More on the plane-wave pseudopotential method: densities and magnetization, metals ans insulators, periodic and non-periodic boundary conditions

P. Giannozzi
Università di Udine and CNR-IOM Democritos, Trieste, Italy

MaX School on Advanced Materials and Molecular Modelling
with Quantum ESPRESSO
May 17-28 2021, ICTP Virtual Meeting

DEMOCRITOS
DEmocritos **MO**deling **C**enter for **R**esearch **I**n a**TO**mistic **S**imulation
CNR-IOM
Istituto **O**fficina dei **M**ateriali

MAX
DRIVING
THE EXASCALE
TRANSITION

Outline

1. About FFT and FFT grids
2. Charge density: symmetrization, augmentation charge
3. Metals: smearing technique
4. Magnetism: Collinear and non-collinear magnetism
5. Systems with reduced periodicity, corrections for periodicity

Fourier Transform

For a *periodic* $f(x)$ function of period L in one dimension:

$$f(x) = \sum_n \widetilde{f}(q_n) e^{iq_n x}, \qquad \widetilde{f}(q_n) = \frac{1}{L} \int f(x) e^{-iq_n x} dx$$

where the q_n:

$$q_n = n\frac{2\pi}{L}, \qquad -\infty < n < \infty$$

are an infinite set of *discrete* values of q.

Discrete Fourier Transform

In the Discrete Fourier Transform, we consider a *finite* set of discrete values for both q and x:

$$\begin{aligned} f(x) &\rightarrow f_m = f(x_m), \qquad x_m = m\frac{L}{N}, \qquad m = 0, .., N-1 \\ \widetilde{f}(q) &\rightarrow \widetilde{f}_n = \widetilde{f}(q_n), \qquad q_n = n\frac{2\pi}{L}, \qquad n = 0, .., N-1 \end{aligned}$$

The *Discrete Fourier Transform* can be written as

$$\begin{aligned} f_m &= \sum_{n=0}^{N-1} \widetilde{f}_n e^{i(2\pi nm/N)} \qquad (x - \text{space}) \\ \widetilde{f}_n &= \frac{1}{N}\sum_{m=0}^{N-1} f_m e^{-i(2\pi nm/N)} \qquad (q - \text{space}) \end{aligned}$$

Note that both q and x are now periodic! The q components of negative value refold into those at the "other side of the box".

Discrete Fourier Transform in 3D

Consider a set of reciprocal-lattice vectors $\mathbf{G}$, centered around the origin:

$$\mathbf{G} = n'_1\mathbf{G}_1 + n'_2\mathbf{G}_2 + n'_3\mathbf{G}_3$$

where $\mathbf{G}_1, \mathbf{G}_2, \mathbf{G}_3$ are the primitive translations that generate the reciprocal lattice;
$n_1 = 0, .., N_1 - 1$, $n'_1 = n_1$ if $n_1 \le N_1/2$, $n'_1 = n_1 - N_1$ otherwise;
same for $n_2 = 0, .., N_2 - 1$ and $n_3 = 0, .., N_3 - 1$.

N_1, N_2, N_3 are the *FFT dimensions* and must be big enough to accommodate all needed Fourier components.

Note that this $\mathbf{G}-$space grid is by construction periodic!

In real space, we consider a grid that spans the unit cell:

$$\mathbf{r} = m_1\frac{\mathbf{R}_1}{N_1} + m_2\frac{\mathbf{R}_2}{N_2} + m_3\frac{\mathbf{R}_3}{N_3}$$

with $m_1 = 0, .., N_1 - 1$, $m_2 = 0, .., N_2 - 1$, $m_3 = 0, .., N_3 - 1$. $\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3$ are the primitive translations that generate the Bravais lattice

Discrete Fourier Transform in 3D (2)

Original Fourier transform:

$$f(\mathbf{r}) = \sum_{\mathbf{G}} \widetilde{f}(\mathbf{G}) e^{i\mathbf{G}\cdot\mathbf{r}} \to f(m_1, m_2, m_3)$$

$$\widetilde{f}(\mathbf{G}) = \frac{1}{\Omega} \int f(\mathbf{r}) e^{-i\mathbf{G}\cdot\mathbf{r}} d\mathbf{r} \to \widetilde{f}(n_1, n_2, n_3)$$

Discretized Fourier Transform:

$$f(m_1, m_2, m_3) = \sum_{n_1, n_2, n_3} \widetilde{f}(n_1, n_2, n_3) e^{i(2\pi n_1 m_1/N_1)} e^{i(2\pi n_2 m_2/N_2)} e^{i(2\pi n_3 m_3/N_3)}$$

$$\widetilde{f}(n_1, n_2, n_3) = \frac{1}{N} \sum_{m_1, m_2, m_3} \widetilde{f}(m_1, m_2, m_3) e^{-i(2\pi n_1 m_1/N_1)} e^{-i(2\pi n_2 m_2/N_2)} e^{-i(2\pi n_3 m_3/N_3)}$$

where $N = N_1 N_2 N_3$. Remember that $\mathbf{G}_i \cdot \mathbf{R}_j = 2\pi\delta_{ij}$.

PW-PP calculations and Discrete Fourier Transform

$$\psi_i(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{G}} c_{\mathbf{k}+\mathbf{G}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \qquad \frac{\hbar^2}{2m}|\mathbf{k}+\mathbf{G}|^2 \le E_{cut}$$

Which grid in **G**-space? We Need to calculate the charge density. In principle (but not in practice):

$$n(\mathbf{G}') = \sum_{\mathbf{G}}\sum_{i,\mathbf{k}} f_{i,\mathbf{k}} c^*_{i,\mathbf{k}+\mathbf{G}} c_{i,\mathbf{k}+\mathbf{G}+\mathbf{G}'}$$

Fourier components $\mathbf{G}'$ with max$(|\mathbf{G}'|) = 2$max$(|\mathbf{G}|)$ appear.
Another case: we need the product of the potential time a wavefunction:

$$(V\psi)(\mathbf{G}) = \sum_{\mathbf{G}'} V(\mathbf{G}-\mathbf{G}')c_{i,\mathbf{k}+\mathbf{G}'}$$

Again, max$(|\mathbf{G}-\mathbf{G}'|) = 2max(|\mathbf{G}|)$. We need a kinetic energy cutoff for the Fourier components of the charge and potentials that is four time larger as the cutoff for the PW basis set:

$$\frac{\hbar^2}{2m}|\mathbf{G}|^2 \le 4E_{cut}$$

In practice such condition may occasionally be relaxed.

Fourier Transform grid

The Fourier Transform grid is thus

$$\mathbf{G} = n'_1\mathbf{G}_1 + n'_2\mathbf{G}_2 + n'_3\mathbf{G}_3$$

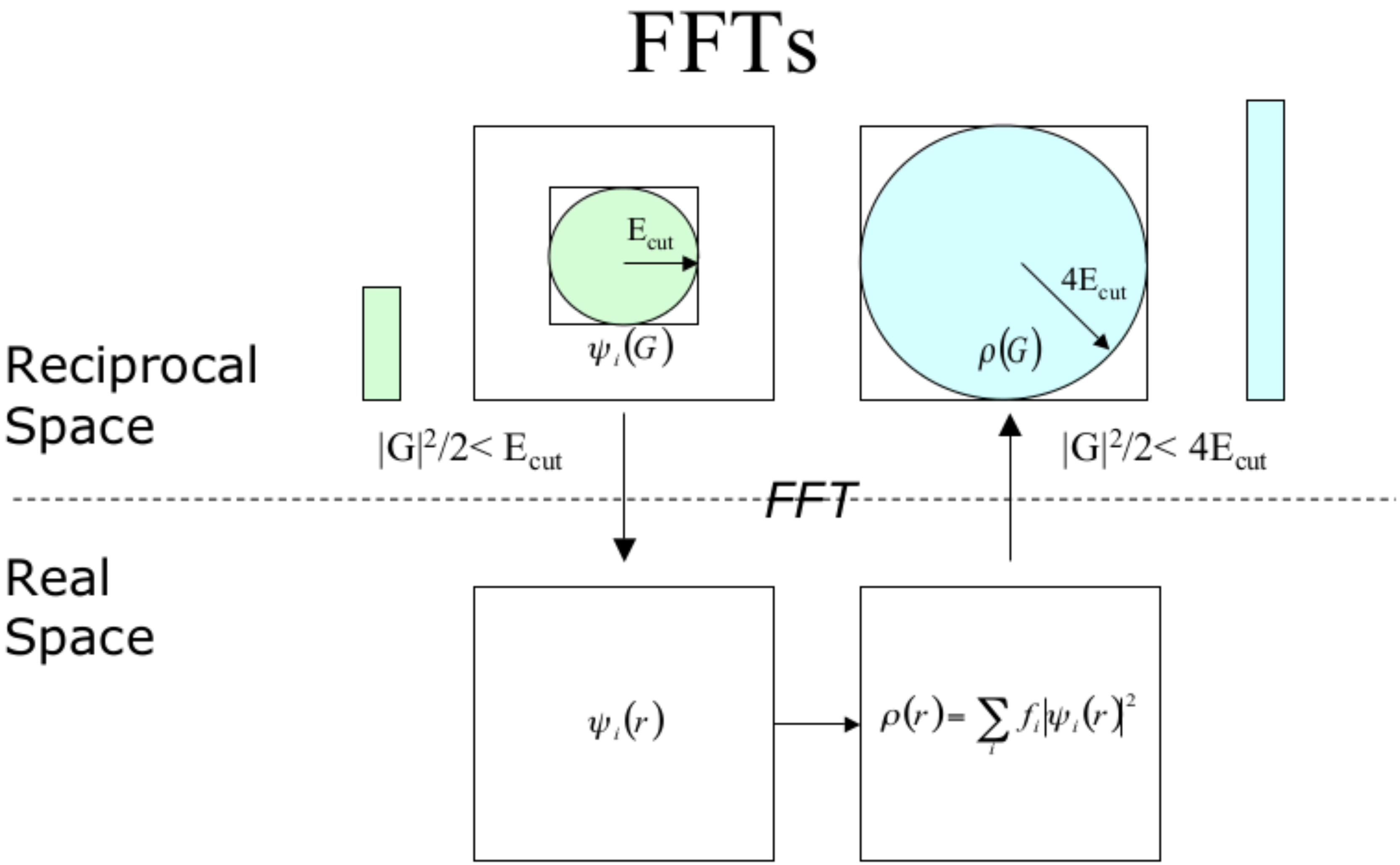
with $n_1 = 0, .., N_1 - 1$, $n_2 = 0, .., N_2 - 1$, $n_3 = 0, .., N_3 - 1$. This grid must be big enough to include all $\mathbf{G}-$vectors up to a cutoff

$$\frac{\hbar^2}{2m}|\mathbf{G}|^2 \leq 4E_{cut}$$

and NOT up to the cutoff of the PW basis set! In general, the grid will also contain “useless” Fourier components (beyond the above-mentioned cutoff, so that $n(\mathbf{G}) = 0, V(\mathbf{G}) = 0$ etc.)

Important: for *ultrasoft* pseudopotentials, a further FT grid may be needed

FFT grid



(Note: $G^2/2$ is the kinetic energy in Hartree atomic units)

Fast Fourier Transform

Computational cost of a conventional Fourier Transform of order n: $T_{CPU} = \mathcal{O}(n^2)$.

Computational cost of a Fast Fourier Transform of order n: $T_{CPU} = \mathcal{O}(n \log n)$.

Difference: enormous in practical applications.

Advantages of the use of FFT in PW-PP calculations: enormous, especially in conjunction with iterative techniques and of the “dual-space” technique

Dual space technique

Basic ingredients: evaluation of products $\phi_i = (H - \epsilon_i)\psi_i$ on trial wavefunctions ψ_i. Same ingredient used in direct minimization, Car-Parrinello, etc.

$$H\psi \equiv (T + \hat{V}_{NL} + V_{loc} + V_H + V_{xc})\psi$$

$(T\psi)$: easy in **G**-space, $T_{CPU} = \mathcal{O}(N)$

$(V_{loc} + V_H + V_{xc})\psi$: easy in **r**-space, $T_{CPU} = \mathcal{O}(N)$

$(\hat{V}_{NL}\psi)$: easy in **G**-space (also in **r**-space) if $\hat{V}$ is written in separable form
$T_{CPU} = \mathcal{O}(mN)$, $m =$number of projectors

FFT is used to jump from real to reciprocal space. Operations are performed where it is easier.

The same technique is used to calculate the charge density from Kohn-Sham orbitals, the exchange-correlation GGA potential from the charge density, etc.: in all cases, we move to the more convenient space to perform the required operations.

Charge density: symmetrization

The code computes an *unsymmetrized* charge density $\tilde{n}(\mathbf{r})$ as:

$$\tilde{n}(\mathbf{r}) = \sum_{\mathbf{k}\in IBZ} \sum_{v} w_{\mathbf{k}} |\psi_{\mathbf{k},v}(\mathbf{r})|^2,$$

where the sum runs on occupied states only and on symmetry-inequivalent k-points only, that belong to the Irreducible Brillouin Zone (IBZ). The symmetry weight $w_{\mathbf{k}}$ is the number of k-points in the *star* of $\mathbf{k}$, that is: equivalent by symmetry to $\mathbf{k}$.

Please note: k-point grids in the IBZ and weights can be either automatically computed or supplied as a list. In the latter case *they must be valid for the symmetry group of the Bravais lattice, not of the crystal*.

The true charge density $n(\mathbf{r})$ is then obtained via *symmetrization*:

$$n(\mathbf{r}) = \frac{1}{N_s} \sum_{n=1}^{N_s} \hat{O}_n \tilde{n}(\mathbf{r}) = \frac{1}{N_s} \sum_{n=1}^{N_s} \tilde{n}(O_n^{-1}\mathbf{r})$$

where $\hat{O}_n$ is the n-th symmetry of the crystal, O_n is the corresponding rotation matrix.

Charge density: symmetrization II

For *non-symmorphic* groups, whose symmetry operations may contain *fractional translations* as well, one generalizes symmetrization as follows:

$$n(\mathbf{r}) = \frac{1}{N_s}\sum_{n=1}^{N_s} \tilde{n}(O_n^{-1}\mathbf{r} - \mathbf{f}_n)$$

where $\mathbf{f}_n$ is the fractional (with respect to lattice vectors) translation for operation n

In practice, the symmetrization in `pw.x` (but not in other codes) is performed in reciprocal space, not in real space:

$$n(\mathbf{G}) = \frac{1}{N_s}\sum_{n=1}^{N_s} \hat{O}_n\tilde{n}(\mathbf{G}) = \frac{1}{N_s}\sum_{n=1}^{N_s} \tilde{n}(O_n^{-1}\mathbf{G})e^{-iO_n^{-1}\mathbf{G}\cdot\mathbf{f}_n}.$$

Please note: the charge density as defined above is adimensional and integrates to the number of electrons in the system, N_{elec}

Charge density: augmentation charge

For USPP (ultrasoft pseudopotentials) and PAW, an *augmentation* term is present:

$$n(\mathbf{r}) = \sum_i \left(|\psi_i(\mathbf{r})|^2 + \sum_{l,m,\mu} \langle\psi_i|\beta_l\rangle Q_{lm,\mu}(\mathbf{r}) \langle\beta_m|\psi_i\rangle \right), \qquad \sum_i \equiv \sum_{\mathbf{k},v},$$

where μ runs over atoms, β_l and Q_{lm} are short-range functions centered on atom μ.

The first term – the square of $|\psi_i|$ – can be easily computed on the same FFT grid used in the calculation of $H\psi$, containing in G-space Fourier components up to a kinetic energy cutoff $E_s = 4E_{cut}$.

Double Grid technique: the augmentation term is computed on a larger (denser in real space) grid containing $\mathbf{G}$ up to a cutoff $E_d > 4E_w$, typically $E_d = 8 \div 12 E_w$.

Real-space grids are not necessarily commensurate, so interpolation from one grid to the other happens in reciprocal space only

Charge density: metals

The most common way to deal with metallic systems, is to introduce a *broadening* of discrete levels. The charge density becomes

$$n(\mathbf{r}) = \sum_i f_i |\psi_i(\mathbf{r})|^2,$$

where the fractional weights (or occupancies) $0 \leq f_i \leq 1$ are given by

$$f_i = \int_{\epsilon<\epsilon_F} \delta\left(\epsilon - \epsilon_i\right) d\epsilon \qquad \sum_i f_i = N(\epsilon_F)$$

and the Fermi energy ϵ_F is determined by the condition $N(\epsilon_F) = N_{elec}$, number of electrons in the system. $\delta(x)$ is the (normalized) broadening function. In principle one might use the Fermi-Dirac function:

$$\delta_{FD}(x) = \frac{1}{e^{x/k_BT} + 1}$$

but it is not convenient: one needs very high T and the distribution has long tails.

Metals: energy

Important notice: the correct way to compute the sum-of-eigenvalue contribution E_{KS} to the DFT energy is

$$E_{KS} = \sum_i \int_{\epsilon<\epsilon_F} \epsilon\delta(\epsilon - \epsilon_i)d\epsilon = \widetilde{E}_{KS} + \delta E_{met}$$

where

$$\widetilde{E}_{KS} = \sum_i \epsilon_i \int_{\epsilon<\epsilon_F} \delta(\epsilon - \epsilon_i)d\epsilon = \sum_i f_i\epsilon_i, \qquad \delta E_{met} = \sum_i \int_{\epsilon<\epsilon_F} (\epsilon - \epsilon_i)\delta(\epsilon - \epsilon_i)d\epsilon$$

and the f_i are the occupancies as previously defined.

The DFT energy computed using E_{KS} is the "variational" one, that is: it is minimized by the ground-state charge density, and forces are its derivatives.

In the output of `pw.x`, δE_{met} is printed (in Ry) as "`smearing contrib.  (-TS)`"

Metals: broadening functions

A simple choice for the broadening function is a normalized Gaussian:

$$\delta(x) = \frac{1}{\sigma\sqrt{\pi}} e^{-x^2/\sigma^2}, \qquad \int \delta(x) = 1.$$

The parameter σ (`degauss`, in Ry) plays the role of a fictitious "temperature" $T = \sigma/k_B$, to which a fictitious "free energy" functional $E_{KS}(\sigma)$ corresponds.

One can show that $E_{KS}(\sigma) \propto \sigma^2$, so large values of σ needed for fast k-point convergence yield a large defference between the "free" and true energy.

Methfessel-Paxton (MP) and Marzari-Vanderbilt (MV) "cold" smearing are devised to make the quadratic σ term vanish so that $E_{KS}(\sigma) \propto \sigma^4$, with much better convergence (see `http://theossrv1.epfl.ch/Main/ElectronicTemperature`)

Note however that:
– With MP, the occupancies f_i may be negative or larger than 1;
– With MV, the occupancies f_i are non-negative but may be larger than 1;
– In both cases, $N(\epsilon)$ is not guaranteed to be a monotonic function;
– As a consequence, the Fermi energy may not be uniquely defined.

Charge density: magnetic case

For the unpolarized (non-magnetic) case, each orbital is doubly occupied and the charge density n is simply

$$n(\mathbf{r}) = 2\sum_i f_i |\psi_i(\mathbf{r})|^2.$$

For the "LSDA" case (a misnomer!), each orbital has either spin up or spin down. We have a charge density n and a (scalar) magnetization m:

$$n(\mathbf{r}) = n_+(\mathbf{r}) + n_-(\mathbf{r}), \qquad m(\mathbf{r}) = n_+(\mathbf{r}) - n_-(\mathbf{r}),$$

where

$$n_\sigma(\mathbf{r}) = \sum_i f_{i,\sigma} |\psi_{i,\sigma}(\mathbf{r})|^2, \qquad \sigma = +, -$$

Note: in `pw.x`, the σ index for orbitals is hidden into the k-point index: k-points are doubled, the first set corresponds to up spin, the second set to down spin

Charge density: non-collinear magnetic case

For the general case on *non-collinear* magnetization, the plane-wave basis set is composed of up and down spinors and the orbitals are spinors:

$$\Psi(\mathbf{r}) = \sum_{n,\sigma} c_{n,\sigma} \frac{e^{i(\mathbf{k}+\mathbf{G}_n)\cdot\mathbf{r}}}{V} \chi_\sigma \equiv \left(\begin{array}{c} \psi_+ \\ \psi_- \end{array} \right), \quad \chi_+ = \left(\begin{array}{c} 1 \\ 0 \end{array} \right), \quad \chi_- = \left(\begin{array}{c} 0 \\ 1 \end{array} \right)$$

In addition to the total charge density n:

$$n(\mathbf{r}) = \sum_i f_i \left(|\psi_{i,+}(\mathbf{r})|^2 + |\psi_{i,-}(\mathbf{r})|^2 \right),$$

there is a (vector) magnetization $\vec{m}$:

$$\vec{m}(\mathbf{r}) = \sum_i f_i \Psi_i^\dagger(\mathbf{r}) \vec{\sigma} \Psi_i(\mathbf{r})$$

where $\vec{\sigma} = (\sigma_x, \sigma_y, \sigma_z)$ and $\sigma_x, \sigma_y, \sigma_z$ are the usual Pauli matrices.

Potentials

The following long-range potential terms are separately divergent for $\mathbf{G} = 0$:

- V_H: electrostatic potentials generated by the electrons
- V_{loc}: local part of the electron-(pseudo)ion interaction

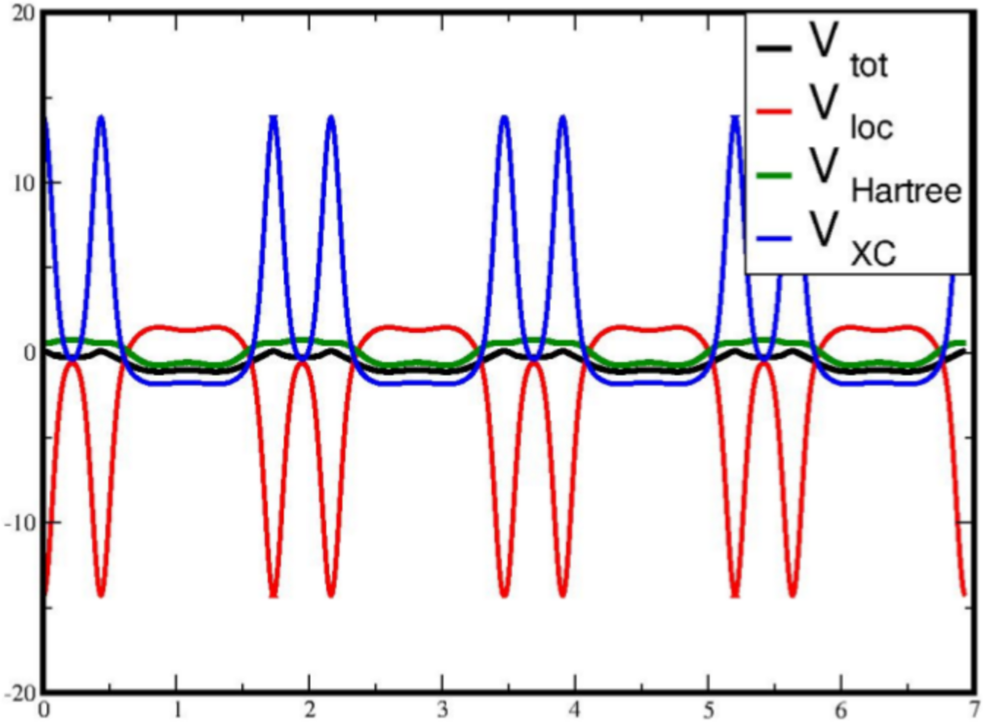
For neutral systems, $(V_H + V_{loc})(\mathbf{G} = 0)$ does *not* diverge.

The remaining potential terms are short range:

- V_{xc}: exchange-correlation potential
- V_{NL}: nonlocal part of the electron-(pseudo)ion interaction

Note: "potentials" are actually multiplied by e and have energy dimensions

An example: potentials in Si



potentials along <111> direction in bulk Si

Energies

The total energy can be written as the sum of the following terms:

$$E = E_{KS} - E_H[n(\mathbf{r})] - \int n(\mathbf{r})V_{xc}(\mathbf{r})d\mathbf{r} + E_{xc}[n(\mathbf{r})] + E_{Ewald}$$

where:

- E_{KS}: sum of one-electron KS energies (as defined earlier), contains electron-ion energy E_{eI}
- E_H: electrostatic electronic energy, with divergent $\mathbf{G} = 0$ contribution removed. This term is present and counted twice in E_{KS} and must be subtracted out.
- The third term is an exchange-correlation energy term to be subtracted out from E_{KS} and to be replaced by E_{xc}: true exchange-correlation energy.
- E_{Ewald}: The Ewald energy is the ion-ion interaction energy (E_{II}) in the presence of a neutralizing background

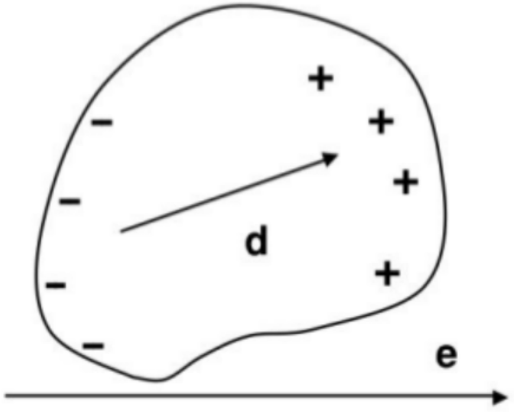
The divergent $\mathbf{G} = 0$ terms in E_H, E_{II}, and E_{eI} cancel out in neutral systems.

Electrostatics in Periodic Boundary Conditions

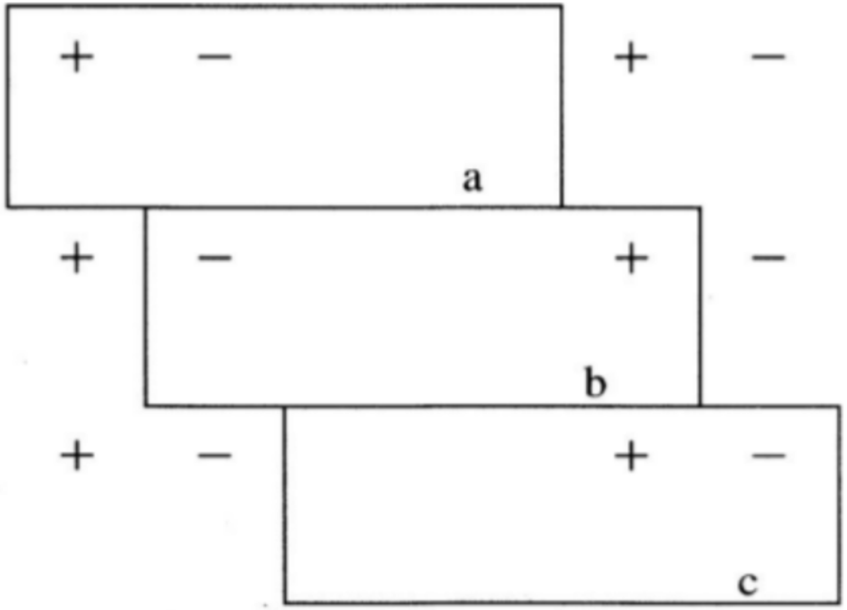
Consequences of periodicity:

- The net charge per unit cell is zero, or else the energy diverges
- All potentials must be lattice-periodic: a macroscopic electric field, which is described by a potential $V(\mathbf{r}) = -e\mathbf{E} \cdot \mathbf{r}$, can not be described in PBC
- The zero of the energy is arbitrary and has no direct relation with the vacuum level "outside" the crystal (there is no "outside")
- The dipole moment per unit cell is in general ill-defined

Dipole moments



Picture from: Electronic Structure: Basic Theory and Practical Methods
R.M. Martin
Cambridge University Press (2008)



Picture from: G. Makov, M.C. Payne PRB **51,** 4014 (1995)

Finite system: $\mathbf{d} = \int n(\mathbf{r})\mathbf{r}d\mathbf{r}$

Periodic system: $\mathbf{d} = ???$

Charged systems

To deal with charged systems in PBC, one treats the $\mathbf{G} = 0$ divergences *as for the neutral system*. Note that:

- The energy is finite but depends upon the specific choice of the $V(\mathbf{G} = 0)$ potential
- Comparison of energies with different N_{elec} for the same system is tricky
- There is no guarantee that structural optimization of charged systems gives reliable results (but no evidence it doesn't either)

Finite systems

For finite systems (e.g. molecules) with PBC, several tricks are available to get rid of spurious interactions with periodic replicas, also for charged systems:

- Correct the energy with an electrostatic model (Makov-Payne)
- Correct both the energy and the potential by cutting off the Coulomb potential in reciprocal space (Martyna-Tuckerman)
- Correct both the energy and the potential by cutting off the Coulomb potential in real space, increasing the dimensions of the cell for the Coulomb potential only (Hockney)

Makov-Payne corrections

Let us assume a molecule in a cubic cell of side L

- Neutral systems
  - If there is no dipole, the energy converges as L^{-5}
  - If there is dipole, the energy converges as L^{-3}
- Charged systems
  - The energy converges no better than L^{-1}

A correction to energy can be written as:

$$E_{corr} = E - \frac{q^2\alpha}{2L} - \frac{2\pi(qQ - d^2)}{3L^2} + \mathcal{O}(L^{-5})$$

where q is the net charge, α the Madelung constant, d the dipole, Q the quadrupole

Polar surfaces: dipole correction

Polar surfaces have a dipole. With PBC and a slab geometry, the dipole produces a fictitious, slowly decaying interaction with periodic replica. Such spurious interaction can be removed by adding a compensating dipole in the empty region. A recent example: potential profile for MoS_2 on Au surfaces (Pedram Khakbaz *et al*, Solid State Electronics, in press).