# $GW$ and BSE methods

Jinyuan Wu

December 27, 2022

## 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 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

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

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 $\omega$ dependence in the potential, and it makes no sense to discuss the poles when we change $\omega$. 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

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

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

# 2 Overview of *GW*

## 2.1 What is *GW*

## 2.2 Deriving formulas

## 2.3 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.

# 3 Accuracy of *GW*

## 3.1 Diagonal or not

We know in the momentum space, we have

$$E^{\mathrm{QP}}_{n\boldsymbol{k}} = E^{0}_{n\boldsymbol{k}} + \Sigma_{n\boldsymbol{k}}(E^{\mathrm{QP}}). \tag{3}$$

Here since $\Sigma$ depends on the corrected propagator, $E^{\mathrm{QP}}_{n\boldsymbol{k}}$ 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 DFT isn't a mean-field approach – it just looks like a mean-field one), so (3) now is

$$E^{\mathrm{QP}}_{\boldsymbol{k}} = E^{\mathrm{KS}}_{\boldsymbol{k}} + \Sigma_{\boldsymbol{k}}(E^{\mathrm{QP}}) - \Sigma^{\mathrm{KS}}. \tag{4}$$

Here $\Sigma^{\mathrm{KS}}$ 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 (4) is an equation about matrices, essentially a single-electron Schrodinger equation. It's first order approximation is

$$E^{\mathrm{QP}}_{n\boldsymbol{k}} = E^{\mathrm{KS}}_{n\boldsymbol{k}} + \left\langle \psi^{\mathrm{KS}}_{n\boldsymbol{k}} \middle| \Sigma(E^{\mathrm{QP}}_{n\boldsymbol{k}}) - \Sigma^{\mathrm{KS}} \middle| \psi^{\mathrm{KS}}_{n\boldsymbol{k}} \right\rangle. \tag{5}$$

## 3.2 Self-consistent or not

There are three iterative schemes. The first is the eigenvalue self-consistent scheme: It's just a self-consistent solver of (5). In this case, we don't need off-diagonal elements, because they are not used in (5). This scheme is mentioned in Section 3.3 in [2]. The second scheme takes the change of eigenvalues into account, and thus iteratively solves (4). In this case we need to take non-diagonal elements seriously [1, 3]. In the third scheme, the form of $\Sigma$ itself is changed: Recall that we need an `epsilon` step to calculate $\epsilon$ and thus the screened interaction potential $W$, and $\Sigma = \mathrm{i}GW$. 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 [4]

### 3.3 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 $G^0W^0$. [5] 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 work flow

### 4.1 Input and output of `pw`

### 4.2 Input and output of `epsilon`

## 5 Standard operation procedures

### 5.1 Insulator

### 5.2 When we get a semimetal in the DFT step but it should be an insulator ...

Naively feeding the semimetal result into BerkeleyGW may result in errors in Section 8.3 and Section 8.4.

## 6 Performance tricks

### 6.1 Choosing cutoff energies wisely

### 6.2 `pseudobands`

## 7 Trouble shooting in QuantumEspresso

### 7.1 `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.

### 7.2 `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.

### 7.3 `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 Error in routine cdiaghg (1052): problems computing cholesky

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

## 8 Trouble shooting in `epsilon` and `sigma`

### 8.1 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.

### 8.2 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.

### 8.3 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.
```

This occurs when BerkeleyGW expects an insulator but gets a metal. TODO: so what to do???

### 8.4 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.

### 8.5 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).

### 8.6 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.
```

## 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 QuantumESPRESSO because its tolerance is very low.

### 9.2 Band structure looks strange

- 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

$$\text{new } z \text{ coordinate} = \frac{\text{old vacuum distance}}{\text{new vacuum distance}} \times \text{old } z \text{ coordinate}. \tag{6}$$

## 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] 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.

[3] 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.

[4] 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.

[5] 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.