tag{2.12}$$

Thus, the effective dielectric constant relating the screened potential and the unscreened potential is

$$\bar{\epsilon}(\boldsymbol{q}, \omega) = \frac{1}{(\epsilon^{-1}(\boldsymbol{q}, \omega))_{\boldsymbol{G}=0, \boldsymbol{G}'=0}}. \tag{2.13}$$

Since we first take the inverse of $\epsilon(\boldsymbol{q}, \omega)$ and then take the $(0, 0)$ matrix element, in $\bar{\epsilon}(\boldsymbol{q}, \omega)$, we have not only $(\epsilon_{00}(\boldsymbol{q}, \omega))^{-1}$ but also contributions from matrix elements

with higher $\boldsymbol{G}$ and $\boldsymbol{G}'$. What we are actually doing here therefore is to integrate out photon modes with non-zero $\boldsymbol{G}$ vectors – or to be accurate, we are integrating over Coulomb interaction lines with non-zero $\boldsymbol{G}$ vectors. Recall that $\epsilon = 1 - v\chi$, and v is the bare Coulomb interaction, we find (2.13) indeed comes from the following diagrammatic resummation:

In an amorphous solid a similar procedure can be performed, but this time the momentum cutoff has to be introduced in a more arbitrary way, since we no longer have first Brillouin zones and hence $\boldsymbol{G}$ vectors.

The next question, then, is in which case we are able to go back to the Clausius-Mossotti relation. [32] says for tightly bound electrons and the cubic lattice, the relation always hold.

If we are only interested in photons whose wave lengths are much longer than the lattice constants, then the dielectric function that relates the screened potential and the unscreened potential, i.e. (2.13), is the one that should appear in the Maxwell's equations about these photons. On the other hand, the dynamics of electrons naturally involve all $\boldsymbol{G}$ vectors, and screening for electrons should always be treated using $\epsilon_{\boldsymbol{G}\boldsymbol{G}'}(\boldsymbol{q}, \omega)$ which contains all $\boldsymbol{G}$ components. Therefore, $\bar{\epsilon}$ defined in (2.13) is often known as the **macroscopic dielectric function**, and the symbol ϵ^{M} is often used to refer to it.

We can repeat this process for complicated structures made of crystalline components as well. Assuming that the electric field is smooth for the atoms in each component, we get the macroscopic dielectric function for each component, and by applying an even stricter constraint and considering only electric field components that are smooth at the length scale of the whole structure, we get the macroscopic dielectric function for the whole structure.

2.3 Absorption

2.3.1 Absorption rate and $\mathrm{Im}\,\epsilon_{\mathbf{r}}$

Absorption is usually modeled by the imaginary part of the macroscopic dielectric function. In principle we can calculate it from $\epsilon_{\boldsymbol{G}\boldsymbol{G}'}(\boldsymbol{q}, \omega)$, but this requires us to consider the influence of excitons, etc. to the dielectric matrix as well, and a much easier way is to calculate it from Fermi golden rule. This is what is done in this section, and since we are only going to deal with the macroscopic dielectric function, the symbol ϵ is used to denote it. We assume the incident light has a determined wave length and polarization. The general form of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{\boldsymbol{k},\sigma} \sqrt{\frac{\hbar}{2\omega_{\boldsymbol{k}\sigma}\epsilon V}} (a_{\boldsymbol{k}\sigma}\hat{\boldsymbol{e}}_{\boldsymbol{k}\sigma} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}-\mathrm{i}\omega_{\boldsymbol{k}\sigma}t} + \mathrm{h.c.}), \tag{2.14}$$

and here we are only considering one $\boldsymbol{k}, \sigma$ mode; from then on we use $\boldsymbol{q}$ to refer to the momentum of the mode, and ω becomes a shorthand for $\omega_{\boldsymbol{k}\sigma}$. Similarly, n and $\hat{\boldsymbol{e}}$ mean $n_{\boldsymbol{q}\sigma}$ and $\hat{\boldsymbol{e}}_{\boldsymbol{q}\sigma}$. The reason for us to use a discrete momentum grid is that normalization is more convenient – and a numerical implementation of light-matter interaction of course is most easily carried out on a discrete momentum grid.

We do the usual real and imaginary refractive index decomposition

$$\sqrt{\epsilon_{\mathrm{r}}} = n + \mathrm{i}\kappa, \tag{2.15}$$

and therefore

$$\epsilon_{\mathrm{r}} = n^2 - \kappa^2 + 2\mathrm{i}n\kappa. \tag{2.16}$$

A beam of light therefore propagates as

$$\boldsymbol{E} \propto \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}\mathrm{e}^{-\mathrm{i}\omega t} \propto \mathrm{e}^{\mathrm{i}(n+\mathrm{i}\kappa)\frac{\omega}{c}r} = \mathrm{e}^{\mathrm{i}\frac{n\omega}{c}r}\mathrm{e}^{-\frac{\kappa\omega}{c}r}. \tag{2.17}$$

This means the intensity is like

$$I \propto |\boldsymbol{E}|^2 \propto \mathrm{e}^{-2\kappa\frac{\omega}{c}r}.$$

If we know α in

$$I = I_0 \mathrm{e}^{-\alpha r}, \tag{2.18}$$

then

$$2\kappa\frac{\omega}{c} = \alpha,$$

and therefore

$$\mathrm{Im}\,\epsilon_{\mathrm{r}} = 2n\kappa = \frac{cn}{\omega}\alpha. \tag{2.19}$$

The next question is what is α. (2.18) means in the distance of $\mathrm{d}r$, the photon number loss is proportional to

$$I(r) - I(r + \mathrm{d}r) = I(r)\alpha\,\mathrm{d}r\,.$$

This means

$$\alpha\,\mathrm{d}r = \frac{\text{photons absorbed per unit volume per second} \cdot \mathrm{d}r}{\text{photons incident per unit area per second}}. \tag{2.20}$$

> **Box 2.1: Capturing decaying of photon with a continuous theory?**
>
> There seems to be a contradiction here: Fermi golden rule looks "discrete": it gives the probability of photon absorption, while $\mathrm{Im}\,\epsilon$ gives us a continuous damping.
>
> Recall how we treat spontaneous radiation using a quantum jump formalism: damping here is introduced by two (related) factors, the first being $\mathrm{Im}\,H_{\mathrm{eff}}$, the second being the quantum jump channels, and once the shapes of the two terms in the master equation are determined, the strengths of which are "equal" to some extent.
>
> Now $\mathrm{Im}\,\epsilon$ is about $\mathrm{Im}\,H_{\mathrm{eff}}$, while Fermi golden rule is about the quantum jump channels, so it's indeed correct to infer $\mathrm{Im}\,\epsilon$ from Fermi golden rule,

using a procedure like, say, comparing the amount of light absorbed calculated from $\mathrm{Im}\,\epsilon$ and from Fermi golden rule. There is no double counting in this procedure: *both* $\mathrm{Im}\,H_\mathrm{eff}$ and Γ are needed for a full account of dissipation. H_eff continuously reduces the possibility to see the system staying in its original state, while quantum jump channels "confirm" that indeed the system decays to a lower energy state when the norm of the wave function has decreased considerably.

In the phenomenological model of spontaneous radiation, we first write down an H_eff, and then find the correct corresponding quantum jump channels to make the theory unitary, while here, we first calculate quantum jump channels (i.e. scattering) and then fit the H_eff according to the strength of scattering.

The numerator is now given by Fermi golden rule. In the below equation we use n to refer to the number of photons. We have

$$\text{photons absorbed per unit volume per second}$$

$$= \frac{1}{V} \cdot \text{photons absorbed per second}$$

$$= \frac{1}{V} \cdot \frac{2\pi}{\hbar} \sum_S |\langle S, n-1|e\boldsymbol{v} \cdot \boldsymbol{A}|0, n\rangle|^2 \delta(\omega - \Omega_S)$$

$$= \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon V} \cdot |\langle S, n-1|\boldsymbol{v} \cdot a_{\boldsymbol{k}\sigma}\hat{\boldsymbol{e}}\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}|0, n\rangle|^2 \delta(\omega - \Omega_S)$$

$$\approx \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon V} \cdot |\langle S, n-1|\boldsymbol{v} \cdot a_{\boldsymbol{k}\sigma}\hat{\boldsymbol{e}}|0, n\rangle|^2 \delta(\omega - \Omega_S)$$

$$= \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon V} \cdot |\boldsymbol{v} \cdot \hat{\boldsymbol{e}}\sqrt{n}\,\langle S, n-1|0, n-1\rangle|^2 \delta(\omega - \Omega_S)$$

$$= \frac{\pi e^2 n}{\omega\epsilon V^2} \sum_S |\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S),$$

where we have used the condition that the momentum of the incident photon is usually very small compared with the momentum of other excitations in a solid. Here the corrected dielectric constant $\epsilon = \epsilon_\mathrm{r}\epsilon_0$ enters the expression because it appears in (2.14). When absorption is strong, we need to replace ϵ_r with n^2. This can be verified by considering the energy flow in the material: when (2.17) is true (or, to be exact, when the electric field is (2.17) plus its complex conjugate: $\boldsymbol{E}$ is a real field and should stay so when we are dealing with energy, etc.; the damping plain wave ansatz (2.17) only works when we solve linear equations), we always have

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{E} \times \left(\frac{\boldsymbol{k}}{\omega} \times \boldsymbol{E}\right) = \frac{\boldsymbol{E}^2}{\mu_0\omega}\boldsymbol{k}, \langle\boldsymbol{S}\rangle = \frac{1}{2}\frac{\boldsymbol{E}_0^2}{\mu_0\omega}\boldsymbol{k}. \tag{2.21}$$

The real part of this is the "real" energy flow, which is now proportional to the imaginary part of $\boldsymbol{k}$, because as is said above, $\boldsymbol{E}$ now is real. We still want this to still be the energy density times the speed of light (c/n), and thus we find the energy density should be

$$w = \frac{1}{2}\epsilon_0 n^2 \boldsymbol{E}_0^2 = \epsilon_0 n^2 \langle\boldsymbol{E}^2\rangle. \tag{2.22}$$

This means we need to replace ϵ_0 in (2.14) with n^2, so that from (2.14) we are able to get (2.22). So we find (below we change the symbol for the photon number back to n_{photon})

$$\text{photons absorbed per unit volume per second} = \frac{\pi e^2 n_{photon}}{\omega \epsilon_0 n^2 V^2} \sum_S |\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S).$$
(2.23)

The denominator is

$$\text{photons incident per unit area per second} = \frac{n_{photon} c/n}{V}.$$

Here we need to put the refractive index into the expression of energy flow; we can verify this by again explicit evaluating $\boldsymbol{E} \times \boldsymbol{H}$ with an electromagnetic plain wave: the wave vector enters the expression because $\boldsymbol{H}$ is proportional to $\nabla \times \boldsymbol{E}$, and its relation with ω contains the speed of light in the material, which is corrected by the refractive index. So putting everything together, we get

$$\alpha = \frac{\pi e^2}{\omega \epsilon_0 n c V} \sum_S |\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S),$$
(2.24)

and from (2.19), we finally get

$$\epsilon_2 := \operatorname{Im} \epsilon_r = \frac{\pi e^2}{\omega^2 \epsilon_0 V} \sum_S |\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S).$$
(2.25)

It can be seen that the refractive index dependence in α is mysteriously canceled by (2.19). This shouldn't be surprising because the dielectric constant has the Feynman diagrammatic meaning of vacuum polarization times the photon propagator, and its imaginary part can be obtained via the optical theorem and is directly proportional to the scattering rate and has no refractive index dependence. The above derivation is based on the picture of matter interacting with photons, so we are working with transverse fields, but since the dipole approximation requires $\boldsymbol{q} \to 0$, where transverse and longitudinal modes can no longer be distinguished, (2.25) is also valid for longitudinal light-matter interaction. Now we see where it comes from: $|\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|$ is essentially the exciton wave function times the transition matrix $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G} = 0)$ divided by iq when $\boldsymbol{q} \to 0$, so (2.25) is just the imaginary part of the exciton propagator times two transition matrices times the Coulomb interaction q/q^2: this is just the definition of $\operatorname{Im} \epsilon^M$!

We see that to directly calculate $\operatorname{Im} \epsilon_r$ is actually easier than to directly calculate α. Indeed, the former is the usual scheme used in *ab initio* calculations.

Alternatively we may want to use the $\boldsymbol{r} \cdot \boldsymbol{E}$ interaction Hamiltonian. From the relation $\boldsymbol{E} = -\partial \boldsymbol{A}/\partial t$, the final expression of $\operatorname{Im} \epsilon_r$ can be directly obtained from (2.25) by removing the $1/\omega^2$ factor (note that the interaction Hamiltonian and hence $\boldsymbol{E}$ is squared in Fermi golden rule) and replacing $\boldsymbol{v}$ by $\boldsymbol{r}$, and we get

$$\operatorname{Im} \epsilon_r = \frac{\pi e^2}{\epsilon_0 V} \sum_S |\langle S|\boldsymbol{r} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S).$$
(2.26)

This expression has the same structure as (2.25); when we ignore electron-hole interaction, this expression gives exactly the same results of (2.25), because now

we have

$$\text{Im}\,\epsilon_{\text{r}}\bigg|_{\text{from }\boldsymbol{r}} = \frac{\pi e^2}{\epsilon_0 V}\sum_c^{\text{emp}}\sum_v^{\text{occ}}|\langle c\boldsymbol{k}|\boldsymbol{r}\cdot\hat{\boldsymbol{e}}|v\boldsymbol{k}\rangle|^2\delta(\omega - \varepsilon_{c\boldsymbol{k}} + \varepsilon_{v\boldsymbol{k}}) \qquad (2.27)$$

and

$$\text{Im}\,\epsilon_{\text{r}}\bigg|_{\text{from }\boldsymbol{v}} = \frac{\pi e^2}{\epsilon_0 \omega^2 V}\sum_c^{\text{emp}}\sum_v^{\text{occ}}|\langle c\boldsymbol{k}|\boldsymbol{v}\cdot\hat{\boldsymbol{e}}|v\boldsymbol{k}\rangle|^2\delta(\omega - \varepsilon_{c\boldsymbol{k}} + \varepsilon_{v\boldsymbol{k}}), \qquad (2.28)$$

and since the electron-hole two-body interaction is ignored, from

$$\boldsymbol{v} = \frac{1}{\text{i}\hbar}[\boldsymbol{r}, H_{\text{single electron}}]$$

we immediately find

$$(\varepsilon_{v\boldsymbol{k}} - \varepsilon_{c\boldsymbol{k}})|\langle c\boldsymbol{k}|\boldsymbol{r}|v\boldsymbol{k}\rangle| = |\langle c\boldsymbol{k}|\boldsymbol{v}|v\boldsymbol{k}\rangle|,$$

and because of the presence of the δ-function, the $1/\omega^2$ factor can be written as $1/(\varepsilon_{v\boldsymbol{k}} - \varepsilon_{c\boldsymbol{k}})^2$, and hence the values $\text{Im}\,\epsilon_{\text{r}}$ under the two choices of the dipole coupling Hamiltonian should be the same. The same happens in the interactive case: now $\boldsymbol{r}$ and $\boldsymbol{v}$ are many-body operators, and the relation between $\boldsymbol{v}$ and $\boldsymbol{r}$ is still $\boldsymbol{v} = [\boldsymbol{r}, H]/\text{i}\hbar$ (the Coulomb two-body term only contains $\boldsymbol{r}$), and from this relation we find

$$|\langle S|\boldsymbol{v}|0\rangle| = (E_0 - E_S)|\langle S|\boldsymbol{r}|0\rangle| = -\Omega_S|\langle S|\boldsymbol{r}|0\rangle|,$$

and hence again the two procedures to calculate $\text{Im}\,\epsilon_{\text{r}}$ result in the same value. The two Hamiltonians $\boldsymbol{r}\cdot\boldsymbol{E}$ and $\boldsymbol{v}\cdot\boldsymbol{A}$ are equivalent, after all.

The relation (2.25) is derived under SI. In Gaussian system of units, the form of ϵ_{r} remains unchanged. This can be seen by

$$\begin{cases} F = \dfrac{(e^{\text{I}})^2}{4\pi\epsilon_0 r^2} = \dfrac{(e^{\text{G}})^2}{r^2}, \quad \boldsymbol{F} = e^{\text{I}}\boldsymbol{E}^{\text{I}} = e^{\text{G}}\boldsymbol{E}^{\text{G}}, \\ \boldsymbol{D}^{\text{I}} = \epsilon_{\text{r}}^{\text{I}}\epsilon_0\boldsymbol{E}^{\text{I}}, \quad \boldsymbol{D}^{\text{G}} = \epsilon_{\text{r}}^{\text{G}}\boldsymbol{E}^{\text{G}}, \\ \boldsymbol{\nabla}\cdot\boldsymbol{D}^{\text{I}} = \rho^{\text{I}}, \quad \boldsymbol{\nabla}\cdot\boldsymbol{D}^{\text{G}} = 4\pi\rho^{\text{G}} \end{cases} \Rightarrow \epsilon_{\text{r}}^{\text{I}} = \epsilon_{\text{r}}^{\text{G}}.$$

So in Gaussian system of units, we have

$$\text{Im}\,\epsilon_{\text{r}} = \frac{4\pi^2 e^2}{\omega^2 V}\sum_S|\langle S|\boldsymbol{v}\cdot\hat{\boldsymbol{e}}|0\rangle|^2\delta(\omega - \Omega_S). \qquad (2.29)$$

Here we replace e^2/ϵ_0 in SI by $4\pi e^2$ in Gaussian units, since the two system of units agree with each other in the unit of force, and we have

$$\frac{e_{\text{SI}}^2}{4\pi\epsilon_0 r^2} = F_{\text{SI}} = F_{\text{Gaussian}} = \frac{e_{\text{Gaussian}}^2}{r^2}.$$

The expression of $\text{Im}\,\epsilon_{\text{r}}$ sees great variance in literature. Eq. (8) in [13] is almost the same as (2.29), except it misses the factor $1/V$. [14] claims the equation should be

$$\epsilon_2 = \text{Im}\,\epsilon = \frac{16\pi^2 e^2}{\omega^2}\sum_S|\langle S|\boldsymbol{v}\cdot\hat{\boldsymbol{e}}|0\rangle|^2\delta(\omega - \Omega_S).$$

The LHS here is shown as it is in [14]; since the system of units used is never explicitly mentioned in this article, we don't know the true meaning of ϵ here; the origin of the 4 factor is also mysterious. This version of (2.29) seems to have an early origin: it appears in [26], which refers us to [1] which doesn't seem to contain something like (2.29). Some of these variances are typos; others are just notational varieties. The procedure used to calculate ϵ_2 when electron-hole interaction is ignored (and Ω_S is just $\varepsilon_c - \varepsilon_v$) can be found in `absp0` in `BSE/absp0.f90` in BerkeleyGW's source code; some comments regarding the correct form of ϵ_2 are attached before the definition of the subroutine. From the codes and the attached comments, it can be seen that the missing of $1/V$ is not correct; half of the factor 4 comes from Rydberg atomic units, where $e^2 = 2$ (and thus the expression $16\pi^2 e^2$ is misleading – it should be $16\pi^2$, since e^2 is already absorbed into the 16 factor); half of the factor 4 comes from spin degeneracy, and when the two spins are considered separately (either when the spin index is treated explicitly as a good quantum number, or in spinor calculation where we have SOC), the 16 factor should be reduced to 8 (note that since the light is assumed to be linearly polarized, the spin of an electron remains the same before and after electron-photon scattering, and therefore the degeneracy factor is 2 instead of 4). So the formula used in BerkeleyGW – which also agrees with (2.29) – is

$$\epsilon_2 = \frac{16\pi^2}{\omega^2 N_{\text{spin-degeneracy}}} \sum_S |\langle S|\boldsymbol{v} \cdot \hat{\boldsymbol{e}}|0\rangle|^2 \delta(\omega - \Omega_S), \qquad (2.30)$$

where $N_{\text{spin-degeneracy}}$ is 1 when there is indeed spin degeneracy and only the orbit parts of the wave functions enter the calculation, and is 2 otherwise. Actually BerkeleyGW works on $\langle S|\boldsymbol{r}|0\rangle$, and the $1/\omega^2$ factor is not present in the source code – $\langle S|\boldsymbol{r}|0\rangle$ is calculated in subroutine `mtxel_m` and `mtxel_v` directly called by `compute_ik_vmtxel` and returned in the output argument `s0` of `mtxel_m` and `mtxel_v` and then passed to `dip%s1k` by `dip%compute_ik_vmtxel`, and then `dip%s1k` is flatten into `dip%s1` by the same subroutine; `dip%s1` is then convoluted with the exciton wave function to calculate `dipoles_r`, the expectation of $\boldsymbol{r}$ under an exciton mode in `diag`, and `dipoles_r` is eventually used to calculate ϵ_2. What is reported as the dipole in `eigenvalues.dat` files is also `s0` (see the subroutine `absp0` which calls `write_eigenvalues_noeh`).

2.4 Electron energy loss spectroscopy (EELS)

Absorption experiments measure energy loss in photon propagation and give us $\text{Im}\,\epsilon^{\text{M}}$. There is another experimental setting that will give use $\text{Im}\,1/\epsilon^{\text{M}}$: electron energy loss spectroscopy. The experimental protocol, in short, is to shed a beam of electrons to the sample and measure how the energies and momenta of the electrons change. The scattering rate is proportional to

$$= \text{Im} \left\{ \cdots \bigcirc\!\!X\!\!\bigcirc \cdots X \cdots \right\} = \text{Im}\left(\frac{1}{\epsilon}\right)_{\boldsymbol{G}=0,\boldsymbol{G}'=0} = \text{Im}\,\frac{1}{\epsilon^{\text{M}}}, \qquad (2.31)$$

where "X" means "exciton", the in-going and out-coming lines in the diagrams are external lines and do not contribute propagators, and we have assumed that the momentum shift of the electron in question is within the first Brillouin zone, and therefore we can choose $\boldsymbol{G} = \boldsymbol{G}' = 0$, and we find that the scattering rate is proportional to the imaginary part of $1/\epsilon^{\mathrm{M}}$.

We already have the imaginary part of ϵ^{M}, and from it the real part can be calculated by Kramers-Kronig relation, so the EELS spectrum should be related to the absorption spectrum.

Chapter 3

DFT

3.1 The discontinuity problem

Several options: [3, 34]
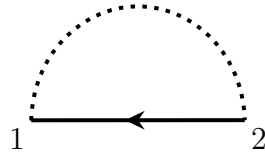
Chapter 4

GW and BSE

4.1   What is GW

4.1.1 GW is screened Hartree-Fock approximation

In short, GW means to consider the Hartree term and the screened Fock term

$$\text{(4.1)}$$

as self-energy diagrams. Thus, apart from the Hartree term, we have

$$\Sigma = \mathrm{i}GW, \tag{4.2}$$

where G is the renormalized Green function, and W is the renormalized (i.e. screened) Coulomb interaction, which is

$$W = \frac{v}{\epsilon} = \quad \cdots\cdots \quad + \quad \cdots \bigcirc \cdots \quad + \quad \cdots \bigcirc \cdots \bigcirc \cdots \quad + \cdots, \tag{4.3}$$

$$\epsilon = 1 - \cdots \bigcirc. \tag{4.4}$$

Compared with the exact Hedin equations, no vertex correction is taken into account here in GW.

It should be noted that the ϵ here is a matrix, but it contains no index of the ordinary 3-vector: it's a *scalar* in the real space, because it connects W and v, both of which are scalar potentials. Sometimes it's called the dielectric matrix, for it has two indices, $\boldsymbol{G}$ and $\boldsymbol{G}'$, that go over all $\boldsymbol{G}$-vectors. The name *dielectric matrix* is also used to refer to the relation between the vectors vector $\boldsymbol{E}$ and $\boldsymbol{D}$, which also has two indices, but the two indices are now spatial indices, and *this* dielectric matrix is a rank-2 tensor. The letter – the ϵ_{ij} – is related to $\epsilon_{\boldsymbol{GG}'}$, but depending on how much microscopic details are included in $\boldsymbol{E}$, the two may be different, between which we have so-called local field correction.

4.1.2 Infinitesimal displacement of time in GW

Suppose 1 means the output state of an electron and 2 the input state (so that $\Sigma(1,2)$ is the self-energy for an electron propagating from 2 to 1, and therefore we follow the quantum mechanic convention that the input state is on the right while the output state is on the left), we actually have

$$\Sigma(1,2) = iW(1^+,2)G(1,2), \tag{4.5}$$

instead of the schematic $\Sigma = iWG$. This expression can be seen in several places, including Wikipedia and [8, 18]. The necessity of this infinitesimal displacement on t_1 can be found in § 4.2.2.

The question, then, is where this displacement on t_1 comes from. My tentative conclusion is that the most first-principle interpretation of the screened Fock diagram is actually $iW(1,2)G(1^-,2)$, where the $-$ superscript comes from (1.17): $e^{-i\omega t}e^{i\omega 0^+} = e^{-i\omega(t^-)}$; since we can leave out the $e^{i\omega 0^+}$ factor in the self-energy (§ 1.2.2), in the time domain, we can adjust the out time in G from 1^- to 1; but since the relation between the out times of W and G is absolute, we also have to push the out time of W slightly forward, and the resulting expression of the GW self-energy becomes $iW(1^+,2)G(1,2)$.

In the frequency domain we can see this more clearly. Assuming we have time reversal symmetry so that the order of $\boldsymbol{r}$ and $\boldsymbol{r}'$ doesn't really matter (§ 1.2.4), we have

$$-i\Sigma(\omega, \boldsymbol{r}, \boldsymbol{r}', \omega) = \quad \overset{\displaystyle \omega'}{\underset{\boldsymbol{r} \quad \overset{\longleftarrow}{\omega+\omega'} \quad \boldsymbol{r}'}{\cdots\cdots}}$$

$$= \int \frac{\mathrm{d}\omega'}{2\pi} iG(\boldsymbol{r}, \boldsymbol{r}', \omega+\omega') e^{i(\omega+\omega')0^+}(-i)W(\boldsymbol{r}, \boldsymbol{r}', \omega'),$$

and since the input and output frequency is ω (as can be seen by energy conservation), according to the discussion in § 1.2.2, we remove the $e^{i\omega 0^+}$ factor on RHS and have

$$\Sigma(\boldsymbol{r}, \boldsymbol{r}', \omega) = i \int \frac{\mathrm{d}\omega'}{2\pi} e^{i\omega'0^+} G(\boldsymbol{r}, \boldsymbol{r}', \omega+\omega') W(\boldsymbol{r}, \boldsymbol{r}', \omega'). \tag{4.6}$$

This is exactly Eq. (4a) in [8].

Here is a tricky point: the sign of the exponent in $e^{i\omega'0^+}$ comes from the sign of ω' dependence in the G factor, not the appearance of ω' in the W factor. This can be seen in Eq. (10) in [20]. Indeed, we know that replacing $W(\omega')$ by $W(-\omega')$ doesn't change the value of any diagram after integrating over all internal variables (§ 1.2.4), so whether there is a minus sign in the W factor makes no difference – thus any sign dependence has to be about the sign of variables involved in the G part instead of the W part. This is consistent with the aforementioned fact that the minus superscript in $W(1^+,2)$, although appearing in the W factor, eventually comes from a rule concerning G.

4.1.3 GW compared with the Hartree-Fock approximation

Note that there *is* screening in self-consistent Hartree-Fock approximation: if we forget about the Fock term, then the Hartree approximation is essentially the same as Thomas-Fermi screening, which considers and only considers screening channels with respect to *density of electrons*, i.e. ring diagrams. Then we add the Fock term, and in the Fock term, there is still screening in the corrected propagator, but there is no screening in the Coulomb interaction line. (On the other hand, in the Hartree term, there shouldn't be any screening in the Coulomb interaction line, or otherwise we have double counting.)

In this perspective, GW is completely natural: the next level of correction is just to correct the Coulomb interaction line, using the same ring diagrams that appear in the self-consistent Hartree term.

4.1.4 From GW to BSE

Suppose we want to find the electromagnetic response of a system. What we want is something looking like $\delta^n n(\boldsymbol{r})/\delta V(\boldsymbol{r}')^n$. With the influence of $V(\boldsymbol{r})$ taken into account, $n(\boldsymbol{r}, t)$ is given by the following diagram:
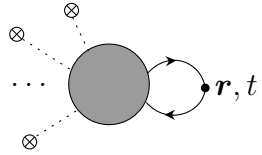
Since the coupling Hamiltonian between the external field with electrons is just ordinary Coulomb interaction, the first vertex a line from an external field encounters is a vertex with one interaction line and two electron lines. The nth-order response, therefore, is given by
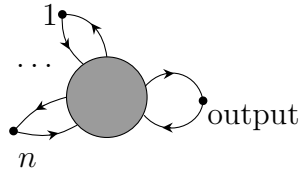
We can easily see that to find the nth-order response, it's necessary to find the nth-order correlation function. So we have two equivalent ways to deal with electromagnetic response: the first is to use Green function equations of motion and find $n(\boldsymbol{r}, t)$ explicitly, and then divide it with V; the other is to calculate all correlation functions; note that if we want not only the response, we shouldn't pair incoming and outgoing lines and fix the coordinates of each two of them to the same value: we should wipe out the solid dots in the last diagram.

Now suppose we want to work with the single-electron Green function G only in the first approach, and the GW self-energy plus the electromagnetic coupling are used to build the effective Hamiltonian. In other words, we are doing **time-dependent adiabatic** GW (for the meaning of "adiabatic", see below). The density $n(\boldsymbol{r}, t)$ is obtained by $G(\boldsymbol{r}, t; \boldsymbol{r}, t)$, or in other words, by connecting the incoming and outgoing lines to the $(\boldsymbol{r}, t)$ point. Note that Σ^{GW} depends on G, and when deciding what happens at $t + \Delta t$, we use $\Sigma^{GW}[G(t)]$. Here in order to

make the Hamiltonian-based time evolution possible, we actually need to do the static COHSEX approximation or something similar to eliminate the retardation effects in Σ^{GW}, or otherwise Σ^{GW} doesn't lead to a well-defined Hamiltonian, and therefore the screened Coulomb interaction line is no longer simply described by RPA – but anyway it relies on G. Also, the locality of the static COHSEX Hamiltonian in time means we need to approximate $G(\boldsymbol{r}, t; \boldsymbol{r}', t')$ by $G(\boldsymbol{r}, t; \boldsymbol{r}', t)$, and therefore we can use an effective single-electron density matrix to replace the Green function. That's why we say the approach is *adiabatic*. The question then is, what's the corresponding first-order response function? What diagrams are included in the latter?

To answer this question, we need to list all diagrams that contain *one and only one* external field line. For the sake of convenience, below we use grey lines to refer to the electron propagator corrected by Σ^{GW} and use black lines to refer to the electron propagator corrected by both Σ^{GW} and electromagnetic coupling. Thus we have

$$ \longrightarrow \ = \ \longrightarrow \ + \ \longrightarrow\!\!\!\longrightarrow \ + \cdots \tag{4.7} $$

In the first-order response theory, we only keep the first two terms. Now the first several terms in the completely corrected Green function are listed below:
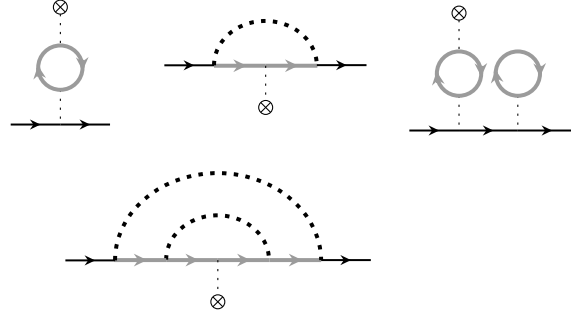
We can then wipe out the $\otimes$ symbol and the interaction line connected to it, and, as is mentioned above, to generalize things a little and separate the two electron lines connected to the interaction line introduced by the external field. The resulting diagrams are listed below:
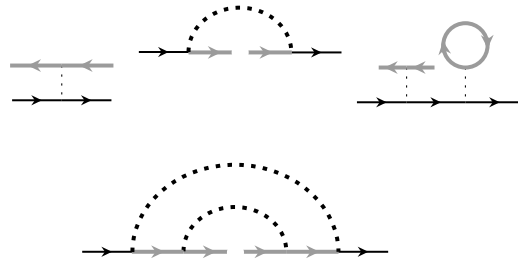
Here we ignore the possibility that we can plant an external field line into a propagator in W: this is equivalent to say that we are using the relation $K = \delta\Sigma^{GW}/\delta G$ to find a BSE kernel from the self energy, and when taking the derivative, W is assumed to be a constant [26].

Collecting the diagrams and summing over them, we find they are simply

The diagrams included here are essentially self-energy diagrams of the two-body Green function; in other words, the diagrams are self-energy diagrams of the electron-hole propagating. The corresponding self-energy diagrams – or, following the more common terminology, **kernel** diagrams – are

$$\blacksquare = \vdots + \rangle \cdots \langle . \tag{4.9}$$

The first term is associated with the screened Fock term in GW, because it comes from erasing one external field line in a Fock diagram; the second term corresponds to the Hartree term in GW. Note that the first term is screened and the second term is not: the reason is the same as the reason in § 4.1.3.

So the conclusion is, if we first do GW correction (without external field) to the electron propagator, and then feed the corrected electron propagator and (4.9) to

$$\bullet = \rightarrow + \blacksquare + \blacksquare\,\blacksquare + \cdots, \tag{4.10}$$

then the electromagnetic response calculated using this method is the same as the weak-field, constant W approximation of the time-dependent GW calculation.

> **Box 4.1: The term 'time-dependent'**
>
> The term "time-dependent" may be confusing: in a frequency-dependent GW calculation (generalized plasmon model, or full-frequency calculation), the time domain $G(t)$ is also time-dependent, but it can't capture electromagnetic response. On the other hand, in "time-dependent GW", the term "time-dependent" means the Hamiltonian of the system is time-dependent, which is equivalent to say coupling with the external field is taken into account.

(4.10) can be written in a way quite similar to single-electron self-energy correction, called the **Bethe-Salpeter equation (BSE)**, although in the context of *ab initio* study, BSE usually refers to (4.10) *plus* (4.9), not just the analytically correct (4.10): the vertex correction, again, is simply ignored.

The corrected two-body Green function i.e. corrected propagator of the electron-hole pair can be seen as generated by an effective two-body Hamiltonian, which contains the kinetic energy of the electron and the hole, as well as terms in (4.9). The renormalized electron-hole pair is called the **exciton**, and the effective two-body Hamiltonian is the Hamiltonian of the exciton. BSE – either in the general meaning or in the meaning of (4.9) – is currently the preferred way for analyzing excitons. The first term in (4.9) is known as the **direct term**, while the second term is known as the **exchange term**; note that the terminology may be kind of confusing, since the exciton exchange term actually comes from the GW Hartree term, and the exciton direct term comes from the GW screened-Fock term. In analytic works on excitons, usually only the direct term is considered, which is probably the reason for the terminology.

Two further remarks are needed on the effective Hamiltonian of excitons. First, we note that (4.9) is only for electron-hole pairs; diagrams for electron-electron pairs are different: there is no exchange term for electron pairs. This is illustrated
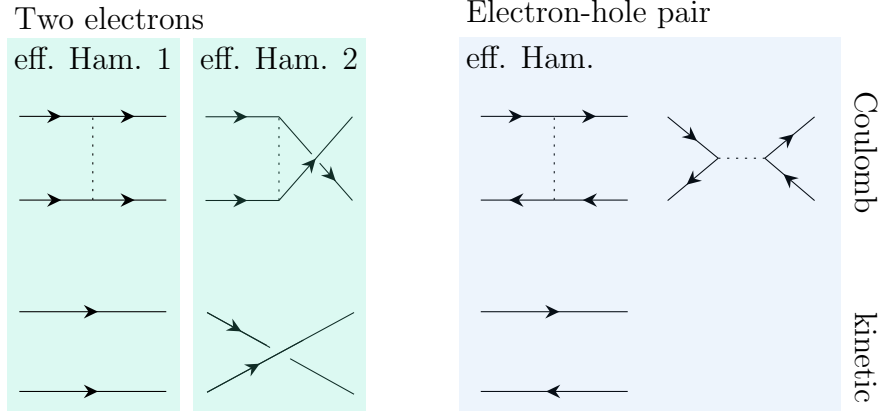
Figure 4.1: Comparison between Feynman diagrams for double electrons and Feynman diagrams for electron-hole pairs. Only the first order perturbation is considered; the interaction line in the direct terms should be screened.

in Fig. 4.1: in the case of two electrons, after doing the contractions, we get t- and u-channels, while in the case of electron-hole pairs, we get the direct term and the exchange term. This can be explicitly verified by calculating $\langle 1, 2|H_{\text{Coulomb}}|3, 4\rangle$ by definition. Since in the former case, the two input particles are of the same species, the two Coulomb interaction diagrams have to form a u-and-t pair, and when we write down the effective Hamiltonian, we only consider the u-channel or the t-channel, but never both, and thus in the effective interaction potential $V(\boldsymbol{r}_1, \boldsymbol{r}_2)$, we only have the direct term, and only one channel – the fact that there are two channels can be automatically revealed once we surround $V(\boldsymbol{r}_1, \boldsymbol{r}_2)$ between a pair of bra-ket. In the latter case, i.e. the case of excitons, we no longer have the u-and-t-pair, and we only have one kinetic term anyway. If we want to model the exciton by an effective potential, $V_{\text{eff}}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ contains both terms in (4.9). The second remark is that the effective Hamiltonian obtained from (4.9) is a *two-body* Hamiltonian, *not* a *many-body* Hamiltonian, that's to say, not an effective interaction vertex: if we want to find an effective interaction vertex, (a) electron modes appearing at in- and out-lines of (4.9) should not be included in the screened interaction line, and (b) the exchange term should not appear. That's what is done in constrained RPA, but not (4.9). Similar phenomena can be observed in trion calculations, where among the two electrons and one hole, the diagrams depicting interaction between the hole and one of the electrons are the BSE diagrams shown in (4.9), while the diagrams about interaction between the electrons are those in the left column in Fig. 4.1.

4.1.5 Remark on non-equilibrium dynamic GW and equilibrium multi-particle formalisms

We have shown that the time-dependent GW method can capture exciton effects to electromagnetic response of the system, because in the weak field limit it is equivalent to BSE. Thus the time-dependent GW method outlined above is also sometimes called the **time-dependent BSE**, although the name is slightly misleading, as if the *four-particle* Green function is the dynamic variable. Both flavors of BSE have been implemented: the exciton Hamiltonian approach is re-

alized by BerkeleyGW; the time-dependent GW approach doesn't seem to have mature open source realization (TODO: really?), but its accuracy has been verified by comparison with the exciton Hamiltonian approach [5].

Most of the time, BSE (time-independent, of course) is preferred, but there are cases time-dependent GW is good to use. Here we provide a demonstration of the equivalence between the non-equilibrium GW formalism and BSE (plus further equilibrium, multi-particle excitation formalisms), and how exciton effects are more conveniently captured by dynamic GW. By calculating $G^<$ using time-dependent GW, we are able to get the ARPES signature that comes from the electron in the exciton created by optical pumping [25, 27], and on the other hand, this of course can be captured by the more traditional Fermi golden rule approach as well. In the lowest order approximation, we assume that there is only one exciton in the system, and the Fermi golden rule contains a $\sum_v |\langle c, v|\text{exciton from light field}\rangle|^2$ factor. By optical theorem we have (note that we don't put Im before the RHS, because we are dealing with $G^<$)

$$\sum_v \left| \times\!\!\!\sim\!\!\!\left(\boxed{X}\right)\!\!\!\begin{array}{c} c \\ v \end{array} \right|^2 \sim \quad c \xleftarrow{} \boxed{X} \xleftarrow{v} \boxed{X} \xleftarrow{} c \quad , \tag{4.11}$$

where X represents the LHS of (4.10) where the kernel in the RHS is (4.9); if we think for a while and expand the RHS by definition, we will soon find that the RHS is just the lowest order terms in the Dyson series of $G^<$ under (time-dependent, non-equilibrium) GW approximation, and hence the consistency between the Fermi golden rule and the formalism based on $G^<$ is verified.[1] We can further see the consistency between the two descriptions. In the $G^<$ formalism, when a strong external light pump is turned on for a while and then turned off, excitons proliferate when probing happens, and in this case we have "self-driven" Floquet effect, caused by the $e^{-iE_{S\boldsymbol{Q}}t}$ factor in $G^<$ [10]. The energy conservation relation

$$E_{S\boldsymbol{Q}} = \omega - \underbrace{\epsilon_{v,\boldsymbol{k}-\boldsymbol{Q}}}_{<0} \tag{4.12}$$

that appears naturally in the LHS in (4.11) is equivalent to the $m = 1$ Floquet effect.

Now we turn to the reason why time-dependent GW is preferred. The first reason is that for ARPES experimental settings with short probe pulses instead of a continuous plane wave, where we don't have a δ function factor, the $G^<$-based RHS of (4.11) still works, while the LHS is no longer correct since Fermi golden rule fails in this case. Another problem can be seen immediately from the aforementioned Floquet phenomena: there can be infinite Floquet quasi-bands in principle, although the effects of most of them are weak, but these signatures are

---

[1] Note that here we are considering contributions of $G^<_{nn}$ to the ARPES spectrum, and we can't have an exciton made by an electron and a hole from the same band mode, in the RHS of (4.11) there are two exciton diagrams. This can be further verified by the expression of the ARPES spectrum in [27], which contains something like $\langle \Psi_n | c^\dagger_{\boldsymbol{k}} U(t_2, t_1) c_{\boldsymbol{k}} | \Psi_n \rangle$, where the exciton wave function appears twice.

absent in Fermi golden rule; and also in Floquet physics we have band renormalization which is also absent in Fermi golden rule.[2] In principle we can capture the effect from some extension of Fermi golden rule: the initial state in Fermi golden rule is actually not an exciton state calculated directly from (equilibrium) BSE, but has already been renormalized by the pump, and may contain exciton modes with different frequencies and multi-particle excitations. Diagrammatically, these high-order Floquet signatures are given by, for example,[3]
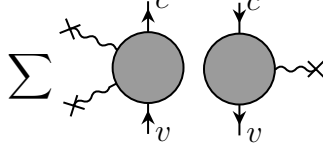
Also, it's in principle possible to have cross terms between different modes, as in

$$\tag{4.13}$$

This is related to band renormalization effects: Floquet corrected state that's closest to $|c_1\rangle$ may also have $|c_2\rangle$ components; note that in ordinary ARPES, since the transition probability takes the form of $\sum_m |\langle \Psi_m^{N-1}|H_{\mathrm{dipole}}|\Psi^N\rangle|^2$, if the dipole interaction Hamiltonian is expanded in the band basis, then the two c modes in the diagram above are the same. The most important effect of band renormalization however is the shift of the dispersion relation, which is not considered in the original formulation of Fermi golden rule, where the energies of the input and output states are assumed to be not influenced by scattering (that's to say, only $\mathrm{Im}\,\Sigma$ is considered, while $\mathrm{Im}\,\Sigma$ is not, similar to the assumptions leading to Boltzmann equation). Correcting the dispersion relation on top of Fermi golden rule is hard: now a single diagram with the form of (4.13) can no longer be regarded as a probability, and we have to sum over all of them to find the rate of electron being driven out. This means we can't just include a δ function to (4.13) intuitively as in Fermi golden rule: we have to sum over all possible $\mathrm{e}^{-\mathrm{i}\Delta E t}$ factors that come with diagrams in the form of (4.13) and then integrate over them, getting a final Fermi golden rule-like δ function.[4] It can be seen that capturing the Floquet effect is hard – although still possible – and therefore the explicit Floquet formalism based on $G^<$ is preferred.

---

[2]The Floquet band renormalization effects are not the only exciton-related band renormalization effects: we can also replace the Coulomb line in the Hartree diagram by a BSE T matrix, letting the two electron lines interact with each other for several times with the screened Coulomb potential. This is equivalent to connecting the in- and out-ends of the hole line in the BSE diagram, instead of connecting the in-ends of the electron and hole lines; the latter is what is treated here. See for example the supplementary materials of [21].

[3]Even when the external pumping is turned off when ARPES measurement happens, since we are dealing with a non-equilibrium state, the external field lines in (4.11) are still there; this is correct, as we can also conceive the excitons in the system as external driving fields.

[4]What is done here is similar to treating a time-independent perturbation as a time-dependent one, and trace the time evolution of a state in the unperturbed basis using time-dependent perturbation theory; if you do the derivation you will find the EOM of one coefficient (or one

4.2 The dielectric matrix ϵ

TODO:

- Truncated Coulomb interaction
- Why we say $\Sigma = V_{\mathrm{xc}} + \Sigma - V_{\mathrm{xc}}$? (see [14] p. 1271)
- What should be symmetric? (11) and (12) in [14]???

4.2.1 Frequency-dependent form

To build concrete expressions from the above floppy arguments, let's apply the rules in § 1.3.3. First, we define

$$q \cdots\!\!\bigcirc\!\!\cdots q = \mathrm{i}\chi(q) \tag{4.14}$$

Here the dotted lines are dummies: they contribute nothing to the value of the diagram. Note that χ is a matrix, with its elements taking the form of $\chi_{\boldsymbol{GG'}}$.

Now we have (every letter in the expression below is a matrix, so pay attention to the order)

$$-\mathrm{i}W(q) = -\mathrm{i}v(q) + (-\mathrm{i}v(q))(\mathrm{i}\chi(q))(-\mathrm{i}v(q)) + \cdots,$$

$$W(q) = v(q) + v(q)\chi(q)v(q) + v(q)\chi(q)v(q)\chi(q)v(q) + \cdots = (1 - v(q)\chi(q))^{-1}v(q).$$

Here we choose the convention that includes the $1/V$ factor into $v(\boldsymbol{q} + \boldsymbol{G})$ to avoid writing $1/V$ for too many times – the normalization is important – see the discussion below about [14]. So we get a concrete expression of ϵ:

$$\epsilon(q) = 1 - v(q)\chi(q), \quad W(q) = \epsilon(q)^{-1}v(q). \tag{4.15}$$

Because

$$(v(q)\chi(q))_{\boldsymbol{GG'}} = \sum_{\boldsymbol{G''}} v(q + \boldsymbol{G''})\delta_{\boldsymbol{GG''}}\chi_{\boldsymbol{G''G}}(q).$$

we have

$$\epsilon_{\boldsymbol{GG'}}(q) = \delta_{\boldsymbol{GG'}} - v(q + \boldsymbol{G})\chi_{\boldsymbol{GG'}}(q). \tag{4.16}$$

Here the inverse of ϵ is matrix inverse, not element-by-element inverse. (13) in [14] is ambiguous: a clearer but more cumbersome notation is

$$[\epsilon_{\boldsymbol{GG'}}(q)]^{-1}_{\boldsymbol{GG'}}v(\boldsymbol{q} + \boldsymbol{G'}).$$

---

lesser Green function) depends on other coefficients/lesser Green functions, and a close-form expression of (4.13) can't be obtained by definition: you have to solve the system of equations which mathematically is equivalent to doing the diagonalization.

Now it comes the evaluation of $\chi_{\boldsymbol{GG}}(q)$. Under the quasiparticle assumption, that the renormalized Green function can still be written as (1.42), We have

$$
i\chi_{\boldsymbol{GG'}}(\boldsymbol{q}, \omega) = q, G \dashrightarrow \underset{n',k}{\overset{n,k+q}{\bigcirc}} \dashrightarrow q, G'
$$

$$
= -\int \frac{\mathrm{d}\omega'}{2\pi} \sum_{\boldsymbol{k}} \sum_{n,n'} \frac{i}{\omega' - \xi_{n'\boldsymbol{k}} + i0^+ \operatorname{sgn}(\xi_{n'\boldsymbol{k}})}
$$

$$
\times \frac{i}{\omega + \omega' - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+ \operatorname{sgn}(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}
$$

$$
\times M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G'}), \tag{4.17}
$$

where the minus sign comes from the closed electron loop, and to make the final expression agree with [14], I choose a rather strange direction of the $\boldsymbol{q}$ arrow. The bands and wave functions are all renormalized ones.

The RHS of the equation can be evaluated straightforwardly: when $\xi_{n'\boldsymbol{k}} < 0$, $\xi_{n,\boldsymbol{k}+\boldsymbol{q}} > 0$, we have

$$
\int \frac{\mathrm{d}\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\boldsymbol{k}} + i0^+ \operatorname{sgn}(\xi_{n'\boldsymbol{k}})} \frac{1}{\omega + \omega' - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+ \operatorname{sgn}(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}
$$

$$
= \int \frac{\mathrm{d}\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\boldsymbol{k}} - i0^+} \frac{1}{\omega + \omega' - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+}
$$

$$
= \frac{2\pi i}{2\pi} \frac{1}{\omega + \xi_{n'\boldsymbol{k}} + i0^+ - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+},
$$

where in the third line we complete the integral contour in the upper plane. Similarly, when $\xi_{n'\boldsymbol{k}} > 0$, $\xi_{n,\boldsymbol{k}+\boldsymbol{q}} < 0$, we have

$$
\int \frac{\mathrm{d}\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\boldsymbol{k}} + i0^+ \operatorname{sgn}(\xi_{n'\boldsymbol{k}})} \frac{1}{\omega + \omega' - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+ \operatorname{sgn}(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}
$$

$$
= \int \frac{\mathrm{d}\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\boldsymbol{k}} + i0^+} \frac{1}{\omega + \omega' - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} - i0^+}
$$

$$
= \frac{2\pi i}{2\pi} \frac{1}{-\omega + \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+ - \xi_{n'\boldsymbol{k}} + i0^+}
$$

When $\xi_{n'\boldsymbol{k}}$ and $\xi_{n,\boldsymbol{k}+\boldsymbol{q}}$ are both positive or negative, the integral is zero. Thus, we find

$$
\chi_{\boldsymbol{GG'}}(\boldsymbol{q}, \omega) = \sum_{\boldsymbol{k}} \sum_{n,n'} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G'})
$$

$$
\times \left(\frac{\theta(-\xi_{n'\boldsymbol{k}})\theta(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}{\omega + \xi_{n'\boldsymbol{k}} - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + i0^+} + \frac{\theta(\xi_{n'\boldsymbol{k}})\theta(-\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}{-\omega + \xi_{n,\boldsymbol{k}+\boldsymbol{q}} - \xi_{n'\boldsymbol{k}} + i0^+} \right).
$$

$$
\tag{4.18}
$$

When the system has time-reversal symmetry – that's to say, we don't have magnetic fields or magnetic disorders or a ferromagnetic order – the above equation can be further simplified. When the system is spinless (when no SOC exists, the

47

electronic problem can be reduced to this circumstance), we have

$$\mathcal{T}\psi_{n\mathbf{k}} := \psi_{n\mathbf{k}}^* = \sum_m U_{nm}(\mathbf{k})\psi_{m,-\mathbf{k}}, \tag{4.19}$$

where U_{mn} is a unitary matrix; thus we have

$$\mathcal{T}^2 = 1 \Rightarrow U(-\mathbf{k})U(\mathbf{k})^* = 1 \Rightarrow U^{-1}(-\mathbf{k}) = U^*(\mathbf{k}). \tag{4.20}$$

Note that $\mathcal{T}$ applies to the U_{nm} factor as well if we apply another $\mathcal{T}$ to (4.19). So now we have

$$\psi_{n\mathbf{k}} = \sum_m U_{nm}^{-1}(-\mathbf{k})\psi_{m,-\mathbf{k}}^* = \sum_m U_{nm}^*(\mathbf{k})\psi_{m,-\mathbf{k}}^*. \tag{4.21}$$

Now we inspect the structure of the first term in (4.18). We have

$$\sum_{n,n'} M_{nn'}(\mathbf{k},\mathbf{q},\mathbf{G})M_{nn'}^*(\mathbf{k},\mathbf{q},\mathbf{G}')\frac{\theta(-\xi_{n'\mathbf{k}})\theta(\xi_{n,\mathbf{k}+\mathbf{q}})}{\omega + \xi_{n'\mathbf{k}} - \xi_{n,\mathbf{k}+\mathbf{q}} + \mathrm{i}0^+}$$

$$= \sum_{n,n'} \int \mathrm{d}^3\mathbf{r}\, \psi_{n,\mathbf{k}+\mathbf{q}}^*(\mathbf{r})\mathrm{e}^{\mathrm{i}(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\psi_{n'\mathbf{k}}(\mathbf{r}) \int \mathrm{d}^3\mathbf{r}'\, \psi_{n,\mathbf{k}+\mathbf{q}}(\mathbf{r}')\mathrm{e}^{-\mathrm{i}(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}'}\psi_{n'\mathbf{k}}^*(\mathbf{r}')$$

$$\times \frac{\theta(-\xi_{n'\mathbf{k}})\theta(\xi_{n,\mathbf{k}+\mathbf{q}})}{\omega + \xi_{n'\mathbf{k}} - \xi_{n,\mathbf{k}+\mathbf{q}} + \mathrm{i}0^+}$$

$$= \sum_{n,n'} \int \mathrm{d}^3\mathbf{r} \int \mathrm{d}^3\mathbf{r}'\, \mathrm{e}^{\mathrm{i}(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}}\mathrm{e}^{-\mathrm{i}(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}'}\frac{\theta(-\xi_{n'\mathbf{k}})\theta(\xi_{n,\mathbf{k}+\mathbf{q}})}{\omega + \xi_{n'\mathbf{k}} - \xi_{n,\mathbf{k}+\mathbf{q}} + \mathrm{i}0^+}$$

$$\times U_{nm}(\mathbf{k}+\mathbf{q})\psi_{m,-\mathbf{k}-\mathbf{q}}(\mathbf{r})U_{n'p}^*(\mathbf{k})\psi_{p,-\mathbf{k}}^*(\mathbf{r})$$

$$\times U_{nr}^*(\mathbf{k}+\mathbf{q})\psi_{r,-\mathbf{k}-\mathbf{q}}^*(\mathbf{r}')U_{n's}\psi_s(\mathbf{k})\psi_{s,-\mathbf{k}}(\mathbf{r}'). \tag{4.22}$$

Recall that the time reversal symmetry is a symmetry of the system, and thus as long as U_{mn}, etc. are non-zero, we have

$$\xi_{n,\mathbf{k}+\mathbf{q}} = \xi_{m,-\mathbf{k}-\mathbf{q}} = \xi_{r,-\mathbf{k}-\mathbf{q}}, \tag{4.23}$$

and

$$\xi_{n'\mathbf{k}} = \xi_{s,-\mathbf{k}} = \xi_{p,-\mathbf{k}}. \tag{4.24}$$

Thus, the propagator factor in the summation does not vary when we sum over n, n', etc., and we can change the band indices Thus we can insert the unitary condition

$$U_{mn}(\mathbf{k})U_{mn'}^*(\mathbf{k}) = \delta_{nn'} \tag{4.25}$$

into the equation and change the band indices in the propagator factor accordingly

and get

$$\sum_{n,n'} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}') \frac{\theta(-\xi_{n'\boldsymbol{k}})\theta(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}{\omega + \xi_{n'\boldsymbol{k}} - \xi_{n,\boldsymbol{k}+\boldsymbol{q}} + \mathrm{i}0^+}$$

$$= \sum_{p,m} \int \mathrm{d}^3\boldsymbol{r} \int \mathrm{d}^3\boldsymbol{r}' \, \mathrm{e}^{\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}} \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}'}$$

$$\times \psi_{m,-\boldsymbol{k}-\boldsymbol{q}}(\boldsymbol{r}) \psi_{p,-\boldsymbol{k}}^*(\boldsymbol{r}) \psi_{m,-\boldsymbol{k}-\boldsymbol{q}}^*(\boldsymbol{r}') \psi_{p,-\boldsymbol{k}}(\boldsymbol{r}')$$

$$\times \frac{\theta(-\xi_{p,-\boldsymbol{k}})\theta(\xi_{m,-\boldsymbol{k}-\boldsymbol{q}})}{\omega + \xi_{p,-\boldsymbol{k}} - \xi_{m,-\boldsymbol{k}-\boldsymbol{q}} + \mathrm{i}0^+}$$

$$= \sum_{n,n'} \int \mathrm{d}^3\boldsymbol{r} \int \mathrm{d}^3\boldsymbol{r}' \, \mathrm{e}^{\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}} \mathrm{e}^{-\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}'}$$

$$\times \psi_{n',-\boldsymbol{k}-\boldsymbol{q}}(\boldsymbol{r}) \psi_{n,-\boldsymbol{k}}^*(\boldsymbol{r}) \psi_{n',-\boldsymbol{k}-\boldsymbol{q}}^*(\boldsymbol{r}') \psi_{n,-\boldsymbol{k}}(\boldsymbol{r}')$$

$$\times \frac{\theta(-\xi_{n,-\boldsymbol{k}})\theta(\xi_{n',-\boldsymbol{k}-\boldsymbol{q}})}{\omega + \xi_{n\boldsymbol{k}} - \xi_{n',\boldsymbol{k}+\boldsymbol{q}} + \mathrm{i}0^+}$$

$$= \sum_{n,n'} M_{nn'}(-\boldsymbol{k}-\boldsymbol{q}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(-\boldsymbol{k}-\boldsymbol{q}, \boldsymbol{q}, \boldsymbol{G}') \times \frac{\theta(-\xi_{n,-\boldsymbol{k}})\theta(\xi_{n',-\boldsymbol{k}-\boldsymbol{q}})}{\omega + \xi_{n,-\boldsymbol{k}} - \xi_{n',-\boldsymbol{k}-\boldsymbol{q}} + \mathrm{i}0^+}.$$

$$(4.26)$$

In the last second line we have renamed the dummy indices to let the stepwise function factors look like their counterparts in (4.18). Note that in this step, we actually have swapped n and n': at first n is related to m, but now m is related to n'. The final step is to redefine $\boldsymbol{k} + \boldsymbol{q}$ to $-\boldsymbol{k}$ and make use of the time reversal symmetry of band energies, and we get

$$\sum_{\boldsymbol{k}} \sum_{n,n'} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}') \times \frac{\theta(-\xi_{n\boldsymbol{k}})\theta(\xi_{n',\boldsymbol{k}+\boldsymbol{q}})}{\omega + \xi_{n\boldsymbol{k}} - \xi_{n',\boldsymbol{k}+\boldsymbol{q}} + \mathrm{i}0^+}$$

$$= \sum_{\boldsymbol{k}} \sum_{n,n'} M_{nn'}(-\boldsymbol{k}-\boldsymbol{q}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(-\boldsymbol{k}-\boldsymbol{q}, \boldsymbol{q}, \boldsymbol{G}') \times \frac{\theta(-\xi_{n,-\boldsymbol{k}})\theta(\xi_{n',-\boldsymbol{k}-\boldsymbol{q}})}{\omega + \xi_{n,-\boldsymbol{k}} - \xi_{n',-\boldsymbol{k}-\boldsymbol{q}} + \mathrm{i}0^+}$$

$$= \sum_{\boldsymbol{k}} \sum_{n,n'} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}') \times \frac{\theta(-\xi_{n,\boldsymbol{k}+\boldsymbol{q}})\theta(\xi_{n',\boldsymbol{k}})}{\omega + \xi_{n,\boldsymbol{k}+\boldsymbol{q}} - \xi_{n',\boldsymbol{k}} + \mathrm{i}0^+}.$$

$$(4.27)$$

So now apart from the denominator, this term already looks exactly the same as the second term in (4.18). Thus, we find time reversal symmetry reduces (4.18) into

$$\chi_{\boldsymbol{G}\boldsymbol{G}'}(\boldsymbol{q}, \omega) = \sum_{\boldsymbol{k}} \sum_{n}^{\mathrm{occ}} \sum_{n'}^{\mathrm{emp}} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}')$$

$$\times \left(\frac{1}{\omega + \xi_{n,\boldsymbol{k}+\boldsymbol{q}} - \xi_{n'\boldsymbol{k}} + \mathrm{i}0^+} + \frac{1}{-\omega + \xi_{n,\boldsymbol{k}+\boldsymbol{q}} - \xi_{n'\boldsymbol{k}} + \mathrm{i}0^+} \right),$$

$$(4.28)$$

where we have recast the stepwise functions into the occupied/empty conditions. For insulators, we can swap the order of the summation over $\boldsymbol{k}$ and the summation over band indices; for metals we can't do this. This form of χ can only be obtained by using time reversal symmetry; this fact is also noted by [30] in its derivation of its (1.40). From the positions of the poles (note that $\xi_{n,\boldsymbol{k}+\boldsymbol{q}} - \xi_{n'\boldsymbol{k}} < 0$) we find

49

that poles with positive real parts have negative imaginary parts, and poles with negative real parts have positive imaginary parts, which agrees with the general properties of time-ordered quantities.

When the system is spin-1/2, i.e. when we are in the so-called spinor mode, we have

$$\mathcal{T}\psi_{n\boldsymbol{k}} := -\mathrm{i}\sigma^y \psi_{n\boldsymbol{k}}^* = \sum_m U_{nm}\psi_{m\boldsymbol{k}}, \tag{4.29}$$

and thus we also have

$$\psi_{n\boldsymbol{k}}^* = \sum_m U_{nm}\psi_{m\boldsymbol{k}}, \tag{4.30}$$

although the definition of U_{nm} should be changed – but it's still unitary; similarly we can build the relation between $U(\boldsymbol{k})$ and $U(-\boldsymbol{k})$ from

$$\mathcal{T}^2 = -1 \Rightarrow U(-\boldsymbol{k})U(\boldsymbol{k}) = -1 \Rightarrow U(-\boldsymbol{k}) = -U^\dagger(\boldsymbol{k}). \tag{4.31}$$

So we have an additional minus one factor; but we have four U factors in (4.22) anyway, and thus the minus one factor contributes nothing to the final result, and (4.28) is still correct when the system has time reversal symmetry.

Note that in [14], the normalization of χ is the same as here, but the definition of $v(\boldsymbol{q})$ doesn't include the $1/V$ factor. So (8) and (10) in [14] can't be both correct, or as we increase the density of the $\boldsymbol{k}$-grid, ϵ becomes larger and larger!

4.2.2 The static limit and the generalized plasmon-pole model

In the $\omega \to 0$ limit, we have

$$\chi_{\boldsymbol{G}\boldsymbol{G}'}(\boldsymbol{q}, \omega = 0) = \sum_{\boldsymbol{k}} \sum_{n,n'} M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}) M_{nn'}^*(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G}')$$
$$\times \left(\frac{\theta(-\xi_{n'\boldsymbol{k}})\theta(\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}{\varepsilon_{n'\boldsymbol{k}} - \varepsilon_{n,\boldsymbol{k}+\boldsymbol{q}}} + \frac{\theta(\xi_{n'\boldsymbol{k}})\theta(-\xi_{n,\boldsymbol{k}+\boldsymbol{q}})}{\varepsilon_{n,\boldsymbol{k}+\boldsymbol{q}} - \varepsilon_{n'\boldsymbol{k}}} \right). \tag{4.32}$$

A frequently used way to obtain some degree of frequency dependence is the **generalized plasmon pole model (GPP)**. The logic is like this. We know $D = \epsilon E$; although in this note ϵ relates the bare *scalar potential* and the screened *scalar potential*, the physical picture is the same. This means $1/\epsilon$ can be seen as the response (in the Feynman diagrammatic sense; I'm not saying $1/\epsilon$ is a response function in linear response theory; indeed its pole structure is the same with time-ordered Green function, not a response function) of E to an externally introduced charge, which looks like this:
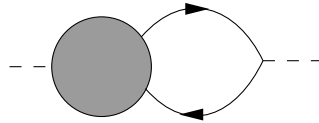
So of course, $1/\epsilon$ can also be seen as unity plus the sum of Green functions of electron density modes. We haven't introduce any approximation yet, so now let's introduce a rather strong approximation: $1/\epsilon - 1$ *only contains the plasmon mode,*

the dispersion relation of which is completely flat. We know this approximation is far from accurate: we always have electron-hole pairs and also excitons besides the plasmon, and the dispersion relation of the plasmon is not strictly flat; but anyway let's just take this as the starting point: in cases where what's important concerning ϵ is only a handful of parameters obtained from it (like ω_{p}), our rather coarse model should work well. In other parts of condensed matter physics, we encounter similar rather coarse models, like the Debye model, which however are (semi-)quantitative.

The structure of $1/\epsilon$, therefore, becomes

$$\frac{1}{\epsilon} - 1 = \text{const.} \times \frac{1}{\omega^2 - \omega_0^2 + \mathrm{i}0^+}, \tag{4.33}$$

where ω_0 is the "plasmon frequency" which is assumed to be the same when we change the plasmon wave vector $\boldsymbol{q}$. This equation can be rewritten into

$$\frac{1}{\epsilon} - 1 = \frac{\Omega^2}{2\tilde{\omega}} \left(\frac{1}{\omega - \tilde{\omega} + \mathrm{i}0^+} - \frac{1}{\omega + \tilde{\omega} - \mathrm{i}0^+} \right). \tag{4.34}$$

This is just the GPP; we call it *generalized* plasmon pole model, because $\tilde{\omega}$ isn't necessarily the plasmon frequency obtained with the free-electron approximation:

$$\omega_{\mathrm{p}} = \sqrt{\frac{ne^2}{m\epsilon_0}}. \tag{4.35}$$

When $\tilde{\omega} = \Omega = \omega_{\mathrm{p}}$, we just get the Drude-like

$$\epsilon = 1 - \frac{\omega_{\mathrm{p}}^2}{\omega^2}. \tag{4.36}$$

The next question is how to find Ω and $\tilde{\omega}$. Of course we need two constraints, and the first is of course obtained from (4.32). The second constraint is the f-sum rule: TODO Putting the two constraints together, we are able to fix $\epsilon(\omega)$ once $\epsilon(0)$ is found by calculating (4.32).

The GPP also justifies why we need to write $G(1,2)W(1^+,2)$: in the frequency domain, this is equivalent to adding a $\mathrm{e}^{\mathrm{i}0^+\omega}$ factor, where ω is the frequency on the Coulomb interaction line. Thus the GW self-energy is proportion to the following factor:

$$\int \mathrm{d}\omega' \, \mathrm{e}^{\mathrm{i}0^+\omega'} \frac{1}{\omega + \omega' - \xi_{n\boldsymbol{k}} + \mathrm{i}0^+ \operatorname{sgn}(\xi_{n\boldsymbol{k}})} \left(\frac{1}{\omega' - \tilde{\omega} + \mathrm{i}0^+} - \frac{1}{\omega' + \tilde{\omega} - \mathrm{i}0^+} \right).$$

When $\xi_{n\boldsymbol{k}} > 0$, we find only the pole of the $1/(\omega' + \tilde{\omega} - \mathrm{i}0^+)$ term contributes to the integral, and the result is proportional to

$$\frac{1}{\omega - \tilde{\omega} - \xi_{n\boldsymbol{k}}},$$

which means the singularity is achieved when $\xi_{n\boldsymbol{k}} = \omega - \tilde{\omega}$. Note that $\xi_{n\boldsymbol{k}}$ may be seen as the energy of the inner electron line, and ω is the input energy, so this means the energy of the inner electron is the input energy minus $\tilde{\omega}$, which is correct according to the physical picture of screening. When $\xi_{n\boldsymbol{k}} < 0$, all terms contribute

to the integral, but now the residue of the pole of $\xi_{n\boldsymbol{k}} = \omega - \tilde{\omega}$ cancels with the residue of the pole of the electron propagator, and the result is proportional to

$$\frac{1}{\xi_{\boldsymbol{k}} - \omega - \tilde{\omega}}.$$

This means $\xi_{\boldsymbol{k}}$ is ω – the input energy – plus $\tilde{\omega}$. But note that in this case $\xi_{n\boldsymbol{k}} < 0$, so the above expression means $\xi_{n\boldsymbol{k}}$ is less negative than the input energy, and thus, the excited hole is closer to the Fermi surface, so the energy of the hole represented by the inner electron propagator is *lower* than $|\omega| = -\omega$, which is the energy of the input hole. In both case, the $\mathrm{e}^{\mathrm{i}0^+\omega}$ factor gives the correct physical picture. The origin of this factor is discussed in § 4.1.2.

4.2.3 The static subspace approach

4.2.4 Truncation

4.3 The self-energy matrix $\Sigma = \mathrm{i}GW$

4.3.1 COHSEX approximation

One way to simply Σ^{GW} is the **Coulomb hole-screened exchange (COHSEX)** approximation. In this approximation, we assume $W(\boldsymbol{r}, t; \boldsymbol{r}', 0)$ doesn't show strong retardation, and therefore in the frequency domain, $W(\boldsymbol{r}, \boldsymbol{r}', \omega)$ shouldn't be far from $W(\boldsymbol{r}, \boldsymbol{r}', 0)$. Therefore, (4.6) is approximately

$$\Sigma(\boldsymbol{r}, \boldsymbol{r}', \omega) \approx \mathrm{i} \int \frac{\mathrm{d}\omega'}{2\pi} \int \mathrm{d}t\, \mathrm{e}^{\mathrm{i}(\omega+\omega')t} G(\boldsymbol{r}, t; \boldsymbol{r}', 0) W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0)$$

$$= \mathrm{i} \int \mathrm{e}^{\mathrm{i}\omega t} \delta(t) G(\boldsymbol{r}, t; \boldsymbol{r}', 0) W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0)$$

$$= \mathrm{i} G(\boldsymbol{r}, t = 0; \boldsymbol{r}', 0) W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0).$$

But note that $G(t = 0)$ is not well-defined: at $t = 0$ there is a stepwise jump (§ 1.2.1). We know Fourier transformation always correct $f(t)$ to $(f(t + 0^+) + f(t - 0^+))/2$, and therefore what we get is essentially

$$\Sigma^{GW}(\boldsymbol{r}, \boldsymbol{r}', \omega) \approx \Sigma^{\mathrm{COHSEX}}(\boldsymbol{r}, \boldsymbol{r}', \omega)$$
$$= \frac{\mathrm{i}}{2}(G(\boldsymbol{r}, t = 0^+; \boldsymbol{r}', 0) + G(\boldsymbol{r}, t = 0^-; \boldsymbol{r}', 0))W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0), \tag{4.37}$$

which doesn't really have ω dependence. The time domain version is

$$\Sigma^{\mathrm{COHSEX}} = \frac{\mathrm{i}}{2}W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0)G(\boldsymbol{r}, t; \boldsymbol{r}', 0)(\delta(t + 0^+) + \delta(t - 0^+)). \tag{4.38}$$

Until now, we still don't see why we call this Coulomb-hole plus screened exchange. However, note that

$$G(\boldsymbol{r}, t = 0^+; \boldsymbol{r}', 0) + G(\boldsymbol{r}, t = 0^-; \boldsymbol{r}', 0)$$
$$= -\mathrm{i}\langle\Psi|\psi(\boldsymbol{r})\psi^\dagger(\boldsymbol{r}) - \psi^\dagger(\boldsymbol{r}')\psi(\boldsymbol{r})|\Psi\rangle$$
$$= -\mathrm{i}(\delta(\boldsymbol{r} - \boldsymbol{r}') - 2\langle\Psi|\psi^\dagger(\boldsymbol{r}')\psi(\boldsymbol{r})|\Psi\rangle),$$

and therefore we have

$$\Sigma^{\mathrm{COHSEX}} = \Sigma^{\mathrm{COH}} + \Sigma^{\mathrm{SEX}}, \tag{4.39}$$

$$\Sigma^{\mathrm{COH}} = \frac{1}{2}\delta(\boldsymbol{r} - \boldsymbol{r}')W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0), \tag{4.40}$$

$$\Sigma^{\mathrm{SEX}} = - \langle\Psi|\psi^\dagger(\boldsymbol{r}')\psi(\boldsymbol{r})|\Psi\rangle \, W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0). \tag{4.41}$$

4.3.2   The exact COHSEX decomposition of Σ

In this section I prove that similar to the case in last section, for the generic GW self-energy, we are still able to split $\Sigma_n^{GW}(\boldsymbol{k}, \omega)$ into two parts, corresponding to the so-called COH and SEX terms above. To make the notation less cumbersome, below I temporarily omit the $\boldsymbol{r}$ and $\boldsymbol{r}'$ variables. Thus the spectral representation of the time-ordered Green function is

$$G(\omega) = \int d\omega' \frac{A(\omega')}{\omega - \omega' + i\operatorname{sgn}(\omega')}. \tag{4.42}$$

Some papers use ω to refer to the "absolute energy" of the electron, which is $\omega + \mu$ in this note; under their convention, we have

$$G(\omega) = \int d\omega' \frac{A(\omega')}{\omega - \omega' + i\operatorname{sgn}(\omega' - \mu)}, \tag{4.43}$$

and the fact that the electron is above the Fermi energy is equivalent to the inequality $\omega > \mu$, not $\omega > 0$. Note that $A(\omega')$ is defined on the real axis and is completely analytic; the denominator in the above equation contains all the singularities of $G(\omega)$.

The GW self-energy then is

$$\Sigma(\omega) = iGW = i \int \frac{d\omega'}{2\pi} e^{-i0^+\omega'} G(\omega - \omega')W(\omega')$$

$$= i \int \frac{d\omega'}{2\pi} e^{-i0^+\omega'} W(\omega') \underbrace{\int d\omega'' \frac{A(\omega'')}{\omega - \omega' - \omega'' + i\operatorname{sgn}(\omega'')}}_{G(\omega - \omega')}. \tag{4.44}$$

Now we know there are poles in the factor $W(\omega')$; this can be done by doing a spectral decomposition for $W(\omega')$. As for the contribution from the $(\omega - \omega' - \omega'')$ denominator of the Green function, we have

$$[\Sigma(\omega)]_{\mathrm{from}\ G}$$

$$= i \int d\omega'' A(\omega'') \int \frac{d\omega'}{2\pi} \frac{e^{-i0^+\omega'} W(\omega')}{\omega - \omega' - \omega'' + i\operatorname{sgn}(\omega'')}$$

$$= i \int d\omega'' A(\omega'')$$

$$\times \begin{cases} 0, & \omega'' > 0, \\ \frac{1}{2\pi} \cdot (-2\pi i)\lim_{\omega' \to \omega - \omega''}(\omega' - \omega - \omega'') \cdot \frac{e^{-i0^+\omega'} W(\omega')}{\omega - \omega' - \omega'' + i\operatorname{sgn}(\omega'')}, & \omega'' \leq 0 \end{cases}$$

$$= i \int_{-\infty}^{0} d\omega'' A(\omega'') \cdot iW(\omega - \omega'')$$

$$= - \int_{-\infty}^{0} d\omega'' A(\omega'')W(\omega - \omega'').$$

Note that the $\mathrm{e}^{-\mathrm{i}0^+\omega'}$ factor means if we want to use the residue theorem and contour integration to evaluate this expression, we should go around the lower half plane; when $\omega'' > \mu$, we get a pole outside of this integral contour, and thus the integral vanishes.

If we are doing Hartree-Fock approximation, and thus W is actually the bare Coulomb interaction v, then this is all we have: W, in this case, contains no pole, and therefore Σ comes completely from the poles of the Green function part, and is just the whole exchange-correlation energy. So when W is the screened Coulomb potential, $[\Sigma]_{\mathrm{from}\ G}$ can be obtained by simply replacing v in Hartree-Fock exchange-correlation energy by $W(\omega)$. Thus we name this term the **screened exchange-correlation energy**, or SEX in short.

Now we move to the part of Σ coming from poles in W. The usual spectral expansion like the follows:

$$W(\omega) = v + \int_0^\infty \mathrm{d}\omega' \frac{2\omega'}{\omega^2 - (\omega' - \mathrm{i}0^+)^2} B(\omega'), \qquad (4.45)$$

from which we find

$$\mathrm{Im}\, W(\omega) = -\pi B(\omega). \qquad (4.46)$$

Here we slightly misuse the notation apply the Re and Im operators only to contributions of poles, i.e. real and imaginary parts that come from the analytic structure with respect to the variable ω; thus, in this section, $\mathrm{Re}\,\Sigma(\boldsymbol{r}, \boldsymbol{r}', \omega)$ is not really real. In free space, since $\mathrm{Re}\,\Sigma(\boldsymbol{k}, \omega)$ is real, and $\mathrm{Re}\,\Sigma(\boldsymbol{r}, \boldsymbol{r}', \omega)$ is its Fourier transform, we find $\mathrm{Re}\,\Sigma(\boldsymbol{r}, \boldsymbol{r}', \omega) \sim \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot(\boldsymbol{r}-\boldsymbol{r}')}$ and obligatory has an imaginary part. Also, after GW correction, we actually have possible non-diagonal terms in $\Sigma_{nn'}(\boldsymbol{k}, \omega)$, and these non-diagonal terms are not necessarily real. Since we assume the system to have time reversal symmetry, we have $W(\boldsymbol{r}, \boldsymbol{r}', \omega) = W(\boldsymbol{r}, \boldsymbol{r}', -\omega)$, and therefore $W(\omega)$ is an even function, and the Kramers–Kronig relation now reads

$$\mathrm{Re}\, W(\omega) = -\frac{1}{\pi} \int_0^\infty \mathrm{d}\omega' \frac{2\omega'}{\omega'^2 - \omega^2} \mathrm{Im}\, W(\omega),$$

and since $W(\omega)$ is to be time-ordered, just like its building blocks, i.e. the time-ordered Green functions, positive poles of $W(\omega)$ should have negative infinitesimal imaginary parts, while negative poles of $W(\omega)$ should have positive infinitesimal imaginary parts, and thus we get (4.45). From the spectral representation (4.45) we have

$$[\Sigma(\omega)]_{\mathrm{from}\ W}$$
$$= \mathrm{i} \int \frac{\mathrm{d}\omega'}{2\pi} W(\omega') \mathrm{e}^{-\mathrm{i}0^+\omega'} \int \mathrm{d}\omega'' \frac{A(\omega'')}{\omega - \omega' - \omega'' + \mathrm{i}\,\mathrm{sgn}(\omega'')}\bigg|_{\mathrm{no\ pole\ from}\ G}$$
$$= \mathrm{i} \frac{-2\pi\mathrm{i} \operatorname{Res}_{\omega'\ \mathrm{in\ lower\ plane}}}{2\pi} \int_0^\infty \mathrm{d}\omega''' \frac{2\omega''' B(\omega''')}{\omega'^2 - (\omega''' - \mathrm{i}0^+)^2} \times$$
$$\int \mathrm{d}\omega'' \frac{A(\omega'')}{\omega - \omega' - \omega'' + \mathrm{i}\,\mathrm{sgn}(\omega'')}$$
$$= \int_0^\infty \mathrm{d}\omega''' \int \mathrm{d}\omega'' \lim_{\omega'\to\omega'''-\mathrm{i}0^+} \frac{(\omega' - \omega''' + \mathrm{i}0^+)}{\omega'^2 - (\omega''' - \mathrm{i}0^+)^2} \frac{2\omega''' A(\omega'') B(\omega''')}{\omega - \omega' - \omega'' + \mathrm{i}\,\mathrm{sgn}(\omega'')}$$
$$= \int_0^\infty \mathrm{d}\omega''' \int \mathrm{d}\omega'' \frac{A(\omega'') B(\omega''')}{\omega - \omega''' - \omega'' + \mathrm{i}\,\mathrm{sgn}(\omega'')}.$$

Here we have omitted the v term in the expression of W, because it contains no pole, and therefore its contribution is already considered in the SEX part. We call the above contribution from poles in W the **Coulomb-hole part**, or COH in short.

In conclusion, we are able to split Σ^{GW} into two parts, namely

$$\Sigma^{GW} = \Sigma^{\text{COH}} + \Sigma^{\text{SEX}}, \tag{4.47}$$

where

$$\Sigma^{\text{SEX}}(\boldsymbol{r}, \boldsymbol{r}', \omega) = -\int_{-\infty}^{0} \mathrm{d}\omega' \, A(\boldsymbol{r}, \boldsymbol{r}', \omega') W(\boldsymbol{r}, \boldsymbol{r}', \omega - \omega'), \tag{4.48}$$

and

$$\Sigma^{\text{COH}}(\boldsymbol{r}, \boldsymbol{r}', \omega) = \int_{0}^{\infty} \mathrm{d}\omega'' \int_{-\infty}^{\infty} \mathrm{d}\omega' \, \frac{A(\boldsymbol{r}, \boldsymbol{r}', \omega') B(\boldsymbol{r}, \boldsymbol{r}', \omega'')}{\omega - \omega' - \omega'' + \mathrm{i}\, \mathrm{sgn}(\omega')}. \tag{4.49}$$

There is a sign difference in (13b) in [20]; but it seems they made a mistake there, since (17b) in [20] agrees well with my result.

When the behavior of the system is Fermi liquid-like and we have well-defined quasiparticles, we can approximately have (1.42) even for interactive Green function, and therefore

$$G(\boldsymbol{r}, \boldsymbol{r}', \omega) = \sum_{n, \boldsymbol{k}} \frac{\phi_{n\boldsymbol{k}}(\boldsymbol{r}) \phi_{n\boldsymbol{k}}^{*}(\boldsymbol{r}')}{\omega - \xi_{n\boldsymbol{k}} + \mathrm{i}0^{+}\, \mathrm{sgn}(\xi_{n\boldsymbol{k}})}, \tag{4.50}$$

$$A(\boldsymbol{r}, \boldsymbol{r}', \omega) = \sum_{n, \boldsymbol{k}} \delta(\omega - \xi_{n\boldsymbol{k}}) \phi_{n\boldsymbol{k}}(\boldsymbol{r}) \phi_{n\boldsymbol{k}}^{*}(\boldsymbol{r}'), \tag{4.51}$$

and hence

$$\Sigma^{\text{SEX}}(\boldsymbol{r}, \boldsymbol{r}', \omega) = -\sum_{n, \boldsymbol{k}}^{\text{occ}} \phi_{n\boldsymbol{k}}(\boldsymbol{r}) \phi_{n\boldsymbol{k}}^{*}(\boldsymbol{r}') W(\boldsymbol{r}, \boldsymbol{r}', \omega - \xi_{n\boldsymbol{k}}), \tag{4.52}$$

$$\Sigma^{\text{COH}}(\boldsymbol{r}, \boldsymbol{r}', \omega) = \sum_{n, \boldsymbol{k}} \phi_{n\boldsymbol{k}}(\boldsymbol{r}) \phi_{n\boldsymbol{k}}^{*}(\boldsymbol{r}') \int_{0}^{\infty} \mathrm{d}\omega' \, \frac{B(\boldsymbol{r}, \boldsymbol{r}', \omega')}{\omega - \xi_{n\boldsymbol{k}} - \omega' + \mathrm{i}\, \mathrm{sgn}(\xi_{n\boldsymbol{k}})}. \tag{4.53}$$

4.3.2.1 Going back to static COHSEX

When we sum over $n, \boldsymbol{k}$, $\xi_{n\boldsymbol{k}}$ varies, and so does $\omega - \xi_{n\boldsymbol{k}}$. This, of course, means Σ^{COH} has frequency dependence. However, when the peak frequency of $B(\omega')$ is much larger than this frequency dependence, we have

$$\Sigma^{\text{COH}}(\boldsymbol{r}, \boldsymbol{r}', \omega) = \sum_{n, \boldsymbol{k}} \phi_{n\boldsymbol{k}}(\boldsymbol{r}) \phi_{n\boldsymbol{k}}^{*}(\boldsymbol{r}') \, \mathrm{P} \int_{0}^{\infty} \mathrm{d}\omega' \, \frac{B(\boldsymbol{r}, \boldsymbol{r}', \omega')}{-\omega'}.$$

The RHS of this equation can be seen by setting ω to zero in (4.45): we have

$$W(\omega = 0) = v + \int_{0}^{\infty} \mathrm{d}\omega' \, \frac{2\omega'}{-\omega'^2 + \mathrm{i}0^{+}} B(\omega'),$$

and thus

$$\Sigma^{\text{COH}}(\boldsymbol{r}, \boldsymbol{r}') = \sum_{n,\boldsymbol{k}} \phi_{n\boldsymbol{k}}(\boldsymbol{r})\phi_{n\boldsymbol{k}}^*(\boldsymbol{r}') \cdot \frac{1}{2} \operatorname{Re}(W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0) - v(\boldsymbol{r}, \boldsymbol{r}'))$$

$$= \frac{1}{2}\delta(\boldsymbol{r} - \boldsymbol{r}') \operatorname{Re}(W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0) - v(\boldsymbol{r}, \boldsymbol{r}')). \tag{4.54}$$

TODO: why in Louie (19) there is a $-v$ term in the COH term

TODO: the name "Coulomb hole"

Similarly, we can replace $\omega - \xi_{n\boldsymbol{k}}$ by zero in (4.52), and thus we get

$$\Sigma^{\text{SEX}}(\boldsymbol{r}, \boldsymbol{r}') = - \sum_{n,\boldsymbol{k}}^{\text{occ}} \phi_{n\boldsymbol{k}}(\boldsymbol{r})\phi_{n\boldsymbol{k}}^*(\boldsymbol{r}') \operatorname{Re} W(\boldsymbol{r}, \boldsymbol{r}', \omega = 0). \tag{4.55}$$

So now, by replacing $\omega - \xi_{n\boldsymbol{k}}$ by zero, or in other words, by getting rid of the frequency dependence, which is the approximation used in (4.3.1), we get (4.40) and (4.41) Note that since we can easily prove that

$$\Sigma^{\text{SEX}}(\boldsymbol{r}, \boldsymbol{r}') = (\Sigma^{\text{SEX}}(\boldsymbol{r}', \boldsymbol{r}))^*,$$

Σ^{SEX} is "real" in the sense of being real in the momentum space (see the start of this section), and we also have

$$\langle \Psi | \psi^\dagger(\boldsymbol{r}')\psi(\boldsymbol{r}) | \Psi \rangle = \phi_{n\boldsymbol{k}}(\boldsymbol{r})\phi_{n\boldsymbol{k}}^*(\boldsymbol{r}'), \tag{4.56}$$

which is also "real" in this sense, and thus (4.55) is the "real part" of (4.41).

TODO: COHSEX for time-dependent GW can't be obtained using the spectral representation techniques in this section?

In conclusion, what is done in § 4.3.1 should better be called *static* COHSEX; even the full GW self-energy can be decomposed in this manner. The static COHSEX approximation only makes use of the $\omega = 0$ component of W, or in other words ϵ: it involves no frequency dependence of screening at all, while GPP (§ 4.2.2) takes some frequency dependence into account, although in a rather crude way.

4.3.3   From dielectric matrix to self-energy

In the last section there are two problems left. First, we need to find an explicit form of the $B(\boldsymbol{r}, \boldsymbol{r}', \omega)$ function; this actually is already done in (4.46), and is reduced to finding the "imaginary part" of the dielectric constant (note that in (4.46) the symbol Im ignores imaginary parts coming from wave functions, etc., and so-called $\operatorname{Im} \epsilon$ isn't necessarily a real number):

$$B(\omega) = -\frac{1}{\pi} \operatorname{Im} \epsilon^{-1}(\omega)v. \tag{4.57}$$

Second, we need to rewrite the equations in the crystal momentum space. We can do this by definition, but a more intuitive way to do so is to notice the fact that although the decomposition of Σ into the COH and SEX terms means the two terms now have *complex analytic structures* different from that of the original iGW, both the COH term and the SEX term still keep the linear contraction

structure shown in (1.50). We label the variables in the screened Fock diagram as the follows:
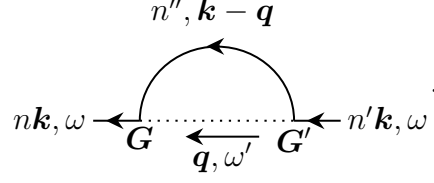
Then, from (4.52), we can immediately find

$$
\begin{aligned}
\langle n\boldsymbol{k}|\Sigma^{\mathrm{SEX}}(\omega)|n'\boldsymbol{k}\rangle = -\sum_{n''}^{\mathrm{occ}} \sum_{\boldsymbol{q},\boldsymbol{G},\boldsymbol{G}'} & M_{n''n}^*(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G}) \\
& \times \epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}(\boldsymbol{q},\omega-\xi_{n'',\boldsymbol{k}-\boldsymbol{q}})v(\boldsymbol{q}+\boldsymbol{G}') \\
& \times M_{n''n'}(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G}'),
\end{aligned}
\tag{4.58}
$$

where we have already integrated out the $m'',\boldsymbol{k}-\boldsymbol{q}$ propagator and therefore the frequency on the Coulomb interaction line is $\omega-\xi_{n'',\boldsymbol{k}-\boldsymbol{q}}$. Similarly from (4.53), we have

$$
\begin{aligned}
\langle n\boldsymbol{k}|\Sigma^{\mathrm{COH}}(\omega)|n'\boldsymbol{k}\rangle = \sum_{n''} \sum_{\boldsymbol{q},\boldsymbol{G},\boldsymbol{G}'} & M_{n''n}^*(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G}) \\
& \times \int_0^\infty \mathrm{d}\omega' \left(-\frac{1}{\pi}\right) \frac{\mathrm{Im}\,\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}(\boldsymbol{q},\omega')v(\boldsymbol{q}+\boldsymbol{G}')}{\omega-\xi_{n\boldsymbol{k}}-\omega'+\mathrm{i}\,\mathrm{sgn}(\xi_{n\boldsymbol{k}})} \\
& \times M_{n''n'}(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G}').
\end{aligned}
\tag{4.59}
$$

The "imaginary part" of the time-ordered dielectric function $\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}$ can be obtained from the advanced and retarded dielectric functions. First, we notice that the difference between the advanced and retarded dielectric functions is just changing the sign of the imaginary parts of the poles, and by Taylor expansion we can explicitly verify that

$$
\frac{1}{a+b\mathrm{i}} - \frac{1}{a-b\mathrm{i}} = 2\mathrm{i}\,\mathrm{Im}\,\frac{1}{a+b\mathrm{i}},
$$

where a,b are considered to be real in the eye of the Im operator. We also know that when $\omega'>0$, $\epsilon(\omega')$ should be same as $\epsilon^{\mathrm{r}}(\omega')$. Thus we find when $\omega'>0$, we always have

$$
\mathrm{Im}\,\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{-1}(\boldsymbol{q},\omega') = \mathrm{Im}[\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{\mathrm{r}}]^{-1}(\boldsymbol{q},\omega') = \frac{[\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{\mathrm{r}}]^{-1}(\boldsymbol{q},\omega') - [\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{\mathrm{a}}]^{-1}(\boldsymbol{q},\omega')}{2\mathrm{i}}.
\tag{4.60}
$$

(4.59) thus becomes

$$
\begin{aligned}
\langle n\boldsymbol{k}|\Sigma^{\mathrm{COH}}(\omega)|n'\boldsymbol{k}\rangle = \frac{\mathrm{i}}{2\pi} \sum_{n''} \sum_{\boldsymbol{q},\boldsymbol{G},\boldsymbol{G}'} & M_{n''n}^*(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G})M_{n''n'}(\boldsymbol{k},-\boldsymbol{q},-\boldsymbol{G}') \\
& \times \int_0^\infty \mathrm{d}\omega' \frac{[\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{\mathrm{r}}]^{-1}(\boldsymbol{q},\omega') - [\epsilon_{\boldsymbol{G}\boldsymbol{G}'}^{\mathrm{a}}]^{-1}(\boldsymbol{q},\omega')}{\omega-\xi_{n\boldsymbol{k}}-\omega'+\mathrm{i}0^+\,\mathrm{sgn}(\xi_{n\boldsymbol{k}})}v(\boldsymbol{q}+\boldsymbol{G}').
\end{aligned}
\tag{4.61}
$$

4.4   Self-consistency modes in GW

We know in the momentum space, we have

$$E_{n\bm{k}}^{\mathrm{QP}} = E_{n\bm{k}}^0 + \Sigma_{n\bm{k}}(E^{\mathrm{QP}}). \tag{4.62}$$

Here since Σ depends on the corrected propagator, $E_{n\bm{k}}^{\mathrm{QP}}$ enters its expression. The cost of GW calculation means we need to first do a DFT calculation and feed this as the input of the GW package (the former usually mysteriously called the "mean field" step, though we may also say GW is a mean-field method, although a more elaborated one; on the other hand, in principle – though of course not in practice – DFT is able to decide everything about the system), so (4.62) now is

$$E_{\bm{k}}^{\mathrm{QP}} = E_{\bm{k}}^{\mathrm{KS}} + \Sigma_{\bm{k}}(E^{\mathrm{QP}}) - \Sigma^{\mathrm{KS}}. \tag{4.63}$$

Here Σ^{KS} is the so-called DFT self-energy, i.e. the Hartree potential plus the exchange-correlation potential. Note that here I don't insert band indices into the equation, because $\Sigma_{\bm{k}}$ may mix different bands together, and (4.63) is an equation about matrices, essentially a single-electron Schrodinger equation.

The problem then is how to solve this highly complicated equation. There are two non-trivial aspects: the first is that the RHS depends on E^{QP}; the second is that the equation is a matrix equation. We are going to discuss different levels of approximations regarding these aspects.

4.4.1   Diagonal G_0W_0

The non-diagonal feature of (4.63) is also what is frequently encountered in single-electron quantum mechanics: a perturbation remixes several eigenstates of the unperturbed system, and when the perturbation is small, perturbation expansion can be used. In the first order approximation, only the eigenvalues are updated; the wave function remains the same. Thus we have

$$E_{n\bm{k}}^{\mathrm{QP}} = E_{n\bm{k}}^{\mathrm{KS}} + \left\langle \psi_{n\bm{k}}^{\mathrm{KS}} \middle| \Sigma(E_{n\bm{k}}^{\mathrm{QP}}) - \Sigma^{\mathrm{KS}} \middle| \psi_{n\bm{k}}^{\mathrm{KS}} \right\rangle. \tag{4.64}$$

The non-diagonal properties are completely ignored here; we still need to deal with the frequency dependence on the RHS. The most naive approximation is

$$E_{n\bm{k}}^{\mathrm{QP},\,0} = E_{n\bm{k}}^{\mathrm{KS}} + \left\langle \psi_{n\bm{k}}^{\mathrm{KS}} \middle| \Sigma(E_{n\bm{k}}^{\mathrm{KS}}) - \Sigma^{\mathrm{KS}} \middle| \psi_{n\bm{k}}^{\mathrm{KS}} \right\rangle. \tag{4.65}$$

This however is usually not good enough; one way to slightly improve the accuracy is outlined in the follows. We have

$$E_{n\bm{k}}^{\mathrm{QP}} \approx E_{n\bm{k}}^{\mathrm{KS}} + \left\langle \psi_{n\bm{k}}^{\mathrm{KS}} \middle| \Sigma(E_{n\bm{k}}^{\mathrm{KS}}) + \frac{\partial \Sigma}{\partial E}\bigg|_{E_{\mathrm{KS}}} (E_{n\bm{k}}^{\mathrm{QP}} - E_{n\bm{k}}^{\mathrm{KS}}) - \Sigma^{\mathrm{KS}} \middle| \psi_{n\bm{k}}^{\mathrm{KS}} \right\rangle$$

$$= E_{n\bm{k}}^{\mathrm{QP},0} + (E_{n\bm{k}}^{\mathrm{QP}} - E_{n\bm{k}}^{\mathrm{KS}}) \cdot \frac{\partial}{\partial E} \left\langle \psi_{n\bm{k}}^{\mathrm{KS}} \middle| \Sigma \middle| \psi_{n\bm{k}}^{\mathrm{KS}} \right\rangle,$$

from which we find

$$E_{n\bm{k}}^{\mathrm{QP},1} = E_{n\bm{k}}^{\mathrm{QP},0} + \frac{\mathrm{d}\Sigma/\mathrm{d}E}{1 - \mathrm{d}\Sigma/\mathrm{d}E}(E_{n\bm{k}}^{\mathrm{QP},0} - E_{n\bm{k}}^{\mathrm{KS}}), \tag{4.66}$$

where $\mathrm{d}\Sigma/\mathrm{d}E$ means $\frac{\partial}{\partial E} \left\langle \psi_{n\bm{k}}^{\mathrm{KS}} \middle| \Sigma \middle| \psi_{n\bm{k}}^{\mathrm{KS}} \right\rangle$ at $E = E_{n\bm{k}}^{\mathrm{KS}}$.

TODO: when doing iterative calculation, V_{H} may no longer be the same between DFT and GW?? So when doing iterative calculation, should we calculate "the GW V_{H}"?

4.4.2 Self-consistent or not

There are three iterative schemes. The first is the eigenvalue self-consistent scheme: It's just a self-consistent solver of (4.64) – self-consistency is needed because the RHS contains E dependence. In this case, we don't need off-diagonal elements, because they are not used in (4.64). This scheme is mentioned in Section 3.3 in [14]. The second scheme takes the change of eigenstates into account, and thus iteratively solves (4.63). In this case we need to take non-diagonal elements seriously [2, 15]. In the third scheme, the form of Σ itself is changed: Recall that we need an `epsilon` step to calculate ϵ and thus the screened interaction potential W, and $\Sigma = iGW$. This in general is not recommended, because we know GW tends to widen the band gap, and sometimes as we iteratively update the band gap, it becomes too large. The origin of this overestimation of band gap is that GW neglects the vertex, so iterative GW only leads us towards the more and more inaccurate way.

The non-self-consistent G_0W_0 calculation proves to be a better choice empirically, if the initial DFT input is of good quality – and here there is another empirical observation that sometimes LDA functional together with G_0W_0 provides better results. Still, the argumentation provided above only explains why iterative GW is bad, but doesn't explain why one-shot GW is good. In other words, we need to know how certain factors in the one-shot GW scheme somehow makes up for the missing vertex correction. Indeed, if we capture the vertex effects, the accuracy can be improved [29]. TODO: physical picture

One possible form of the vertex is the electron-hole interaction, which is calculated by solving the BSE. Now an empirical fact is LDA tends to give the same band gap as BSE, leading to a pretty good one-shot approximation.

The question, then, is why LDA in some cases works as well as BSE. The reason for this is because of the relation between the derivative discontinuity in DFT and electron-hole interaction kernel TODO: the relation with [24]

4.5 The exciton BSE kernel

4.5.1 Frequency dependence

Something similar to the COHSEX decomposition in § 4.3.2 can be done in BSE. For example, consider the single scattering term in (4.10). It reads

$$
\int \frac{d\omega_e}{2\pi} \int \frac{d\omega_e'}{2\pi} \frac{i}{\omega_e - \xi_{\bm{k}_e} + i0^+} \frac{i}{\omega_e - \omega - \xi_{\bm{k}_h} - i0^+} \\
\times (-i)K(\omega_e - \omega_e') \frac{i}{\omega_e' - \xi_{\bm{k}_e'} + i0^+} \frac{i}{\omega_e' - \omega - \xi_{\bm{k}_h'} - i0^+}, \tag{4.67}
$$

given that we have well-defined quasiparticles (an assumption we made in (4.50)), and then again the pole structure is very close to the pole structure in Σ^{GW}, where we have poles from the Green functions and from the BSE kernel. For the contribution from the Green functions, around the poles that have contribution we find $\omega_e = \omega + \xi_{\bm{k}_e}$ and similarly $\omega_{e'} = \omega + \xi_{\bm{k}_{e'}}$; the order of magnitude of $\omega_e - \omega_e'$ is therefore the order of magnitude of energy variance in electron (or hole) states

on the same band with large weights in the exciton wave function; this usually isn't quite large (typically, the bottom of a valley or the top of a peak), compared to the plasmon frequency that defines the pole of $K(\omega)$, which is, say, $5\,\text{eV}$. So in the terms from the Green function holes, $K(\omega)$ is roughly a constant and can be replaced by $K(\omega = 0)$. This fact immediately in turn means that terms from the poles of the BSE kernel $K(\omega)$ can be ignored: when we are around these poles, $\omega_\text{e} - \omega_\text{e}' \simeq \omega_\text{p} \gg \xi$, and the electron propagators are then suppressed to zero. The conclusion is, although BSE is in principle much more complicated than GW, it's simpler than GW in one aspect: the static approximation of the "self-energy" – the exciton kernel in BSE – is almost always true in BSE, while in GW, it (i.e. static COHSEX; recall that in § 4.3.1, what we're doing is replacing $W(\omega)$ by $W(0)$, which is the same as the approximation made here) is almost always quantitatively wrong.

Interestingly, when the frequency-dependence is considered, the BSE kernel is sometimes written as a weird form which contains $1/\Delta E$ factors besides the screened Coulomb potential W [26]. This form is seen as early as [31], although in this paper, the weird $1/\Delta E$ factor comes together with ϵ. A much later appearance of the weird form of the BSE kernel is in [33],[5] which cites [26] as a reference; another reference the paper cites is the book [7]; the problem is that book never mentions such a weird form. [33] has an important difference with [26]: in the former, the ΔE looks like $E_c' - E_v$ or $E_c - E_v'$, and if this is true, it can actually be derived from the equation above by letting the integrations over ω_e and ω_e' take different routes. I still don't quite get why this should be the case.

See also Quasiparticle and Optical Properties of Carrier-Doped Monolayer MoTe2 from First Principles

4.5.2 The Tamm-Damcoff approximation

4.6 Accuracy of GW

GW proves to be accurate enough for most weakly-correlated systems. TODO: any counter-examples? Recently, it's also applied successfully to systems like polymers, nano-wires and molecules.

4.7 On so-called failure of GW and convergence issues

Some (weak-correlated, of course) materials are claimed to be impossible to be characterized correctly using GW, or at least G^0W^0. [28] refutes such a claim, at least for ZnO. The root for this seems to be poor convergence test: people often use insufficient number of bands, etc.

See https://www.nersc.gov/assets/Uploads/ConvergenceinBGW.pdf

---

[5]Its definition of W is slightly different from mine: somehow one unscreened v term is included.

4.8   Building the exciton kernel

One issue surrounding the Tamm-Dancoff approximation is why it is an approximation at all: the fact that we have de-excitation terms in the "exciton wave function" might be surprising, since in Fermi liquid theory we know there is absolutely no electron occupation above the Fermi energy (blurriness only appears in the momentum distribution), and thus the Fermi sea is the ground state of the system and it seems we can't expect to see de-excitation terms, since the latter comes from $\psi(\boldsymbol{r})\psi^{\dagger}(\boldsymbol{r}')$[6] creating an electron-hole pair that happens to be a component of the interactive ground state. The point here is in Fermi liquid theory, the ground state is *not* a Fock state, although in the eyes of single-electron operators, the ground state looks identical to a Fock state (when we are in the $T \to 0$ limit and the uncertainty of the momentum isn't serious). The clear-cut Fermi energy, above which the expected number of electron is zero, can be seen by taking the imaginary part of the Lehmann representation of the single-electron Green function; but then this doesn't mean the ground state or excited states is any kind of Fock state with the wave function attached to each state being $\langle \Omega | \psi(\boldsymbol{r}) | \text{excited state} \rangle$.[7]

To see this clearly, we first note that the "single-electron wave function" in Lehmann representation of the single-electron Green function also contains contributions from the valence bands, which, if the ground state were a Fock state, would vanish, just like how $\chi_S(\boldsymbol{r}, \boldsymbol{r}')$ contains the B terms. In the below expression, which is a part of $\langle \Omega | \psi(\boldsymbol{r}) | \text{single-electron} \rangle$ and therefore is a part of the

---

[6]Recall that the electron field is a complex field, and therefore $\psi(\boldsymbol{r})$ only contains annihilation operators; in relativistic QFT it also contains creation operators of positrons but they are always ignored in condensed matter physics.

[7]The fact that an exciton mode can't be one hundred percent understood as an electron-like quasiparticle and a hole-like quasiparticle bound together, eventually, comes from the fact that we're dealing with emergent behaviors of a many-body system after all, even if the system is weakly correlated, and hence quasiparticles are not truly elementary: rather, both (electron- or hole-like) quasiparticles and excitons are both branches in the spectrum of the condensed matter system, and none of the two in principle can be said to be more "fundamental". If anything, we should say the *bare* electrons that make up the system are more fundamental, but in condensed matter physics we're only interested in interactively renormalized electrons i.e. quasiparticles anyway.

This observation actually has an unexpected generalization. Suppose now we turn our attention to particle physics. Could it be that in the real theory of everything, we should also say that what we conceive as fundamental particles (like electrons and quarks) and what we conceive as composite particles (atoms, for example) are both emergent phenomena of something deeper, and there are circumstances where we can't really say one is more fundamental than another? Xiao-gang Wen proposes that maybe the universe is a string-net liquid: a so-called fundamental fermion is an end of an open string, and composite particles are collective behaviors of a bunch of open strings. If this is true and if we somehow are able to see the universe in the basis of string-nets, it seems the only thing that makes a fundamental fermion differ from a composite particle is that the two have different energies. Here the string net plays the role of bare electrons, which are there but no one directly sees them. The point, then, is that it's possible that the basic degrees of freedom of the world are things that are quite different from what we can see, and it's possible that what we conceive as more elementary in particle physics are merely the high-energy part of the world's spectrum. This radical emergent perception of physics differs quite a lot from the reductionist perspective, the latter contending that in principle, the basic degrees of freedom of the theory of everything can be found by doing experiments that can intuitively be understood as "breaking matter up".

"wave function" of a renormalized single-electron state (not a single-hole state), all energy level diagrams represent Fock states. Since the ground state is interactively renormalized, it of course contains more than one Fock state component, which, intuitively speaking, means it contains spontaneously generated electron-hole pairs. So now we can see it's possible that ψ kills the electron above the Fermi sea in the ket part of the expression, and the resulting state is just the non-interactive Fermi sea state; but it's also possible that the ψ operator kills one electron in the occupied state, and the resulting state now has to combine with the first term (that's proportional to the strength of interaction g) in the bra part:

$$\left(g \left\langle \vcenter{\hbox{\includegraphics{diag1}}} \right| + \left\langle \vcenter{\hbox{\includegraphics{diag2}}} \right| + \cdots \right) \psi \left(\left| \vcenter{\hbox{\includegraphics{diag3}}} \right\rangle + \cdots \right). \tag{4.68}$$

The first possibility gives us the unrenormalized wave function of the occupied state above the Fermi sea, and the second possibility gives us the unrenormalized wave function of an occupied state in the Fermi sea; so we find that in the so-called interactive single-electron wave function $\langle\Omega|\psi(\boldsymbol{r})|\text{single-electron}\rangle$ on the conduction band, we do have contributions from the valence band.

Now we can use $\langle\Omega|\psi(\boldsymbol{r})|\text{single-hole}\rangle$ to construct Fock states (which contain non-interactive Fermi seas) under which contributions from the $g^n(n \geq 1)$ terms in (4.68) cancel each other, and do single-electron calculations and this will not lead to any problem, because these "single-electron wave functions" appear in the Lehmann representation of the single-electron Green function, which is the source of all single-electron quantities. But the effect of Coulomb interaction is not just how $\phi_{n\boldsymbol{k}}$'s (or even ϕ_n's, since the momentum actually is ill-defined – although this is not a serious problem for Fermi liquid) in a Fock state should be modified, but also that either the ground state or the excited states are no longer Fock states any more after Coulomb Interaction is turned on. This difference in the structure of the eigenstates after Coulomb interaction can be seen by excitons. In the "exciton wave function" $\langle\Omega|\psi^\dagger(\boldsymbol{r})\psi(\boldsymbol{r}')|\text{exciton}\rangle$, just like the case in (4.68), the exciton many-body state vector contains a small component of the non-interactive Fermi sea, while $\langle\Omega|$ contains a small component of a non-interactive electron-hole pair, and the contribution of these two components to the exciton wave function clearly contains $\phi_v(\boldsymbol{r}')$ and $\phi_c^*(\boldsymbol{r})$, so it seems the valence band wave function is playing the role of the electron part in the exciton while the conduction band wave function is playing the role of the hole part in the exciton, and hence the B terms in $\chi_S(\boldsymbol{r},\boldsymbol{r}')$.[8]

We do note that these "de-excitation" terms are always there in the four-point electron Green function: for example, when there is no Coulomb interaction, we have

$$\chi^0_{12,34} = \delta_{13}\delta_{24}\frac{f_1 - f_2}{\omega + E_1 - E_2 + \mathrm{i}0^+},$$

---

[8]From this explanation we also see an important difference between Goldstone diagrams for Brillouin-Wigner perturbation theory and the approach here: the "wave functions" calculated in the former formalism via Goldstone diagrams are indeed the many-body wave function components under a Fock basis, and therefore it's $\langle 0|\psi^\dagger(\boldsymbol{r})\psi(\boldsymbol{r}')|\text{exciton}\rangle$, and therefore all the conduction-to-valence transitions are absent. In calculations of optical responses etc., however, the "wave function" used in this note, namely $\langle\Omega|\psi^\dagger(\boldsymbol{r})\psi(\boldsymbol{r}')|\text{exciton}\rangle$, is more useful, because the system resides at $|\Omega\rangle$ and not $|0\rangle$ before the perturbation starts, and therefore the response always takes the form of $\langle\Omega|\text{evolution and measurement}|\Omega\rangle$, which eventually is reduced to a polynomial of things like $\langle\Omega|\psi^\dagger(\boldsymbol{r})\psi(\boldsymbol{r}')|\text{exciton}\rangle$.

and clearly we have $\omega < 0$ poles. But here the de-excitation transitions are not coupled to the excitation transitions, while when Coulomb interaction is turned on, the de-excitation transitions are coupled to the excitation transitions (although the coupling is usually small for excitons). In principle, valence-to-valence transitions or conduction-to-conduction transitions are also possible, but since each term in BSE contains the free electron-hole propagator, which, under the Fock basis where electron orbitals are renormalized by the single-electron self-energy, contains the $f_2 - f_1$ factor shown above, the exciton wave function doesn't really contain these transitions.

Following the same logic, it can be seen that when we are dealing with a *plasmon*,

Chapter 5

The QuantumESPRESSO-BerkeleyGW ecosystem

5.1 Overview of the pipeline

TODO: whether and how BerkeleyGW supports HSE06. See https://groups.google.com/a/berkeleyg

Note that the division of labor is different in the *GW* step and the BSE step. The `sigma` program doesn't really do diagonalization, so building Σ and finding quasiparticle energies are done in one step, which is implemented in `sigma`. On the other hand, diagonalization *is* needed for BSE: this can be seen from the physical picture of the exciton as a pair of electrons and holes constantly changing their momenta. So building the kernel – counterpart of Σ – is done in one step (`kernel`), while diagonalizing it is done in another step (`absorption`). Of course, for "*GW* for exciton" – where the Coulomb line is replaced by, say, the effective interaction line induced by phonons – the division of labor seen in the *GW* part of BerkeleyGW should be used; but this is not implemented in BerkeleyGW anyway.

5.2 Relativistic effects

BerkeleyGW supports SOC calculation, with the so-called fully relativistic mode and the scalar relativistic mode; I say "so-called" because all the calculations are still done in the framework of Schrodinger equation, with relativistic effects being introduced by adding perturbative terms like $\boldsymbol{L} \cdot \boldsymbol{S}$ or $\boldsymbol{p}^4$ [6]. On the other hand, when it comes to calculating the inner structure of atoms (which, in condensed matter physics, is usually seen when we generate our own pseudopotentials), being "fully relativistic" means using *Dirac equation*.

5.3   Input and output of `pw`

5.4   The `epsilon` step

5.4.1   Procedure and speed

What `epsilon` does, as its name implies, is to calculate ϵ – and since ϵ is used to find W, we need ϵ^{-1}. The relative equations are (8-10) in [14]; relative discussions can be found in § 4.2. There are three steps in `epsilon`:

1. Calculate $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$;

2. Summing over $n, n', \boldsymbol{k}$;

3. Finding ϵ^{-1}.

With a fixed accuracy requirement, the time cost of first step is $\sim N^3 \log N$, where N is the number of atoms per unit cell. The $\boldsymbol{k}$ and $\boldsymbol{q}$ points are given by the $\boldsymbol{q}$-grid given in `epsilon.inp` and the $\boldsymbol{k}$-grid in the wave function files, so they are fixed are not a part of the scaling. With the cutoff energy fixed, the size of the $\boldsymbol{G}$-grid is proportional to V, which is in turn proportional to N (the distance between atoms is roughly fixed, and therefore the more atoms we have, the larger the unit cell is). With a fixed accuracy standard, the required numbers of occupied bands and empty bands are all $\sim N$, so the number of matrix elements of $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$ scales as N^3. For each matrix element, we need to calculate $\langle n, \boldsymbol{k} + \boldsymbol{q} | \mathrm{e}^{\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}} | n'\boldsymbol{k} \rangle$. Note that the matrix inside contains only $\boldsymbol{r}$, and the expression therefore can be evaluated as

$$\int \mathrm{d}^3 \boldsymbol{r} \, \phi^*_{n,\boldsymbol{k}+\boldsymbol{q}}(\boldsymbol{r}) \mathrm{e}^{\mathrm{i}(\boldsymbol{q}+\boldsymbol{G})\cdot\boldsymbol{r}} \phi_{n'\boldsymbol{k}}(\boldsymbol{r}),$$

and the scale of the calculation needed is proportional to V and again N. In practice, we use the $\boldsymbol{G}$ representation to calculate the matrix element, and again the time cost is $\sim N$. (Note that the two estimations are equivalent: by saying the time cost is proportional to V, we implicitly imply the absolute spatial resolution is fixed, which, in other words, means we fix the cutoff energy.) So naively, the time cost is $\sim N^4$. Fortunately the matrix element $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$ with fixed $n, n', \boldsymbol{G}$ can be evaluated by fast Fourier transformation, so eventually, the time cost scales like $N^3 \log N$.

The time cost of the second step, in a serial program, scales like N^4. We sum over n and n', each of which has $\sim N$ values. And we need to calculate $\chi_{\boldsymbol{G}\boldsymbol{G}'}$, where the values of $\boldsymbol{G}$ and $\boldsymbol{G}'$ are all roughly proportional to N. So the final scaling of the time cost is N^4. This however can be parallelized, and eventually, in a well optimized parallelized package, the time cost scales like N^2.

The time cost of the third step – the matrix inversion step – scales like N^3.

5.4.2 Divergence problems when $q \to 0$

When calculating $\epsilon_{GG'}(q, \omega = 0)$, we notice that when $G = G' = 0$, the matrix element diverges as $q \to 0$. For an insulator, we have

$$\epsilon_{00}(q \to 0, \omega = 0) \propto -\frac{1}{q^2}\chi_{00}(q \to 0, \omega = 0)$$

$$\propto \sum_{n \text{ occupied}, n' \text{ empty}} -\frac{1}{q^2}|\langle nk|1 + \mathrm{i}(q + G) \cdot r + \cdots|n'k\rangle|^2 \quad (5.1)$$

$$\propto \text{const.} \times \frac{q^2}{q^2}.$$

Here the first term in the Taylor expansion of $\mathrm{e}^{\mathrm{i}(q+G)\cdot r}$ vanishes because of orthogonality conditions. We see $\epsilon_{00}(q \to 0, \omega = 0)$ has a definite value.

For a metal, some bands are both occupied and empty, so we can no longer use the orthogonality conditions, and $\epsilon_{00}(q \to 0, \omega = 0)$ scales like C/q^2. TODO: whether this is a numerical artifact or not To decide the constant C, very find description of the Fermi surface is needed, so we need a very fine k-grid. On the other hand, we don't really need many bands, because for the metal $\epsilon_{00}(q \to 0, \omega = 0)$, most of the relevant transitions are inter-band ones.

5.4.3 Frequency dependence of ϵ

The $\epsilon_{GG'}(q, \omega)$ matrix has frequency dependence. The fastest way to handle this is to calculate the $\omega = 0$ case only and then find the plasmon frequency, and then, the $\epsilon_{GG'}(q, \omega)$ curve can be fitted using sum rules. TODO

The spectral function – used in ARPES simulation – can't be accurately obtained just by the above mentioned method. Thus, **full frequency** treatments are needed. A trick called **static subspace** can speed up the process. TODO

5.4.4 Console output

The console output of `epsilon` has the following structure:

1. Initialization

2. Iterating over the q-grid. For each q-point,

   (a) The output starts with something like

   ```
   ================================================================================
     ↪
    13:59:21    Dealing with q =    0.000000   0.500000
       ↪ 0.000000      5 / 35
   ================================================================================
     ↪

   This is a regular non-zero q-point.
   ```

   (b) Then we can see lines about `Rank of the polarizability matrix`, `BLACS processor grid`, and `Number of k-points in the irreducible BZ(q) (nrk)`.

(c) Now we enter the first step mentioned in § 5.4.1. The start line looks like

```
Started calculation of matrix elements with 324
    ↪ transition(s) at 13:59:21.
```

(d) TODO: what's the corresponding step of the following line:

```
Started building polarizability matrix with 320 processor
    ↪ (s) at 13:59:30.
```

5.4.5 Other output files

The `epsilon` program generates the files mentioned in this page. Note that `eps0mat.h5` is *not* ϵ from DFT orbitals! In steps after `epsilon`, screening obtained from DFT orbitals are not used; we only use ϵ from *GW* orbitals. The file `eps0mat.h5` is the ϵ around the Γ point. Thus, in principle, we can calculate `eps0mat.h5` and `epsmat.h5` in two runs, and indeed this is the case when we deal with a metallic system (§ 7.2.4).

5.5 Absorption

5.5.1

5.6 Systems of units

There are several system of units used in QuantumESPRESSO-BerkeleyGW ecosystem. The choice is often well-motivated: the cutoff energy, for example, is always in Ry; we note that

$$1\,\mathrm{Ry} = \frac{\hbar^2}{2ma_0^2}, \tag{5.2}$$

and therefore if $\boldsymbol{G} = \boldsymbol{n} \cdot 1/a_0$, we have

$$\frac{\hbar^2 \boldsymbol{G}^2}{2m} \le E_{\mathrm{cut}} \Leftrightarrow \left(\frac{2\pi a_0}{a}\right)^2 n_x^2 + \left(\frac{2\pi a_0}{b}\right)^2 n_y^2 + \left(\frac{2\pi a_0}{c}\right)^2 n_z^2 \le E_{\mathrm{cut}}/\mathrm{Ry}. \tag{5.3}$$

Chapter 6

Tight-binding models

TODO: from ab initio to models

6.1 Wannier functions and tight-binding models

For an insulator, we can construct exponentially decaying Wannier functions if and only if the Chern numbers are zero, and thus all insulators with inversion symmetry have exponentially decaying Wannier functions [9]

[23] introduces **maximally localized Wannier functions**. Although we usually say Wannier functions are Fourier transformations of Bloch functions, the fact that we have a gauge freedom in the global phase of each Bloch state – which may depend on $\boldsymbol{k}$ – means Wannier functions are strongly non-unique. We want Wannier functions to be as localized as possible.

The standard of localization is the **localization functional**

$$\Omega = \sum_n (\langle \boldsymbol{r}^2 \rangle_n - \langle \boldsymbol{r} \rangle_n^2). \tag{6.1}$$

It's possible to use density overlap or Coulomb self-interaction or something else to decide how localized a wave function is. The gauge-dependent part of the functional is

$$\tilde{\Omega} = \sum_n \sum \text{TODO} \tag{6.2}$$

It contains a derivative $\boldsymbol{\nabla}_{\boldsymbol{k}}$, and the usual numerical format is used.

Suppose we choose J bands (**frozen window**) and a set of smooth trial orbitals $g_n(\boldsymbol{r})$. We define

$$|\phi_{n\boldsymbol{k}}\rangle = \sum_{m=1}^{J} |\psi_{m\boldsymbol{k}}\rangle \, \langle \psi_{m\boldsymbol{k}} | g_n \rangle . \tag{6.3}$$

Now the arbitrary phase factors all cancel out. Picking up g_n requires chemical intuition: if, say, a band mainly consists of electrons from a sp^2 chemical bond, then we should let g_n to be the sp^2 orbitals.

6.2 The cRPA approach to obtain effective models

Suppose we already have identified the bands that cause strongly correlated effects and are now ready to write an effective theory for them. Below, I use the subscript d for the bands considered and r for the rest. Then we have [4]

$$W = (1 - W_r P_d)^{-1} W_r, \quad W_r = (1 - v P_r)^{-1} v, \tag{6.4}$$

where P_d is the polarization *within* the d subspace, and P_r is the polarization involving at least one state in the r space (and may also involve a state in d subspace). The exact form of P_d is hard to obtain, because the d bands are strongly correlated and therefore usual Feynman diagram resummation schemes fail for them, but r bands are dispersive and extended enough for a diagrammatic resummation scheme like GW to work. Thus, the part concerning only r electrons in P_r can be reliably obtained using, say, RPA, which can be done using existing *ab initio* codes; the hopping between d bands and r bands is more tricky, TODO: why RPA for them is acceptable

6.3 Interaction channels

Several interaction channels with clear physical pictures can be identified from the tight-binding (screened) Coulomb repulsion TODO: $1/2N$ or ?

$$H = \sum_{\sigma, \sigma'} \sum_{m,n,r,s} U_{mnrs} c_{m\sigma}^\dagger c_{n\sigma'}^\dagger c_{r\sigma'} c_{s\sigma'}. \tag{6.5}$$

The **direct density-density interaction** is

$$\sum_{m,n} \underbrace{U_{mnnm}}_{V_{mn}} n_m n_n, \quad n_m = \sum_\sigma c_{m\sigma}^\dagger c_{n\sigma}. \tag{6.6}$$

Specifically, we have **on-site repulsion**

$$\sum_m U_{mmmm} c_{m\sigma}^\dagger c_{m\sigma'}^\dagger c_{m\sigma'} c_{m\sigma} = \sum_m \underbrace{U_{mmmm}}_{U} n_{m\uparrow} n_{m\downarrow}. \tag{6.7}$$

This is just the Hubbard-U term. If we can attribute the influence of other interaction channels to band structure self-energy correction, we just get the **Hubbard models**

$$H = -\sum_{m,n,\sigma} t_{mn} c_{m\sigma}^\dagger c_{n\sigma} + U \sum_m n_{m\uparrow} n_{m\downarrow}. \tag{6.8}$$

This deceptively simple model takes people more than half of a century to solve – and probably more, because even today, there is still no generalized way to find its ground state even on a 2D lattice.

The **exchange interaction** is

$$\sum_{m \neq n, \sigma, \sigma'} U_{mnmn} c_{m\sigma}^\dagger c_{n\sigma'}^\dagger c_{r\sigma'} c_{s\sigma'} = -2 \sum_{m \neq n} \underbrace{U_{mnmn}}_{J_{mn}} \left(\boldsymbol{S}_m \cdot \boldsymbol{S}_n + \frac{1}{4} n_m n_n \right). \tag{6.9}$$

Here we have used the formula

$$\boldsymbol{\sigma}_{\alpha\beta} \cdot \boldsymbol{\sigma}_{\gamma\delta} = 2\delta_{\alpha\beta}\delta_{\beta\gamma} - \delta_{\alpha\beta}\delta_{\delta\gamma}. \tag{6.10}$$

We call this the exchange interaction, because the way we find it is mathematically identical to splitting the expectation of Coulomb interaction into the direct and the exchange terms using first quantization. The exchange interaction supports the first Hund's rule: if spins align with each other, the total energy can be reduced (it's more likely that $U_{mnmn} > 0$ because Coulomb interaction is repulsive), and therefore the total spin is to be maximized.

Chapter 7

Carrying out calculations

List of technical issues:

- Band mode of BerkeleyGW
- Wannier "oscillation"
-

7.1 Details in installation

7.1.1 QuantumESPRESSO

What I write below is about QuantumESPRESSO 6.7Max; the names of compilation options and tags may change, but the overall idea is the same.

To enable OpenMP, the option `--enable-openmp` has to be added when we run `./configure`. MPI parallelization however can't be obtained with adding this tag only: we need a MPI-wrapped compiler, as well as appropriately linked libraries like ScaLAPACK. The last time I tried this on Cori, the key point is to load module `impi` (i.e. IntelMPI) before running `./configure`. It seems following this documentation page complicates matters (but probably they will update this page, and it then becomes useful; anyway, what to be remembered is we need MPI compilers to get the fully parallelized version).

Whether an MPI compiler is present can be checked by looking at `make.inc`; alternatively, you may pay attention to whether the message `WARNING: parallel and serial compiler are the same` appears in the output of `make`.

7.2 Standard operation procedures

7.2.1 Avoid data pollution

The follows are strongly recommended whenever modifying files:

- Read out loud the *directory* of the file being modified. Are you mistakenly modify files in another directory?
- When copying file from a folder to another, similarly read out loud the name of the source and the target.

- If a file is going to be overwritten, make sure it's really no longer needed. Whenever you have doubts, make backups.

-

7.2.2 Finding the structure

7.2.2.1 From open data

Sometimes the atomic positions are given with the help of Wyckoff positions: an example can be found in [22]. In this case, the number of atomic positions given is *less* than the total numbers of atoms in a primitive unit cell. The difference between the two can be found according to the multiplicity of the Wyckoff positions involved – in the case of [22], all atomic positions given are in the 2a position, the multiplicity of which is two, and 6 atomic positions are given, so we have $2 \times 6 = 12$ atoms per primitive cell. But then the primitive cell mentioned in the experimental paper may be a primitive cell in the bulk while it's possible that we only want a layer in the bulk material (i.e. we are dealing with the monolayer version), and then we should only pick the atoms that are close enough to each other in the z direction.

7.2.2.2 Comparing existing structures for the same material

One good idea is to first find some high symmetry points (the inversion center, the reflection mirror, etc.) and then compare the relative relations of these positions in the two structures.

7.2.3 Insulator DFT+GW+BSE

7.2.3.1 The DFT stage

1. Do a `scf` calculation in `1-scf`.

2. Do a `bands` calculation in `2.1-wfn`. This step includes:

   (a) Create a `the-suffix-you-set.save` folder in `2.1-wfn`, and link `data-file-schema.xml` and `charge-density.dat` from `1-scf` into this folder. These are files required for a `bands` calculation.

   (b) Run

   ```
   data-file2kgrid.py --kgrid nx ny nz the-
       ↪ suffix-you-set.save/data-file-schema.
       ↪ xml kgrid.inp
   ```

   to create `kgrid.inp`, which describes how to create a $\mathbf{k}$-grid with size `nx ny nz`. This can't be done with options in QuantumESPRESSO's `KPOINTS` section because QuantumESPRESSO and BerkeleyGW have different tolerance for symmetry.

   (c) Run

   ```
   kgrid.x kgrid.inp kgrid.out kgrid.log
   ```

   to obtain `kgrid.out`. The content in `kgrid.out` will be used as the `KPOINTS` section for the input file of `pw`.

(d) Preparing the `bands.in` file, which is the input file of `pw.x`. Do the following checklist:

- Whether `calculation` is `bands`.
- Whether `pseudo_dir` is correct.
- Whether `nbnd` is set to, say, 1000.
- Whether `lspinorb = .true.` and `noncolin = .true.` are set for an SOC run.

(e) Run `pw2bgw.x` in `2.1-wfn`. Do the following checklist:

- This step should be done with *exactly the same* parallelization setting with `pw.x`.
- The `wfng_nk1`, `wfng_nk2`, `wfng_nk3` parameters should be set to `nx`, `ny`, `nz` mentioned above. (This item needs double check especially when `pw2bgw.inp` comes from another run.)
- Whether `rhog_flag` is `.true.`.
- Whether `vxc_flag` is `.true.`.
- Whether `wfng_flag` is `.true.`.

3. Do a `bands` calculation in `2.2-wfnq`. The steps are similar to `2.1-wfn`:

(a) Linking files from `1-scf`.

(b) Run

```
data-file2kgrid.py --kgrid nx ny nz --qshift
    ↪ qx qy qz the-suffix-you-set.save/data-
    ↪ file-schema.xml kgrid.inp
```

to get the `kgrid.inp` file. Here `qx qy qz` is a small displacement used to regularize Coulomb interaction at $q = 0$. A common choice is `0 0 0.001`; when dealing with a 2D material, choose `0 0.001 0`, because with `cell_slab_truncation` open in the `epsilon.x` step, non-zero z components of k-points are forbidden.

(c) Run `kgrid.x`.

(d) Preparing the `bands.in` file.

(e) Run `pw2bgw.x`. Do the following checklist:

- This step should be done with *exactly the same* parallelization setting with `pw.x`.
- The `wfng_nk1`, `wfng_nk2`, `wfng_nk3` parameters should be set to `nx`, `ny`, `nz` mentioned above.
- The `wfng_dk1`, `wfng_dk2`, `wfng_dk3` parameters should be set to `wfng_nk1` × `qx`, etc. (If `kshift` is used, it also should be added to `wfng_dk1`, etc.)

   Note that this doesn't mean the displacement imposed to the k-grid is `wfng_nk1` × `qx`: the displacement is still `qx qy qz`. Here the `wfng_dk1`, `wfng_dk2`, `wfng_dk3` are conventional parameters used in Monkhorst-Pack grids, and `wfng_dk1 = 0.5` means the grid is shifted towards x direction by half a *grid step* – and therefore the displacement is 0.5×1 / `wfng_nk1` in the crystal coordinates. Now we understand why we need to set `wfng_dk1` to `wfng_nk1` × `qx`. Indeed, below is a part of the header of a `WFN` file in `2.2-wfnq`:

```
k-grid:   24  24   1
k-shifts:      0.000000     0.024000     0.000000
[ifmin = lowest occupied band, ifmax = highest
    ↪ occupied band, for each spin]
   Index         Coordinates (crystal)
      ↪ Weight   Number of G-vectors      ifmin
      ↪ ifmax
      1    0.000000     0.001000     0.000000
           ↪ 0.001736                   36275          1
           ↪         120
```

It can be seen that the first k-point is displaced 0.001 in the y direction, and the k-shifts parameter corresponding to the y direction is 0.024; since the size of the grid in the y direction is 24, the displacement instructed by the latter is $0.024/24 = 0.001$, exactly the displacement recorded in the first k-point.

7.2.3.2   The GW stage

1. Do a epsilon calculation in 1-epsilon. The steps are listed below:

   (a) Linking files. Come to 1-epsilon and do the follows:

   ```
   ln -sf ../2.1-wfn/WFN
   ln -sf ../2.2-wfnq/WFN ./WFNq
   ```

   (b) Prepare epsilon.inp. Do the follow checklist:
   - Whether we are setting qpoints instead of kpoints.
   - Whether there is an end line of the qpoints block.
   - Whether each line of the qpoints block is in the format (see here)

     ```
     qx qy qz 1 is_q0
     ```

   - Especially, whether the line corresponding to the Γ point has is-q0 = 1.
   - Don't forget to replace the coordinates of the Γ point with the displaced point in 2.2-wfnq.
   - If we are dealing with a 2D material, add cell_slab_truncation.
   - Set epsilon_cutoff to, say, 10; the exact value is to be decided by convergence tests.
   - Set number_bands to the highest *total* number of bands allowed by degeneracy_check.x (§ 8.6.4).

2. Do a sigma calculation in 2-sigma. The steps are listed below:

   (a) Link necessary files:

   ```
   ln -sf ../2.1-wfn/vxc.dat
   ln -sf ../2.1-wfn/RHO
   ln -sf ../2.1-wfn/WFN ./WFN_inner
   ln -sf ../1-epsilon/epsmat.h5
   ln -sf ../1-epsilon/eps0mat.h5
   ```

   (b) Prepare sigma.inp. Do the following checklist:
   - Whether we are setting kpoints instead of qpoints. (This time it's not qpoints!)

- Whether there is an `end` line of the `kpoints` block.
- Whether each line of the `kpoints` block is in the format

```
kx ky kz 1
```

- If we are dealing with a 2D material, add `cell_slab_truncation`.
- Set `number_bands` to the same value in `epsilon.inp`.
- Set `band_index_min` and `band_index_max`. The bands between the two are corrected by (4.62), and others are not.

7.2.3.3 The BSE stage, with $Q = 0$

The `epsilon` run used in a BSE run may have a lower cutoff energy and number of bands; that's because in exciton physics, the distance between the electron and the hole usually isn't much smaller than the lattice constants (the exciton radius is actually $\sim$ the Bohr radius), and therefore $\epsilon_{GG'}$ only matters when G and G' are small; thus we don't need a large cutoff energy, and consequently we don't need a large number of bands.

1. Do a `kernel` calculation in `3-bse`. The steps are listed below:

   (a) Link necessary files.

   ```
   ln -sf ../1-epsilon/WFN ./WFN_co
   ln -sf ../1-epsilon/epsmat.h5
   ln -sf ../1-epsilon/eps0mat.h5
   ```

   (b) Prepare `kernel.inp`. Do the following checklist:
   - Whether the following lines are there, if the coarse grid only contains the irreducible 1BZ:

   ```
   use_symmetries_coarse_grid
   ```

   Otherwise *do not* turn on this option, as it may lead to possible errors in the dipole matrix elements.
   - Whether `number_val_bands` and `number_cond_bands` are specified.
   - If we are dealing with a 2D material, whether `cell_slab_truncation` is added.

2. Do an `absorption` step.

   (a) Build a denser k-grid and repeat `2.1-wfn` and `2.2-wfnq` steps.
   (b) Link necessary files.

   ```
   ln -sf ../2.1-wfn/WFN ./WFN_co
   ln -sf ../2.2-wfnq/WFN ./WFNq_co
   ln -sf ..2.1-wfn-dense/WFN ./WFN_fi
   ln -sf ..2.2-wfnq-dense/WFN ./WFNq_fi
   ln -sf ../1-epsilon/epsmat.h5 ./
   ln -sf ../1-epsilon/eps0mat.h5 ./
   ln -sf ../3-kernel/bsemat.h5 ./
   ```

   The `bsedmat` and `msexmat` files are missing if HDF5 format is used to generate the output files – the documentation of BerkeleyGW says we need the files but it's wrong: When `bsemat.h5` is present, the program runs well.

(c) Prepare `absorption.inp`. Do the following checklist:
  - Whether the following lines are in `absorption.inp`, *depending on whether the coarse and fine grids only contain the irreducible 1BZ, or otherwise there may be errors in the dipole matrix elements:*

    ```
    use_symmetries_fine_grid
    use_symmetries_coarse_grid
    ```

  - Whether

    ```
    number_cond_bands_coarse
    number_cond_bands_fine
    number_val_bands_coarse
    number_val_bands_fine
    ```

    are specified.
  -

7.2.3.4 The BSE stage, $Q \neq 0$

7.2.4 Metal DFT+GW+BSE

For metals, the q-displacement technique can no longer be used. The working procedure now is

1. Do a `scf` calculation in `1-scf`.

2. Do a `bands` calculation in `2.1-wfn`.

3. Do `2.2-wfn0` with a *finer* k-grid, still *without* any `qshift` displacement.

4. Do a `epsilon` calculation in `1-epsilon`. The steps are listed below:

   (a) Link files according to

   ```
   ln -sf ../2.1-wfn/WFN
   ln -sf ../2.1-wfn/WFN ./WFNq
   ```

   The `2.2-wfnq` step is not needed, because we are not going to deal with the Γ point in this step.

   (b) Preparing `epsilon.inp`. Do the following checklist:
     - Whether we are setting `eqpoints` instead of `kpoints`.
     - Make sure the Γ point is *not* included in the `qpoints` block.
     - Whether each line of the `qpoints` block is in the format

       ```
       qx qy qz 1 0
       ```

5. Do a `epsilon` calculation in `1.2-epsilon0`.

   (a) Link files according to

   ```
   ln -sf ../2.2-wfn0/WFN
   ln -sf ../2.2-wfn0/WFN ./WFNq
   ```

   Now the outputs of the `2.1-wfn` step is not used.

   (b) Prepare `epsilon.inp`. Do the following checklist:
     - Whether we are setting `eqpoints` instead of `kpoints`.
     - The *only point* included in the `qpoints` block should be the non-zero k-point with smallest length in the k-grid used in `2.2-wfn0`.

- Whether the Γ point is in the format

```
qx qy qz 1 2
```

7.2.5 Hartree-Fock calculation

The `sigma` code can also be used to do Hartree-Fock calculation. No `epsilon` run is needed for this kind of calculation. The standard operation procedure is like the follows:

1. `WFN_inner`, `vxc.dat` and `RHO` need to be linked; `epsmat.h5` and `eps0mat.h5` are not needed.

2. The `sigma.inp` file should be changed in the following way:
   - Add a line: `frequency_dependence -1`
   - Copy an insulator-style `qpoints` to `sigma.inp`: the `is_q0` flag for the Γ point with displacement should be set to `1`; the `is_q0` flag for other points is `0`.
   - Add a `qgrid` line, giving the size of the full first Brillouin zone, like `10 10 1`.

7.2.6 Band plot

7.2.6.1 DFT level: k-path

To plot the bands, just do a `bands` calculation where the `K_POINTS` block is in `crystal_b` mode.

The results can be processed by the `bands.x` utility. An example of the input file:

```
&BANDS
prefix = 'WTe2'
outdir = './'
filband = 'WTe2_bands.dat'
lsigma(1) = .true.
lsigma(2) = .true.
lsigma(3) = .true.
lsym = .false.
/
```

Here the `lsigma` options are only available for a full SOC calculation. A script used to read the output file can be found here.

Note that when plotting the band structure, the Fermi energy is to be found in `scf.out` or in a `nscf` run. It won't appear in a `bands` run.

7.2.6.2 *GW* level: `inteqp`

The *GW* level bands can be obtained by the `inteqp` program. Note that this shouldn't be used for a system considered metallic by DFT (§ 8.9.4). The standard operation procedure is listed here:

1. Go to `4-path` – the folder responsible for the DFT level calculation – and perform a `pw2bgw` run needed to create a `WFN` file. Note that if we are not dealing with a k-grid, `wfng_nk1`, `wfng_nk2`, and `wfng_nk3` should be skipped.

2. Create a folder within `1-epsilon/`, and go into it.

3. Link the necessary files:

```
ln -sf ../../2.1-wfn/WFN ./WFN_co
ln -sf ../../4-path/WFN ./WFN_fi
cp ../../1-epsilon/eqp1.dat ./eqp_co.dat
```

4.

7.2.6.3 *GW* level: using `WFN_outer`

It's possible to do the same thing in § 7.2.6.1 with BerkeleyGW: we can

7.2.6.4 BSE level

7.2.7 Wannier functions and tight-binding models

Tight-binding models are a more compact way to show the band structure.

7.2.7.1 `wannier90` for DFT

1. The first step is to perform a fresh QuantumESPRESSO run. `wannier90` requires a full Monkhorst–Pack grid, and to create a `bands.in` file, (don't forget to add the `utility` folder to the `PATH` environment variable) just prepare a `bands.in` file without the `K_POINTS` block and run

```
kmesh.pl 20 20 1 >> bands.in
```

Replace `20 20 1` by the k-grid size you want. The number of bands in this DFT run *shouldn't* be too large, or otherwise `pw2wannier90` will be extremely slow.

2. Prepare a `prefix.win` file, where `prefix` is the QuantumESPRESSO prefix used. Examples of this input file can be found by searching "wannier90 2007 workshop" and downloading the tutorial files coming together with the web page about this event – currently it's here. Do the following checklist:

   - Is `num_bands` plus the number of excluded bands the same as `nbnd` in the QuantumESPRESSO run (§ 8.7.2)? If you don't want all valence bands in the DFT run, always decide how many bands to include *before* doing any `wannier90.x` run; changing parameters in `prefix.win` after a `wannier90.x` run may lead to inconsistencies.

3. Run

```
wannier90.x -pp prefix.win
```

to get `prefix.nnkp`.

4. Run `pw2wannier90.x`. The input file looks like

```
&inputpp
outdir     =  './'
prefix     =  'WTe2'
seedname   =  'WTe2'
scdm_proj  =  .true.
write_amn  =  .true.
write_mmn  =  .true.
write_unk  =  .true.
/
```

This step should finish in several minutes. If it takes too long, reduce `nbnd` in the QuantumESPRESSO step and redo the whole procedure. This step generates an `.eig` file.

5. Run

```
wannier90.x prefix.win
```

7.2.7.2 GW level: `sig2wan`

The `sig2wan` code can extract a Wannier interpolated band structure from output files of the `sigma` step. An example can be found here. The routine is almost the same as `wannier90` for DFT, except for one thing: after running `pw2wannier90.x`, we need to replace the `perfix.eig` file generated by `pw2wannier90.x` by the `prefix.eig` file generated by `sig2wan`.

7.2.8 Self-consistent GW

Self-consistent schemes in GW are listed in § 4.4.2. In this section, I talk about how to do them.

7.2.8.1 Energy self-consistent calculation in GW

The `epsilon` code has an option called `eqp_corrections`, which takes the

7.2.8.2 Eigenstate self-consistent calculation in GW

Eigenstate self-consistent GW – in other words, diagonalizing Σ instead of taking into account only the diagonal elements – is realized by `scGWtool.py`. The procedure is listed below:

1. Perform a DFT run, with *non-diagonal* VXC in the `pw2bgw` step. An example of the part of `pw2bgw.in` concerning VXC:

   ```
   vxcg_flag = .true.
   vxc_diag_nmax    = 1000
   vxc_offdiag_nmax = 1000
   ```

   Note that here we are using VXC instead of `vxc.dat` as the output format of V_{xc}. Whenever weird errors occur with the `pw2bgw` step, redo the `bands` step and do `pw2bgw` in exactly the same parallelization environment.

2. Run `sigma` with the options

   ```
   sigma_matrix -1 0
   dont_use_vxcdat
   ```

and rename the output file `sigma_hp.log` to `sigma_hp_col.log`.

3. Run `sigma` with the options

   ```
   sigma_matrix -2 0
   dont_use_vxcdat
   ```

   and rename the output file `sigma_hp.log` to `sigma_hp_row.log`. TODO: understand what's going on in this step; the explanation involving "the lower triangle and the upper triangle" doesn't seem to agree with the official documentation

4. Rename (or link, or whatever) the `WFN_inner.h5` file used in the last steps as `wfn_old.h5`. If the last two steps are not done with the HDF5 format (i.e. they are done with the Fortran binary format), a `wfn2hdf.x` run is needed.

5. Run `scGWtools.py` to build and diagonalize the quasiparticle Hamiltonian. The output files are `vxc_new.dat` and `wfn_new.h5`.

6. Rename the output files into `vxc.dat` and `WFN.h5`. Run `sigma` with `use_-wfn_hdf5` and remove the `dont_use_vxcdat` option.

7.2.9 Topological invariants with `z2pack`

The error `ValueError:  The given WCC are not degenerate Kramers pairs at the edges of the surface` occurs because one topological band and a trivial band appear in the `bands` option of `z2pack.tb.System`.

7.2.10 Band projection

One thing to keep in mind: If you want to do a DOS calculation for all k's in the first Brillouin zone, *never* do band projection after a k-path run; or otherwise the DOS output will be only about states on the k-path. Also, DOS calculation is extremely sensitive to both energy grid and k-grid.

7.2.11 Finite momentum BSE

1. Prepare the wave function files. Note that finite Q calculation doesn't support symmetry reduction: the `WFN_co` and `WFNq_co` files have to contain the full k-grid.

   TODO: could `epsilon` contain reduced k-grid?

2.

7.3 Performance tricks

7.3.1 Parallelization

When virtual thread is not enabled, we have

$$\text{\# of MPI tasks} \times \text{\# of OpenMP threads} = \text{\# of nodes} \times \text{\# of cores per node}. \tag{7.1}$$

It's a good idea to have the number of MPI tasks divisible by the number of nodes, since grouping CPUs from different nodes under one MPI task means we need to let threads from different nodes communicate with each other frequently, which is highly .

Increasing the number of OpenMP threads per MPI task gives an MPI task more resources. but if the number of nodes used is fixed, this also means the number of MPI tasks is reduced; problems like insufficient memory usually can't be solved by decreasing the number of MPI tasks and allocate more resource to one MPI task: on the contrary, we need to increase the number of MPI tasks so that the burden on each MPI task is smaller.

QuantumESPRESSO uses a parallelization technique called momentum pool: $\boldsymbol{k}$-points are grouped into "pools" which are parallelized over; $\boldsymbol{k}$-points in the same pool however are evaluated one by one.

7.3.2 Choosing cutoff energies wisely

The cutoff energies, especially the one in the GW step, of course should be large enough, but not too large: if in a benchmark test, a smaller cutoff energy gives almost the same result compared to a higher cutoff energy, then the smaller cutoff energy should be used unless we have reasons against this practice.

7.3.3 `pseudobands`

One observation is that the exact dispersion relations of high-energy bands required in GW aren't that important: these bands are there mainly because we need to make the normalization correct (TODO: citation needed). This fact can be exploited to speed up calculation.

Fig. 7.1 shows one way to do this: we need two parameters, E_1 and λ. We leave all bands whose average energies over $\boldsymbol{k}$ points are below E_1 untouched (we say they are protected). Then we divide bands whose average energies are above E_1 according to the following procedure. Bands whose average values are between E_1 and $E_1 + \lambda E_1$ are placed into block 1; Bands whose average values are between $E_2 = E_1 + \lambda E_1$ and λE_2 are placed into block 2 ... Repeating this until every band is placed in one block. Then, each block is replaced by an averaged band, whose energy is the average of all bands in that block, and whose wave function is the sum of all bands in that block. Note that E_1 starts from the (quite arbitrary) energy zero-point, instead of the Fermi energy, and therefore it's required to check the magnitude of the Fermi energy and make E_1 much larger than that. Using `pseudobands` thus breaks the norm conserving condition. Therefore, in `espilon.inp`, we need to add `dont_check_norms`. This method seems to be mentioned in [12,16], where they mention something called "stochastic pseudobands". The meaning of the term "stochastic" is not clear. Some publications, like [11,19], are cited as the source of this technique, but they never mention a word about `pseudobands`, let alone why summing over wave functions makes any sense. [17] mention a similar method, but it seems to involve no summation among wave functions.

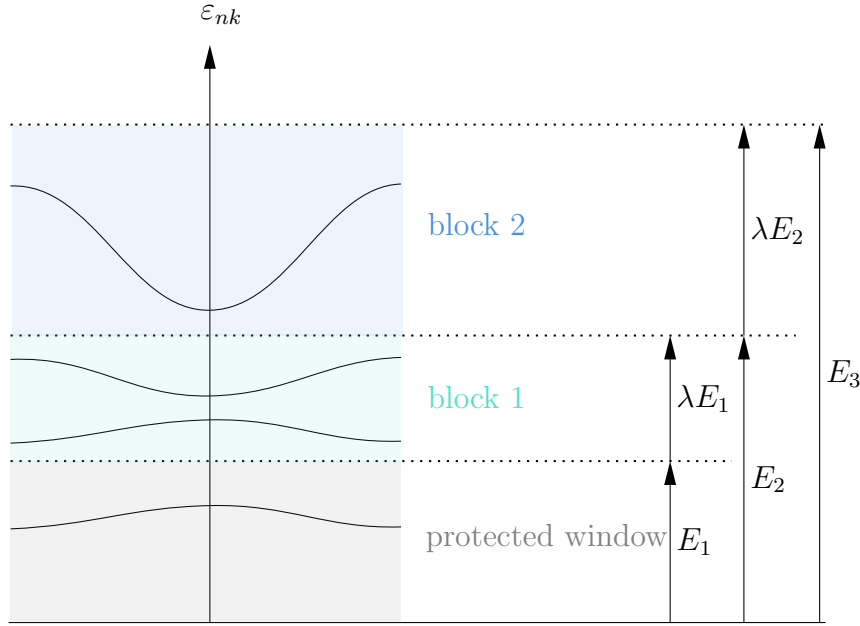The usage of `pseudobands` is shown below:

```
wfn2hdf.x BIN WFN WFN.h5
```

Figure 7.1: One way to average over high-energy bands. Note that the x axis represents the energy zero-point, and not μ. Which block a band is in is decided by its average value.

```
pseudobands.py WFN.h5 WFNo.h5 0.7 0.02
hdf2wfn.x BIN WFNo.h5 WFNo
```

Here `0.7` is E_1 (in Rydberg), and `0.02` is λ.

Details in the code:

- `en_min` and `en_max` are not $E_{\min}$ or $E_{\max}$: they are just the average energy of bands with the first spin direction. It seems `h5py` transposes the input HDF5 file so make it closer to the C/Python convention.

- `nb_fixed`: indices of bands that are to be averaged over

7.3.4 "Manual parallelization"

Sometimes a complete `2.1-wfn` or `sigma` or `epsilon` run just takes too much time; one way to overcome the problem is to manually divide the $\boldsymbol{k}$- or $\boldsymbol{q}$-points and concatenate the output files. Some utilities have been provided by BerkeleyGW.

7.3.4.1 2.1-wfn

From § 7.2.3.2, we know that what is needed in the DFT stage is `WFN` (and `WFNq` or fine-grained `WFN` for metallic calculation), `RHO` and `vxc.dat`. Each of them can be calculated with manual parallelization over Brillouin zone sample points.

1. Do a `pw2bgw` run in `1-scf` folder to get the complete `RHO`;
2. Divide the $\boldsymbol{k}$-points generated by `kgrid.x` into several parts, and prepare QuantumESPRESSO input files for each part.
3. Run QuantumESPRESSO and `pw2bgw` for each part in parallel.

4. Use `wfnmerge.x` to merge `WFN` files. TODO: pw2bgw setting

5. Use `cat vxc.dat-1 vxc.dat-2 > vxc.dat` to create the complete `vxc.dat` file.

7.4   Convergence tests

Here, the term **convergence test** means to make sure the parameters that control the size of the problem are large enough. The parameters included are:

- The $\boldsymbol{k}$- (and therefore $\boldsymbol{q}$-)grid this is defined in the `2.1-wfn` step, which influences the accuracy of the dielectric matrix;

- The cutoff energy in DFT, which influences the accuracy of DFT, and therefore is relevant in the accuracy of both $\varepsilon_{n\boldsymbol{k}}$ and $c_{n\boldsymbol{k}}(\boldsymbol{G}, \sigma)$, hence the accuracy of $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$;

- The number of empty bands given to `epsilon`;

- The cutoff energy in `epsilon`, or in other words the screened cutoff energy;

- The number of bands in CH summation (4.61), i.e. the number of bands in `sigma`.

TODO: problem: do high bands require more $\boldsymbol{G}$ vectors, or do low bands require more $\boldsymbol{G}$ vectors?

List of theoretical problems:

- Wavefunction cutoff only affects bare exchange significantly, can be treated separately; large $\boldsymbol{G}$'s only contribute to bare exchange.

-

More theoretical reading is needed to understand https://www.nersc.gov/assets/Uploads/Conver

7.5   Third-party tools

7.5.1   Running Python 2 scripts

Some auxiliary scripts are written in Python 2. If you need to run them on a cluster where no NumPy is installed for Python 2, you need to create an Anaconda environment to run these scripts:

```
conda create -n wfn-tools python=2.7
conda activate wfn-tools
conda install numpy==1.15.0
```

Chapter 8

Trouble shooting

8.1 Unexpected units

8.1.1 Band energy output of `pw.x`

When invoked in the `bands` or `nscf` modes, `pw.x` gives a list of $\boldsymbol{k}$-points and the corresponding band energies. The unit of the $\boldsymbol{k}$ points is the momentum space version of `alat`, which is *not* based on the crystal coordinates.

8.2 Trouble shooting in MPI

8.2.1 `srun: fatal: Can not execute`

This error may come from an error in compilation, but it can also occur because the directory to the program is misspelt, or an environment variable involved in the directory is not defined, etc.

8.2.2 `error parsing parameters`

This sometimes occurs because you misspell the name of a program to be `mpirun`ed.

8.2.3

8.2.4 Each process is run serially and doesn't communicate with others

A possibility is when compiling the program to be launched in parallel, you choose the serial version (§ 7.1.1).

8.2.5 `Insufficient virtual memor`

Try to reduce the number of MPI processes per node and increase the number of OpenMP threads per MPI process.

8.3 Trouble shooting in Python

8.3.1 AttributeError: 'Dataset' object has no attribute 'value'

This error comes from an update of `hdf5` which deprecated the `.value` attribute. To solve the problem, either downgrade the library or change the statements using `.value`. This error is known in `scGWtool.py`.

8.4 Trouble shooting in QuantumEspresso

8.4.1 Fatal error in PMPI_Comm_free: Invalid communicator

This happens sometimes when doing `2.1-wfn`. It's possible that this comes from some sort of out-of-memory issue. This of course occurs when the resource allocated is not enough, but sometimes it also occurs from wrong settings in input files of the softwares being run. One example: `calculation` is accidentally set to `scf` in a QuantumESPRESSO run when it should be `bands`.

8.4.2 Intel MKL FATAL ERROR: Cannot load symbol MKLMPI_- Get_wrappers.

8.4.3 Program frozen

Check whether too much resource is given to a simple task.

8.4.4 Electron convergence not achieved

This is related to smearing, TODO

8.4.5 gamma_only and noncolin not allowed

This sometimes is due to forgetting to set up the `K_POINTS` block; in this case Quantum ESPRESSO automatically starts a Γ-point only calculation, which is not compatible with SOC.

8.4.6 Error in routine allocate_fft (1): wrong ngms

I'm still not quite sure what causes this error, but it seems to be related to parallelization: in a run with 2240 MPI tasks, the error occurred, but when I used 320 MPI tasks, the error disappeared. The error can occur with `pw.x` or `bands.x`.

8.4.7 Error reading attribute index : expected integer , found *

This error occurs when we use a pseudopotential that is obtained by converting another pseudopotential in a different format (see here). Usually we don't need to "correct" it.

8.4.8 cdiaghg (159): eigenvectors failed to converge

Usually by changing `diagonalization` to `cg`, this can be solved; `cg` is more stable but much slower.

8.4.9 Error in routine cdiaghg (1052): problems computing cholesky

This also seems to be a convergence problem that can be solved by changing `diagonalization` to `cg`.

8.4.10 Error in routine set_occupations (1): smearing requires a vaklue for gaussian broadening (degauss)

This happens whenever smearing is used but the smearing parameter is not set. Note that this is *not* restricted to Gaussian smearing: all smearing schemes are controlled by the `degauss` parameter, and when this parameter is not set, the error occurs.

8.4.11 Error in routine splitwf (36197): wrong size for pwt

Usually this occurs when `pw2bgw` is redone (`pw2bgw` deletes intermediate files, making another `pw2bgw` run impossible). This also appears in cases similar to § 8.4.6. A complete `bands-pw2bgw` run has to be redone.

8.4.12 Error in routine PW2BGW(19):input pw2bow

Usually this occurs when something else happens between a `bands` run and a `pw2bgw` run for it. A complete `bands-pw2bgw` run has to be redone.

8.4.13 Error in routine PW2BGW (19): input_pw2bgw

This can occur when you misspell one option in `pw2bgw.in`.

8.4.14 stress for hybrid functionals not available with pools

As the error message implies, turning `tstress` to `.false.` (or simply deleting anything about this option) solves the problem.

8.4.15 Error in routine projwave (1): Cannot project on zero atomic wavefunctions!

This occurs when running `projwfc.x`. It occurs when atomic wave functions are not shipped together with the pseudopotentials used. This is the case for some types of pseudopotentials, including ONCV TODO: full list.

8.4.16 Error in routine diropn (3): wrong record length

This sometimes occur when too many MPI processes are located for a small number of k-points.

8.4.17 S matrix not positive definite

This usually happens when there are too many empty bands compared with the cutoff energy. The cutoff energy should be increased. If this is what is needed in BerkeleyGW, `parabands` may be one way to work around it. The k-grid being too dense can also be a reason.

8.4.18 Error in routine c_bands (1): too many bands are not converged

As the error message indicates, there are too many bands not converged. The reasons and solutions may be one or more of the follows:

- The diagonalization algorithm is not numerically stable; try to change `diagonalization` to `cg`.
- The parameter `nbnd` falling below the highest occupation can also lead to the error; this happens when dealing with multilayer materials and forgetting to set `nbnd` according to the number of layers currently in use.
- The convergence thresholds are too high.

8.4.19 Error in routine checkallsym (2): not orthogonal operation

This appears when the

8.4.20 some processors have no G-vectors for symmetrization

This happens frequently on Perlmutter. The reason is usually that too many MPI tasks are launched for a small system.

8.4.21 there are processes with no planes. Use pencil decomposition

This seems to be an indication of too many MPI tasks.

8.4.22 Error in routine angle_rot (1): problem with the matrix

In QuantumESPRESSO 7.0 and 7.1, this error occurs when we use `verbosity = 'high'`. When verbosity is high, `pw.x` checks whether the rotational matrices identified as symmetric operations are indeed rotational operations. It seems that when the crystal has hexagonal symmetries, the internal representation of the

rotational symmetries is in crystal coordinates, while criteria checking whether a matrix is a rotation operation works in Cartesian coordinates.

8.4.23 Error in routine sym_rho_init_shell (2): lone vector

This can be solved by slightly increasing `ecutrho` (for example we may want to set it to $4.1 \times$ `ecutwfc`) or by turning off the symmetry. The error occurs when numerical errors in symmetrization make some G vectors missing.

8.5 Trouble shooting in parabands

8.5.1 failed to find G-vector for k-point

A rather weird origin of this error is using momentum pool in the input QuantumESPRESSO calculation: when momentum pool is used, the error occasionally appears, and it doesn't appear immediately after `parabands` is started, but after several k points having been successfully finished.

8.6 Trouble shooting in epsilon and sigma

8.6.1 Floating point exception

This error sometimes occurs because the GPU version is run on CPU nodes.

8.6.2 ERROR: find_efermi: nband out of bounds

Don't be confused by the function name `find_efermi`: sometimes it's just because the total number of bands in `epsilon.inp` etc. exceeds the number of bands supplied in the `WFN` file. This can happen when the pseudopotential or the crystal structure is changed, and after pseudobands processing, the number of pseudobands varies.

8.6.3 WARNING: checkbz: unfolded BZ from epsilon.inp has missing q-points

In metallic GW this message is bond to appear, because the Γ point is not calculated in `1-epsilon`, but in `1-epsilon/epsilon_0`.

8.6.4 Selected number of bands breaks degenerate subspace.

Run `degeneracy_check.x WFN` to see degeneracy-allowed number of bands. This error occurs when one band in a degeneracy subspace is considered but others are not. Also, the `band_index_min` and `band_index_max` parameters shouldn't be too close to `vxc_diag_min` and `vxc_diag_max`, or the error occurs.

8.6.5 WFN ifmin/ifmax fields are inconsistent

The full message is

```
WFN ifmin/ifmax fields are inconsistent:
 - there is a valence state above the middle energy
 - there is a conduction state below the middle energy
Possible causes are:
(1) Your k-point sampling is too coarse and cannot resolve the
    ↪  Fermi energy.
 Try to carefully inspect your mean-field energies, and
    ↪ consider using a finer
k-grid.
(2) You are using eqp.dat and the QP corrections change the
    ↪ character of some s
tates
 from valence<->conduction. In this case, you should use
    ↪ another mean-field the
ory
 that gives the same ground state as your GW calculation.
(3) You are running inteqp, but you are either shifting the
    ↪ Fermi energy or usi
ng
 restricted transformation.
```

The direct causes of this error are already listed above. But it takes some time to see what is the deeper reason, and how to solve it:

- Sometimes when the `occupation` option in the `2.1-wfn` and `2.2-wfnq` steps is not correct. If `fixed` is used for a metal, for example, some positions that should be a part of a hole Fermi pocket are occupied by electrons, and therefore the highest occupied state has higher energy than the lowest unoccupied state. This usually means the smearing scheme needs to be changed.

- When you shift the occupation in a `WFN` file (§ 8.9.9), but forget to change the `ifmax` dataset correspondingly, this error will also occur, regardless of how you change the `eqp.dat`.

- When you shift the energy levels near the Fermi surface in a `WFN` file according to, say, self-consistent GW, or equivalently, replace the bands near the Fermi surface by their counterparts in a self-consistent GW calculation, by mindful of the overall energy zero point shift: if the Fermi energies in the self-consistently updated bands and the rest of the bands are not consistent, a higher conduction band may have lower energies in the update `WFN` or `eqp.dat`, and when it's even lower than the highest valence band, the error is reported.

8.6.6 Segmentation fault: address not mapped to object at address

The root of this error differs from case to case.

If we see

```
q-pt      2: Head of Epsilon        =   NaN      NaN
q-pt      2: Epsilon(2,2)           =   NaN      NaN
```

usually this means a "divided-by-zero" error occurs. This may occur when we incorrectly use the insulator procedure to calculate a metallic system (as long as DFT thinks the system is metallic, the error has the potential to occur, regardless of whether the system is metallic after GW correction), often regardless of the smearing type.

8.6.7 `eqpcor mean-field energy mismatch`

This error happens when we try to do an eigenvalue self-consistent calculation, and `epsilon` finds the DFT energies given in `eqp.dat` are different from the energies in `WFN`.

This sometimes is a technical problem (the Rydberg energy definitions used in QuantumESPRESSO and BerkeleyGW are slightly different), and can be solved by increasing `TOL_eqp` in the source code of BerkeleyGW. The error may also be reported when the DFT energies in `eqp.dat` are mistakenly changed (we should only change the column corresponding to the corrected energy).

A k-point appearing in `eqp.dat` twice can also cause the problem. In the `eqpcor` subroutine, once the difference between the DFT energy in `eqp.dat` is verified to be the same as the DFT energy `WFN` at given n and k, the GW corrected value in `eqp.dat` is assigned to the array element `eqp(ib,kk,is)` recording the DFT energy from `WFN` in the `eqpcor` routine at n, k. Now if the same k-point appears in `eqp.dat` again, when comparing the DFT energy in `eqp.dat` and the DFT energy in `WFN` – represented by `eqp(ib,kk,is)` – we are actually comparing the DFT energy and the GW energy, which are obviously not close to each other. And the error occurs.

8.6.8 `ERROR: occupations (ifmax field) inconsistent between WFN and WFNq files.`

```
ERROR: occupations (ifmax) inconsistent between WFN and WFNq
    ↪ files.
Remember that you should NOT use WFNq for metals and graphene.
```

8.6.9 `ERROR: Unexpected characters were found while reading the value for the keyword`

This usually happens when the input file contains a line like

```
number_bands = 148
```

while the correct format is

```
number_bands   148
```

8.6.10 `forrtl: severe (24): end-of-file during read, unit -5, file Internal List-Directed Read`

This means there is a syntax error in the input file. It usually happens when one of the follows happens:

- We should write `qpoints` but actually write `kpoints` (or the opposite).

- The `is_q0` flag is not set for a `qpoints` block or is accidentally set for a `kpoints` block.

- When both `qpoints` and `kpoints` are needed (in Hartree-Fock calculation), only the `qpoints` block is given.

8.6.11 ERROR: Inconsistent screening, trunction, or q0 vector

The full error message may be

```
ERROR: the input truncation flag indicates that the Coloumb
    ↪ interaction v(q0)
diverges for q0->0. However, you have q0 exactly zero.
You should always specify a *nonzero* q0->0 vector unless you
    ↪ have 0D
truncation, i.e., spherical or box truncation.


ERROR: Inconsistent screening, trunction, or q0 vector
```

This arises when we are dealing with a metal but forget to add `screening_metal` to `sigma.inp`.

```
ERROR: cannot use metallic screening with q0 = 0.
You should either specify a nonzero q0->0 vector or use
    ↪ another screening flag.
```

8.6.12 cannot use metallic screening with q=0

As the name implies, when doing a metallic calculation, we need to make sure that the $q \to 0$ point is the smallest *non-zero* point in the dense k-grid (§ 7.2.4). Note that if we accidentally set $q = 0$ in the `epsilon_0` step, *no* error will occur in this step (indeed, as long as the q-point given is within the k-grid of the WFN file, no error will be generated) – but then in the `sigma` step, the error occurs.

8.6.13 ERROR: genwf mpi: No match for rkq point

8.6.14 ERROR: Missing bands in file eqp_co.dat

This happens when doing `inteqp`. The origin of the error is just its name implies. Note that this error sometimes occurs for unnoticed reasons:

- When the pseudopotential is changed, the number of bands may change, because the orbitals included in the pseudopotentials change. Now if `band_-index_max` and `band_index_min` are set according to the old pseudopotentials, and the Fermi energy happens to fall outside of the range between the two, the error occurs.

8.6.15 `forrtl: severe (71): integer divide by zero`

If this happens after `Calculation parameters:`, it may arise from mistakenly setting `band_index_max` to a value smaller than `band_index_min`. In this case, the number of bands considered will be zero, and thus a divide-by-zero error occurs when `sigma.x` tries to distribute computational loads.

8.6.16 `ERROR: screened Coulomb cutoff is bigger than epsilon cutoff`

This happens in the `sigma` step. When the cutoff energies of `epsmat.h5` and `eps0mat.h5` are different (this may happen when you want to reuse existing `eps0mat.h5` for a metallic calculation), this may occur.

8.6.17 `ERROR: Incorrect kinetic energies in epsmat.`

This happens in the `sigma` step, usually when `epsmat.h5` or `eps0mat.h5` is not appropriately generated. Note that even when it's `esp0mat.h5` is broken, the error message still contains `in epsmat` instead of `in eps0mat`.

8.6.18 `ERROR: Bad Screening Options`

In `sigma`, we may see a warning message like this: `You have q0vec=0 but a metal!!`, and then the error in the caption occurs. This is due to

8.6.19 `_int_malloc: Assertion '(unsigned long) (size) >= (unsigned long) (nb)' failed.`

This occurs when running `sigma` in the Hartree-Fock mode. The reason seems to be inappropriate setting of `number_bands`: it shouldn't be greater than `band_index_max`. Somehow mysteriously, when such an inappropriate configuration does happen, a warning message like "the number of bands is changed to that is needed for calculating exchange" will be given, but the error still occurs.

8.6.20 `IndexError: list index out of range`

This happens when running `scGWtool.py`. When this happens in the `Reading sigma_hp files and constructing Hqp sub-space matrices` step, it's likely to be caused by missing lines in `sigma_hp_row.log` and/or `sigma_hp_col.log`. This could happen when we divide k points into several jobs, and when concatenating `sigma_hp_col`s generated, accidentally delete some lines containing data.

For example, `readlines` in Julia doesn't consider a final, empty line in a file to be a line (and that's self-consistent because saying that there is a final, empty line at the end of a file is equivalent to say that the final ends with `n`, which however should be regarded as a part of the second last line). But in ordinary text editors the final, empty line is always displayed. Thus it's very easy to make mistakes when measuring the size of the footer part of `sigma_hp.log`, and sometimes the last line of the last k-block in the file is deleted, resulting in the error.

8.7 Trouble shooting in `wannier90` and `pw2wannier90`

8.7.1 `MPIDI_CRAY_init: GPU_SUPPORT_ENABLED is requested, but GTL library is not linked`

This, as the error message implies, happens because a GPU program is run on a pure CPU environment. On Perlmutter, the default Wannier90 module is supposed to be run on a pure CPU environment, but when the `gpu` module is loaded, it may want to use GPU acceleration and fail. Loading the `cpu` module solves the problem.

8.7.2 `w90_wannier90_readwrite_read: mismatch in WTe2.eig`

When reading a `.eig` file, `wannier90` doesn't really look at the band index and k-point index. Rather, it completely relies on `num_bands` and `mp_grid`. Thus, if the number of `nbnd` in the QuantumESPRESSO run – which decides the number of bands in the `.eig` file – isn't `num_bands`, `wannier90` finds bands in the `.eig` file messed up, and the error occurs.

8.7.3 `WTe2.amn has not the right number of bands`

Sometimes the `.amn` file only contains `num_wann` bands, while the `.eig` file contains `num_bands` bands, and when the two values are different, the error occurs. Redoing the whole procedure with a fixed `prefix.win` file can solve this problem. TODO: what's really going on here

8.7.4 `forrtl: severe (174): SIGSEGV, segmentation fault occurred`

TODO: occurs when `num_bands` > `num_wann`

8.7.5 `too many projections to be used without selecting a subset`

When there are more projection wave functions specified in the `projections` block (they can be found in the `prefix.nnkp` file) than the desired number of Wannier functions, this error occurs. It means you need to tighten the constraint in the `projections` block, or add a `select_projections` statement in the `.win` file (that's so-called "selecting a subseth").

8.7.6 `Direct lattice mismatch`

As the error message says, it's due to mismatch between the crystal constants in DFT and in `prefix.win`. Note that it's possible that the crystal structure in `bands.in` is correct, but the crystal structure in `scf.in` is wrong; in that case, the latter overrides the former, and the error message becomes hard to understand.

8.7.7 `Unable to satisfy B1 with any of the first 36 shells`

This error comes together with the following text: `Unable to satisfy B1 with any of the first 36 shells Your cell might be very long, or you may have an irregular MP grid Try increasing the parameter search_shells in the win file (default=12)`. I'm not sure why it appears, but increasing the parameter `search_shells` does work.

8.8 Trouble shooting in `kernel` and `absorption`

8.8.1 `ERROR: Inconsistent symmetry treatment of the fine and shifted grids with the momentum operator`

8.8.2 `Fatal error in PMPI_Waitall:  Request pending due to failure`

This MPI issue is sometimes an indication of too much IO: for example, if it appears after `Polarization:  b3 Sum rule (BSE)`, maybe the eigenvector file of all excitons is simply too large for the supercomputer.

8.8.3 `could not get a validated dataspace from file_space_- id`

The full message looks like this:
```
Started interpolating BSE kernel with 32 block(s) at 13:16:41.
HDF5-DIAG: Error detected in HDF5 (1.10.4) MPI-process 0:
  #000: H5Dio.c line 185 in H5Dread(): could not get a
    ↪ validated dataspace from file_space_id
  major: Invalid arguments to routine
  minor: Bad value
  #001: H5S.c line 254 in H5S_get_validated_dataspace():
    ↪ selection + offset not within extent
  major: Dataspace
  minor: Out of range
```

This happens whenever certain states referred to in an `absorption` calculation are absent in the `bsemat.h5` file supplied. A common source of the problem is that the number of bands given in `bsemat.h5` is smaller than the number of bands required in `absorption`. Another possible cause of the error is that the `extended_kernel` flag is turned on when the `bsemat.h5` file is not generated with this flag.

8.8.4 `ERROR: Momentum and Finite_q are incompatible`

This happens when we are doing a finite Q calculation but is using the `use_- momentum` option. BerkeleyGW expects `use_velocity` whenever finite Q is specified.

8.8.5 `epscopy: read illegal ng from epsmat`

This happens when the DFT G-grid used in calculating `epsmat.h5` is not consistent with the G-grid provided in `WFN_co`. TODO: is it possible that different `ecutrho` also lead to the error?

8.8.6 `WARNING:Degeneracies at Fermi level`

The literal cause of this warning is quite simple: there is at least one energy that is quite close to the Fermi energy. The reason might be that the wave function file is broken.

8.9 Checklist for unexpected results

Sometimes the calculation ends successfully, but the result seems strange. Below we discuss some frequently encountered cases, together with possible causes and solutions.

8.9.1 Band symmetry higher than the space group shown at the beginning of `bands.out`

- Usually this is because of an approximate symmetry, which is ignored by QuantumESPRESSO because its tolerance is very low.
- Are all steps in the DFT calculation using the same crystal structure?

8.9.2 Band structure looks very far from the literature

- Check the crystal structure: if it comes from relaxation, does it converges?
- For 2D materials, when we change the vacuum distance and use crystal coordinates for atomic positions at the same time, always double check whether we scale the atomic positions correctly. The formula is

$$\text{new } z \text{ coordinate} = \frac{\text{old vacuum distance}}{\text{new vacuum distance}} \times \text{old } z \text{ coordinate}. \quad (8.1)$$

- Are all steps in the DFT calculation using the same crystal structure?
- Is the Fermi energy correct? Sometimes we change the band structure but forget to change the Fermi energy used to plot bands.

8.9.3 Band plot is empty

- Are there enough bands? If `nbnd` is not set for an insulator, no conduction band will be considered.
- Is the Fermi energy correct? If the Fermi energy is set too high (which may come from, say, wrong unit), then naturally there is no band in the plot.
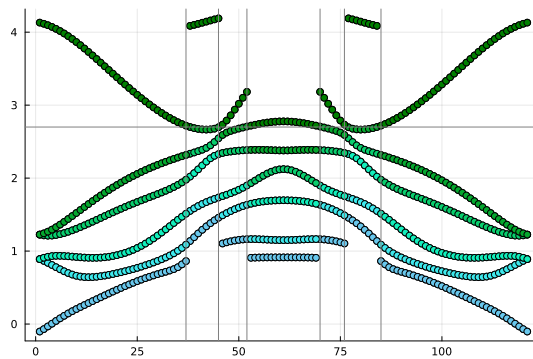
Figure 8.1: Example of how `inteqp` misidentifies DFT-level bands: the data is from the DFT column of `eqp.dat`; the colors of points are about the band index.

8.9.4 Band plot is not continuous

If the system is metallic in the DFT level (regardless of what GW says about the material, and the screening model used) and `inteqp` is used, this is expected: somehow, `inteqp` assumes the system is an insulator, and thus states above the Fermi energy (but not too far from it) has to be in one band. Thus, if, say, the 120th band has more than one intersection points with the Fermi energy level, its part below φ_F will be recognized as the 119th band in the eyes of `inteqp`. This can be seen by plotting the DFT column of the `eqp.dat` output of `inteqp` (Fig. 8.1).

Another type of band plot non-continuity is described in [8] and can be solved by COHSEX??? TODO

A further type of non-continuity comes from mistakenly linking a wrong `epsmat.h5` or `eps0mat.h5` file. This sometimes displaces some segments of bands below a certain band index downwards or upwards.

8.9.5 Band plot is flat

This may happen when the k-path is obtained from a software designed for 3D systems, while we are actually dealing with a monolayer.

8.9.6 Too many k points in the result of `inteqp`

Check the symmetry setting. No symmetry should be applied to the k-path in the fine-grained wave function file input to `inteqp`.

8.9.7 The size of band gap

DFT is infamous for underestimating the band gap.

- Use hybrid functionals like HSE.
- Let GW correct the band structure. TODO: but how? How to avoid the error in § 8.6.5?

8.9.8 SOC effects are too strong

When SOC effects are much stronger than expected:

- Are you using relativistic pseudopotentials for a non-SOC run? This is *not* correct (and unfortunately QuantumESPRESSO never tells us when this happens).
- TODO:

8.9.9 When we get a semimetal in the DFT step but it should be an insulator

This is similar to § 8.9.7.

Note that naively feeding the semimetal result into BerkeleyGW while still using the insulator procedure may result in errors in § 8.6.6.

One way to solve the problem is to manually move the conduction bands and the valence bands away from each other. Naively using the `eqp_correction` option and shifting bands near the Fermi surface away from each other leads to § 8.6.5, precisely because "QP corrections" (i.e. the energy shift manually added by me) change the character of some states from valence to conduction. The `mf_header/kpoints/occ` and `mf_header/kpoints/ifmax` datasets in the `WFN.h5` file have to be modified accordingly. A procedure to do so can be found here. TODO: should this be done in `2-sigma`?

The problem with this method is we don't have `eqp_correction` in `kernel`. TODO

8.9.10 The band plot seems reasonable but the band gap is strange

When doing convergence tests, you may find changing a parameter somehow leads to a rather large change in the band gap. This may be because you accidentally use `WFN` with `number_bands` set for `WFNo` in one of the instances. This essentially leads to an under-converged problem: suppose we have 200 bands in `WFNo`, but there are 1000 bands in `WFN`, and if we set `number_bands` to 200 while using `WFN`, then of course the calculation is severely under-converged.

8.9.11 The DOS curve is too smooth

This usually can be solved by setting `occupations` to `tetrahedra` in the `bands` step, and using a smaller `degauss` in the `projwfc` step. Note that when doing so, the `smearing` and `degauss` options should not be set, or otherwise by default the calculation will be done with gaussian smearing.

8.9.12 The Wannier-interpolated band structure looks weird

- A denser k-grid always improves the accuracy. `20 20 1` is a safe choice; a smaller k-grid may create wiggles in the band plot.

8.9.13 The exciton wave function looks strange

$A_{cv\boldsymbol{k}}^{SQ}$ can be retrieved from the `eigenvalues.h5` file generated by `absorption`. Sometimes the $\sum_{c,v} |A_{cv\boldsymbol{k}}^{SQ}|^2$ heatmap looks rather weird: we may get asymmetric signals, or maybe all of the weight falls on several points.

Asymmetric results sometimes are normal: we may have a degeneracy space, and the basis chosen by `absorption` happens to lose the symmetry we expect. Note that due to numerical errors, it's possible that the energies assigned to the eigenstates have small differences from each other. We need to compare how the energy difference is compared with the energy difference between clearly non-degenerate states to decide whether the states are in one degeneracy space.

Sometimes weird exciton weight heatmap comes from real technical issues. To name a few:

- The `bsemat.h5` file used is broken.
- Smearing is used in the DFT step and some conduction states are mistakenly considered to be filled by `pw2bgw.x`.

8.9.14 Excitons that should be dark become bright

When the `kernel` step is done with `use_symmetries_coarse_grid`, but the `WFN_co` file contains the full 1BZ, and `no_symmetries_coarse_grid` is the option used in the `absorption` step, typically we find that states that are supposed to be dark become bright, although the exciton wave functions are correct.

A note on what it means by the exciton wave function being correct. Because of the random phase factor in DFT diagonalization, the phase of $A_{cv\boldsymbol{k}}^{SQ}$ often seem random. If we're sure that the lowest exciton mode has a 1s character, the random phase factors can be extracted as $A_{cv\boldsymbol{k}}^{SQ}/|A_{cv\boldsymbol{k}}^{SQ}|$, and hence both the dipole matrix elements and the higher exciton modes can be corrected by this phase factor. After this step typically we find that the exciton wave functions look like what we expect in hydrogen-like systems, and the dipole matrix element is smooth with respect to $\mathbf{k}$.

When the `kernel` step is done with `use_symmetries_coarse_grid`, but the `WFN_co` file contains the full 1BZ, the exciton wave functions are still correct, but the dipole matrix elements are not smooth with respect to $\mathbf{k}$. Instead we find that the dipole matrix element, when c and v are fixed, look fragmented: the heatmap consists of a serious of fragments, in each of which $d_{cv\mathbf{k}}$ is smooth, but then we see non-smooth jumps between different fragments (Fig. 8.2).

The reason is, when `use_symmetries_coarse_grid` is turned on, the phase convention of the kernel matrix (on the coarse $\mathbf{k}$ grid) is also symmetric, which is inconsistent with the phase covention of the actual `WFN_fi` file. Therefore when `no_symmetries_coarse_grid` is used in the `absorption` step, phase factors from the $\mathbf{k}_{\mathrm{co}}$ points in the `bsemat.h5` matrix and the $\mathbf{k}_{\mathrm{co}}$ points in `WFN_co` used to interpolate the BSE kernel do not cancel each other. The result is the phase difference between a $\mathbf{k}_{\mathrm{co}}$ point in the `WFN_co` file and its symmetrically equivalent point that overwrites the former in the `kernel` step is passed to the interpolated BSE kernel on the fine grid. This is why we have fragments in Fig. 8.2: each fragment corresponds to a group of fine $\mathbf{k}$ points that are closest to a single $\mathbf{k}_{\mathrm{co}}$
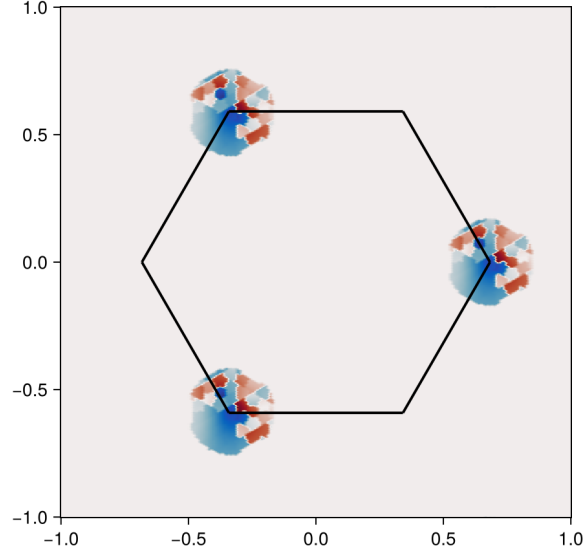
Figure 8.2: An instance of fragmented dipole matrix element

point. This does not make the BSE calculation itself wrong, because besides the interpolated BSE kernel, there is no other term that has phase factors in the eigenvalue problem, but it does make the phase convention in the BSE calculation different from that in `WFN_fi`.

8.9.15 The `absorption_b{1,2,3}_eh.dat` files are empty

This seems to happen in the following scenario: `extended_kernel` is used,[1] and in `kernel` the flag `screening_metal` is wrongly turned on, while in `absorption` the flag `screening_semiconductor` is turned on. Not sure what's going on here.

---

[1] In the two steps of BSE, this option should be consistently used; but following this rule alone doesn't help to resolve the problem.

Bibliography

[1] B Adolph, VI Gavrilenko, K Tenelsen, F Bechstedt, and R Del Sole. Nonlocality and many-body effects in the optical properties of semiconductors. *Physical Review B*, 53(15):9797, 1996.

[2] Irene Aguilera, Christoph Friedrich, Gustav Bihlmayer, and Stefan Blügel. G w study of topological insulators bi 2 se 3, bi 2 te 3, and sb 2 te 3: Beyond the perturbative one-shot approach. *Physical Review B*, 88(4):045206, 2013.

[3] Xavier Andrade and Alán Aspuru-Guzik. Prediction of the derivative discontinuity in density functional theory from an electrostatic description of the exchange and correlation potential. *Physical review letters*, 107(18):183002, 2011.

[4] F Aryasetiawan, T Miyake, and R Sakuma. 7 the constrained rpa method for calculating the hubbard u from first-principles. *The LDA+ DMFT approach to strongly correlated materials*, 2011.

[5] Claudio Attaccalite, M Grüning, and A Marini. Real-time approach to the optical properties of solids and nanostructures: Time-dependent bethe-salpeter equation. *Physical Review B*, 84(24):245110, 2011.

[6] Bradford A Barker, Jack Deslippe, Johannes Lischner, Manish Jain, Oleg V Yazyev, David A Strubbe, and Steven G Louie. Spinor g w/bethe-salpeter calculations in berkeleygw: Implementation, symmetries, benchmarking, and performance. *Physical Review B*, 106(11):115127, 2022.

[7] Friedhelm Bechstedt. *Many-body approach to electronic excitations*. Springer, 2016.

[8] J Arjan Berger, Pierre-François Loos, and Pina Romaniello. Potential energy surfaces without unphysical discontinuities: The coulomb hole plus screened exchange approach. *Journal of Chemical Theory and Computation*, 17(1):191–200, 2020.

[9] Christian Brouder, Gianluca Panati, Matteo Calandra, Christophe Mourougane, and Nicola Marzari. Exponential localization of wannier functions in insulators. *Physical review letters*, 98(4):046402, 2007.

[10] Y-H Chan, Diana Y Qiu, Felipe H da Jornada, and Steven G Louie. Giant self-driven exciton-floquet signatures in time-resolved photoemission spectroscopy of mos _2 from time-dependent gw approach. *arXiv preprint arXiv:2302.01719*, 2023.

[11] Mauro Del Ben, H Felipe, Gabriel Antonius, Tonatiuh Rangel, Steven G Louie, Jack Deslippe, and Andrew Canning. Static subspace approximation for the evaluation of g 0 w 0 quasiparticle energies within a sum-over-bands approach. *Physical Review B*, 99(12):125128, 2019.

[12] Mauro Del Ben, H Felipe, Andrew Canning, Nathan Wichmann, Karthik Raman, Ruchira Sasanka, Chao Yang, Steven G Louie, and Jack Deslippe. Large-scale gw calculations on pre-exascale hpc systems. *Computer Physics Communications*, 235:187–195, 2019.

[13] R Del Sole and Raffaello Girlanda. Optical properties of semiconductors within the independent-quasiparticle approximation. *Physical Review B*, 48(16):11789, 1993.

[14] Jack Deslippe, Georgy Samsonidze, David A. Strubbe, Manish Jain, Marvin L. Cohen, and Steven G. Louie. Berkeleygw: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Computer Physics Communications*, 183(6):1269–1289, 2012.

[15] Sergey V Faleev, Mark Van Schilfgaarde, and Takao Kotani. All-electron self-consistent g w approximation: Application to si, mno, and nio. *Physical review letters*, 93(12):126406, 2004.

[16] Weiwei Gao, H Felipe, Mauro Del Ben, Jack Deslippe, Steven G Louie, and James R Chelikowsky. Quasiparticle energies and optical excitations of 3c-sic divacancy from g w and g w plus bethe-salpeter equation calculations. *Physical Review Materials*, 6(3):036201, 2022.

[17] Weiwei Gao, Weiyi Xia, Xiang Gao, and Peihong Zhang. Speeding up GW Calculations to Meet the Challenge of Large Scale Quasiparticle Predictions. *Scientific Reports*, 6(1):36849, 2016.

[18] Lars Hedin. New method for calculating the one-particle green's function with application to the electron-gas problem. *Physical Review*, 139(3A):A796, 1965.

[19] Linda Hung, H Felipe, Jaime Souto-Casares, James R Chelikowsky, Steven G Louie, and Serdar Öğüt. Excitation spectra of aromatic molecules within a real-space g w-bse formalism: Role of self-consistency and vertex corrections. *Physical Review B*, 94(8):085125, 2016.

[20] Mark S Hybertsen and Steven G Louie. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Physical Review B*, 34(8):5390, 1986.

[21] Yi Lin, Yang-hao Chan, Woojoo Lee, Li-Syuan Lu, Zhenglu Li, Wen-Hao Chang, Chih-Kang Shih, Robert A Kaindl, Steven G Louie, and Alessandra Lanzara. Exciton-driven renormalization of quasiparticle band structure in monolayer mos 2. *Physical Review B*, 106(8):L081117, 2022.

[22] Arthur Mar, Stephane Jobic, and James A Ibers. Metal-metal vs tellurium-tellurium bonding in wte2 and its ternary variants tairte4 and nbirte4. *Journal of the American Chemical Society*, 114(23):8963–8971, 1992.

[23] Nicola Marzari, Arash A Mostofi, Jonathan R Yates, Ivo Souza, and David Vanderbilt. Maximally localized wannier functions: Theory and applications. *Reviews of Modern Physics*, 84(4):1419, 2012.

[24] John P Perdew, Robert G Parr, Mel Levy, and Jose L Balduz Jr. Density-functional theory for fractional particle number: derivative discontinuities of the energy. *Physical Review Letters*, 49(23):1691, 1982.

[25] E Perfetto, D Sangalli, A Marini, and G Stefanucci. First-principles approach to excitons in time-resolved and angle-resolved photoemission spectra. *Physical Review B*, 94(24):245303, 2016.

[26] Michael Rohlfing and Steven G Louie. Electron-hole excitations and optical spectra from first principles. *Physical Review B*, 62(8):4927, 2000.

[27] Avinash Rustagi and Alexander F Kemper. Photoemission signature of excitons. *Physical Review B*, 97(23):235310, 2018.

[28] Bi-Ching Shih, Yu Xue, Peihong Zhang, Marvin L Cohen, and Steven G Louie. Quasiparticle band gap of zno: High accuracy from the conventional g 0 w 0 approach. *Physical review letters*, 105(14):146401, 2010.

[29] M. Shishkin, M. Marsman, and G. Kresse. Accurate quasiparticle spectra from self-consistent gw calculations with vertex corrections. *Phys. Rev. Lett.*, 99:246403, Dec 2007.

[30] Catalin-Dan Spataru. *Electron excitations in solids and novel materials*. PhD thesis, University of California, Berkeley, 2004.

[31] G Strinati. Effects of dynamical screening on resonances at inner-shell thresholds in semiconductors. *Physical Review B*, 29(10):5718, 1984.

[32] Nathan Wiser. Dielectric constant with local field effects included. *Phys. Rev.*, 129:62–69, Jan 1963.

[33] Xiao Zhang, Joshua A Leveillee, and André Schleife. Effect of dynamical screening in the bethe-salpeter framework: Excitons in crystalline naphthalene. *arXiv preprint arXiv:2302.07948*, 2023.

[34] Xiao Zheng, Aron J Cohen, Paula Mori-Sánchez, Xiangqian Hu, and Weitao Yang. Improving band gap prediction in density functional theory from molecules to solids. *Physical review letters*, 107(2):026403, 2011.