

Manual for DFT+ GW +BSE and the related matters

Jinyuan Wu

February 24, 2023

Contents

1 Preliminaries	4
1.1 Second quantization	4
1.1.1 For electrons	4
1.2 Details in diagrammatics	5
1.2.1 Infinitesimals	5
1.2.2 The expression of Σ and W	6
1.2.3 About “antiparticles”	7
1.2.4 The direction of momentum on the interaction line	8
1.3 “Quantitative” Feynman rules	10
1.3.1 In the free space	10
1.3.2 In a crystal	11
1.3.3 Diagrammatic components as tensors	14
1.4 Quasiparticles	14
1.4.1 Spectral density	15
2 Experimental characterization methods	16
2.1 Light-matter interaction	16
2.1.1 Interaction Hamiltonian	16
2.1.2 Dipole approximation	16
2.1.3 How to capture absorption	17
3 <i>GW</i> and BSE	20
3.1 What is <i>GW</i>	20
3.1.1 <i>GW</i> is screened Hartree-Fock approximation	20
3.1.2 Infinitesimal displacement of time in <i>GW</i>	20
3.1.3 <i>GW</i> compared with the Hartree-Fock approximation	21
3.1.4 BSE with the same order of approximation	21
3.2 The dielectric matrix ϵ	22
3.2.1 Frequency-dependent form	22
3.2.2 The static limit and the generalized plasmon-pole model	23
3.2.3 The static subspace approach	24
3.3 The self-energy matrix Σ	24
3.3.1 COHSEX approximation	24
3.3.2 Diagonal or not	25
3.3.3 Self-consistent or not	25
3.4 From <i>GW</i> to BSE	26
3.5 Accuracy of <i>GW</i>	28
3.6 On so-called failure of <i>GW</i> and convergence issues	28
4 The QuantumESPRESSO-BerkeleyGW ecosystem	29
4.1 Overview of the pipeline	29
4.2 Relativistic effects	29
4.3 Input and output of <code>pw</code>	29
4.4 The <code>epsilon</code> step	29
4.4.1 Procedure and speed	29
4.4.2 Divergence problems when $\mathbf{q} \rightarrow 0$	30
4.4.3 Frequency dependence of ϵ	30

4.4.4	Console output	30
4.4.5	Other output files	31
4.5	Systems of units	31
5	Tight-binding models	32
5.1	Wannier functions and tight-binding models	32
5.2	The cRPA approach to obtain effective models	32
6	Standard operation procedures	34
6.1	Details in installation	34
6.1.1	QuantumESPRESSO	34
6.2	Standard operation procedures	34
6.2.1	Avoid data pollution	34
6.2.2	Finding the structure	34
6.2.2.1	From open data	34
6.2.2.2	Comparing existing structures for the same material	35
6.2.3	Insulator DFT+GW+BSE	35
6.2.3.1	The DFT stage	35
6.2.3.2	The GW stage	36
6.2.3.3	The BSE stage	37
6.2.4	Metal DFT+GW+BSE	37
6.2.5	Band plot	38
6.2.5.1	DFT level: k -path	38
6.2.5.2	GW level: inteqp	38
6.2.5.3	GW level: using WFN_outer	39
6.2.5.4	BSE level	39
6.2.6	Wannier functions and tight-binding models	39
6.2.6.1	wannier90 for DFT	39
6.2.6.2	GW level: sig2wan	39
6.2.7	Self-consistent GW	39
6.2.7.1	Energy self-consistent calculation in GW	40
6.2.7.2	Eigenstate self-consistent calculation in GW	40
6.2.8	Topological invariants with z2pack	40
6.2.9	Band projection	40
6.3	Performance tricks	40
6.3.1	Parallelization	40
6.3.2	Choosing cutoff energies wisely	40
6.3.3	pseudobands	41
6.4	Convergence tests	41
7	Trouble shooting	42
7.1	Unexpected units	42
7.1.1	Band energy output of pw.x	42
7.2	Trouble shooting in MPI	42
7.2.1	srun: fatal: Can not execute	42
7.2.2	error parsing parameters	42
7.2.3	Each process is run serially and doesn't communicate with others	42
7.2.4	nsufficient virtual memor	42
7.3	Trouble shooting in Python	42
7.3.1	AttributeError: 'Dataset' object has no attribute 'value'	42
7.4	Trouble shooting in QuantumEspresso	43
7.4.1	Intel MKL FATAL ERROR: Cannot load symbol MKLMPI_Get_wrappers.	43
7.4.2	Program frozen	43
7.4.3	Error in routine allocate_fft (1): wrong ngms	43
7.4.4	Error reading attribute index : expected integer , found *	43
7.4.5	cdiaghg (159): eigenvectors failed to converge	43
7.4.6	Error in routine cdiaghg (1052): problems computing cholesky	43
7.4.7	Error in routine set_occupations (1): smearing requires a vnlue for gaussian broadening (degauss)	43

7.4.8	Error in routine splitwf (36197): wrong size for pwt	43
7.4.9	Error in routine PW2BGW(19):input pw2bow	43
7.4.10	Error in routine PW2BGW (19): input_pw2bgw	43
7.4.11	stress for hybrid functionals not available with pools	43
7.4.12	Error in routine projwave (1): Cannot project on zero atomic wavefunctions!	44
7.4.13	Error in routine diropn (3): wrong record length	44
7.5	Trouble shooting in epsilon and sigma	44
7.5.1	WARNING: checkbz: unfolded BZ from epsilon.inp has missing q-points	44
7.5.2	Selected number of bands breaks degenerate subspace.	44
7.5.3	WFN ifmin/ifmax fields are inconsistent	44
7.5.4	Segmentation fault: address not mapped to object at address	45
7.5.5	eqpcor mean-field energy mismatch	45
7.5.6	ERROR: occupations (ifmax field) inconsistent between WFN and WFNq files.	45
7.5.7	ERROR: Unexpected characters were found while reading the value for the keyword	45
7.5.8	forrtl: severe (24): end-of-file during read, unit -5, file Internal List-Directed Read	45
7.5.9	ERROR: Inconsistent screening, truncation, or q0 vector	45
7.5.10	cannot use metallic screening with q=0	46
7.5.11	ERROR: genwf mpi: No match for rkq point	46
7.5.12	ERROR: Missing bands in file eqp_co.dat	46
7.5.13	forrtl: severe (71): integer divide by zero	46
7.5.14	ERROR: screened Coulomb cutoff is bigger than epsilon cutoff	46
7.5.15	ERROR: Incorrect kinetic energies in epsmat.	46
7.6	Trouble shooting in wannier90 and pw2wannier90	46
7.6.1	w90_wannier90_readwrite_read: mismatch in WTe2.eig	46
7.6.2	WTe2.amn has not the right number of bands	46
7.6.3	forrtl: severe (174): SIGSEGV, segmentation fault occurred	46
7.7	Trouble shooting in kernel and absorption	47
7.7.1	ERROR: Inconsistent symmetry treatment of the fine and shifted grids with the momentum operator	47
7.8	Checklist for unexpected results	47
7.8.1	Band symmetry higher than the space group shown at the beginning of bands.out	47
7.8.2	Band structure looks very far from the literature	47
7.8.3	Band plot is empty	47
7.8.4	Band plot is not continuous	48
7.8.5	The size of band gap	48
7.8.6	SOC effects are too strong	48
7.8.7	When we get a semimetal in the DFT step but it should be an insulator	48
7.8.8	The band plot seems reasonable but the band gap is strange	48

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

$$\{c_i, c_j^\dagger\} = \delta_{ij}, \quad (1.1)$$

and Fock states are given by

$$|n_1, n_2, \dots\rangle = (c_1^\dagger)^{n_1} (c_2^\dagger)^{n_2} \dots |0\rangle. \quad (1.2)$$

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

$$H_0 = \sum_{i=1}^N h_i \quad (1.3)$$

is to be rewritten as

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

and a two-particle Hamiltonian

$$H_I = \frac{1}{2} \sum_{i \neq j=1}^N V_{ij} = \sum_{\text{pair } i, j} V_{ij} \quad (1.5)$$

is to be rewritten as

$$H_I = \frac{1}{2} \sum_{\alpha, \beta, \delta, \gamma} \langle \alpha \beta | V | \delta \gamma \rangle c_\alpha^\dagger c_\beta^\dagger c_\delta c_\gamma. \quad (1.6)$$

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

$$c_\alpha^\dagger = \sum_{\tilde{\alpha}} \langle \tilde{\alpha} | \alpha \rangle c_{\tilde{\alpha}}^\dagger \quad (1.7)$$

and

$$c_\alpha = \sum_{\tilde{\alpha}} \langle \alpha | \tilde{\alpha} \rangle c_{\tilde{\alpha}}. \quad (1.8)$$

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

$$\psi^\dagger(\mathbf{x}) = \sum_{\alpha} \langle \mathbf{x} | \alpha \rangle c_{\alpha}^{\dagger} \quad (1.9)$$

and $\psi(\mathbf{x})$ 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

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

Under this

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

$$\begin{aligned} \text{---} \xrightarrow{k} \text{---} &:= \mathcal{T} \langle c_{\mathbf{k}}(t - 0^+) c_{\mathbf{k}}^\dagger(0) \rangle \\ &= \int \frac{d\omega}{2\pi} e^{-i\omega(t-0^+)} \underbrace{\frac{i}{\omega - \xi_{\mathbf{k}} + i0^+ \text{sgn}(\xi_{\mathbf{k}})}}_{iG_0(\omega, \mathbf{k})} = \int \frac{d\omega}{2\pi} e^{-i\omega t} e^{i\omega 0^+} iG_0(\omega, \mathbf{k}). \end{aligned} \quad (1.11)$$

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 $e^{i\omega 0^+}$ factor; some may *define* that iG is the correct propagator, and the $e^{i\omega 0^+}$ factor or the small time displacement is embedded into the definition of G .

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

$$\int \frac{d\omega}{2\pi} \frac{i}{\omega - \xi_{\mathbf{k}} + i0^+ \text{sgn}(\xi_{\mathbf{k}})}.$$

This integral is not zero, but we want it to be zero when $\xi_{\mathbf{k}} > 0$, so we have to add a $e^{i\omega 0^+}$ 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

$$\int_{|\omega|=R \gg 1} e^{i\omega 0^+} \frac{d\omega}{2\pi} \frac{i}{\omega - \xi_{\mathbf{k}} + i0^+ \text{sgn}(\xi_{\mathbf{k}})} = 0.$$

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. $\omega \rightarrow 0$) limit of the photon propagator, which has $\omega^2 - \mathbf{q}^2 + i0^+$ as the denominator, and we get

$$V(q) = \frac{4\pi e^2}{\mathbf{q}^2 - \omega^2 - i0^+}. \quad (1.12)$$

Also, for screened Coulomb interaction, the correct retardation is important, because now something looking like (1.12) appears again.

The 1^+ in $\Sigma(1,2) = iW(1^+,2)G(1,2)$ seems to be from the infinitesimal in the electronic Green function instead of the infinitesimal in the Coulomb interaction line (Section 3.1.2).

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 \mathbf{r} or \mathbf{k} are not considered.

The self-energy correction is visualized as the follows:

$$\text{double line with arrow} = \text{single line with arrow} + \text{single line with arrow} \times \text{grey circle} , \quad (1.13)$$

and from it we have

$$iG = iG_0 + iG_0 iG \times \text{grey circle} .$$

It's then a good idea to define

$$-i\Sigma = \text{grey circle} , \quad (1.14)$$

because in this case, we have

$$G = G_0 + GG_0\Sigma, \quad (1.15)$$

and therefore

$$\underbrace{\omega - \xi_{\mathbf{k}}^0}_{1/G_0} = \underbrace{\omega - \xi_{\mathbf{k}}}_{1/G} + \Sigma, \quad (1.16)$$

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 (1.11)) out. After taking these factors into account, we find (1.15) becomes

$$e^{i\omega 0^+} iG = e^{i\omega 0^+} iG_0 + e^{i\omega 0^+} iG_0 \times \text{grey circle} \times e^{i\omega 0^+} iG. \quad (1.17)$$

We can eliminate the common $e^{i\omega 0^+}$ 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 (1.14) 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 (1.14), if any, can also be removed.

Similarly, we define the corrected interaction line as

$$-iW = \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \end{array}, \quad (1.18)$$

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

$$W = \frac{e^2}{r} =: v. \quad (1.19)$$

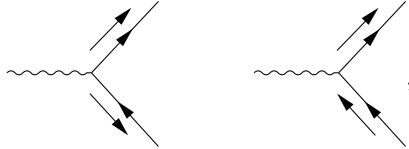
1.2.3 About “antiparticles”

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 e^{-iHt} 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 e^{-iHt} , and this is visualized as an “antiparticle” external line with an outward momentum line. So here, the “particle-antiparticle transformation” is just swapping $c_{\mathbf{k}}$ and $c_{\mathbf{k}}^\dagger$ – 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

$$\int_{-\infty}^{\infty} e^{i\omega t} dt \mathcal{T} \langle c_{\mathbf{k}}(t) c_{\mathbf{k}}^\dagger(0) \rangle = \frac{i}{\omega - \epsilon_{\mathbf{k}} + \mu}, \quad (1.20)$$

which can be straightforwardly obtained by looking at

$$H = \sum_{\mathbf{k}} \epsilon_{\mathbf{k}} c_{\mathbf{k}}^\dagger c_{\mathbf{k}} - \mu N \quad (1.21)$$

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 \mathbf{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 \mathbf{k} , \mathbf{k} has to be under the Fermi surface in the ground state, so $\omega_{\mathbf{k}} = \epsilon_{\mathbf{k}} - \mu < 0$, and the point $\omega = \omega_{\mathbf{k}}$ 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 (\mathbf{k} over Fermi surface), antiparticles never appear; when antiparticles appear (\mathbf{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

$$b_{\mathbf{k}} = \begin{cases} c_{\mathbf{k}}, & \epsilon_{\mathbf{k}} > \mu, \\ c_{\mathbf{k}}^{\dagger}, & \epsilon_{\mathbf{k}} < \mu, \end{cases} \quad (1.22)$$

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

$$\int_{-\infty}^{\infty} e^{i\omega t} dt \mathcal{T} \langle b_{\mathbf{k}}(t) b_{\mathbf{k}}^{\dagger}(0) \rangle = \frac{i}{\omega - \mu + \epsilon_{\mathbf{k}}}, \quad (1.23)$$

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

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

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

$$\begin{aligned} E &= \sum_{\mathbf{k}' \neq \mathbf{k}, \epsilon_{\mathbf{k}'} < \mu} \epsilon_{\mathbf{k}'} - \mu(N-1) - \left(\sum_{\epsilon_{\mathbf{k}'} < \mu} \epsilon_{\mathbf{k}'} - \mu N \right) \\ &= \mu - \epsilon_{\mathbf{k}} > 0. \end{aligned} \quad (1.24)$$

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 \mathbf{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 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

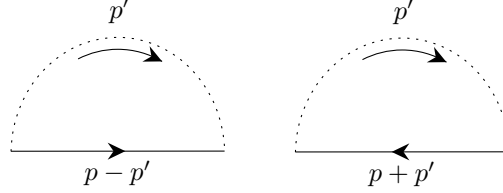
$$W(\omega) = \int dt e^{i\omega t} W(t),$$

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

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

and

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

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 $W(\mathbf{p}, \omega)$ with $W(-\mathbf{p}, -\omega)$, the physical rules of the system *always* remains the same, as long as we have translational symmetries for time and space (if \mathbf{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 $W(\mathbf{p}, \omega)$ only makes sense in

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

By manipulating the variables, we easily find

$$\begin{aligned} S_{\text{int}} &= \sum_{p_1, p_2, q} c_{p_1-q}^\dagger c_{p_2+q}^\dagger W(-q) c_{p_2} c_{p_1} \\ &= \sum_{p_1, p_2, q} c_{p_2+q}^\dagger c_{p_1-q}^\dagger W(-q) c_{p_1} c_{p_2} \\ &= \sum_{p_1, p_2, q} c_{p_1+q}^\dagger c_{p_2-q}^\dagger W(-q) c_{p_2} c_{p_1}, \end{aligned}$$

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 $W(\mathbf{r}, \mathbf{r}', \omega)$ equals to $W(\mathbf{r}', \mathbf{r}, -\omega)$. We don’t need to add a minus sign to \mathbf{r} or \mathbf{r}' , but we need to swap them, so that

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

while

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

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

$$\psi(\mathbf{x}, t) \xrightarrow{T} i\psi^\dagger(\mathbf{x}, -t), \quad (1.25)$$

and therefore

$$\begin{aligned} \mathcal{T} \langle \psi(\mathbf{x}, t) \psi^\dagger(\mathbf{x}', 0) \rangle &\xrightarrow{T} \mathcal{T} \langle i\psi^\dagger(\mathbf{x}, -t) i\psi(\mathbf{x}', 0) \rangle \\ &= -\mathcal{T} \langle \psi^\dagger(\mathbf{x}, -t) \psi(\mathbf{x}', 0) \rangle \\ &= \mathcal{T} \langle \psi(\mathbf{x}', 0) \psi^\dagger(\mathbf{x}, -t) \rangle. \end{aligned} \quad (1.26)$$

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

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 \mathbf{r} and \mathbf{r}' ! Let's have a look at what it means to change t to $-t$:

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

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 \mathbf{r} and \mathbf{r}' untouched and changing t to $-t$ is equivalent to keeping t untouched and swap \mathbf{r} and \mathbf{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 $(\mathbf{r}, \mathbf{r}', \omega)$ representation, if we take the first representation of the time reversal operation, we have

$$W(\mathbf{r}, \mathbf{r}', \omega) \xrightarrow{T} W(\mathbf{r}', \mathbf{r}, \omega), \quad (1.27)$$

while if we take the second representation, we get

$$W(\mathbf{r}, \mathbf{r}', \omega) \xrightarrow{T} W(\mathbf{r}, \mathbf{r}', -\omega). \quad (1.28)$$

Note that $W(\mathbf{r}, \mathbf{r}', \omega)$ is always equivalent to $W(\mathbf{r}', \mathbf{r}, -\omega)$, but the former is only equivalent to $W(\mathbf{r}', \mathbf{r}, \omega)$ 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

TODO: n interaction line $\Rightarrow n$ independent momenta, not all of which are on Coulomb interaction lines

A large benefit of this formalism is it lifts the burden to worry about normalization concerning δ -functions: no δ -function appears in the rules now!

$$v(\mathbf{q}) = \frac{4\pi e^2}{q^2}. \quad (1.29)$$

1.3.2 In a crystal

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

- The propagator is no longer $1/(\omega - \mathbf{k}^2/2m)$, but $\propto \sum_{\mathbf{k}} \psi_{n\mathbf{k}}(\mathbf{r}) \psi_{n\mathbf{k}}^*(\mathbf{r}')/(\omega - \xi_{n\mathbf{k}})$;
- The crystal momentum \mathbf{k} appears above. To carry out Feynman diagrammatic calculations in an actual computer, a cutoff on the density of \mathbf{k} points is needed (essentially, a cutoff on the size of the system V). This means we should replace $\int d^3\mathbf{k}/2\pi$ with properly normalized $\sum_{\mathbf{k}}$, where \mathbf{k} goes over the discretely infinite \mathbf{k} -grid.
- Usually we limit all momenta to the **first Brillouin zone (1BZ)**, and therefore a sum over an unbounded momentum variable \mathbf{k} or \mathbf{q} has to be recast into $\sum_G \sum_{\mathbf{k}} f(\mathbf{k} + \mathbf{G})$.

Below I describe a set of Feynman rules that are compatible with the notation in [13]. 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

$$\int \frac{d\omega_1}{2\pi} \cdots \int \frac{d\omega_n}{2\pi} \int d^3\mathbf{r}_1 \cdots \int d^3\mathbf{r}_{2n}. \quad (1.30)$$

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

$$c_{n\mathbf{k}}^\dagger = \int d^3\mathbf{r} \underbrace{e^{i\mathbf{k}\cdot\mathbf{r}} u_{n\mathbf{k}}(\mathbf{r})}_{\psi_{n\mathbf{k}}} \psi^\dagger(\mathbf{r}), \quad \{c_{n\mathbf{k}}, c_{n'\mathbf{k}'}^\dagger\} = \delta_{nn'} \delta_{\mathbf{k}\mathbf{k}'}, \quad (1.31)$$

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

$$\int d^3\mathbf{r} \psi_{n\mathbf{k}}^*(\mathbf{r}) \psi_{n'\mathbf{k}'}(\mathbf{r}) = \delta_{nn'} \delta_{\mathbf{k}\mathbf{k}'}, \quad (1.32)$$

and

$$G^0(\mathbf{r}, \mathbf{r}', \omega) = \sum_{n, \mathbf{k}} \frac{\psi_{n\mathbf{k}}(\mathbf{r}) \psi_{n\mathbf{k}}^*(\mathbf{r}')}{\omega - \xi_{n\mathbf{k}} + i0^+ \text{sgn}(\omega)}. \quad (1.33)$$

In the free-electron case, we expect

$$\psi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V}} e^{i(\mathbf{k} + \mathbf{G}_n) \cdot \mathbf{r}}, \quad (1.34)$$

and therefore (note that the only possibility for $\mathbf{G}_n + \mathbf{k}$ to be equal to $\mathbf{G}_{n'} + \mathbf{k}'$ is for \mathbf{G}_n to be $\mathbf{G}_{n'}$ and \mathbf{k} to be \mathbf{k}')

$$\text{LHS of (1.32)} = \frac{1}{V} \int d^3\mathbf{r} e^{i\mathbf{r} \cdot (-\mathbf{G}_n - \mathbf{k} + \mathbf{G}_{n'} + \mathbf{k}')} = \frac{1}{V} V \delta_{\mathbf{G}_n + \mathbf{k}, \mathbf{G}_{n'} + \mathbf{k}'} = \delta_{nn'} \delta_{\mathbf{k}\mathbf{k}'},$$

and

$$\begin{aligned} \text{RHS of (1.33)} &= \frac{1}{V} \sum_{n, \mathbf{k}} e^{i(\mathbf{k} + \mathbf{G}_n) \cdot (\mathbf{r} - \mathbf{r}')} = \frac{1}{V} \sum_{\text{unbounded } \mathbf{k}} e^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} \\ &= \int \frac{d^3\mathbf{k}}{(2\pi)^3} e^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} =: \text{LHS of (1.33)}, \end{aligned}$$

so (1.32) and (1.33) 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:

$$M = \int d^3\mathbf{r} \int d^3\mathbf{r}' iG(\mathbf{r}_4, \mathbf{r}, \omega_1) iG(\mathbf{r}_3, \mathbf{r}', \omega_2) iG(\mathbf{r}', \mathbf{r}_2, \omega_3) iG(\mathbf{r}, \mathbf{r}_1, \omega_4) (-i)v(\mathbf{r} - \mathbf{r}'). \quad (1.35)$$

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 (1.33) into the above expression, and using the definition

$$v(\mathbf{r} - \mathbf{r}') = \int \frac{d^3\mathbf{p}}{(2\pi)^3} e^{i\mathbf{p}\cdot(\mathbf{r}-\mathbf{r}')} v(\mathbf{p}), \quad (1.36)$$

we have

$$\begin{aligned} M &= \psi_4(\mathbf{r}_4)\psi_3(\mathbf{r}_3)\psi_2^*(\mathbf{r}_2)\psi_1^*(\mathbf{r}_1) \\ &\times \int \frac{d^3\mathbf{q}}{(2\pi)^3} \sum_{1,2,3,4} \frac{i}{\omega_1 - \xi_4 + i0^+ \text{sgn}(\omega_1)} \frac{i}{\omega_1 - \xi_3 + i0^+ \text{sgn}(\omega_2)} \\ &\times \frac{i}{\omega_1 - \xi_2 + i0^+ \text{sgn}(\omega_3)} \frac{i}{\omega_1 - \xi_1 + i0^+ \text{sgn}(\omega_4)} \\ &\times \int d^3\mathbf{r} \psi_4^*(\mathbf{r})\psi_1(\mathbf{r})e^{i\mathbf{q}\cdot\mathbf{r}} \int d^3\mathbf{r}' \psi_3^*(\mathbf{r}')\psi_2(\mathbf{r}')e^{-i\mathbf{q}\cdot\mathbf{r}'} v(\mathbf{q}), \end{aligned} \quad (1.37)$$

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

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

and eventually we get

$$\begin{aligned} M &= \psi_4(\mathbf{r}_4)\psi_3(\mathbf{r}_3)\psi_2^*(\mathbf{r}_2)\psi_1^*(\mathbf{r}_1) \\ &\times \frac{1}{V} \sum_{\mathbf{G}} \sum_{\mathbf{q}} \sum_{1,2,3,4} \frac{i}{\omega_1 - \xi_4 + i0^+ \text{sgn}(\omega_1)} \frac{i}{\omega_1 - \xi_3 + i0^+ \text{sgn}(\omega_2)} \\ &\times \frac{i}{\omega_1 - \xi_2 + i0^+ \text{sgn}(\omega_3)} \frac{i}{\omega_1 - \xi_1 + i0^+ \text{sgn}(\omega_4)} \\ &\times \langle 4 | e^{i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}} | 1 \rangle \langle 3 | e^{-i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}} | 2 \rangle v(\mathbf{q} + \mathbf{G}). \end{aligned} \quad (1.38)$$

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

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

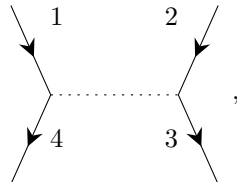
$$\overrightarrow{n, k} = \frac{i}{\omega - \xi_{n\mathbf{k}} + i0^+ \text{sgn}(\omega)} =: iG_{n\mathbf{k}}^0(\omega), \quad (1.39)$$

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

- For each Coulomb interaction line, write

$$\dots\dots\dots q, \mathbf{G} = -i \frac{1}{V} v(\mathbf{q} + \mathbf{G}) \quad (1.40)$$

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 $\langle \text{out} | e^{\pm i(\mathbf{q} + \mathbf{G})} | \text{in} \rangle$:

$$\begin{array}{c} n', k \\ \searrow \\ \text{---} \xrightarrow{q, \mathbf{G}} \text{---} \\ \nearrow \\ n, k - q \end{array} = \langle n, \mathbf{k} - \mathbf{q} | e^{-i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n', \mathbf{k} \rangle, \quad (1.41)$$

and

$$\begin{array}{c} n', k \\ \searrow \\ \text{---} \xleftarrow{q, \mathbf{G}} \text{---} \\ \nearrow \\ n, k + q \end{array} = \langle n, \mathbf{k} + \mathbf{q} | e^{i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n', \mathbf{k} \rangle. \quad (1.42)$$

The sign is decided by the direction of \mathbf{q} . The direction of \mathbf{q} doesn't matter: we *don't* sum over the two directions. The value of \mathbf{q} is determined by the values of free momentum variables; if \mathbf{q} is not selected as a free momentum variable, it shouldn't be integrated over.

Now we compare the above Feynman rules with the momentum space Feynman rules in Section 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 $\int d^3\mathbf{q}$ by $\sum_{\mathbf{G}} \sum_{\mathbf{q}}$.

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 \mathbf{q} . In [13], the $1/V$ factor comes together with $v(\mathbf{q} + \mathbf{G})$. Indeed, they define

$$v(\mathbf{q} + \mathbf{G}) = \frac{4\pi e^2}{V|\mathbf{q} + \mathbf{G}|^2}. \quad (1.43)$$

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

$$\begin{array}{c} n, k \\ \curvearrowright \\ \text{---} \text{---} \text{---} \\ \curvearrowleft \\ m, q - k \end{array} = \int \frac{d\omega}{2\pi} \sum_{m,n} iG_n^0(\mathbf{k}, \omega) iG_m^0(\mathbf{q} - \mathbf{k}, \omega_0 - \omega), \quad (1.44)$$

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

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

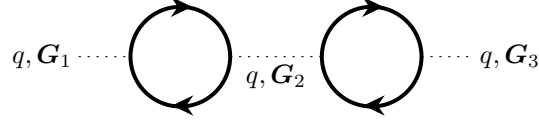
$$H = \sum_{1,2,3,4} c_4^\dagger c_3^\dagger H_{4321} c_2 c_1, \quad (1.45)$$

$$H_{4321} = \frac{1}{2} \int d^3\mathbf{r} \int d^3\mathbf{r}' \langle 4 | \mathbf{r} \rangle \langle 3 | \mathbf{r}' \rangle \langle \mathbf{r}' | 2 \rangle \langle \mathbf{r} | 1 \rangle v(\mathbf{r} - \mathbf{r}'), \quad (1.46)$$

Now again invoking the Fourier transformation from $v(\mathbf{r} - \mathbf{r}')$ to $v(\mathbf{q})$, and discretize the integration over \mathbf{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 \mathbf{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 \mathbf{G}_n , but once the crystal potential and the self-energy is added, electrons can be scattered from one \mathbf{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 \mathbf{G} vectors.

Thus, the \mathbf{G} - and n -summations mentioned in Section 1.3.2 can be recast into matrix forms:

$$\longrightarrow \overset{k}{\longrightarrow} = iG_{\mathbf{k}}^0(\omega) := i[\delta_{nn'}G_{n\mathbf{k}}^0(\omega)]_{nn'}, \quad (1.47)$$

$$\cdots \overset{q}{\cdots} = -iv(q) := -i[\delta_{\mathbf{G}\mathbf{G}'}v(\mathbf{q} + \mathbf{G})]_{\mathbf{G}\mathbf{G}'}, \quad (1.48)$$

and

$$\begin{array}{c} k \\ \swarrow \\ \text{---} \overset{q}{\longleftarrow} \text{---} \\ \searrow \\ k+q \end{array} = M := [\langle n, \mathbf{k} + \mathbf{q} | e^{i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n' \mathbf{k} \rangle]_{nn' \mathbf{G}}, \quad (1.49)$$

$$\begin{array}{c} k \\ \swarrow \\ \text{---} \overset{q}{\longrightarrow} \text{---} \\ \searrow \\ k+q \end{array} = M^* = [\langle n' \mathbf{k} | e^{-i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n, \mathbf{k} + \mathbf{q} \rangle]_{nn' \mathbf{G}}. \quad (1.50)$$

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

In the context of *ab initio* calculations, the term quasiparticle usually means renormalized band electrons and holes. Bosonic modes are just called “excitations”, and non-electron-like fermionic modes like spinons are simply absent (we are studying weakly correlated systems, anyway), 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 $\varepsilon_{n\mathbf{k}}$ of quasiparticles. The correction to the $\varepsilon_{n\mathbf{k}}$ 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 $\langle \text{quasiparticle } n, \mathbf{k} | \text{ground state} \rangle \neq 1$, 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

$$E = \varepsilon_{n\mathbf{k}} n_{n\mathbf{k}} + f n n. \quad (1.51)$$

The relation between diagrammatic components and the ε , f functions are given in [12].

1.4.1 Spectral density

Chapter 2

Experimental characterization methods

2.1 Light-matter interaction

2.1.1 Interaction Hamiltonian

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

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

and therefore the full light-matter interaction Hamiltonian is

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

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

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

Note that here \mathbf{p} is the original momentum operator, i.e. $-i\nabla$, 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 $V(\mathbf{r}, \mathbf{r}')$ is not diagonal in the coordinate representation. If we stick to writing the Hamiltonian in (\mathbf{x}, \mathbf{p}) , then essentially this means we have \mathbf{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 \mathbf{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

$$H_{\text{light-matter}} = -\mathbf{d} \cdot \mathbf{E}. \quad (2.4)$$

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

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

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

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

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

$$\varphi = 0, \quad \nabla \cdot \mathbf{A} = 0, \quad (2.7)$$

and this means

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t}. \quad (2.8)$$

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 $\frac{\partial}{\partial t} \nabla \cdot \mathbf{E} = 0$, 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 $V(\mathbf{r}, \mathbf{r}')$ into $V(\mathbf{r}, \mathbf{p})$, and with appropriate normal ordering, we have

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

Therefore the single-body Hamiltonian after electromagnetic coupling is

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

where

$$\dot{\mathbf{r}} = \mathbf{v} = \frac{1}{i\hbar}[\mathbf{r}, H] = \frac{\mathbf{p}}{m} + \frac{1}{i\hbar} \underbrace{[\mathbf{r}, V]}_{\neq 0}. \quad (2.9)$$

So we get 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.

2.1.3 How to capture absorption

Absorption is usually modeled by the imaginary part of ϵ , and it's usually calculated by Fermi golden rule. In this section, we assume the incident light has a determined wave length and polarization. The general form of \mathbf{A} is

$$\mathbf{A} = \sum_{\mathbf{k}, \sigma} \sqrt{\frac{\hbar}{2\omega_{\mathbf{k}\sigma}\epsilon_0 V}} (a_{\mathbf{k}\sigma} \hat{\mathbf{e}}_{\mathbf{k}\sigma} e^{i\mathbf{k} \cdot \mathbf{r} - i\omega_{\mathbf{k}\sigma} t} + \text{h.c.}), \quad (2.10)$$

and here we are only considering one \mathbf{k}, σ mode; from then on we use \mathbf{q} to refer to the momentum of the mode, and ω becomes a shorthand for $\omega_{\mathbf{q}\sigma}$. Similarly, n and $\hat{\mathbf{e}}$ mean $n_{\mathbf{q}\sigma}$ and $\hat{\mathbf{e}}_{\mathbf{q}\sigma}$. 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

$$\sqrt{\epsilon_r} = n + i\kappa, \quad (2.11)$$

and therefore (TODO: metal)

$$\epsilon_r = n^2 - \kappa^2 + 2in\kappa \approx n^2 + 2i\kappa. \quad (2.12)$$

Here we assume ϵ_r is not far from one, and therefore n is close to one and κ is small. A beam of light therefore propagates as

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

This means the intensity is like

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

If we know α in

$$I = I_0 e^{-\alpha r}, \quad (2.13)$$

then

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

and therefore

$$\text{Im } \epsilon_r = 2\kappa = \frac{c}{\omega} \alpha. \quad (2.14)$$

The next question is what is α . (2.13) means in the distance of dr , the photon number loss is proportional to

$$I(r) - I(r + dr) = I(r) \alpha dr.$$

This means

$$\alpha dr = \frac{\text{photons absorbed per unit volume per second} \cdot dr}{\text{photons incident per unit area per second}}. \quad (2.15)$$

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 $\text{Im } \epsilon$ 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 $\text{Im } H_{\text{eff}}$, 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 $\text{Im } \epsilon$ is about $\text{Im } H_{\text{eff}}$, while Fermi golden rule is about the quantum jump channels, so it’s indeed correct to infer $\text{Im } \epsilon$ from Fermi golden rule, using a procedure like, say, comparing the amount of light absorbed calculated from $\text{Im } \epsilon$ and from Fermi golden rule. There is no double counting in this procedure: *both* $\text{Im } H_{\text{eff}}$ 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:

$$\begin{aligned}
& \text{photons absorbed per unit volume per second} \\
&= \frac{1}{V} \cdot \text{photons absorbed per second} \\
&= \frac{1}{V} \cdot \frac{2\pi}{\hbar} \sum_S |\langle S, n-1 | e\mathbf{v} \cdot \mathbf{A} | 0, n \rangle|^2 \delta(\omega - \Omega_S) \\
&= \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon_0 V} \cdot |\langle S, n-1 | \mathbf{v} \cdot a_{\mathbf{k}\sigma} \hat{\mathbf{e}} e^{i\mathbf{k}\cdot\mathbf{r}} | 0, n \rangle|^2 \delta(\omega - \Omega_S) \\
&\approx \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon_0 V} \cdot |\langle S, n-1 | \mathbf{v} \cdot a_{\mathbf{k}\sigma} \hat{\mathbf{e}} | 0, n \rangle|^2 \delta(\omega - \Omega_S) \\
&= \frac{1}{V} \frac{2\pi}{\hbar} \sum_S \frac{\hbar e^2}{2\omega\epsilon_0 V} \cdot |\mathbf{v} \cdot \hat{\mathbf{e}} \sqrt{n} \langle S, n-1 | 0, n-1 \rangle|^2 \delta(\omega - \Omega_S) \\
&= \frac{\pi e^2 n}{\omega\epsilon_0 V^2} \sum_S |\langle S | \mathbf{v} \cdot \hat{\mathbf{e}} | 0 \rangle|^2 \delta(\omega - \Omega_S),
\end{aligned}$$

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. The denominator is

$$\text{photons incident per unit area per second} = \frac{nc}{V}.$$

So putting everything together, we get

$$\alpha = \frac{\pi e^2}{\omega\epsilon_0 c V} \sum_S |\langle S | \mathbf{v} \cdot \hat{\mathbf{e}} | 0 \rangle|^2 \delta(\omega - \Omega_S), \quad (2.16)$$

and

$$\text{Im } \epsilon_r = \frac{\pi e^2}{\omega^2 \epsilon_0 V} \sum_S |\langle S | \mathbf{v} \cdot \hat{\mathbf{e}} | 0 \rangle|^2 \delta(\omega - \Omega_S). \quad (2.17)$$

The relation (2.17) is derived under SI. In Gaussian system of units, the form of ϵ_r remains unchanged. This can be seen by

$$\begin{cases} F = \frac{(e^I)^2}{4\pi\epsilon_0 r^2} = \frac{(e^G)^2}{r^2}, & \mathbf{F} = e^I \mathbf{E}^I = e^G \mathbf{E}^G, \\ \mathbf{D}^I = \epsilon_r^I \epsilon_0 \mathbf{E}^I, & \mathbf{D}^G = \epsilon_r^G \mathbf{E}^G, \\ \nabla \cdot \mathbf{D}^I = \rho^I, & \nabla \cdot \mathbf{D}^G = 4\pi \rho^G \end{cases} \Rightarrow \epsilon_r^I = \epsilon_r^G.$$

So in Gaussian system of units, we have

$$\text{Im } \epsilon_r = \frac{4\pi^2 e^2}{\omega^2 V} \sum_S |\langle S | \mathbf{v} \cdot \hat{\mathbf{e}} | 0 \rangle|^2 \delta(\omega - \Omega_S). \quad (2.18)$$

The expression of $\text{Im } \epsilon_r$ sees great variance in literature. Eq. (8) in [9] is almost the same as (2.18), except it misses the factor $1/V$. [10] claims the equation should be

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

The LHS here is shown as it is in [10]; 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 4π factor is also mysterious. This version of (2.18) seems to have an early origin: it appears in [17], which refers us to [1] which doesn't seem to contain something like (2.18).

Chapter 3

GW and BSE

3.1 What is GW

3.1.1 GW is screened Hartree-Fock approximation

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


(3.1)

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

$$\Sigma = iGW, \quad (3.2)$$

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

$$W = \frac{v}{\epsilon} = \dots + \dots \bigcirc \dots + \dots \bigcirc \bigcirc \dots + \dots, \quad (3.3)$$

$$\epsilon = 1 - \dots \bigcirc \dots \quad (3.4)$$

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. The *dielectric matrix* used to connect vector \mathbf{E} and \mathbf{D} is related to $\epsilon_{\mathbf{G}\mathbf{G}'}$ (TODO: how?) but they are not simply equal.

3.1.2 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

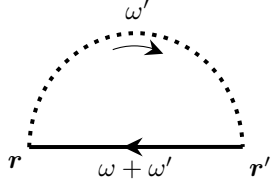
$$\Sigma(1, 2) = iW(1^+, 2)G(1, 2), \quad (3.5)$$

instead of the schematic $\Sigma = iWG$. This expression can be seen in several places, including Wikipedia and [6, 12]. The necessity of this infinitesimal displacement on t_1 can be found in Section 3.2.2.

The question, then, is where this displacement on t_1 comes from. TODO: verify My tentative conclusion is that the original interpretation of the screened Fock diagram is actually

$iW(1,2)G(1^-,2)$, where the $-$ superscript comes from (1.11); since we can leave out the $e^{i\omega 0^+}$ factor in the self-energy (Section 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 \mathbf{r} and \mathbf{r}' doesn't really matter (Section 1.2.4), we have

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


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

$$\Sigma(\mathbf{r}, \mathbf{r}', \omega) = i \int \frac{d\omega'}{2\pi} e^{i\omega' 0^+} G(\mathbf{r}, \mathbf{r}', \omega + \omega') W(\mathbf{r}, \mathbf{r}', \omega'). \quad (3.6)$$

This is exactly Eq. (4a) in [6].

Here is a tricky point: the sign of the exponent in $e^{i\omega' 0^+}$ depends on the sign of ω' dependence in the G factor, not the sign of ω' in the W factor. This can be seen in Eq. (10) in [13]. Indeed, we know that replacing $W(\omega')$ by $W(-\omega')$ doesn't change the value of any diagram after integrating over all internal variables (Section 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 .

3.1.3 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.

3.1.4 BSE with the same order of approximation

TODO:

- Truncated Coulomb interaction
- Why we say $\Sigma = V_{xc} + \Sigma - V_{xc}$? (see [10] p. 1271)
- What should be symmetric? (11) and (12) in [10]???

3.2 The dielectric matrix ϵ

3.2.1 Frequency-dependent form

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

$$q \cdots \cdots \bigcirc \cdots \cdots q = i\chi(q) \quad (3.7)$$

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 $\chi_{\mathbf{G}\mathbf{G}'}$.

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

$$-iW(q) = -iv(q) + (-iv(q))(i\chi(q))(-iv(q)) + \cdots,$$

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

So we get a concrete expression of ϵ :

$$\epsilon(q) = 1 - v(q)\chi(q), \quad W(q) = \epsilon(q)^{-1}v(q). \quad (3.8)$$

Because

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

we have

$$\epsilon_{\mathbf{G}\mathbf{G}'}(q) = \delta_{\mathbf{G}\mathbf{G}'} - v(q + \mathbf{G})\chi_{\mathbf{G}\mathbf{G}'}(q). \quad (3.9)$$

Here the inverse of ϵ is matrix inverse, not element-by-element inverse. (13) in [10] is ambiguous: a clearer but more cumbersome notation is

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

Now it comes the evaluation of $\chi_{\mathbf{G}\mathbf{G}'}(q)$. Under the quasiparticle assumption, that the renormalized Green function can still be written as (1.33), We have

$$\begin{aligned} i\chi_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega) &= q, G \cdots \cdots \bigcirc \cdots \cdots q, G' \\ &= - \int \frac{d\omega'}{2\pi} \sum_{\mathbf{k}} \sum_{n, n'} \frac{i}{\omega' - \xi_{n'\mathbf{k}} + i0^+ \text{sgn}(\xi_{n'\mathbf{k}})} \frac{i}{\omega + \omega' - \xi_{n, \mathbf{k}+\mathbf{q}} + i0^+ \text{sgn}(\xi_{n, \mathbf{k}+\mathbf{q}})} \\ &\quad \times M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) M_{nn'}^*(\mathbf{k}, \mathbf{q}, \mathbf{G}'), \end{aligned} \quad (3.10)$$

where the minus sign comes from the closed electron loop, and to make the final expression agree with [10], I choose a rather strange direction of the \mathbf{q} arrow. The bands and wave functions are all renormalized ones.

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

$$\begin{aligned} &\int \frac{d\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\mathbf{k}} + i0^+ \text{sgn}(\xi_{n'\mathbf{k}})} \frac{1}{\omega + \omega' - \xi_{n, \mathbf{k}+\mathbf{q}} + i0^+ \text{sgn}(\xi_{n, \mathbf{k}+\mathbf{q}})} \\ &= \int \frac{d\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\mathbf{k}} - i0^+} \frac{1}{\omega + \omega' - \xi_{n, \mathbf{k}+\mathbf{q}} + i0^+} \\ &= \frac{2\pi i}{2\pi} \frac{1}{\omega + \xi_{n'\mathbf{k}} + i0^+ - \xi_{n, \mathbf{k}+\mathbf{q}} + i0^+}, \end{aligned}$$

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

$$\begin{aligned} & \int \frac{d\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\mathbf{k}} + i0^+ \operatorname{sgn}(\xi_{n'\mathbf{k}})} \frac{1}{\omega + \omega' - \xi_{n,\mathbf{k}+\mathbf{q}} + i0^+ \operatorname{sgn}(\xi_{n,\mathbf{k}+\mathbf{q}})} \\ &= \int \frac{d\omega'}{2\pi} \frac{1}{\omega' - \xi_{n'\mathbf{k}} + i0^+} \frac{1}{\omega + \omega' - \xi_{n,\mathbf{k}+\mathbf{q}} - i0^+} \\ &= \frac{2\pi i}{2\pi} \frac{1}{-\omega + \xi_{n,\mathbf{k}+\mathbf{q}} + i0^+ - \xi_{n'\mathbf{k}} + i0^+} \end{aligned}$$

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

$$\begin{aligned} \chi_{GG'}(\mathbf{q}, \omega) &= \sum_{\mathbf{k}} \sum_{n,n'} M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) M_{nn'}^*(\mathbf{k}, \mathbf{q}, \mathbf{G}') \\ &\times \left(\frac{\theta(-\xi_{n'\mathbf{k}})\theta(\xi_{n,\mathbf{k}+\mathbf{q}})}{\omega + \xi_{n'\mathbf{k}} - \xi_{n,\mathbf{k}+\mathbf{q}} + i0^+} + \frac{\theta(\xi_{n'\mathbf{k}})\theta(-\xi_{n,\mathbf{k}+\mathbf{q}})}{-\omega + \xi_{n,\mathbf{k}+\mathbf{q}} - \xi_{n'\mathbf{k}} + i0^+} \right). \end{aligned} \quad (3.11)$$

TODO: occupied, or empty:

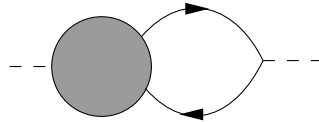
Note that in [10], the normalization of χ is the same as here, but the definition of $v(\mathbf{q})$ doesn't include the $1/V$ factor. So (8) in [10] is wrong: it leads to a ridiculous result: as we increase the density of the \mathbf{k} -grid, χ becomes larger and larger!

3.2.2 The static limit and the generalized plasmon-pole model

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

$$\begin{aligned} \chi_{GG'}(\mathbf{q}, \omega = 0) &= \sum_{\mathbf{k}} \sum_{n,n'} M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G}) M_{nn'}^*(\mathbf{k}, \mathbf{q}, \mathbf{G}') \\ &\times \left(\frac{\theta(-\xi_{n'\mathbf{k}})\theta(\xi_{n,\mathbf{k}+\mathbf{q}})}{\varepsilon_{n'\mathbf{k}} - \varepsilon_{n,\mathbf{k}+\mathbf{q}}} + \frac{\theta(\xi_{n'\mathbf{k}})\theta(-\xi_{n,\mathbf{k}+\mathbf{q}})}{\varepsilon_{n,\mathbf{k}+\mathbf{q}} - \varepsilon_{n'\mathbf{k}}} \right). \end{aligned} \quad (3.12)$$

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 introduced 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

$$\frac{1}{\epsilon} - 1 = \text{const.} \times \frac{1}{\omega^2 - \omega_0^2 + i0^+}, \quad (3.13)$$

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

$$\frac{1}{\epsilon} - 1 = \frac{\Omega^2}{2\tilde{\omega}} \left(\frac{1}{\omega - \tilde{\omega} + i0^+} - \frac{1}{\omega + \tilde{\omega} - i0^+} \right). \quad (3.14)$$

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:

$$\omega_p = \sqrt{\frac{ne^2}{m\epsilon_0}}. \quad (3.15)$$

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

$$\epsilon = 1 - \frac{\omega_p^2}{\omega^2}. \quad (3.16)$$

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 (3.12). 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 (3.12).

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 $e^{i0^+\omega}$ factor, where ω is the frequency on the Coulomb interaction line. Thus the GW self-energy is proportion to the following factor:

$$\int d\omega' e^{i0^+\omega'} \frac{1}{\omega + \omega' - \xi_{n\mathbf{k}} + i0^+ \text{sgn}(\xi_{n\mathbf{k}})} \left(\frac{1}{\omega' - \tilde{\omega} + i0^+} - \frac{1}{\omega' + \tilde{\omega} - i0^+} \right).$$

When $\xi_{n\mathbf{k}} > 0$, we find only the pole of the $1/(\omega' + \tilde{\omega} - i0^+)$ term contributes to the integral, and the result is proportional to

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

which means the singularity is achieved when $\xi_{n\mathbf{k}} = \omega - \tilde{\omega}$. Note that $\xi_{n\mathbf{k}}$ 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 $\xi_{n\mathbf{k}} < 0$, all terms contribute to the integral, but now the residue of the pole of $\xi_{n\mathbf{k}} = \omega - \tilde{\omega}$ cancels with the residue of the pole of the electron propagator, and the result is proportional to

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

This means $\xi_{n\mathbf{k}}$ is ω – the input energy – plus $\tilde{\omega}$. But note that in this case $\xi_{n\mathbf{k}} < 0$, so the above expression means $\xi_{n\mathbf{k}}$ 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 $e^{i0^+\omega}$ factor gives the correct physical picture. The origin of this factor is discussed in Section 3.1.2.

3.2.3 The static subspace approach

3.3 The self-energy matrix Σ

3.3.1 COHSEX approximation

One way to simply Σ^{GW} is the **Coulomb hole-screened exchange (COHSEX)** approximation. In this approximation, we assume $W(\mathbf{r}, t; \mathbf{r}', 0)$ doesn't show strong retardation, and therefore in the frequency domain, $W(\mathbf{r}, \mathbf{r}', \omega)$ shouldn't be far from $W(\mathbf{r}, \mathbf{r}', 0)$. Therefore, (3.6) is approximately

$$\begin{aligned} \Sigma(\mathbf{r}, \mathbf{r}', \omega) &\approx i \int \frac{d\omega'}{2\pi} \int dt e^{i(\omega+\omega')t} G(\mathbf{r}, t; \mathbf{r}', 0) W(\mathbf{r}, \mathbf{r}', \omega = 0) \\ &= i \int e^{i\omega t} \delta(t) G(\mathbf{r}, t; \mathbf{r}', 0) W(\mathbf{r}, \mathbf{r}', \omega = 0) \\ &= i G(\mathbf{r}, t = 0; \mathbf{r}', 0) W(\mathbf{r}, \mathbf{r}', \omega = 0). \end{aligned}$$

But note that $G(t = 0)$ is not well-defined: at $t = 0$ there is a stepwise jump (Section 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

$$\begin{aligned}\Sigma^{GW}(\mathbf{r}, \mathbf{r}', \omega) &\approx \Sigma^{\text{COHSEX}}(\mathbf{r}, \mathbf{r}', \omega) \\ &= \frac{i}{2}(G(\mathbf{r}, t = 0^+; \mathbf{r}', 0) + G(\mathbf{r}, t = 0^-; \mathbf{r}', 0))W(\mathbf{r}, \mathbf{r}', \omega = 0),\end{aligned}\quad (3.17)$$

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

$$\Sigma^{\text{COHSEX}} = \frac{i}{2}W(\mathbf{r}, \mathbf{r}', \omega = 0)G(\mathbf{r}, t; \mathbf{r}', 0)(\delta(t + 0^+) + \delta(t - 0^+)). \quad (3.18)$$

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

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

and therefore we have

$$\Sigma^{\text{COHSEX}} = \Sigma^{\text{COH}} + \Sigma^{\text{SEX}}, \quad (3.19)$$

$$\Sigma^{\text{COH}} = \frac{1}{2}\delta(\mathbf{r} - \mathbf{r}')W(\mathbf{r}, \mathbf{r}', \omega = 0), \quad (3.20)$$

$$\Sigma^{\text{SEX}} = - \langle \Psi | \psi^\dagger(\mathbf{r}') \psi(\mathbf{r}) | \Psi \rangle W(\mathbf{r}, \mathbf{r}', \omega = 0). \quad (3.21)$$

3.3.2 Diagonal or not

We know in the momentum space, we have

$$E_{n\mathbf{k}}^{\text{QP}} = E_{n\mathbf{k}}^0 + \Sigma_{n\mathbf{k}}(E_{n\mathbf{k}}^{\text{QP}}). \quad (3.22)$$

Here since Σ depends on the corrected propagator, $E_{n\mathbf{k}}^{\text{QP}}$ 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; on the other hand, in principle – though of course not in practice – DFT is able to decide everything about the system), so (3.22) now is

$$E_{\mathbf{k}}^{\text{QP}} = E_{\mathbf{k}}^{\text{KS}} + \Sigma_{\mathbf{k}}(E_{\mathbf{k}}^{\text{QP}}) - \Sigma^{\text{KS}}. \quad (3.23)$$

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 $\Sigma_{\mathbf{k}}$ may mix different bands together, and (3.23) is an equation about matrices, essentially a single-electron Schrodinger equation. Its first order approximation is

$$E_{n\mathbf{k}}^{\text{QP}} = E_{n\mathbf{k}}^{\text{KS}} + \langle \psi_{n\mathbf{k}}^{\text{KS}} | \Sigma(E_{n\mathbf{k}}^{\text{QP}}) - \Sigma^{\text{KS}} | \psi_{n\mathbf{k}}^{\text{KS}} \rangle. \quad (3.24)$$

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} ”?

3.3.3 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 (3.24). In this case, we don't need off-diagonal elements, because they are not used in (3.24). This scheme is mentioned in Section 3.3 in [10]. The second scheme takes the change of eigenstates into account, and thus iteratively solves (3.23). In this case we need to take non-diagonal elements seriously [2, 11]. 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

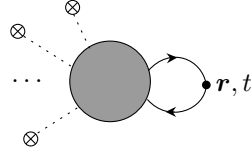
LDA functional together with G_0W_0 provides better results. Still, the argumentation provided above only explains why iterative GW is bad, but doesn't explain why one-shot GW is good. In other words, we need to know how certain factors in the one-shot GW scheme somehow makes up for the missing vertex correction. Indeed, if we capture the vertex effects, the accuracy can be improved [19]. 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.

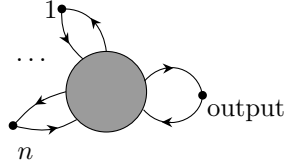
The question, then, is why LDA in some cases works as well as BSE. The reason for this is because of the relation between the derivative discontinuity in DFT and electron-hole interaction kernel TODO: the relation with [16]

3.4 From GW to BSE

Suppose we want to find the electromagnetic response of a system. What we want is something looking like $\delta^n n(\mathbf{r})/\delta V(\mathbf{r}')^n$. With the influence of $V(\mathbf{r})$ taken into account, $n(\mathbf{r}, t)$ 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 n th-order response, therefore, is given by



We can easily see that to find the n th-order response, it's necessary to find the n th-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 $n(\mathbf{r}, t)$ 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.

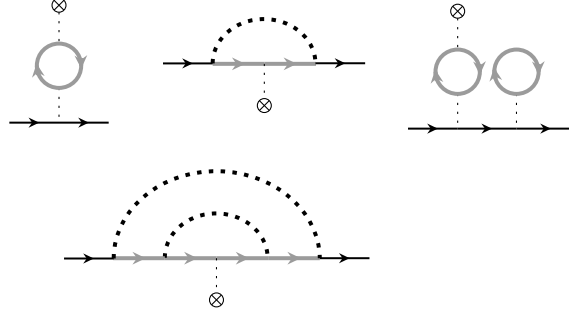
Now suppose we want to work with the single-electron Green function G only in the first approach, and the GW self-energy plus the electromagnetic coupling are used to build the effective Hamiltonian. In other words, we are doing **time-dependent adiabatic GW** (for the meaning of “adiabatic”, see below). The density $n(\mathbf{r}, t)$ is obtained by $G(\mathbf{r}, t; \mathbf{r}, t)$, or in other words, by connecting the incoming and outgoing lines to the (\mathbf{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 $G(\mathbf{r}, t; \mathbf{r}', t')$ by $G(\mathbf{r}, t; \mathbf{r}', t)$, 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}

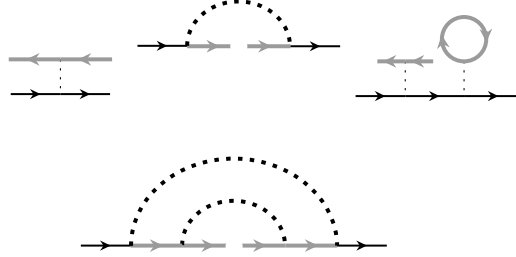
and electromagnetic coupling. Thus we have

$$\text{---} \text{---} \text{---} = \text{---} \text{---} \text{---} + \text{---} \text{---} \text{---} \text{---} \otimes + \dots \quad (3.25)$$

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:



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

$$\begin{array}{c} \text{connected} \\ \text{to external field} \end{array} \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} \begin{array}{c} \text{connected} \\ \text{to response} \end{array} = \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \dots + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \dots \quad (3.26)$$

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

$$\begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} = \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} \quad (3.27)$$

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 Section 3.1.3.

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 (3.27) to

$$\begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} = \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \begin{array}{c} \text{---} \text{---} \text{---} \\ \text{---} \text{---} \text{---} \end{array} + \dots, \quad (3.28)$$

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

Box 3.1: 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.

(3.28) 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 (3.28) *plus* (3.27): 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 (3.27). 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 (3.27) – is essentially analyzing excitons.

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**. Both flavors of BSE have been implemented: the exciton Hamiltonian approach is realized 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 [4].

3.5 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.

3.6 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 . [18] 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>

Chapter 4

The QuantumESPRESSO-BerkeleyGW ecosystem

4.1 Overview of the pipeline

TODO: whether and how BerkeleyGW supports HSE06. See <https://groups.google.com/a/berkeleygw.org/g/help/c/I>

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, so building the kernel – counterpart of Σ – is done in one step (`kernel`), while diagonalizing it is done in another step (`absorption`).

4.2 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 $\mathbf{L} \cdot \mathbf{S}$ or p^4 [5]. 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*.

4.3 Input and output of pw

4.4 The epsilon step

4.4.1 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 [10]; relative discussions can be found in Section 3.2. There are three steps in `epsilon`:

1. Calculate $M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G})$;
2. Summing over n, n', \mathbf{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 \mathbf{k} and \mathbf{q} points are given by the \mathbf{q} -grid given in `epsilon.inp` and the \mathbf{k} -grid in the wave function files, so they are fixed and not a part of the scaling. With the cutoff energy fixed, the size of the \mathbf{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 $M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G})$ scales as N^3 .

For each matrix element, we need to calculate $\langle n, \mathbf{k} + \mathbf{q} | e^{i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r}} | n', \mathbf{k} \rangle$. Note that the matrix inside contains only \mathbf{r} , and the expression therefore can be evaluated as

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

and the scale of the calculation needed is proportional to V and again N . In practice, we use the \mathbf{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 $M_{nn'}(\mathbf{k}, \mathbf{q}, \mathbf{G})$ with fixed n, n', \mathbf{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 $\chi_{\mathbf{G}\mathbf{G}'}$, where the values of \mathbf{G} and \mathbf{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 .

4.4.2 Divergence problems when $\mathbf{q} \rightarrow 0$

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

$$\begin{aligned} \epsilon_{00}(\mathbf{q} \rightarrow 0, \omega = 0) &\propto -\frac{1}{q^2} \chi_{00}(\mathbf{q} \rightarrow 0, \omega = 0) \\ &\propto \sum_{n \text{ occupied}, n' \text{ empty}} -\frac{1}{q^2} |\langle n\mathbf{k} | 1 + i(\mathbf{q} + \mathbf{G}) \cdot \mathbf{r} + \dots | n'\mathbf{k} \rangle|^2 \\ &\propto \text{const.} \times \frac{q^2}{q^2}. \end{aligned} \quad (4.1)$$

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

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

4.4.3 Frequency dependence of ϵ

The $\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega)$ 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 $\epsilon_{\mathbf{G}\mathbf{G}'}(\mathbf{q}, \omega)$ 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

4.4.4 Console output

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

1. Initialization
2. Iterating over the \mathbf{q} -grid. For each \mathbf{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 Section 4.4.1. 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.
```

4.4.5 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 (Section 6.2.4).

4.5 Systems of units

TODO

Chapter 5

Tight-binding models

TODO: from ab initio to models

5.1 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 [7]

[15] 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 \mathbf{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**

$$\Omega = \sum_n (\langle \mathbf{r}^2 \rangle_n - \langle \mathbf{r} \rangle_n^2). \quad (5.1)$$

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

$$\tilde{\Omega} = \sum_n \sum_n \text{TODO} \quad (5.2)$$

It contains a derivative $\nabla_{\mathbf{k}}$, and the usual numerical format is used.

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

$$|\phi_{n\mathbf{k}}\rangle = \sum_{m=1}^J |\psi_{m\mathbf{k}}\rangle \langle \psi_{m\mathbf{k}} | g_n \rangle. \quad (5.3)$$

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.

5.2 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 [3]

$$W = (1 - W_r P_d)^{-1} W_r, \quad W_r = (1 - v P_r)^{-1} v, \quad (5.4)$$

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

Chapter 6

Standard operation procedures

List of technical issues:

- Band mode of BerkeleyGW
- Wannier “oscillation”
-

6.1 Details in installation

6.1.1 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`.

6.2 Standard operation procedures

6.2.1 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.
-

6.2.2 Finding the structure

6.2.2.1 From open data

Sometimes the atomic positions are given with the help of Wyckoff positions: an example can be found in [\[14\]](#). 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 [14], 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.

6.2.2.2 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.

6.2.3 Insulator DFT+GW+BSE

6.2.3.1 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 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 $\mathbf{q} = 0$. 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 $0.5 \times 1 / \text{wfng_nk1}$ 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.

6.2.3.2 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`.
- 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` (Section 7.5.2).

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 (3.22), and others are not.

6.2.3.3 The BSE stage

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:

```
use_symmetries_coarse_grid
```

- 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 \mathbf{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/bsema.h5 ./
```

The `bsema` 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 `bsema.h5` is present, the program runs well.

- (c) Prepare `absorption.inp`. Do the following checklist:

- Whether the following lines are in `absorption.inp`:

```
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.

-

6.2.4 Metal DFT+GW+BSE

For metals, the \mathbf{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* \mathbf{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
```

6.2.5 Band plot

6.2.5.1 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.

6.2.5.2 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 (Section 7.8.4). 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.

6.2.5.3 GW level: using WFN_outer

It's possible to do the same thing in Section 6.2.5.1 with BerkeleyGW: we can

6.2.5.4 BSE level

6.2.6 Wannier functions and tight-binding models

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

6.2.6.1 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 (Section 7.6.1)? 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.

5. Run

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

6.2.6.2 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**.

6.2.7 Self-consistent GW

Self-consistent schemes in *GW* are listed in Section 3.3.3. In this section, I talk about how to do them.

6.2.7.1 Energy self-consistent calculation in *GW*

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

6.2.7.2 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_vxc.dat
```

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_vxc.dat
```

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_vxc.dat` option.

6.2.8 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`.

6.2.9 Band projection

6.3 Performance tricks

6.3.1 Parallelization

TODO: Do more MPI tasks already result in faster speed?

6.3.2 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.

6.3.3 pseudobands

```
wfn2hdf.x BIN WFN WFN.h5
pseudobands.py WFN.h5 WFN.h5 0.7 0.02
hdf2wfn.x BIN WFN.h5 WFN
```

Using `pseudobands` breaks the norm conserving condition. Therefore, in `epsilon.inp`, we need to add `dont_check_norms`.

TODO: how it works; maybe [8] may provide some hints.

6.4 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 \mathbf{k} - (and therefore \mathbf{q} -)grid this is defined in the 2.1-wfn step;
- The cutoff energy in DFT;
- The number of empty bands given to `epsilon`;
- The cutoff energy in `epsilon`;
- TODO: what's number in CH summation? Possibly just `number_bands`.

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

List of theoretical problems:

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

More theoretical reading is needed to understand <https://www.nersc.gov/assets/Uploads/ConvergenceinBGW.pdf>

Chapter 7

Trouble shooting

7.1 Unexpected units

7.1.1 Band energy output of `pw.x`

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

7.2 Trouble shooting in MPI

7.2.1 `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.

7.2.2 `error parsing parameters`

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

7.2.3 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 (Section [6.1.1](#)).

7.2.4 `nsufficient virtual memor`

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

7.3 Trouble shooting in Python

7.3.1 `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`.

7.4 Trouble shooting in QuantumEspresso

7.4.1 Intel MKL FATAL ERROR: Cannot load symbol MKLMPI_Get_wrappers.

7.4.2 Program frozen

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

7.4.3 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`.

7.4.4 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.

7.4.5 `cdiaghg (159): eigenvectors failed to converge`

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

7.4.6 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`.

7.4.7 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.

7.4.8 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 Section 7.4.3. A complete `bands-pw2bgw` run has to be redone.

7.4.9 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.

7.4.10 Error in routine `PW2BGW (19): input_pw2bgw`

This can occur when you misspell one option in `pw2bgw.in`.

7.4.11 `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.

7.4.12 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.

7.4.13 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.

7.5 Trouble shooting in epsilon and sigma

7.5.1 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.

7.5.2 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.

7.5.3 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 (Section 7.8.7), but forget to change the `ifmax` dataset correspondingly, this error will also occur, regardless of how you change the `eqp.dat`.

7.5.4 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.

7.5.5 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).

7.5.6 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.
```

7.5.7 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
```

7.5.8 forrtl: severe (24): end-of-file during read, unit -5, file Internal List-Directed Read

This usually happens when we should write `qpoints` but actually write `kpoints` (or the opposite).

7.5.9 ERROR: Inconsistent screening, truncation, 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 OD
truncation, i.e., spherical or box truncation.
```

```
ERROR: Inconsistent screening, truncation, 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.
```

7.5.10 cannot use metallic screening with $q=0$

As the name implies, when doing a metallic calculation, we need to make sure that the $q \rightarrow 0$ point is the smallest *non-zero* point in the dense k -grid (Section 6.2.4). 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.

7.5.11 ERROR: genwf mpi: No match for rkq point

7.5.12 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.

7.5.13 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.

7.5.14 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.

7.5.15 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 `eps0mat.h5` is broken, the error message still contains `in epsmat` instead of `in eps0mat`.

7.6 Trouble shooting in wannier90 and pw2wannier90

7.6.1 `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.

7.6.2 `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

7.6.3 `forrtl`: severe (174): SIGSEGV, segmentation fault occurred

TODO: occurs when `num_bands > num_wann`

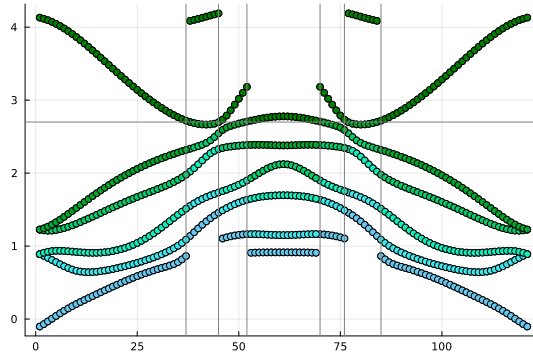


Figure 7.1: Example of how `inteqp` misidentifies DFT-level bands: the data is from the DFT column of `eqp.dat`; the colors of points are about the band index.

7.7 Trouble shooting in kernel and absorption

7.7.1 ERROR: Inconsistent symmetry treatment of the fine and shifted grids with the momentum operator

7.8 Checklist for unexpected results

Sometimes the calculation ends successfully, but the result seems strange. Below are some checklists.

7.8.1 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?

7.8.2 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

$$\text{new } z \text{ coordinate} = \frac{\text{old vacuum distance}}{\text{new vacuum distance}} \times \text{old } z \text{ coordinate.} \quad (7.1)$$

- 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.

7.8.3 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.

7.8.4 Band plot is not continuous

TODO: it's \mathbf{k} -points that are shuffled, or the band indices, or both???

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` (Figure 7.1).

Another type of band plot non-continuity is described in [6] and can be solved by COHSEX???

TODO

7.8.5 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 Section 7.5.3?

7.8.6 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:

7.8.7 When we get a semimetal in the DFT step but it should be an insulator

This is similar to Section 7.8.5.

Note that naively feeding the semimetal result into BerkeleyGW while still using the insulator procedure may result in errors in Section 7.5.4.

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 Section 7.5.3, 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

7.8.8 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 `WFn0` in one of the instances. This essentially leads to an under-converged problem: suppose we have 200 bands in `WFn0`, 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.

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] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] R Del Sole and Raffaello Girlanda. Optical properties of semiconductors within the independent-quasiparticle approximation. *Physical Review B*, 48(16):11789, 1993.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.

- [14] 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.
- [15] 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.
- [16] 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.
- [17] Michael Rohlfing and Steven G Louie. Electron-hole excitations and optical spectra from first principles. *Physical Review B*, 62(8):4927, 2000.
- [18] 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.
- [19] 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.