6. Order-N methods

6.1 Basic considerations

The computation time and necessary resources (memory) in *ab initio* calculation increase rapidly with the size of problem N (number of atoms or electrons, basis size). The scaling with N is governed by several factors and is generally $\sim N^3$ in both DFT and HF-based realizations. Depending on technical realization of the algorithm, the scaling may be worse than that. In the following, we'll discuss physical origins of the scaling with N, and the ways to improve it.

In traditional implementation of the DFT, like most of those discussed in Chapter 3, one has to solve Kohn–Sham equations (3.16), where one-electron functions are expanded over a fixed basis set of size, say, Q, Eq. (4.2). This results in a generalized eigenvalue problem, Eq. (4.3). So at some point we have to perform diagonalization that is a $\sim Q^3$ procedure, at least for algorithms so far known. Typically, the basis size Q is much larger than the number of electrons N.[37] It is not necessary to diagonalize the secular equation matrix in full – it suffices to know the N lowest eigenvalues and corresponding eigenvectors, necessary to reconstruct the density along Eq. (2.19). For this, iterative diagonalization techniques are known that scale, in the best case, as $\sim N^2 Q$. Even if this may be much better tan $\sim Q^3$, the scaling with the number of atoms is still at least $\sim N^3$.

It is important to realize that the limitation $\sim N^3$ is not purely technical. Even if more efficient diagonalization scheme appear, there is an essential intrinsic constant that actually sets this $\sim N^3$ limit. namely, this is the condition that the solutions, i.e., eigenvectors $C_{\alpha p}$ in the expansion (4.2), must be orthogonal for all $\alpha \leq N$. As the system grows, each wavefunction is represented by a longer eigenvector, $\sim N$. There are more occupied states, that brings in one more $\sim N$ factor. Each eigenvector must be orthogonalized to all others, that results in the total $\sim N^3$ scaling, on the very best.

What are then the arguments that better scaling can exist at all? From physical point of view, there seem to be some reasons to argue that the scaling as slow as linear is possible. Indeed, it is known – at least from chemistry – that electronic properties are to a large extent governed by near neighborhood. There are many justifications for different types of systems – indeed, cluster methods are used to simulate crystalline or amorphous solids in view of solid experience that details of electronic structure, local charge density, magnetic moments etc. converge reasonably fast to perfect bulk values as the cluster size grows. A somehow more sensitive property is total energy per atom; the following figure from a review paper by Goedecker[38] shows how rapidly it converges to its asymptotic (bulk) value with the cluster size.

[37] As we discussed earlier, that depends very much on the basis set used. In compact tight-binding bases, there are usually just several (few) basis functions per electron; in plane-wave schemes there can be hundreds.

[38] Stefan Goedecker, *Linear scaling electronic structure methods*, Rev. Mod. Phys. **71**, 1085–1123 (1999); http://arXiv.org/abs/cond-mat/9806073.

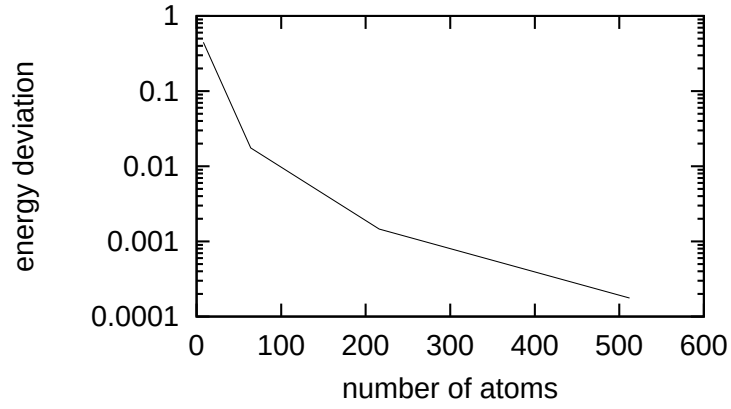


Fig. 1 of Goedecker: The deviation of the total energy per silicon atom from its asymptotic bulk value as a function of the size of the periodic volume in which it is embedded. The calculation was done with a tight-binding scheme using exact diagonalization.

Another manifestation of the dominance of short-range interactions is the behaviour of the density matrix depending on the separation between particles. The density matrix was introduced in Eq. (2.20). In the Hartree-Fock approximation, it is expressed via one-electron orbitals as in Eq. (2.22):

$$\gamma(\mathbf{r},\mathbf{r}') = \sum_{\alpha} \varphi_\alpha^*(\mathbf{r})\,\varphi_\alpha(\mathbf{r}')\,.$$

The summation is straightforward for fully polarized uniform electron gas:

$$\sum_{\alpha} \;\to\; \int_0^{k_\mathrm{F}} dN \;\to\; \frac{V}{8\pi^3} \int_0^{k_\mathrm{F}} k^2 dk \sin\vartheta_k d\vartheta_k d\phi_k$$

(for non-spin-polarized case, see Sec. 2.4). Hence

$$\begin{aligned}
\gamma(\mathbf{r},\mathbf{r}') &= \frac{V}{8\pi^3}\, 2\pi \int_0^{k_\mathrm{F}} k^2 dk \int_0^{\pi} \sin\vartheta\, d\vartheta\, e^{ik|\mathbf{r}-\mathbf{r}'|\cos\vartheta} \\
&= \frac{V}{2\pi^2} \int_0^{k_\mathrm{F}} k dk\, \frac{e^{ik|\mathbf{r}-\mathbf{r}'|} - e^{-ik|\mathbf{r}-\mathbf{r}'|}}{2i|\mathbf{r}-\mathbf{r}'|} = \frac{V}{2\pi^2|\mathbf{r}-\mathbf{r}'|} \int_0^{k_\mathrm{F}} k\, dk\, \sin(k|\mathbf{r}-\mathbf{r}'|) \\
&= \frac{V}{2\pi^2} \left[\frac{\sin(k_\mathrm{F}|\mathbf{r}'-\mathbf{r}|)}{|\mathbf{r}-\mathbf{r}'|^3} - \frac{k_\mathrm{F}\cos(k_\mathrm{F}|\mathbf{r}-\mathbf{r}'|)}{|\mathbf{r}-\mathbf{r}'|^2} \right] \\
&= \frac{V}{2\pi^2}\, \frac{k_\mathrm{F}^2}{|\mathbf{r}-\mathbf{r}'|}\, j_1\left(k_\mathrm{F}|\mathbf{r}-\mathbf{r}'|\right) \qquad (6.1)
\end{aligned}$$

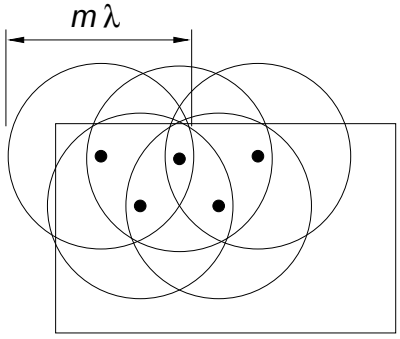
with $k_\mathrm{F}^3/(3\pi^2) = \rho$, see (2.35).

This HF asymptotics is essentially preserved in "true" metallic systems, as inter-electron correlations influence the specific short-range behaviour but not long-range decay. One can show that the decay of $\gamma(\mathbf{r},\mathbf{r}')$ with increasing $|\mathbf{r}-\mathbf{r}'|$ becomes even faster (i.e., exponential) in two cases, with the presence of an insulating band gap and when the Fermi distribution function is smeared out (for instance, due to the introduction of electronic temperature). The latter trick is sometimes used in order to control, or force, the decay of the density matrix. This option will be addressed later on.

Walter Kohn[39] discussed these issues of the dominance of short-range interactions and the decay of the density matrix with distance on a somehow more systematic basis, and he introduced a concept of "nearsightedness" of *equilibrium* system consisting of *many* quantum mechanical particles moving in an external potential *without* long-range (e.g., electric) fields, in the following way:

Let $F(\mathbf{r}_1, \ldots, \mathbf{r}_\nu)$ be a *static* property (for example: charge density, pair correlation function etc.) depending on coordinates of ν particles, all within a restricted volume of linear dimension $\sim \lambda$, a typical de Broglie wavelength in the ground state wavefunction. Then, at fixed chemical potential μ, a change of the external potential $\Delta v(\mathbf{r}')$, *no matter how large*, has a small effect on F, provided only that $\Delta v(\mathbf{r}')$ is limited to a *distant* region, i.e., for all $\mathbf{r}'$, $|\mathbf{r} - \mathbf{r}'| \gg \lambda$. Thus, F does not "see" $\Delta v(\mathbf{r}')$.

Kohn gives the following explanations to this principle.
1. The principle is a manifestation of wave-mechanical destructive interference.
2. It is essential to have many particles (not necessarily interacting).
3. There are exceptions: noninteracting bosons below critical point, systems with translationally invariant long-range order, like Wigner crystal in a torus.
4. If Δv includes long-range electric fields (unscreened Coulomb, as in ionic crystals), this must be added self-consistently to the effective potential. In other words, a Coulomb field from, say, an extra charge certainly will have effect on the many-particle system in question. But this effect will be, primarily, a rigid shift of potential over the whole region of dimension λ, that would shift the chemical potential there. However in the formulation above it was assumed that the chemical potential must be fixed. The principle says that, once the effect of bare Coulomb field is considered by a correspondingly shifted chemical potential, as determined in a self-consistency procedure, the variations Δv on top of bare Coulomb field won't affect the property F.

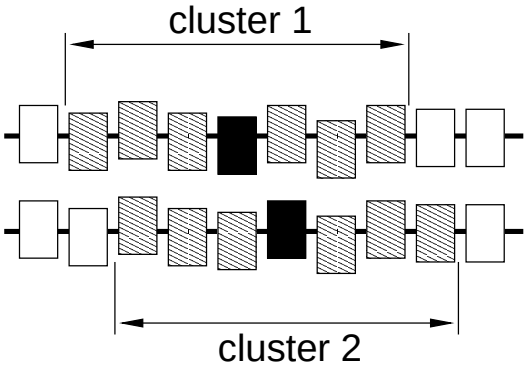


W. Kohn argued further on that the possibility of $O(N)$ methods immediately follows from the principle of "nearsightedness". For a system enclosed in a large volume Ω, one can introduce a system of overlapping smaller volumes, $\omega \sim (m\lambda)^3$ "where m is, say 100" – writes W. Kohn. One can consider each subvolume independently, because of nearsightedness, and enclose it in a hard wall boundary. The computational effort is proportional to the number of subvolumes $\sim N$, albeit probably with a quite high prefactor.

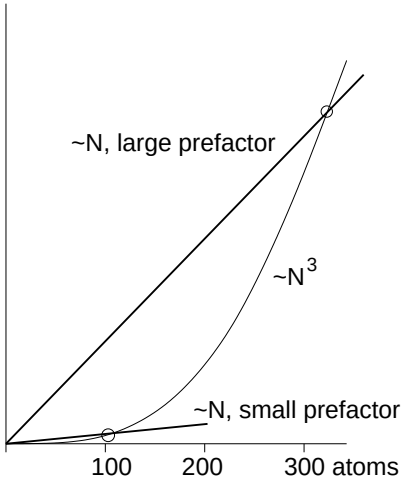
A quite straightforward implementation of this program would be applying a cluster method with open boundary conditions, centered on one atom after another (or selecting small subvolumes. One can imagine calculation of, say, electronic structure of DNA base by base (shown schematically as black boxes in the next figure), including in each step only certain number of neighboring structure elements in such "cluster" (shaded boxes)

[39] W. Kohn, *Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms*, Phys. Rev. Lett. **76**, 3168–3171 (1996).

and discarding more distant ones (white boxes):



The disadvantage is clearly that the computational effort, although linearly scalable with number of "clusters" considered, has so large prefactor that it hardly is affordable. Generally, when comparing $O(N^3)$ with $O(N)$ methods one is always a "crossover" of computational load (CPU time, memory requirements). The point in developing more sophisticated order-N methods is to shift the crossover point towards, say, 100 atoms or less, that would only make these methods practically interesting.



Most of workable schemes now in use deal with density matrix in one or another representation. The "most general" definition (2.20) won't be used in the following, because many-body wavefunction is not specified in density functional theory. However, we'll rely on the following expressions for kinetic and potential energies of electron systems, both being expectation values of a single-particle operators:

$$E_{\rm kin} = -\frac{\hbar^2}{2m}\int \nabla^2_{\mathbf{r}}\,\gamma(\mathbf{r},\mathbf{r}')\Big|_{\mathbf{r}=\mathbf{r}'}\,d\mathbf{r}'\,, \tag{6.2}$$

$$E_{\rm pot} = \int \gamma(\mathbf{r},\mathbf{r})\,v_{\rm eff}(\mathbf{r})\,d\mathbf{r}\,, \tag{6.3}$$

where $\gamma(\mathbf{r},\mathbf{r}) = \rho(\mathbf{r})$ according to Eq. (2.21). The above formulae follow from Eq. (3.4) under consideration that kinetic energy *is* a single-particle operator; the potential energy of Coulomb interaction in many-body system is two-particle operator, but *effective* interaction in the Kohn-Sham formalism is a single-particle operator all right,

$$v_{\rm eff}(\mathbf{r}) = \int \frac{\rho(\mathbf{r}')\,d\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|} + u^{\rm ext}(\mathbf{r}) + V_{\rm XC}(\mathbf{r})\,. \tag{6.4}$$

As long as we build on the Kohn-Sham approach with no two-particle operators explicitly appearing, we can quite consistently construct $\gamma(\mathbf{r},\mathbf{r}')$ as in the HF formalism, see

Eq. (2.22). The obvious generalization in terms of fixed basis functions will be [compare (4.5)]:

$$\gamma(\mathbf{r},\mathbf{r}') = \sum_{pq} \left[\sum_{\alpha=1}^{N(\mathrm{occ.})} C^*_{\alpha q}\, C_{\alpha p} \right] \chi^*_q(\mathbf{r}')\, \chi_p(\mathbf{r}),. \tag{6.5}$$

The main reason behind these manipulations is to reformulate the Kohn-Sham approach in such way as to avoid matrix diagonalization and/or the need to orthogonalize eigenvectors, that are essentially $\sim N^3$ algorithms. Actually, in many cases we do not care so much about individual eigenvectors (band structure, that becomes messy in a large system anyway) as about total energy and density matrix. The first allows us to perform dynamical calculations, and the second – to calculate expectation values of all single-particle operators, according to Eq. (3.3):

$$\langle F \rangle = \mathrm{Tr}(\hat{F}\gamma)\,.$$

The Kohn-Sham total energy of Eq. (3.17)

$$E_{\mathrm{tot}} = \sum_{\alpha=1}^{N} \varepsilon_\alpha - \frac{e^2}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}' - \int V_{\mathrm{XC}}(\mathbf{r})\, \rho(\mathbf{r})\, d\mathbf{r} + E'_{\mathrm{XC}}[\rho]$$

can be casted into

$$E_{\mathrm{tot}} = \underbrace{E_{\mathrm{BS}}}_{\substack{\equiv \sum_\alpha \varepsilon_\alpha \\ \text{(band-structure energy)}}} - \underbrace{E_{\mathrm{DC}}}_{\substack{\text{(double-counted terms)}}}. \tag{6.6}$$

$E_{\mathrm{BS}} = E_{\mathrm{kin}} + E_{\mathrm{pot}}$ in terms of (6.2) and (6.3). Indeed, with $\gamma(\mathbf{r},\mathbf{r}') = \sum_\alpha \varphi^*_\alpha(\mathbf{r}')\,\varphi_\alpha(\mathbf{r})$ (6.2) and (6.3) become

$$\begin{aligned}
E_{\mathrm{kin}} + E_{\mathrm{pot}} &= -\frac{\hbar^2}{2m}\int \nabla^2_{\mathbf{r}'} \sum_{\alpha=1}^{N} \varphi^*_\alpha(\mathbf{r}')\,\varphi_\alpha(\mathbf{r})\Big|_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}' + \\
&\quad + \sum_{\alpha=1}^{N} \int \varphi^*_\alpha(\mathbf{r}')\, v_{\mathrm{eff}}(\mathbf{r})\, \varphi_\alpha(\mathbf{r})\Big|_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}' = \\
&= \sum_{\alpha=1}^{N} \int d\mathbf{r}\, \varphi^*_\alpha(\mathbf{r}) \left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) \right] \varphi_\alpha(\mathbf{r}) = \\
&= \text{[compare to Eq. (3.16)]} \quad \sum_{\alpha=1}^{N} \varepsilon_\alpha\,.
\end{aligned}$$

The way it is written,

$$E_{\mathrm{BS}} = \mathrm{Tr}\,(\hat{\gamma}\,\hat{\mathcal{H}})\,.$$

This is obvious in the coordinate representation, or, one can transform to the preselected basis of χ_q:

$$E_{\mathrm{BS}} = \overline{\sum_{\alpha=1}^{N}} \underline{\int d\mathbf{r}} \sum_{p=1}^{Q} \overline{C^*_{\alpha q}}\, \underline{\chi^*_q(\mathbf{r}) \left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\mathrm{eff}}(\mathbf{r}) \right]} \sum_{p=1}^{Q} \overline{C_{\alpha p}}\, \underline{\chi_p(\mathbf{r})} = \sum_{\substack{p=1\\q=1}}^{Q} D_{pq}\, H_{qp}\,,$$

where H_{qp} incorporates underlined terms, and D_{pq} – those with the bars above.

Double-counting terms can be similarly casted in a form expressed via a density matrix. Let's add to the potential energy a yet undefined potential $U(\mathbf{r})$ and find out what it must be, in order to arrive at the correct expression for the total energy, with the double-counting terms subtracted.

$$\begin{aligned} E_{\rm tot} &= \sum_{\alpha=1}^{N} \int d\mathbf{r}\, \varphi_\alpha^*(\mathbf{r}) \left[-\frac{\hbar^2}{2m}\nabla^2 + U(\mathbf{r}) \right] \varphi_\alpha(\mathbf{r}) = \\ &= \sum_{\alpha=1}^{N} \int d\mathbf{r}\, \varphi_\alpha^* \left[-\frac{\hbar^2}{2m}\nabla^2 + v_{\rm eff}(\mathbf{r}) \right] \varphi_\alpha(\mathbf{r}) + \\ &\quad + \sum_{\alpha=1}^{N} \int d\mathbf{r}\, \varphi_\alpha^* \left[U(\mathbf{r}) - v_{\rm eff}(\mathbf{r}) \right] \varphi_\alpha(\mathbf{r}) \\ &= \sum_{\alpha=1}^{N} \varepsilon_\alpha - \int d\mathbf{r} \left[v_{\rm eff}(\mathbf{r}) - U(\mathbf{r}) \right] . \end{aligned}$$

Comparing to Eq. (3.17) yields

$$v_{\rm eff}(\mathbf{r}) - U(\mathbf{r}) = \frac{e^2}{2} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + V_{\rm XC}(\mathbf{r}) - \underbrace{\frac{\delta E_{\rm XC}[\rho]}{\delta \rho}}_{V_{\rm XC}} ,$$

hence

$$U(\mathbf{r}) = v_{\rm XC}(\mathbf{r}) - \frac{e^2}{2} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} , \quad \text{and}$$

$$\begin{aligned} E_{\rm tot} &= -\frac{\hbar^2}{2m} \int \nabla_{\mathbf{r}}^2\, \gamma(\mathbf{r},\mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} \\ &\quad + \int \gamma(\mathbf{r},\mathbf{r}) \left[\frac{e^2}{2} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}' + u^{\rm ext}(\mathbf{r}) + V_{\rm XC}(\mathbf{r}) \right] d\mathbf{r} , \end{aligned}$$

without reference to band energies. It is important for the following that this expression doe not contain anymore the results of diagonalization, ε_α nor $C_{\alpha q}$, if we address the density matrix D_{pq} later on *directly*, without reference to its representation as a sum over eigenvectors.

Since we want to abandon direct reference to ε_α whenever possible, we face a problem how to populate the states numbered by α. Conventionally we kept trace on those $\alpha \leq N$ corresponding to the lowest ε_α. In other formulation, the necessary sums may run over *all* states, weighted by the step function (=1 for $\alpha \leq N$, =0 for $\alpha > N$). Now a trick will be that we introduce instead of this Fermi distribution function,

$$F(\varepsilon) = \frac{1}{e^{\frac{\varepsilon-\mu}{kT}} + 1} , \tag{6.7}$$

corresponding to a certain chemical potential μ and, probably to a certain electronic temperature T. It was mentioned earlier that this may be also advantageous to force

better spatial decay of the density matrix. (Note however that this temperature has nothing to do with "lattice" temperature introduced in dynamical simulations). With the Fermi function thus introduced, the density matrix and other properties (energy etc.) can be expressed via sums over all basis functions:

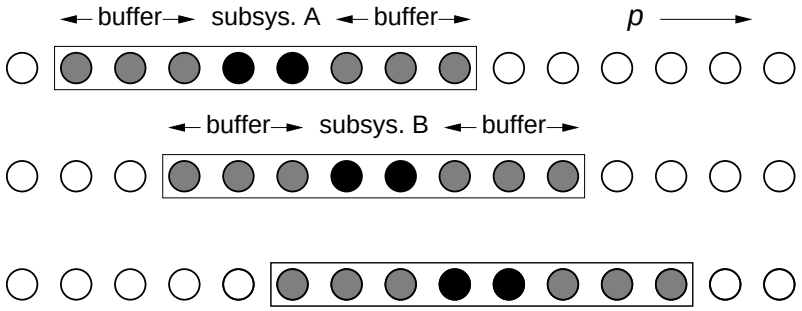
$$\begin{aligned} \gamma(\mathbf{r},\mathbf{r}') &= \sum_{\alpha=1}^{Q} F(\varepsilon_\alpha)\,\varphi_\alpha^*(\mathbf{r})\,\varphi_\alpha(\mathbf{r}')\,, \\ D_{pq} &= \sum_{\alpha=1}^{Q} F(\varepsilon_\alpha)\,C_{\alpha q}^*\,C_{\alpha p}\,, \\ E_{\rm BS} &= \sum_{p,q=1}^{Q} D_{pq}\,H_{qp} = \sum_{\alpha=1}^{Q} F(\varepsilon_\alpha)\,\varepsilon_\alpha\,. \end{aligned} \tag{6.8}$$

After this general introduction, we review several methods that allow in principle order-N scaling in some detail.

6.2 Divide-and-Conquer

We begin with a historically important order-N method, proposed by Yang and Zhao in 1991–1995, the "divide-and-conquer" (D&C) method. It performs better than merely superposition of free clusters, because D&C allows interaction between subsystems. Therefore the "clusters" can be kept reasonably small.

The system is split into several subsystems, as already discussed earlier, and each subsystem is surrounded by its buffer, like e.g. for the linear chain:



For subsystem A, the eigenvalue problem is solved (e.g., by diagonalization) in terms of $\mathcal{H}$ and $\mathcal{S}$ matrices between pairs of functions within the subsystem *and* its buffer:

$$\sum_p \left[H_{qp}^A - \varepsilon_\alpha^A S_{qp}^A\right] C_{\alpha p}^A = 0\,. \tag{6.9}$$

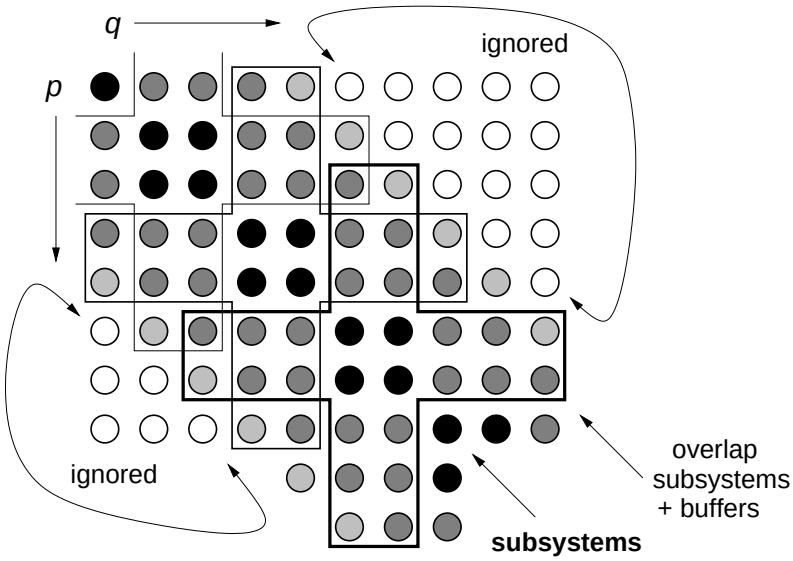
The size of these matrices is $(N^A \times N^A)$, counting the basis functions within the system + buffer, therefore the related diagonalization amounts to (hopefully moderate) $\sim (N^A)^3$. In order to compute the total energy, we need density matrix for the whole system, computed from individual D_{pq}^A using *partition matrix* d_{pq}^A such that

$$\sum_A d_{pq}^A = 1\,, \qquad D_{pq}^{\rm global} = \sum_A d_{pq}^A\,D_{pq}^A\,. \tag{6.10}$$

The partition scheme is subject to an (arbitrary, up to the sum rule above) choice, e.g.,

$$d_{pq}^{A} = \begin{cases} 0 & \text{if} \quad p \in \text{buffer and } q \in \text{buffer}, \\ \frac{1}{2} & \text{if} \quad p \in A \text{ and } q \in \text{buffer, or vice versa} \\ 1 & \text{if} \quad p \in A \text{ and } q \in A, \end{cases}$$

or more sophisticated. For the case of linear chain as shown above, the (two-dimensional) density matrix would be reconstructed from the overlap of individual D_{pq}^{A} as follows:



The global density matrix, by virtue of Eq. (6.10), is

$$\sum_{A} d_{pq}^{A} = 1\,, \qquad D_{pq}^{\text{global}} = \sum_{A} d_{pq}^{A} \sum_{\alpha=1}^{Q} F(\varepsilon_\alpha)\, C_{\alpha q}^{A*}\, C_{\alpha p}^{A}\,. \tag{6.11}$$

The value of the chemical potential μ appearing in $F(\varepsilon)$ is shared by all subsystems; this allows for the charge transfer between them. μ must be self-consistently determined from the normalization condition,

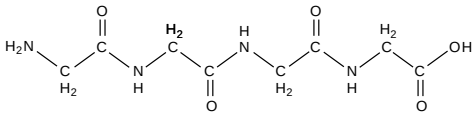
$$\sum_{p,q=1}^{Q^{\,\text{global}}} \underbrace{D_{pq}}_{\substack{\text{contains} \\ F(\varepsilon)}} S_{qp} = N\,. \tag{6.12}$$

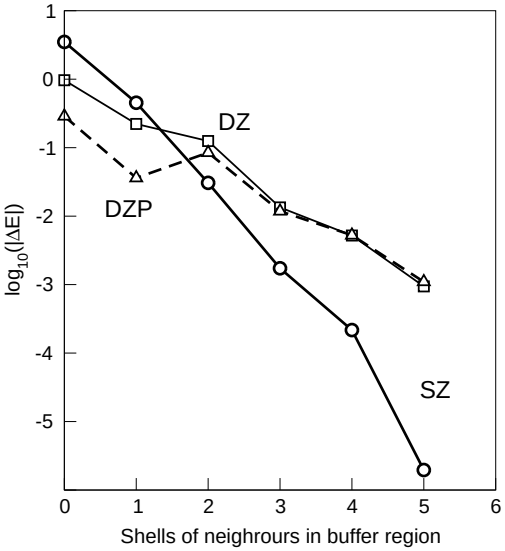
that is identical to

$$\sum_{pq} \sum_{\alpha=1}^{N} C_{\alpha q}^{*}\, C_{\alpha p} \int \chi_q^{*}(\mathbf{r})\, \chi(\mathbf{r})\, d\mathbf{r} = \sum_{\alpha=1}^{N} \int d\mathbf{r} \underbrace{\sum_{q}^{Q} C_{\alpha q}^{*} \chi_q^{*}(\mathbf{r})}_{\varphi_\alpha^{*}(\mathbf{r})} \underbrace{\sum_{p}^{Q} C_{\alpha p} \chi_p(\mathbf{r})}_{\varphi_\alpha(\mathbf{r})} = N\,.$$

The convergence of results with the buffer size may look like in the following figure,

reproduced from the review by Ordejón[40] and presenting the results of Yang[41] for a tetrapeptide molecule (shown below), with different buffer sizes (varying from 0 to 5 neighbouring groups in the polypeptide chain) and different choice of bases: "single-zeta" (SZ), "double-zeta" (DZ) and "double-zeta including polarization orbitals" (DZP). The basis size increases from SZ (99) to DZ (181) to DZP (308), the number in brackets giving the total number of basis functions in the whole molecule. The accuracy typically becomes acceptable after already just several shells of neighbors (=buffer size) included. However, the method is not variational, i.e., the error is not necessarily reduced while using larger (or better) basis set. The Hellmann-Feynman forces are nevertheless available.





Results from Table 1 of Yang (1991): absolute value of errors in the total energies for a tetrapeptide molecule for the D&C approach, as compared with the Kohn-Sham result for the whole molecule with the corresponding choice of basis.

6.3 Fermi operator expansion

The starting point of this method is the Fermi distribution function (6.7), which, as a function of ε, is approximated by a polynomial expansion. The temperature T enters as an extrenal parameter; from physical considerations, the case $T \to 0$ is of most interest, but keeping the temperature small non-zero is advantageous for getting better polynomial approximation, as well as for achieving better spatial localization of the density matrix. The idea behind the method is to avoid the diagonalization and the selection of N lowest states. Instead, one fixes the chemical potential μ in advance (and corrects it afterwards, if the total number of electrons comes out wrongly), and the calculation of the density matrix is done iteratively, from a "reasonable" starting guess. The Fermi distribution function can be generalized to a "Fermi operator"

$$F(\mathcal{H}) = \frac{1}{e^{\frac{\mathcal{H}-\mu}{kT}} + 1} \,,$$

which, for $T = 0$, reduces to merely $\mathcal{H}$ if acting on occupied states, and zero for the states whose eigenvalues are beyond μ.

Casting the Kohn-Sham equation in the form

$$F(\mathcal{H})|\varphi_\alpha\rangle = F(\varepsilon_\alpha)|\varphi_\alpha\rangle \,, \tag{6.13}$$

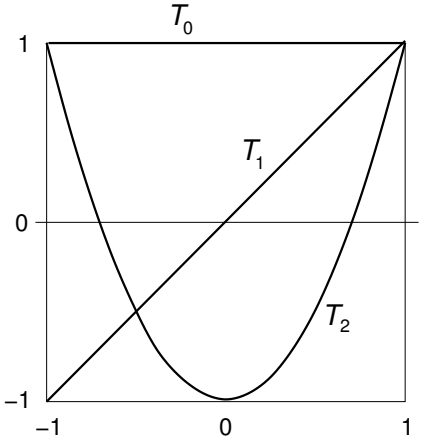
[40] Pablo Ordejón, *Order-N tight-binding methods for electronic-structure and molecular dynamics*, Comput. Mater. Sci. **12, 157–191 (1998).**

[41] Weitao Yang, *Direct calculation of electron density in density-functional theory: Implementation for benzene and a tetrapeptide*, Phys. Rev. A **44**, 7823 (1991).

we'll now use instead of the the "true" Fermi operator $F(\varepsilon)$ its polynomial approximation $P(\varepsilon)$:

$$P(\mathcal{H})|\varphi_\alpha\rangle = P(\varepsilon_\alpha)|\varphi_\alpha\rangle\,. \tag{6.14}$$

The polynomial approximation $P(\mathcal{H})$ can be taken of any appropriate form; it is advantageous to use Tchebyshev polynomials T_j for better numerical stability:



$$\begin{aligned} T_0 &= 1 \\ T_1 &= x \\ T_2 &= 2x^3 - 1 \\ T_3 &= 4x^3 - 3x \\ &\cdots \\ T_{j+1}(x) &= 2x\,T_j(x) - T_{j-1}(x)\,. \end{aligned}$$

Then the polynomial approximation $P(\mathcal{H})$ reads:

$$P(\mathcal{H}) = \frac{C_0\mathcal{I}}{2} + \sum_{j=1}^{n_{\rm poly}} C_j\,T_j(\mathcal{H})\,. \tag{6.15}$$

This allows, instead of direct diagonalization, to construct "Tchebyshev matrices" (which will sum up to give the density matrix) iteratively, as follows:

$$\begin{aligned} |t_l^0\rangle &= |e_l\rangle \\ |t_l^1\rangle &= \mathcal{H}|e_l\rangle \\ &\cdots \\ |t_l^{j+1}\rangle &= 2\mathcal{H}|t_l^j\rangle - |t_l^{j-1}\rangle\,. \end{aligned}$$

Here $|t_l^j\rangle$ is the l'th column of the j'th Tchebyshev matrix, $|e_l\rangle$ is a starting guess, typically a l'th column of the unit matrix. On constructing all t_l^j's up to $j = n_{poly}$ by matrix multiplication, one can recover the l'th column of the density matrix:

$$|D_l\rangle = \frac{C_0}{2}|t_l^0\rangle + \sum_{j=1}^{n_{poly}} C_j|t_l^j\rangle\,. \tag{6.16}$$

Note that the chemical potential μ was fixed at the beginning. If this would result in a wrong number of electrons (obtained from the trace of the density matrix), μ can be corrected by

$$\Delta\mu = \frac{\Delta N_{\rm el}}{{\rm Tr}\,[P'(\mathcal{H})]}\,,$$

where $P\prime$ is the derivative of the Tchebyshev polynomial P chosen to represent the Fermi distribution, by the virtue of

$$\begin{aligned} N_{\rm el} = {\rm Tr}\,[F] &\rightarrow {\rm Tr}\,[P]\,, \\ \frac{dN_{\rm el}}{d\mu} &\rightarrow {\rm Tr}\left[\frac{dP}{d\mu}\right]\,. \end{aligned}$$

Without further assumptions, the method is of the order N^2. Indeed, for M_b basis functions, the density matrix is $M_b \times M_b$, i.e. one has to calculate M_b full columns. For the calculation of each column, one performs $n_{\rm poly}$ matrix $\times$ vector multiplications, each costs $M_b\, n_H$, assuming $\mathcal{H}$ is a sparse matrix with n_H non-zero elements per row (or column). The total effort is then

$$\sim M_b^2 \underbrace{n_{\rm poly}\, n_H}_{\substack{\text{independent}\\ \text{on the}\\ \text{system size}}} .$$

The linear scaling is recovered if one introduces a localization region for each column of the density matrix, outside of which the elements are negligibly small. Denoting the number of such non-vanishing elements in each column of the density matrix by $M_{\rm loc}$, we arrive at the scaling $\sim M_b\, M_{\rm loc}\, n_{\rm poly}\, n_H$, that is linear.

6.4 Variational methods

We consider now two different approaches which both avoid the diagonalization and directly optimize total energy in terms of either elements of the density matrix, or individual orbitals. The total energy, for our further purposes, will be casted as

$$\begin{aligned} E_{\rm tot} &= \mathrm{Tr}\left(\hat{\gamma}\,\hat{\mathcal{H}}\right) , \\ \text{or} \quad E_{\rm tot} &= \sum_{\alpha=1}^{N} \langle\varphi_\alpha|\mathcal{H}|\varphi_\alpha\rangle . \end{aligned}$$

In order to make te function to be minimized insensitive to the rigid shift of the potential, we substract μN from $E_{\rm tot}$ and thus define the *grand potential* Ω,

$$\Omega = \mathrm{Tr}\,(\hat{\gamma}\mathcal{H}) - \mu N = \mathrm{Tr}\left[\hat{\gamma}(\mathcal{H} - \mu\,\hat{I})\right] , \tag{6.17}$$

$$\text{or} \quad \Omega = \sum_{\alpha=1}^{N} \langle\varphi_\alpha|\mathcal{H} - \mu\,\hat{I}|\varphi_\alpha\rangle . \tag{6.18}$$

Under constant shift, the potential energy increases by $\Delta U{\cdot}N$; in order to conserve the number of electrons, μ must also shift upwards and thus Ω remains constant. One can achieve minimization of Ω only by applying non-uniform potential, that would lead to a redistribution of charge density, reducing the total energy in a "non-trivial" way. In the following, we shall refer to

$$\mathcal{H}' \equiv \mathcal{H} - \mu\,\hat{I}$$

as "shifted Hamiltonian". The minimization of either (6.17) in terms of density matrix, or of (6.18) in terms of orbitals is subject to constraints that are automatically satisfied in the course of a conventional diagonalization procedure, but we have to build in explicitly if we envisage to abandon the diagonalization. These constraints are, on the orbitals, that they must be orthonormal,

$$\langle\varphi_\alpha|\varphi_\beta\rangle = \delta_{\alpha\beta} ,$$

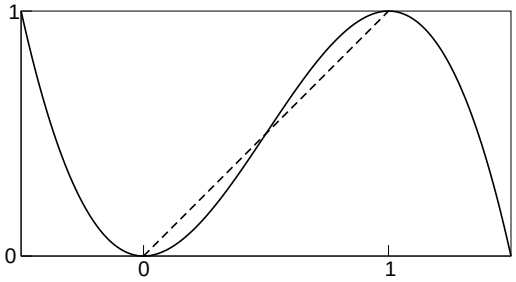
and on the density matrix – that it has the property of *idempotency*, [42] i.e.

$$\hat{\gamma}^2 = \hat{\gamma}\,. \tag{6.19}$$

Moreover, it must conserve the number of electrons:

$$N = \mathrm{Tr}(\hat{\gamma})\,. \tag{6.20}$$

In a conventional energy-minimization scheme one proceeds iteratively; in each step, the gradient of Ω with respect to either γ, or $\{\varphi_\alpha\}$, is calculated, the variables are moved in the direction of gradients, and the constraints imposed. A disadvantage is that imposing directly orthonormality or idempotency constraints is an over-linear operation. An alternative would be, instead of imposing constraints explicitly in each iteration, to gradually influence the variables in such a way that the constraints would become ever better satisfied in the course of iterations. Several functionals which do not require explicit imposition of constraints have been developed. Li, Nunes and Vanderbilt [43] used a *purification* transformation of the density matrix via the McWeeny function:



$$\tilde{\gamma} = 3\gamma^2 - 2\gamma^3\,. \tag{6.21}$$

If eigenvalues of γ are close to either 0, or 1, the eigenvalues of $\tilde{\gamma}$ become even close to the same values. The Li–Nunes–Vanderbilt (LNV) grand potential is then, instead of $\mathrm{Tr}[\hat{\gamma}(\mathcal{H} - \mu\hat{I}]$,

$$\Omega^{\mathrm{LNV}} = \mathrm{Tr}\left[\left(3\hat{\gamma}^2 - 2\hat{\gamma}^3\right)\left(\mathcal{H} - \mu\hat{I}\right)\right]\,, \tag{6.22}$$

and no constraints imposed during the minimization. On the convergency, $3\gamma^2 - 2\gamma^3 = \gamma$ by the virtue of $\gamma^2 = \gamma$, hence $\Omega^{\mathrm{LNV}} = \Omega^{\mathrm{true}}$.

The gradient of Ω^{LNV} can be found using the following differentiation rules for the trace of a matrix:

$$\text{for}\quad M = \sum_{ij} A_{ij}B_{ji}, \qquad \frac{\partial M}{\partial A_{\alpha\beta}} = B_{\beta\alpha}\,;$$

$$\text{for}\quad M = \sum_{ijk} A_{ki}A_{ij}B_{jk}, \qquad \frac{\partial M}{\partial A_{\alpha\beta}} = \sum_{k}\left(B_{\beta k}A_{k\alpha} + A_{\beta k}B_{k\alpha}\right)\,;$$

$$\text{for}\quad M = \sum_{ijkl} A_{ij}A_{jk}A_{kl}B_{li}, \qquad \frac{\partial M}{\partial A_{\alpha\beta}} = \sum_{k}\left(A_{\beta k}A_{kl}B_{l\alpha} + A_{\beta k}B_{ki}A_{i\alpha} + B_{\beta k}A_{ki}A_{i\alpha}\right)\,,$$

hence

$$\frac{\partial\Omega^{\mathrm{LNV}}}{\partial\hat{\gamma}} = 3\left(\hat{\gamma}\mathcal{H}' + \mathcal{H}'\hat{\gamma}\right) - 2\left(\hat{\gamma}^2\mathcal{H}' + \hat{\gamma}\mathcal{H}'\hat{\gamma} + \mathcal{H}'\hat{\gamma}^2\right)\,, \tag{6.23}$$

[42] for instance, in the coordinate representation it follows immediately from the definition of the density matrix, Eq.(2.20),

$$\gamma(\mathbf{r},\mathbf{r}') = \int \gamma(\mathbf{r},\mathbf{r}'')\,\gamma(\mathbf{r}'',\mathbf{r})\,d\mathbf{r}''\,.$$

[43] X.-P. Li, R. W. Nunes, and David Vanderbilt, *Density-matrix electronic-structure method with linear system-size scaling*, Phys. Rev. B **47**, 10891–10894 (1993).

with $\mathcal{H}' \equiv \mathcal{H} - \mu\hat{I}$. Since $\mathcal{H}\varphi = \varepsilon\varphi$ and $\hat{\gamma}\varphi = f(\varepsilon)\varphi$,

$$\mathcal{H}\hat{\gamma}\varphi = \varepsilon f(\varepsilon)\varphi = \hat{\gamma}\mathcal{H}\varphi \quad \text{and hence} \quad [\hat{\gamma}, \mathcal{H}'] = 0,$$

then

$$\frac{\partial \Omega^{\mathrm{LNV}}}{\partial \hat{\gamma}} = 6\gamma\mathcal{H}' - 6\gamma^2\mathcal{H}' = 0$$

for true γ, i.e., $\gamma = \gamma^2$.

Since the LNV functional is a cubic polynomial in all its degrees of freedom, and cubic polynomial may have only one local minimum, there is exactly one local minimum in a multidimensional minimization, which can be easily found. That means that the LNV functional is a "well behaved" one.

A complication may arise that, unless μ is set to a "correct" band gap from the beginning, the minimization may end up with a wrong number of electrons. To overcome this, one must simply foresee on option to adjust μ in the course of minimization (see the discussion to this point at the end of Section 6.3.

The forces on atoms are available as

$$\frac{d\Omega}{d\mathbf{R}_\alpha} = \frac{\partial\Omega}{\partial\gamma}\frac{\partial\gamma}{\partial\mathbf{R}_\alpha} + \frac{\partial\Omega}{\partial\mathcal{H}}\frac{\partial\mathcal{H}}{\partial\mathbf{R}_\alpha}.$$

$\frac{\partial\Omega}{\partial\gamma}$ vanishes at the solution, and the previous formula simplifies to

$$\frac{d\Omega}{d\mathbf{R}_\alpha} = \frac{\partial\Omega}{\partial\mathcal{H}}\frac{\partial\mathcal{H}}{\partial\mathbf{R}_\alpha} = \mathrm{Tr}\left[\left(3\gamma^2 - 2\gamma^3\right)\frac{\partial\mathcal{H}}{\partial\mathbf{R}_\alpha}\right]. \tag{6.24}$$

Now we come back to a possibility to minimize the functional Ω in a different way – not via the optimization of density as in Eq. (6.17) but via optimization of wavefunctions as in Eq. (6.18). Remember that we want to allow an *unconstrained* variation of $|\varphi_\alpha\rangle$, in order to avoid the orthonormalization which is an order-N^3 procedure. Therefore $|\varphi_\alpha\rangle$ will remain not necessarily orthogonal underway and will only become orthonormalized in the course of optimization. So there will exist an overlap

$$S_{\alpha\beta} = \langle\varphi_\alpha|\varphi_\beta\rangle\,, \tag{6.25}$$

and the "band structure" energy takes the form

$$E_{\mathrm{BS}} = \sum_{\alpha,\beta=1}^{N} (\mathcal{S}^{-1})_{\alpha\beta}\langle\varphi_\beta|\mathcal{H}|\varphi_\alpha\rangle\,. \tag{6.26}$$

Indeed, as has been introduced earlier,

$$E_{\mathrm{BS}} = \sum_{\alpha=1}^{N}\varepsilon_\alpha = \sum_{\alpha=1}^{N}\langle\varphi_\alpha|\mathcal{H}|\varphi_\alpha\rangle$$

for an orthonormal basis. For an arbitrary basis, it holds anyway

$$\mathcal{H}|\varphi_\alpha\rangle = \varepsilon_\alpha|\varphi_\alpha\rangle \quad \Rightarrow \quad \langle\varphi_\beta|\mathcal{H}|\varphi_\beta\rangle = \varepsilon_\alpha\underbrace{\langle\varphi_\beta|\varphi_\alpha\rangle}_{S_{\beta\alpha}}$$

hence

$$\varepsilon_\alpha = \sum_\beta \left(\mathcal{S}^{-1}\right)_{\alpha\beta} \langle\varphi_\beta|\mathcal{H}|\varphi_\alpha\rangle \quad \text{and} \quad E_{\rm BS} = \sum_{\alpha,\beta=1}^{N} \left(\mathcal{S}^{-1}\right)_{\alpha\beta} \langle\varphi_\beta|\mathcal{H}|\varphi_\alpha\rangle \,.$$

For the density matrix defined as

$$\hat{\gamma} = \sum_{\alpha,\beta=1}^{N} |\varphi_\alpha\rangle \left(\mathcal{S}^{-1}\right)_{\alpha\beta} \langle\varphi_\beta| \,,$$

$$\langle\varphi_i|\hat{\gamma}|\varphi_j\rangle = \sum_{ij} \underbrace{\langle\varphi_i|\varphi_\alpha\rangle}_{S_{i\alpha}} \left(\mathcal{S}^{-1}\right)_{\alpha\beta} \underbrace{\langle\varphi_\beta|\varphi_j\rangle}_{S_{\beta j}} = \sum_\beta \delta_{i\beta} S_{\beta j} = S_{ij} \,, \quad \text{as it should be.}$$

With (6.26), the grand potential Ω (6.18) reads

$$\Omega = \sum_{\alpha,\beta=1}^{N} \left(\mathcal{S}^{-1}\right)_{\alpha\beta} \underbrace{(H_{\beta\alpha} - \mu S_{\beta\alpha})}_{\equiv\, H'_{\beta\alpha}\,,\ \text{shifted Hamiltonian.}} \tag{6.27}$$

The problem with minimizing this functional is that, no matter how localized the orbitals are, the inversion of $\mathcal{S}$ is *not* an order-N operation and results in a *not sparse* matrix. In order to get rid of $(\mathcal{S}^{-1})$, Mauri *et al.*[44] proposed to use a Taylor expansion,

$$\left(\mathcal{S}^{-1}\right) = \sum_{n=0}^{\infty} \left(\hat{I} - \mathcal{S}\right)^n \,. \tag{6.28}$$

Indeed, expanding $1/x$ about $x = 1$ yields

$$\frac{1}{x} = 1 - (x-1) + \frac{2}{2!}(x-1)^2 - \frac{6}{3!}(x-1)^3 + \ldots = 1 + (1-x) + (1-x)^2 + \ldots$$

Truncating this series at certain n leaves us with the n-th order approximation to the true overlap^{-1},

$$Q^{(n)} = \sum_{n=0}^{m} \left(\hat{I} - \mathcal{S}\right) \,, \tag{6.29}$$

and

$$\Omega \Rightarrow \sum_{\alpha,\beta=1}^{N} Q^{(m)}_{\alpha\beta} \left(H_{\beta\alpha} - \mu S_{\beta\alpha}\right) \,. \tag{6.30}$$

Now the functional Ω can be obtained by merely matrix multiplication. In the Taylor expansion (6.29), it suffices to take the lowest non-trivial term ($m = 1$), if we would show that this results in the correct minimum value of the functional. Let's try it.

$$Q^{(1)} = 2\hat{I} - \mathcal{S} \,;$$

[44] Francesco Mauri, Giulia Galli, and Roberto Car, *Orbital formulation for electronic-structure calculations with linear system-size scaling*, Phys. Rev. B **47**, 9973–9976 (1993).

the variation of $|\varphi_\alpha\rangle$ will be, in practice, a variation of their expansion coefficient in a fixed basis system, like in Eq. (4.2).

$$\begin{aligned}\Omega &= \sum_{\alpha,\beta}^{N} (2\delta_{\alpha\beta} - S_{\alpha\beta})\, H'_{\beta\alpha} = \\ &= 2\sum_{\alpha=1}^{N}\sum_{p,q=1}^{Q} C^*_{\alpha q} H'_{pq} C_{\alpha q} - \sum_{\alpha,\beta=1}^{N}\sum_{p,q=1}^{Q} C^*_{\alpha p} H'_{pq} C_{\beta q} \underbrace{\sum_{r} C^*_{\alpha r} C_{\beta r}}_{=S_{\alpha\beta}} . \end{aligned} \tag{6.31}$$

In the orthonormal basis ($\sum_r C^*_{\alpha r} C_{\beta r} \equiv S_{\alpha\beta} = \delta_{\alpha\beta}$) (6.31) reduces to

$$\Omega = \sum_{\alpha=1}^{N}\sum_{p,q=1}^{Q} C^*_{\alpha p} H'_{pq} C_{\alpha q} = \sum_{\alpha=1}^{N} \varepsilon_\alpha - \mu N \,,$$

as it should be. In the general (i.e. not necessarily orthonormal) case, the differentiation of (6.31) yields:

$$\frac{\partial\Omega}{\partial C_{\gamma s}} = 2\sum_{p=1}^{Q} C^*_{\gamma p} H'_{ps} - \sum_{\alpha=1}^{N}\sum_{p=1}^{Q} C^*_{\alpha p} H'_{ps} \sum_{q=1}^{Q} C^*_{\alpha q} C_{\gamma q} - \sum_{\alpha=1}^{N}\sum_{p,q=1}^{Q} C^*_{\alpha p} H'_{pq} C_{\gamma q} C^*_{\alpha s} \,. \tag{6.32}$$

Using

$$\sum_q H_{pq} C_{\gamma q} = \varepsilon_\gamma C_{\gamma p} \,, \quad \sum_p C^*_{\gamma p} H_{pq} = \varepsilon_\gamma C^*_{\gamma q}$$

and hence

$$\sum_q H'_{pq} C_{\gamma q} = (\varepsilon_\gamma - \mu)\, C_{\gamma p} \,, \quad \sum_p C^*_{\gamma p} H'_{pq} = (\varepsilon_\gamma - \mu)\, C^*_{\gamma q} \,,$$

eq. (6.32) reduces to:

$$\begin{aligned}\frac{\partial\Omega}{\partial C_{\gamma s}} &= 2(\varepsilon_\gamma - \mu) C^*_{\gamma s} - \sum_{\alpha=1}^{N} (\varepsilon_\alpha - \mu)\, C^*_{\alpha s} \sum_{q=1}^{Q} C^*_{\alpha q} C_{\gamma q} - \sum_{\alpha=1}^{N}\sum_{p=1}^{Q} C^*_{\alpha p} (\varepsilon_\gamma - \mu)\, C_{\gamma p} C^*_{\alpha s} \\ &= 2(\varepsilon_\gamma - \mu) C^*_{\gamma s} - \sum_{\alpha=1}^{N} C^*_{\alpha s} \underbrace{\sum_{p=1}^{Q} C^*_p\, C_{\gamma p}}_{\substack{=\delta_{\alpha\gamma} \\ \text{for exact solution,} \\ \text{when orthonormality} \\ \text{is restored}}} (\varepsilon_\alpha + \varepsilon_\gamma - 2\mu) \\ &= 2(\varepsilon_\gamma - \mu) C^*_{\gamma s} - C^*_{\gamma s} (2\varepsilon_\gamma - 2\mu) = 0 \,. \end{aligned}$$

We have shown that the derivative $\partial\Omega/\partial C_{\gamma s}$ of the form (6.32), which is valid for *any* $C_{\alpha p}$ in the course of minimization (i.e. also away from the orthonormality) really vanishes if the orthonormality is enforced. In other words, the extremum $\partial\Omega/\partial C_{\gamma s} = 0$ corresponds to the situation when eigenvectors are orthonormal.

Differentiating (6.32), the second derivative $\frac{\partial^2}{\partial C_{\gamma s}^2}$ is:

$$\begin{aligned}\frac{\partial^2}{\partial C_{\gamma s}^2} &= -\sum_{\alpha=1}^{N}\sum_{p=1}^{Q} C_{\alpha p}^* H'_{ps} C_{\alpha s}^* - \sum_{\alpha=1}^{N}\sum_{p=1}^{Q} C_{\alpha p}^* H'_{ps} C_{\alpha s}^* \\ &= -2\sum_{\alpha=1}^{N} \underbrace{\sum_{p=1}^{Q} C_{\alpha p}^* H'_{ps}}_{\substack{=(\varepsilon_\alpha-\mu)C_{\alpha s}^* \\ \text{at the exact solution}}} C_{\alpha s}^* \,.\end{aligned}$$

$\varepsilon_\alpha - \mu \leq 0$ for the lowest states, hence $\frac{\partial^2 \Omega}{\partial C_{\gamma s}^2} > 0$ at the exact solution, corresponding to the ground state (i.e., when N lowest Kohn-Sham levels are occupied). Therefore, the above formalism presents a *variational* approach.

Once we reduce our computational task to *unconstrained minimization* of the functional Ω (6.17/6.18) in one or another representation (in terms of either density matrix elements, or expansion coefficient over the orbital basis), the matrix multiplication remains the main time-consuming step, so it is advantageous to keep it fast, using sparse matrices. For maintaining the representation of either the density matrix, or of the orbitals $|\varphi_\alpha\rangle$ in a sparse form, it is essential to use well localized basis functions. Possible choices are:
– Gaussian-type orbitals (sparseness enforced by discarding overlaps smaller than a certain cutoff value);
– generalized Wannier functions (centered somewhere at the chemical bonds; falldown exponential for insulators and power-low for metals);
– *ad hoc* strictly localized functions, as e.g. Sankey-Niklewski functions[45]

It is essential for the control over the accuracy that the method to use is variational. In this case the localization constraint slightly increases the total energy, but keeping trace on the latter helps to choose an acceptable degree of localization for the basis functions. More details can be found in the review by Ordejón cited above.

[45] Otto F. Sankey and David J. Niklewski, *Ab initio multicenter tight-binding model for molecular-dynamics simulations and other applications in covalent systems*, Phys. Rev. B **40**, 3979–3995 (1989).