GW and BSE methods

Jinyuan Wu

January 5, 2023

1 Preliminaries

1.1 Diagrammatics

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

1.1.1 Infinitesimals

Note that here we need to add some convergence factors. The first is about the value of the propagator to ensure that when $t = 0$, $\mathcal{T} \langle c(t)c^\dagger(0) \rangle$ is the particle number (so that if we evaluate the tadpole diagram, we get the Hartree term), the contribution of an electron line is actually

$$
\begin{aligned}
\mathcal{T} \langle c_{\boldsymbol{k}}(t)c_{\boldsymbol{k}}^\dagger(0) \rangle &:= \mathcal{T} \langle c_{\boldsymbol{k}}(t - 0^+)c_{\boldsymbol{k}}^\dagger(0) \rangle \\
&= \int \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{-\mathrm{i}\omega(t-0^+)} \underbrace{\frac{\mathrm{i}}{\omega - \xi_{\boldsymbol{k}} + \mathrm{i}0^+ \operatorname{sgn}(\xi_{\boldsymbol{k}})}}_{\mathrm{i}G_0(\omega, \boldsymbol{k})} = \int \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{e}^{\mathrm{i}\omega 0^+} \mathrm{i}G_0(\omega, \boldsymbol{k}).
\end{aligned} \tag{1}
$$

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

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

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

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

Another mini-regularization is when necessary, for a real space interaction line – screened or unscreened – we should assume the "out-time" is the "in-time" plus 0^+, because the Coulomb interaction isn't really spontaneous and there is a small time retardation. In the frequency space, we need to assume that there is an infinite amount of energy on the interaction line,

For bare Coulomb interaction this is rarely needed, because we don't have ω dependence in the potential, and it makes no sense to discuss the poles when we change ω. It does make sense to talk about retardation in the relativistic origin of Coulomb interaction: the Coulomb interaction is mediated by virtual photons, and is therefore proportional to the off-shell (i.e. $\omega \to 0$) limit of the photon propagator, which has $\omega^2 - \boldsymbol{q}^2 + \mathrm{i}0^+$ as the denominator, and we get

$$
V(q) = \frac{4\pi e^2}{\boldsymbol{q}^2 - \omega^2 - \mathrm{i}0^+}. \tag{2}
$$

For screened Coulomb interaction, however, the correct retardation is important, because now something looking like (2) appears again.

1.1.2 The position of imaginary units

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

The self-energy correction is visualized as the follows:

$$\tag{3}$$

and from it we have

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

It's then a good idea to define

$$-\mathrm{i}\Sigma = \quad \bigcirc \quad , \tag{4}$$

because in this case, we have

$$G = G_0 + GG_0\Sigma, \tag{5}$$

and therefore

$$\underbrace{\omega - \xi_{\boldsymbol{k}}^0}_{1/G_0} = \underbrace{\omega - \xi_{\boldsymbol{k}}}_{1/G} + \Sigma, \tag{6}$$

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

Similarly, we define the corrected interaction line as

$$-\mathrm{i}W = \quad \bigg\} \quad , \tag{7}$$

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

$$W = \frac{e^2}{r}. \tag{8}$$

1.1.3 About "antiparticles"

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

Here the problem lies on what it means to be the antiparticle of a kind of particle. In particle physics, when we do the particle-antiparticle transformation to an electron state whose polarization is one of the Dirac basis, a ψ^+ particle is flipped into a ψ^- particle and vice versa. In condensed matter physics, the label of ψ^+ and ψ^- (or ϕ and χ as people often call them:

$\psi = (\phi, \chi))$ is no longer there: the χ modes in the Dirac field have already been integrated out. So electrons in condensed matter physics don't really have antiparticles in the context of high energy physics.

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

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

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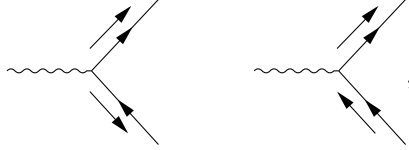
because now it's possible to annihilate a preexisting electron in the ground state, but for internal lines, momentum labels can still be directly attached to the internal lines. This can be also seen by reckoning how Feynman rules are derived: the series we obtain by expanding $\mathrm{e}^{-\mathrm{i}Ht}$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

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

$$\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t\, \mathcal{T}\, \langle c_{\boldsymbol{k}}(t)c^{\dagger}(0)\rangle = \frac{\mathrm{i}}{\omega - \epsilon_{\boldsymbol{k}} + \mu}, \tag{9}$$

which can be straightforwardly obtained by looking at

$$H = \sum_{\boldsymbol{k}} \epsilon_{\boldsymbol{k}} c_{\boldsymbol{k}}^{\dagger} c_{\boldsymbol{k}} - \mu N \tag{10}$$

without doing any calculation.

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

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

$$b_{\boldsymbol{k}} = \begin{cases} c_{\boldsymbol{k}}, & \epsilon_{\boldsymbol{k}} > \mu, \\ c_{\boldsymbol{k}}^{\dagger}, & \epsilon_{\boldsymbol{k}} < \mu, \end{cases} \tag{11}$$

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

$$\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t\, \mathcal{T}\, \langle b_{\boldsymbol{k}}(t)b^{\dagger}(0)\rangle = \frac{\mathrm{i}}{\omega - \mu + \epsilon_{\boldsymbol{k}}}, \tag{12}$$

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

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

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

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

$$\begin{aligned} E &= \sum_{\boldsymbol{k}' \neq \boldsymbol{k}, \epsilon_{\boldsymbol{k}'} < \mu} \epsilon_{\boldsymbol{k}'} - \mu(N-1) - \left(\sum_{\epsilon_{\boldsymbol{k}'} < \mu} \epsilon_{\boldsymbol{k}'} - \mu N \right) \\ &= \mu - \epsilon_{\boldsymbol{k}} > 0. \end{aligned} \tag{13}$$

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 of position, where the electron and the position 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 we can always view a part of the Fermi ball as electrons)

1.1.4 Normalization

$$\int_{-\infty}^{\infty} e^{i\omega t} \, dt \tag{14}$$

1.2 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.

1.3 Light-matter interaction

1.3.1 Interaction Hamiltonian

The non-relativistic minimal coupling is

$$H = \frac{(\boldsymbol{p} - q\boldsymbol{A})^2}{2m} = \frac{(\boldsymbol{p} + e\boldsymbol{A})^2}{2m}, \tag{15}$$

and therefore the full light-matter interaction Hamiltonian is

$$H_{\text{light-matter}} = \frac{e\boldsymbol{p} \cdot \boldsymbol{A}}{m} + \frac{e^2 \boldsymbol{A}^2}{2m}. \tag{16}$$

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

$$H_{\text{light-matter}} = \frac{e\boldsymbol{p} \cdot \boldsymbol{A}}{m}. \tag{17}$$

Note that here $\boldsymbol{p}$ is the original momentum operator, i.e. $-i\boldsymbol{\nabla}$.

The dipole approximation is

$$H_{\text{light-matter}} = -\boldsymbol{d} \cdot \boldsymbol{E}. \tag{18}$$

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

$$H_{\text{light-matter}} = q\varphi \approx \text{const} + q\boldsymbol{r} \cdot \boldsymbol{\nabla}\varphi \simeq \underbrace{q\boldsymbol{r}}_{-\boldsymbol{d}} \cdot \boldsymbol{E}. \tag{19}$$

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

Another light-matter interaction Hamiltonian is

$$H_{\text{light-matter}} = e\boldsymbol{v} \cdot \boldsymbol{A}. \tag{20}$$

The problem with this Hamiltonian is, when the potential term is not $V(\boldsymbol{r})$, we have

$$\boldsymbol{v} = \frac{1}{\mathrm{i}}[\boldsymbol{r}, H] = \frac{\boldsymbol{p}}{m} + \frac{1}{\mathrm{i}}\underbrace{[\boldsymbol{r}, V]}_{\neq 0}. \tag{21}$$

TODO: then why it works?

1.3.2 How to capture absorption

Absorption is usually modeled by the imaginary part of ϵ, and it's usually calculated by Fermi golden rule. There seems to be a contradiction here: Fermi golden rule looks "discrete", while $\operatorname{Im}\epsilon$ 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 $\operatorname{Im} H_{\text{eff}}$, 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 $\operatorname{Im}\epsilon$ is about $\operatorname{Im} H_{\text{eff}}$, while Fermi golden rule is about the quantum jump channels, so it's indeed correct to infer $\operatorname{Im}\epsilon$ from Fermi golden rule, using a procedure like, say, comparing the amount of light absorbed calculated from $\operatorname{Im}\epsilon$ and from Fermi golden rule. There is no double counting in this procedure: *both* $\operatorname{Im} H_{\text{eff}}$ 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.

2 Overview of GW

2.1 What is GW

In short, GW means
$$\Sigma = \mathrm{i}GW, \tag{22}$$
where G is the renormalized Green function, and W is the renormalized (i.e. screened) Coulomb interaction.

2.2 Discussion: what's missing in the Hartree-Fock approximation, then?

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.

2.3 Deriving formulas

TODO:

- Truncated Coulomb interaction
- Derive

$$\text{Im}\,\epsilon = \frac{16\pi^2 e^2}{\omega^2} \sum_S |\hat{\boldsymbol{e}} \cdot \langle 0|\boldsymbol{v}|S\rangle|^2 \delta(\omega - \Omega^S). \tag{23}$$

- Why we say $\Sigma = V_{\text{xc}} + \Sigma - V_{\text{xc}}$? (see [3] p. 1271)
-

3 Accuracy of GW

3.1 Coverage

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.

3.2 Diagonal or not

We know in the momentum space, we have

$$E_{n\boldsymbol{k}}^{\text{QP}} = E_{n\boldsymbol{k}}^0 + \Sigma_{n\boldsymbol{k}}(E^{\text{QP}}). \tag{24}$$

Here since Σ depends on the corrected propagator, $E_{n\boldsymbol{k}}^{\text{QP}}$ 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 (24) now is

$$E_{\boldsymbol{k}}^{\text{QP}} = E_{\boldsymbol{k}}^{\text{KS}} + \Sigma_{\boldsymbol{k}}(E^{\text{QP}}) - \Sigma^{\text{KS}}. \tag{25}$$

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 $\Sigma_{\boldsymbol{k}}$ may mix different bands together, and (25) is an equation about matrices, essentially a single-electron Schrodinger equation. It's first order approximation is

$$E_{n\boldsymbol{k}}^{\text{QP}} = E_{n\boldsymbol{k}}^{\text{KS}} + \left\langle \psi_{n\boldsymbol{k}}^{\text{KS}} \middle| \Sigma(E_{n\boldsymbol{k}}^{\text{QP}}) - \Sigma^{\text{KS}} \middle| \psi_{n\boldsymbol{k}}^{\text{KS}} \right\rangle. \tag{26}$$

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 (26). In this case, we don't need off-diagonal elements, because they are not used in (26). This scheme is mentioned in Section 3.3 in [3]. The second scheme takes the change of eigenvalues into account, and thus iteratively solves (25). In this case we need to take non-diagonal elements seriously [1,4]. 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 $\Sigma = \text{i}GW$. 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.

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 [5]

3.4 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. [6] 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

4 The QuantumESPRESSO-BerkeleyGW ecosystem

4.1 Overview of the pipeline

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 Input and output of `pw`

4.3 `epsilon`

4.3.1 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 5.2).

4.3.2 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 [3]. There are three steps in `epsilon`:

1. Calculate $M_{nn'}(\boldsymbol{k}, \boldsymbol{q}, \boldsymbol{G})$;
2. Summing over $n, n', \boldsymbol{k}$;
3. Finding ϵ^{-1}.

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

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

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

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

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

4.3.3 Divergence problems

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

$$
\begin{aligned}
\epsilon_{\boldsymbol{00}}(\boldsymbol{q} \to \boldsymbol{0}, \omega = 0) &\propto -\frac{1}{\boldsymbol{q}^2} \chi_{\boldsymbol{00}}(\boldsymbol{q} \to \boldsymbol{0}, \omega = 0) \\
&\propto \sum_{n \text{ occupied}, n' \text{ empty}} -\frac{1}{\boldsymbol{q}^2} |\langle n\boldsymbol{k}|1 + \mathrm{i}(\boldsymbol{q}+\boldsymbol{G}) \cdot \boldsymbol{r} + \cdots |n'\boldsymbol{k}\rangle|^2 \qquad (27) \\
&\propto \text{const.} \times \frac{\boldsymbol{q}^2}{\boldsymbol{q}^2}.
\end{aligned}
$$

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

For a metal, some bands are both occupied and empty, so we can no longer use the orthogonality conditions, and $\epsilon_{\boldsymbol{00}}(\boldsymbol{q} \to \boldsymbol{0}, \omega = 0)$ scales like const./$\boldsymbol{q}^2$.

4.4 Systems of units

TODO

5 Standard operation procedures

5.1 Insulator

5.1.1 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 $\boldsymbol{k}$-grid with size `nx ny nz`. This can't be done with options in QuantumESPRESSO's KPOINTS section because QuantumESPRESSO and BerkeleyGW have different tolerance for symmetry.

   (c) Run

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

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

(d) Preparing the `bands.in` file, which is the input file of `pw.x`. Do the following checklist:
  - Whether `calculation` is `bands`.
  - Whether `pseudo_dir` is correct.
  - Whether `nbnd` is set to, say, `1000`.
  - Whether `lspinorb = .true.` and `noncolin = .true.` are set for an SOC run.
(e) Run `pw2bgw.x` in `2.1-wfn`. Do the following checklist:
  - This step should be done with *exactly the same* parallelization setting with `pw.x`.
  - The `wfng_nk1`, `wfng_nk2`, `wfng_nk3` parameters should be set to `nx`, `ny`, `nz` mentioned above. (This item needs double check especially when `pw2bgw.inp` comes from another run.)
  - Whether `rhog_flag` is `.true.`.
  - Whether `vxc_flag` is `.true.`.
  - Whether `wfng_flag` is `.true.`.

3. Do a `bands` calculation in `2.2-wfnq`. The steps are similar to `2.1-wfn`:
   (a) Linking files from `1-scf`.
   (b) Run

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

   to get the `kgrid.inp` file. Here `qx qy qz` is a small displacement used to regularize Coulomb interaction at $q = 0$. A common choice is `0 0 0.001`; when dealing with a 2D material, choose `0 0.001 0`, because with `cell_slab_truncation` open in the `epsilon.x` step, non-zero z components of k-points are forbidden.
   (c) Run `kgrid.x`.
   (d) Preparing the `bands.in` file.
   (e) Run `pw2bgw.x`. Do the following checklist:
     - This step should be done with *exactly the same* parallelization setting with `pw.x`.
     - The `wfng_nk1`, `wfng_nk2`, `wfng_nk3` parameters should be set to `nx`, `ny`, `nz` mentioned above.
     - The `wfng_dk1`, `wfng_dk2`, `wfng_dk3` parameters should be set to `wfng_nk1` × `qx`, etc. (If `kshift` is used, it also should be added to `wfng_dk1`, etc.)
       Note that this doesn't mean the displacement imposed to the k-grid is `wfng_nk1` × `qx`: the displacement is still `qx qy qz`. Here the `wfng_dk1`, `wfng_dk2`, `wfng_dk3` are conventional parameters used in Monkhorst-Pack grids, and `wfng_dk1 = 0.5` means the grid is shifted towards x direction by half a *grid step* – and therefore the displacement is 0.5×1 / `wfng_nk1` in the crystal coordinates. Now we understand why we need to set `wfng_dk1` to `wfng_nk1` × `qx`. Indeed, below is a part of the header of a `WFN` file in `2.2-wfnq`:

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

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

5.1.2 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 8.1).

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 (24), and others are not.

5.1.3 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:

2. Do a `absorption` step.

(a) Prepare `absorption.inp`. Do the following checklist:
- Whether the following lines are in `absorption.inp`:

```
use_symmetries_fine_grid
use_symmetries_coarse_grid
```

5.2 Metal

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

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

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

3. Do `2.2-wfn0` with a *finer* $\boldsymbol{k}$-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 $\boldsymbol{k}$-point with smallest length in the $\boldsymbol{k}$-grid used in `2.2-wfn0`.
   - Whether the Γ point is in the format

   ```
   qx qy qz 1 2
   ```

5.3 Band plot

5.3.1 DFT level

The Fermi energy is to be found in `scf.out`. It won't appear in `bands.out`.

5.3.2 *GW* level

The *GW* level bands can be obtained by the `inteqp` program. 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 $\boldsymbol{k}$-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.

5.3.3 BSE level

6 Performance tricks

6.1 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.

6.2 `pseudobands`

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

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

TODO: how it works; maybe [2] may provide some hints.

7 Trouble shooting in QuantumEspresso

7.1 Program frozen

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

7.2 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.

7.3 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 cdiaghg (159): eigenvectors failed to converge

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

7.5 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`.

7.6 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 same error occurs..

7.7 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 7.2. A complete `bands-pw2bgw` run has to be redone.

7.8 Error in routine PW2BGW(19):input pw2bow

Usually this occurs when something else happens between a `bands` run and a `pw2bgw` run for it. A complete `bands-pw2bgw` run has to be redone.

8 Trouble shooting in `epsilon` and `sigma`

8.1 Selected number of bands breaks degenerate subspace.

Run `degeneracy_check.x WFN` to see degeneracy-allowed number of bands. This error occurs when one band in a degeneracy subspace is considered but others are not. Also, the `band_-index_min` and `band_index_max` parameters shouldn't be too close to `vxc_diag_min` and `vxc_-diag_max`, or the error occurs.

8.2 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.
```

This occurs 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.

8.3 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 the smearing type is `code` or `gaussian`; changing the smearing type to, say, `fermi-dirac` may solve the problem.

8.4 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).

8.5 ERROR: occupations (ifmax field) inconsistent between WFN and WFNq files.

```
ERROR: occupations (ifmax) inconsistent between WFN and WFNq files.
Remember that you should NOT use WFNq for metals and graphene.
```

8.6 `ERROR: Unexpected characters were found while reading the value for the keyword`

This usually happens when the input file contains a line like

```
number_bands = 148
```

while the correct format is

```
number_bands    148
```

8.7 `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).

8.8 `ERROR: Inconsistent screening, trunction, or q0 vector`

The full error message may be

```
ERROR: the input truncation flag indicates that the Coloumb interaction v(q0)
diverges for q0->0. However, you have q0 exactly zero.
You should always specify a *nonzero* q0->0 vector unless you have 0D
truncation, i.e., spherical or box truncation.


ERROR: Inconsistent screening, trunction, or q0 vector
```

This arises when we are dealing with a metal but forget to add `screening_metal` to `sigma.inp`.

```
ERROR: cannot use metallic screening with q0 = 0.
You should either specify a nonzero q0->0 vector or use another screening
    ↪ flag.
```

8.9 `cannot use metallic screening with q=0`

8.10 `ERROR: genwf mpi:  No match for rkq point`

9 Checklist for unexpected results

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

9.1 Band symmetry higher than the space group shown at the beginning of `bands.out`

- Usually this is because of an approximate symmetry, which is ignored by QuantumE-SPRESSO because its tolerance is very low.
- Are all steps in the DFT calculation using the same crystal structure?

9.2 Band structure looks very far from the literature

- Check the crystal structure: if it comes from relaxation, does it converges?
- For 2D materials, when we change the vacuum distance and use crystal coordinates for atomic positions at the same time, always double check whether we scale the atomic positions correctly. The formula is

$$\text{new } z \text{ coordinate} = \frac{\text{old vacuum distance}}{\text{new vacuum distance}} \times \text{old } z \text{ coordinate}. \tag{28}$$

- 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.

14

9.3 Band plot is empty

- Are there enough bands? If `nbnd` is not set for an insulator, no conduction band will be considered.
- Is the Fermi energy correct? If the Fermi energy is set too high (which may come from, say, wrong unit), then naturally there is no band in the plot.

9.4 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 8.2?

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

This is similar to Section 9.4.

Note that naively feeding the semimetal result into BerkeleyGW may result in errors in Section 8.2 and Section 8.3. TODO: what to do then

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.