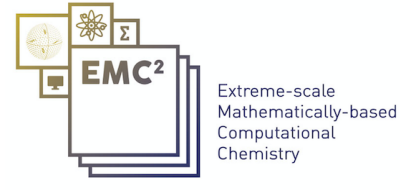
Analysis of SCF and minimization algorithms for electronic structure

Éric Cancès, **Gaspard Kemlin**, Antoine Levitt

gaspard.kemlin@enpc.fr
PhD student with É. Cancès & A. Levitt,
CERMICS, ENPC & Inria Paris, team MATHERIALS

EMC2 Seminar, October 16th 2020



Electronic structure

Our goal: compute the ground state energy of the many-body Hamiltonian H_e for a given system with N electrons

$$\min \left\{ \langle \Psi | H_e | \Psi \rangle \;\middle|\; \Psi = (\psi_i) \in \mathsf{L}^2(\mathbb{R}^3, \mathbb{C})^N,\ \langle \psi_i | \psi_j \rangle = \delta_{ij} \right\}.$$

$\rightsquigarrow$ too hard to solve !

1. approximation;
2. discretization;
3. resolution. $\leftarrow$ we focus on this

General framework

General form of the energy

$$E(P) := \mathrm{Tr}\left(H_0 P\right) + E_{\mathrm{nl}}(P),$$

where

- H_0 is the core Hamiltonian;
- E_{nl} models the electron-electron interaction depending on the model.

Examples: Kohn-Sham DFT, Hartree-Fock, Gross-Pitaevskii,...

General framework

- **Constrained minimization:**

$$\inf_{P \in \mathcal{M}_N} E(P) := \mathrm{Tr}\,(H_0 P) + E_{\mathrm{nl}}(P), \tag{1}$$

$$\mathcal{M}_N := \left\{ P \in \mathbb{R}^{N_b \times N_b} \mid P = P^*,\ \mathrm{Tr}(P) = N,\ P^2 = P \right\}.$$

- **Euler-Lagrange equations:**

$$\begin{cases} (H_0 + \nabla E_{\mathrm{nl}}(P))\phi_i = \varepsilon_i \phi_i,\ \varepsilon_1 \leqslant \cdots \leqslant \varepsilon_N \\ \phi_i^* \phi_j = \delta_{ij}, \\ P = \displaystyle\sum_{i=1}^{N} \phi_i \phi_i^*. \end{cases} \tag{2}$$

$\rightsquigarrow E_{\mathrm{nl}} = 0$: linear eigenvalue problem.

Mathematical framework

- $\mathcal{H} := \left(\mathbb{R}^{N_b \times N_b}_{\text{sym}}, \|\cdot\|_{\text{F}}\right)$ endowed with the Frobenius scalar product $\langle A, B\rangle_{\text{F}} := \text{Tr}(AB)$;
- $H(P) = \nabla E(P)$ and $K(P) := \Pi_P \nabla^2 E(P) \Pi_P$;
- Π_P is the orthogonal projection on $\mathcal{T}_P\mathcal{M}_N$:

$$\Pi_P(X) = PX(1-P) + (1-P)XP.$$

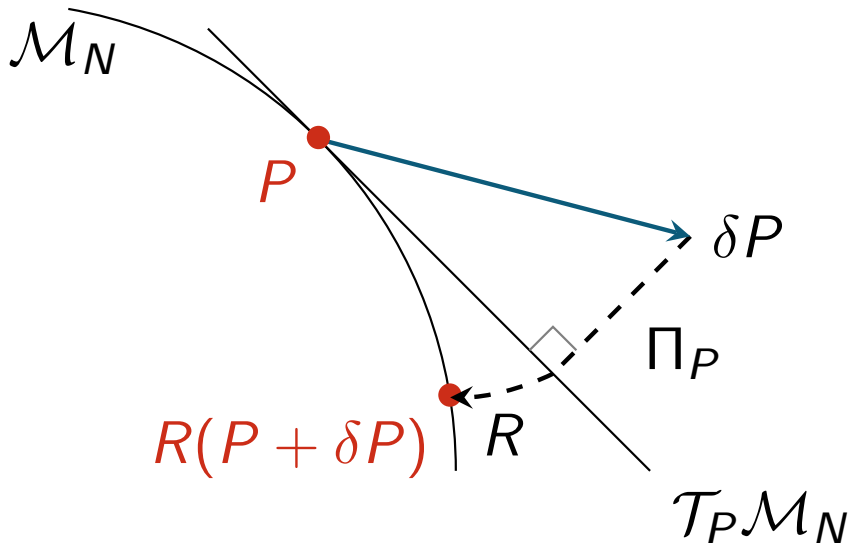
In the decomposition $\mathcal{H} = \text{Ran}(P) \oplus \text{Ran}(1-P)$, we have:

$$P = \begin{bmatrix} 1_N & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{T}_P\mathcal{M}_N := \left\{ X = \begin{bmatrix} 0 & \times \\ \times^* & 0 \end{bmatrix} \right\}.$$

Mathematical framework

- $R : \mathcal{H} \to \mathcal{M}_N$ is a retraction s.t.

$$R(P + \delta P) = P + \Pi_P \delta P + O(\delta P^2) \quad \text{for} \quad P \in \mathcal{M}_N.$$



Mathematical framework

Assumption 1 $E_{\mathrm{nl}} : \mathcal{H} \to \mathbb{R}$ is twice continuously differentiable, and thus so is E.

Assumption 2 $P_* \in \mathcal{M}_N$ is a nondegenerate local minimizer of (1) in the sense that there exists some $\eta > 0$ such that, for $P \in \mathcal{M}_N$ in a neighborhood of P_*, we have

$$E(P) \geqslant E(P_*) + \eta \|P - P_*\|_{\mathrm{F}}^2.$$

Let $H_* := H(P_*)$ and $K_* := K(P_*)$.

Approach

Study, analyze and compare two algorithms,

	SCF	**Gradient descent**
Classical	Damped SCF	Project gradient
With memory	Anderson	LBFGS

Table: Classes of algorithms

$\rightsquigarrow$ which one is better, why ?

Approach

Study, analyze and compare two algorithms,

	SCF	**Gradient descent**
Classical	Damped SCF	Project gradient
With memory	Anderson	LBFGS

Table: Classes of algorithms

$\rightsquigarrow$ which one is better, why ?

First order condition

The first-order optimality condition is $\Pi_{P_*}(H_*) = 0$, which gives

$$\boxed{P_* H_* (1 - P_*) = (1 - P_*) H_* P_* = 0}.$$

- $[H_*, P_*] = 0 \Rightarrow H_*$ and P_* can be codiagonalized;
- if $(\phi_i)_{1 \leqslant i \leqslant N_b}$ is an o.n.b. of eigenvectors of H_* ordered by increasing eigenvalues, then $P_* = \sum_{i \in \mathcal{I}} \phi_i \phi_i^*$, with $\mathcal{I}$ the set of occupied orbitals;
- $\mathcal{I} \subset \{1, \ldots, N_b\}$ and $|\mathcal{I}| = N$:
  - $\mathcal{I} = \{1, \ldots, N\}$: *Aufbau* principle;
  - $\mathcal{I} = \{1, \ldots, N\}$ and $\varepsilon_N < \varepsilon_{N+1}$: strong *Aufbau* principle.

Second order condition

The second order optimality condition reads

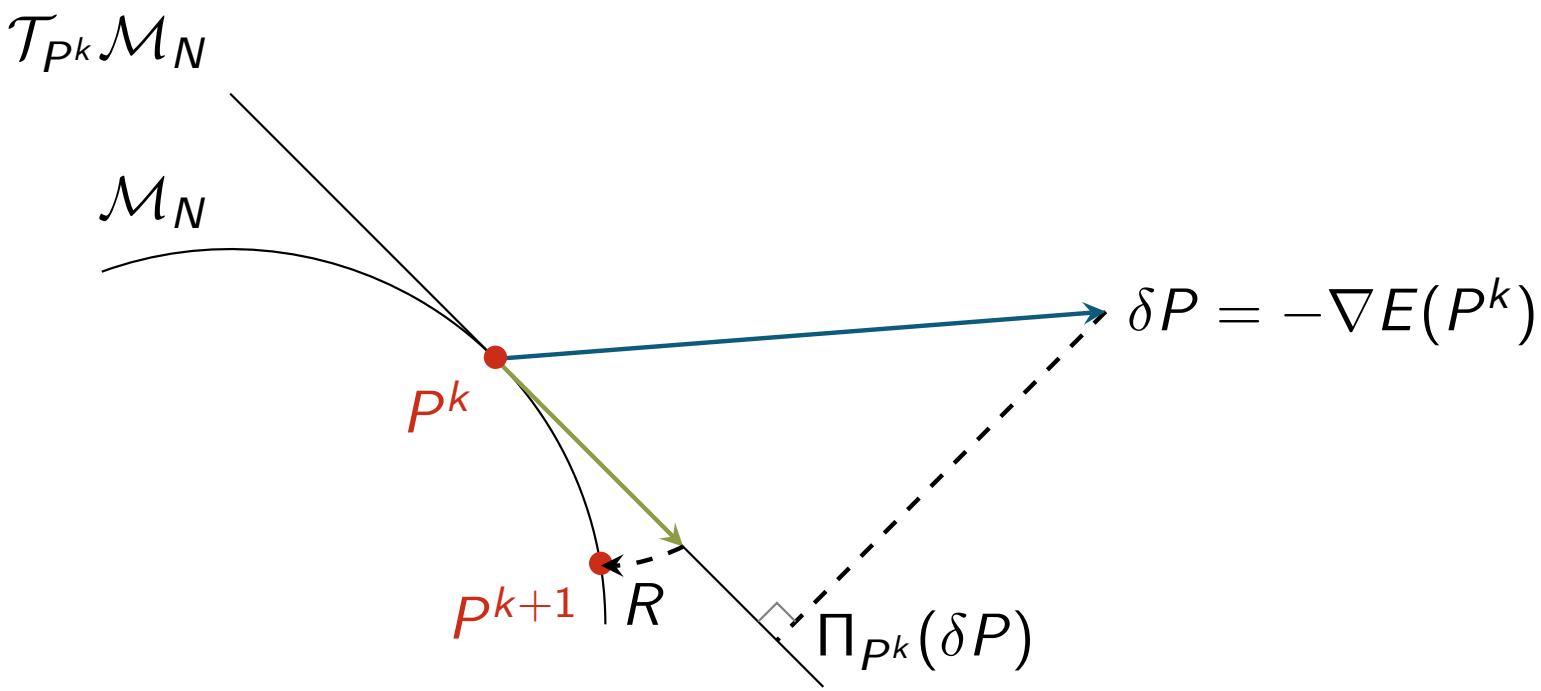
$$\boxed{\forall\, X \in \mathcal{T}_{P_*}\mathcal{M}_N,\ \langle X, (\Omega_* + K_*)X\rangle_{\mathsf{F}} \geqslant \eta \|X\|_{\mathsf{F}}^2}.$$

- the operator $\Omega_* : \mathcal{T}_{P_*}\mathcal{M}_N \to \mathcal{T}_{P_*}\mathcal{M}_N$ is defined by, for $i \in \mathcal{I}$ and $a \notin \mathcal{I}$,

$$(\Omega_* X)_{ia} = (\varepsilon_a - \varepsilon_i)X_{ia} \quad \text{and} \quad (\Omega_* X)_{ai} = (\varepsilon_a - \varepsilon_i)X_{ai},$$

  so that the gap $\nu = \min_{a\notin\mathcal{I}} \varepsilon_a - \max_{i\in\mathcal{I}} \varepsilon_i$ is the smallest eigenvalue of Ω_*;
- $\Omega_* + K_*$ can be interpreted as the Hessian of the energy on the manifold, Ω_* represents the influence of the curvature.

Projected gradient algorithm



Data: $P^0 \in \mathcal{M}_N$

while *convergence not reached* **do**

$$P^{k+1} := R\left(P^k - \beta \Pi_{P^k} \nabla E(P^k)\right);$$

end

Convergence

Theorem (Classical result)

Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 1 and 2 with P_* a local minimizer of (1). Then, if $P^0 \in \mathcal{M}_N$ is close enough to P_*, the iterations

$$P^{k+1} := R\left(P^k - \beta \Pi_{P^k} \nabla E(P^k)\right)$$

linearly converge to P_* for $\beta > 0$ small enough, with asymptotic rate $r(1 - \beta J_{\text{grad}})$ where $J_{\text{grad}} := \Omega_* + K_*$.

Proof: Apply lemma to

$$f : \left| \begin{array}{ccc} \mathcal{M}_N & \to & \mathcal{M}_N \\ P & \mapsto & R\left(P - \beta \Pi_P \left(\nabla E\left(P\right)\right)\right) \end{array} \right.$$

and show that $r\left(\mathrm{d}f(P_*)\right) < 1$:

1. the differential at P_* on the tangent plane $\mathcal{T}_{P_*}\mathcal{M}_N$ is

$$\mathrm{d}f(P_*) = 1 - \beta\left(K_* + \Omega_*\right);$$

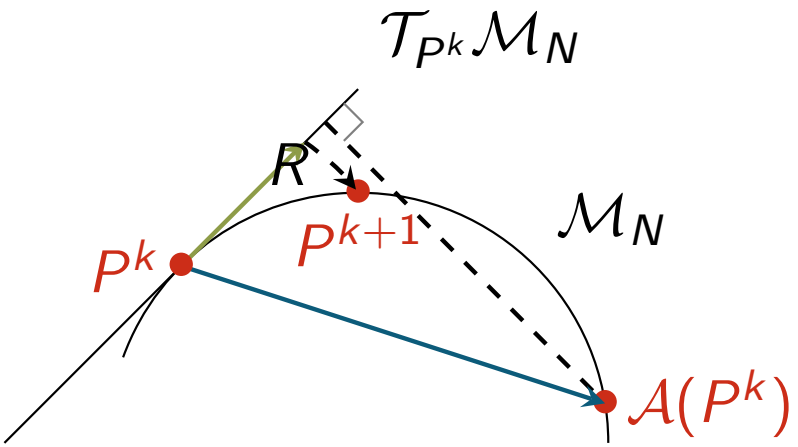
2. recall the second order condition:

$$K_* + \Omega_* \geqslant \eta > 0,$$

therefore, for β small enough, $r(\mathrm{d}f(P_*)) < 1$.

□

SCF algorithm



$$\mathcal{A}(P^k) = \sum_{i=1}^{N} \phi_i^k \left(\phi_i^k\right)^*$$

Data: $P^0 \in \mathcal{P}_N$

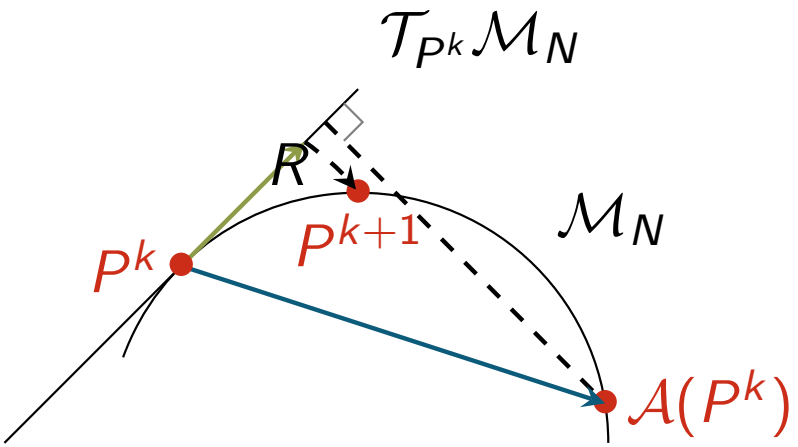
while *convergence not reached* **do**

solve $\begin{cases} H(P^k)\phi_i^k = \varepsilon_i^k \phi_i^k, \quad \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_N^k < \varepsilon_{N+1}^k \\ \left(\phi_i^k\right)^* \phi_j^k = \delta_{ij}, \end{cases}$;

$P^{k+1} := R\left(P^k + \beta \Pi_{P^k}\left(\mathcal{A}(P^k) - P^k\right)\right)$;

end

SCF algorithm



$$\mathcal{A}(P^k) = \sum_{i=1}^{N} \phi_i^k \left(\phi_i^k\right)^*$$

Data: $P^0 \in \mathcal{P}_N$

while *convergence not reached* **do**

$$\text{solve} \begin{cases} H(P^k)\phi_i^k = \varepsilon_i^k \phi_i^k, \quad \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_N^k < \varepsilon_{N+1}^k \\ \left(\phi_i^k\right)^* \phi_j^k = \delta_{ij}, \end{cases} ;$$

$$P^{k+1} := R\left(P^k + \beta \Pi_{P^k}\left(\mathcal{A}(P^k) - P^k\right)\right);$$

end

Convergence

Theorem

Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 1 and 2 with P_* a local minimizer of (1). Assume that P_* satisfies the strong Aufbau *principle*

$$\mathcal{A}(P_*) = P_* \text{ and } \nu := \varepsilon_{N+1} - \varepsilon_N > 0.$$

Then, for $\beta > 0$ small enough and $P^0 \in \mathcal{M}_N$ close enough to P_*, the iterations

$$P^{k+1} := R\left(P^k + \beta \Pi_{P^k}\left(\mathcal{A}(P^k) - P^k\right)\right)$$

linearly converge to P_*, with asymptotic rate $r(1 - \beta J_{\mathrm{SCF}})$ where $J_{\mathrm{SCF}} := 1 + \Omega_*^{-1} K_*$.

Proof: Apply lemma to

$$f : \left| \begin{array}{ccc} \mathcal{M}_N & \to & \mathcal{M}_N \\ P & \mapsto & R\left(P + \beta \Pi_P \left(\mathcal{A}(P) - P\right)\right) \end{array} \right.$$

and show that $r\left(\mathrm{d}f(P_*)\right) < 1$:

1. compute the differential of $\mathcal{A}$ on $\mathcal{T}_{P_*}\mathcal{M}_N$ with a perturbation method: $\mathrm{d}\mathcal{A}(P_*) = -\Omega_*^{-1} K_*$;
2. the differential at P_* on $\mathcal{T}_{P_*}\mathcal{M}_N$ is

$$\mathrm{d}f(P_*) = 1 - \beta(1 + \Omega_*^{-1} K_*);$$

3. $1 + \Omega_*^{-1} K_* \sim \Omega_*^{-1/2}(\Omega_* + K_*)\Omega_*^{-1/2}$ which has real positive eigenvalues by the second order condition and therefore, for β small enough, $r(\mathrm{d}f(P_*)) < 1$.

□

Comparing the Jacobian matrices

Both algorithms have Jacobian matrices of the form $1 - \beta J$ with

- Gradient descent: $J_{\text{grad}} = \Omega_* + K_*$ is self-adjoint for $\langle \cdot, \cdot \rangle_{\text{F}}$;
- SCF: $J_{\text{SCF}} = 1 + \Omega_*^{-1} K_*$ is self-adjoint for $\langle \Omega_* \cdot, \cdot \rangle_{\text{F}}$.

Hence

- in the linear regime, the SCF can be seen as a matrix splitting method for the gradient descent;
- the smaller the gap, the more difficult the convergence of the SCF.

Comparing the Jacobian matrices

Fastest convergence: eigenvalues of $1 - \beta J$ as close to 0 as possible. If λ_{min} and λ_{max} are the smallest and largest eigenvalues of J, the optimal step is $\beta_* = 2/(\lambda_{\text{min}} + \lambda_{\text{max}})$ and the rate of convergence is

$$r = \frac{\kappa - 1}{\kappa + 1} \quad \text{with} \quad \kappa = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}}.$$

Comparing the Jacobian matrices

$$J_{\text{grad}} = \Omega_* + K_*$$

$$\kappa(J_{\text{grad}}) \leqslant \frac{\|\Omega_*\|_{\text{op}} + \|K_*\|_{\text{op}}}{\eta}$$

- no relationship between η and the gap ν;
- the smaller η, the slower the convergence;
- the bigger $\|\Omega_*\|_{\text{op}} = \varepsilon_{N_b} - \varepsilon_1$, the slower the convergence (solved by preconditioning).

$$J_{\text{SCF}} = 1 + \Omega_*^{-1} K_*$$

$$\kappa(J_{\text{SCF}}) \leqslant \frac{1 + \nu^{-1} \|K_*\|_{\text{op}}}{\tilde{\eta}}$$

with $\tilde{\eta}$ independent of N_b (unif. coerc. assumption, often valid in practice)

- the smaller the gap ν, the slower the convergence (consistent with well-known issues).

A toy model with tunable gap

We consider the 2×2 real matrices P such that $P^2 = P = P^*$ and $\mathrm{Tr}(P) = 1$. Let

$$E_\varepsilon(P) := \mathrm{Tr}\left(\left(P - \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 0 \end{bmatrix}\right)^2\right).$$

Then, the gap $\nu(\varepsilon)$ behaves

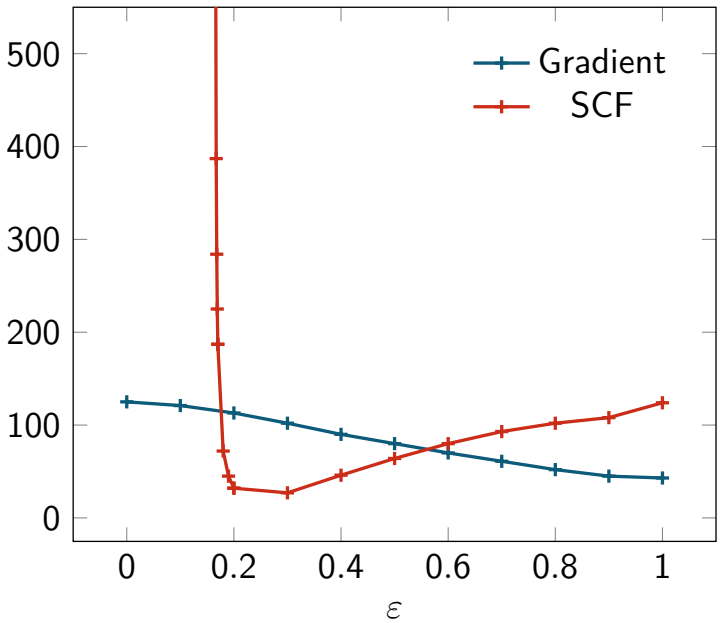
$$\nu(\varepsilon) \sim_{\varepsilon \to 0} 4\varepsilon^2$$

$\rightsquigarrow$ when $\varepsilon \to 0$, the gap goes to 0: suitable model to study the influence of the gap.

A toy model with tunable gap

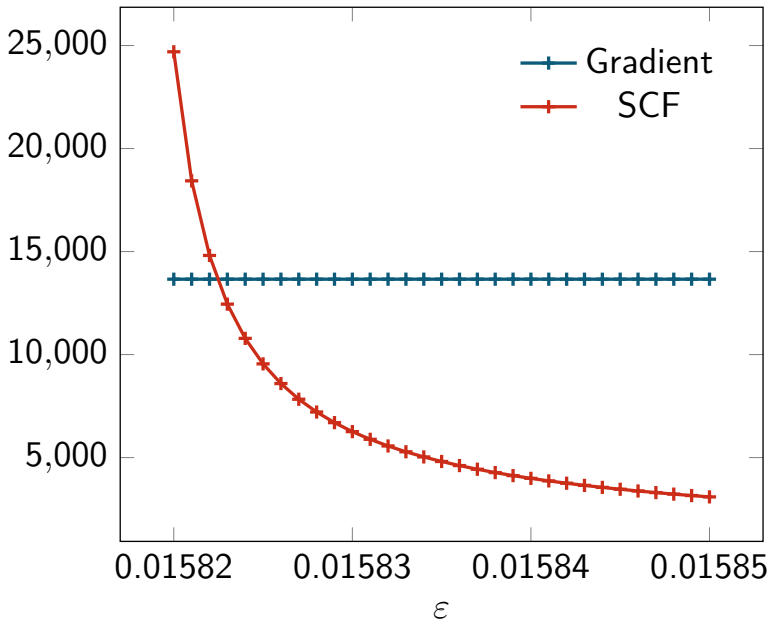
$1 + \Omega_*^{-1} K_*$ has a single eigenvalue $1 + \frac{2}{\nu(\varepsilon)} \approx_{\varepsilon \to 0} 1 + \frac{1}{2\varepsilon^2}$

$\Rightarrow$ convergence for $\beta < 4\varepsilon^2$ and for fixed β, $\varepsilon_{\text{crit}} = \sqrt{\beta/4}$.



$\beta = 10^{-1}$

$\varepsilon_{\text{crit}} \approx 0.158$



$\beta = 10^{-3}$

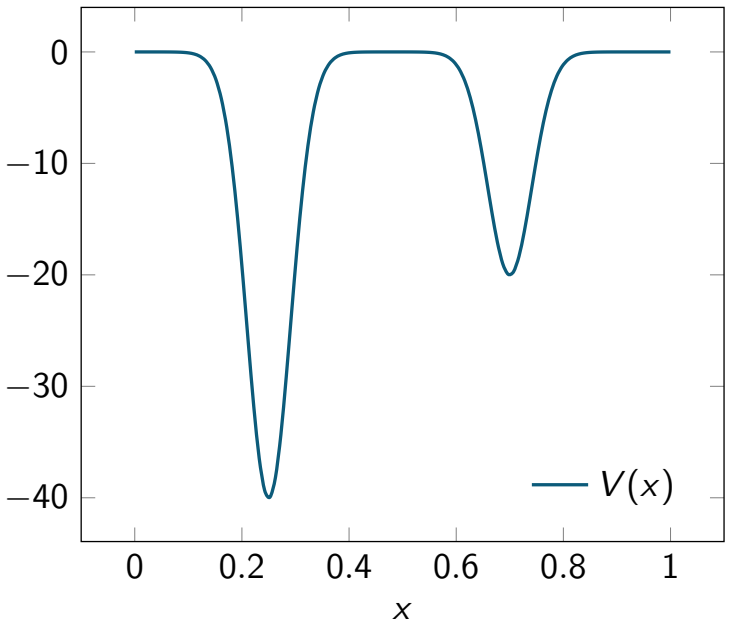
$\varepsilon_{\text{crit}} \approx 0.0158$

Local convergence for 1D Gross-Pitaevskii equation

We look at the periodic 1D GP equation on $[0, 1]$:

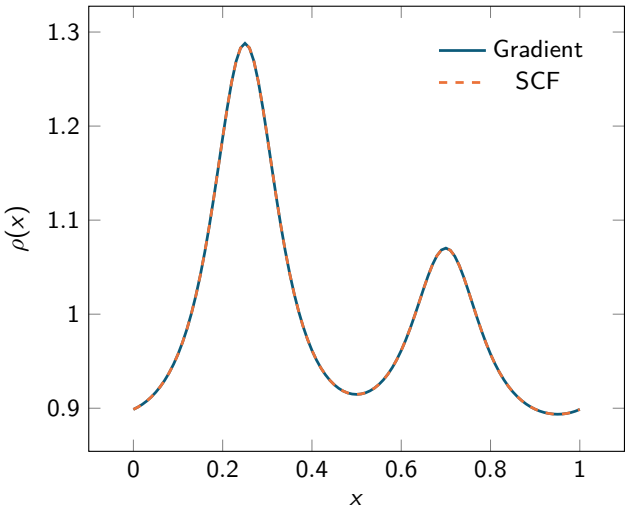
$$-\frac{1}{2}\Delta\phi_i + V\phi_i + \alpha\rho\phi_i = \varepsilon_i\phi_i, \quad \int_0^1 \phi_i\phi_j = \delta_{ij}, \quad \rho = \sum_{i=1}^{N} |\phi_i|^2,$$

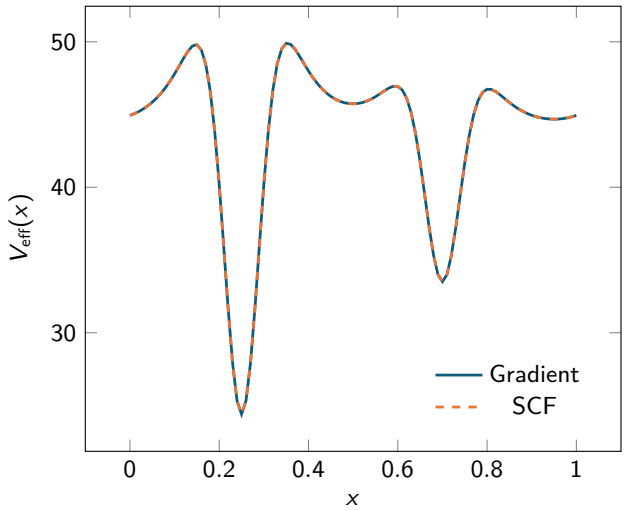
with V the following smooth potential



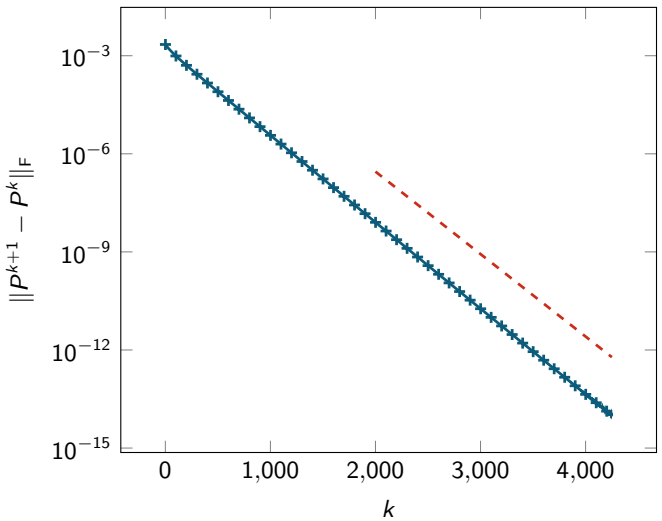
and then we discretize it with finite differences.

$N = 1$, $\alpha = 50$

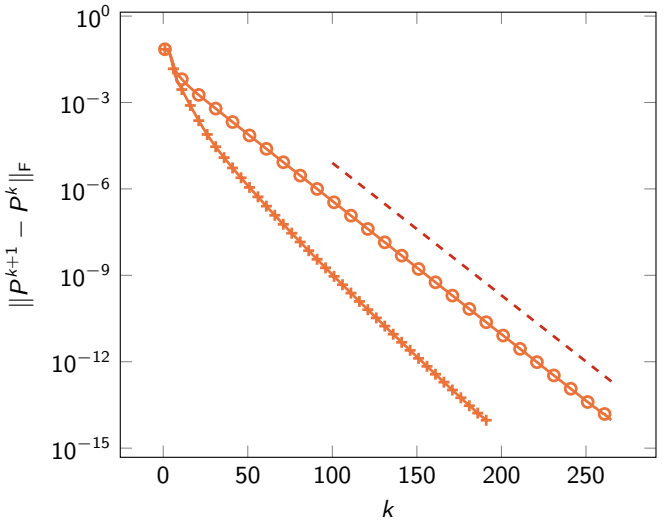




Gradient $\beta = 5 \cdot 10^{-5}$ (core start) asymptotic rate 0.994
slope log of spectral radius $r(1 - \beta J_{\text{grad}}) = 0.994$



SCF $\beta = 0.1$ (core start) asymptotic rate 0.89
SCF $\beta = 0.1$ (random start) asymptotic rate 0.90
slope log of spectral radius $r(1 - \beta J_{\text{SCF}}) = 0.90$



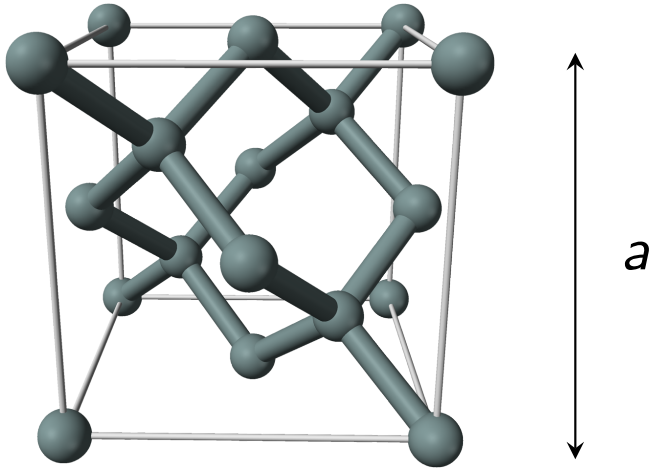
DFTK

Plane-wave basis `Julia` package for KS-DFT under a pseudo-potential approximation. It is developed by M. F. Herbst and A. Levitt at CERMICS.

More details on `https://dftk.org`.

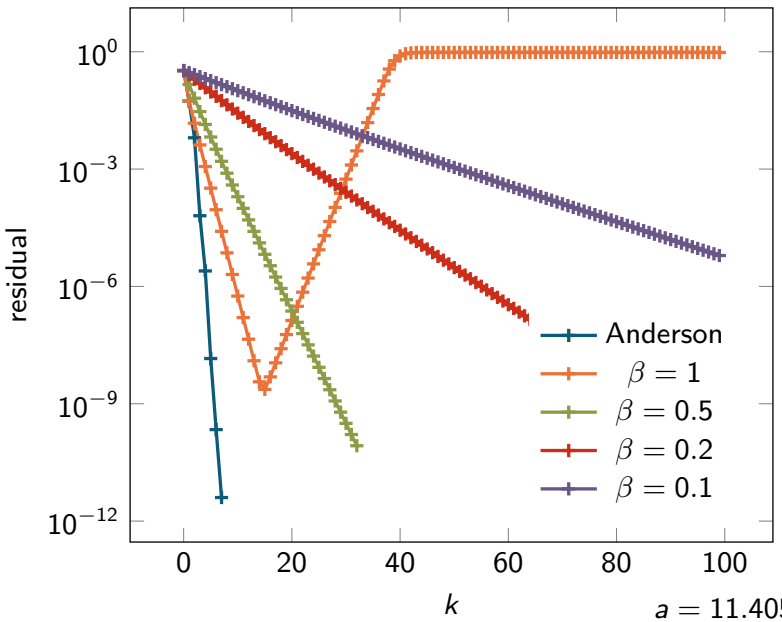
KS-DFT for a Silicon crystal

- LDA approximation;
- GTH pseudopotentials;
- cutoff energy = 30 Ha;
- Γ-only Brillouin zone.

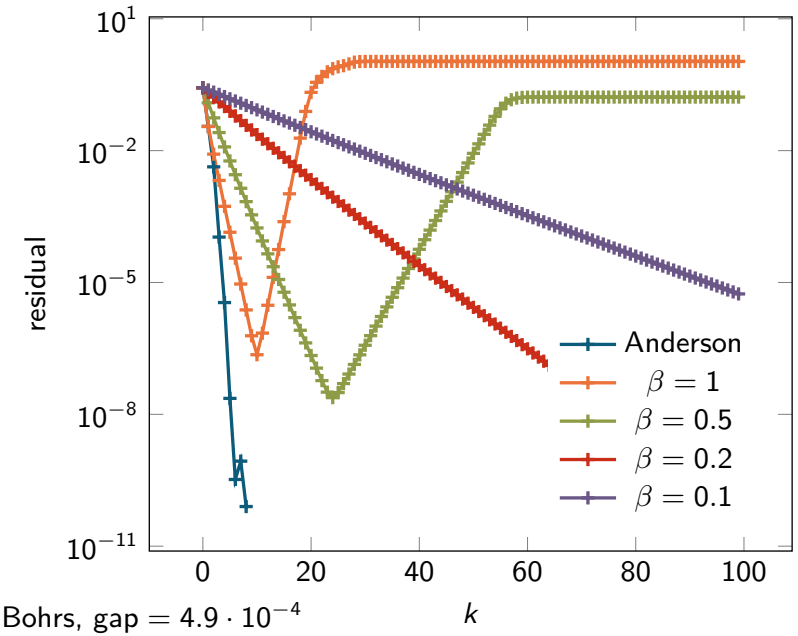


We vary the lattice constant a (which reduces the gap) and study the convergence for different steps.

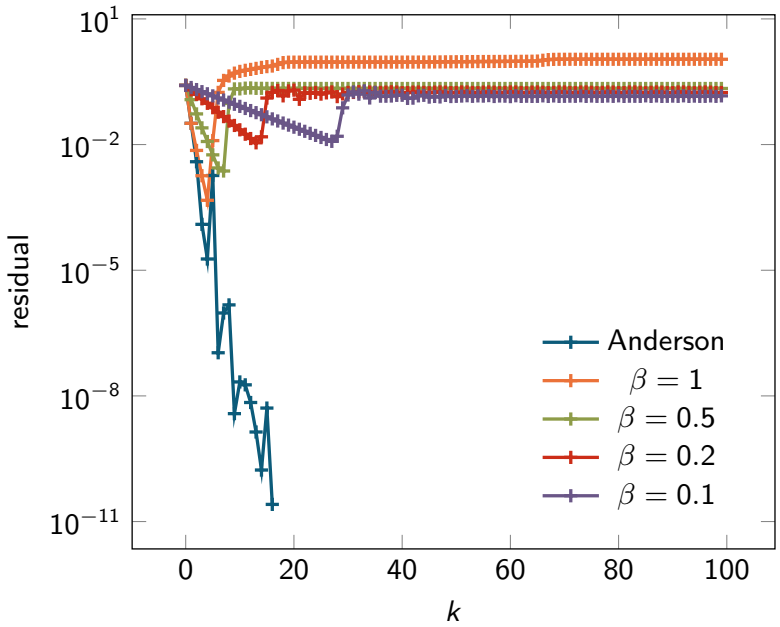
$a = 10.26$ Bohrs, gap $= 7.8 \cdot 10^{-2}$



$a = 11.2$ Bohrs, gap $= 1.8 \cdot 10^{-2}$



$a = 11.405$ Bohrs, gap $= 4.9 \cdot 10^{-4}$



Conclusion

- Both algorithms converge locally with different rates:
  - the SCF is sensitive to the gap;
  - the gradient is sensitive to the spectral radius of the Hamiltonian.
- In practice, which one should be preferred ? It depends on the convergence rate but also on the cost of each step.
  - quantum chemistry: computation of $H(P)$ is limiting $\rightsquigarrow$ both algorithms are of roughly equal cost; we tend to prefer the SCF for *Aufbau* solutions and the gradient otherwise;
  - condensed-matter: P and $H(P)$ are not store explicitly and the SCF is performed with block solvers $\rightsquigarrow$ gradient methods should more efficient and more robust (the step can be chosen to minimize the energy). In practice, SCF is more used for its tricks that are known to work for metallic problems often met in condensed-matter.

Ongoing works

This framework is useful to connect error and residual: in the linear regime

$$\underbrace{P - P_*}_{\text{error}} \approx (\Omega_* + K_*)^{-1} \underbrace{[P, [P, H(P)]]}_{\text{residual}}.$$

$\rightsquigarrow$ finding norms for which the operator norm above is not too big can be useful to derive good *a posteriori* estimators. We aim to implement this in DFTK at some point.

Ongoing works with Éric Cancès, Geneviève Dusson and Antoine Levitt.

Lemma

Let $f : \mathcal{M}_N \to \mathcal{M}_N$ and $P_* \in \mathcal{M}_N$ s.t. $f(P_*) = P_*$ and $r(\mathrm{d}f(P_*)) < 1$.
Then, for P^0 close enough to P_*, the fixed point iteration $P^{k+1} = f(P^k)$ linearly converges to P_* with asymptotic rate $r(\mathrm{d}f(P_*))$, in the sense that for all $\theta > 0$ there exists $C_\theta > 0$ s.t.

$$\left\| P^k - P_* \right\| \leqslant C_\theta \left(r(\mathrm{d}f(P_*)) + \theta \right)^k \left\| P^0 - P_* \right\|.$$

Details on $\mathrm{d}\mathcal{A}(P_*)$: $\mathcal{A} = A \circ H$ where, $A = \mathbf{1}_{(-\infty,\varepsilon_N]}$ and, by the strong *Aufbau* principle, there exists a contour $\mathcal{C}$ in the complex plane enclosing the lowest N eigenvalues of H_* such that, for H close to H_*,

$$A(H) = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H} \mathrm{d}z.$$

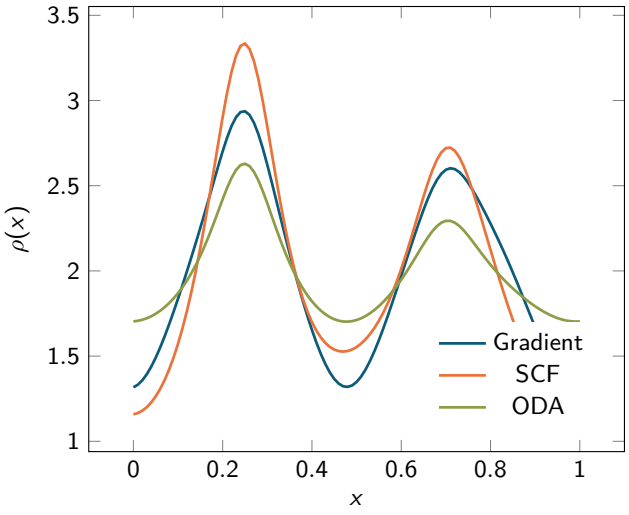
$$\begin{aligned}
\forall\ h \in \mathcal{H}, \quad \mathrm{d}A(H_*)h &= \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H_*} h \frac{1}{z - H_*}\ \mathrm{d}z \\
&= \sum_{k=1}^{N_b} \sum_{l=1}^{N_b} \left(\frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \varepsilon_k} h_{kl} \frac{1}{z - \varepsilon_l}\ \mathrm{d}z \right) \phi_k \phi_l^*, \\
&= \sum_{i=1}^{N} \sum_{a=N+1}^{N_b} \frac{1}{\varepsilon_i - \varepsilon_a} \left(h_{ia} \phi_i \phi_a^* + h_{ai} \phi_a \phi_i^* \right) = -\Omega_*^{-1} \Pi_{P_*} h,
\end{aligned}$$

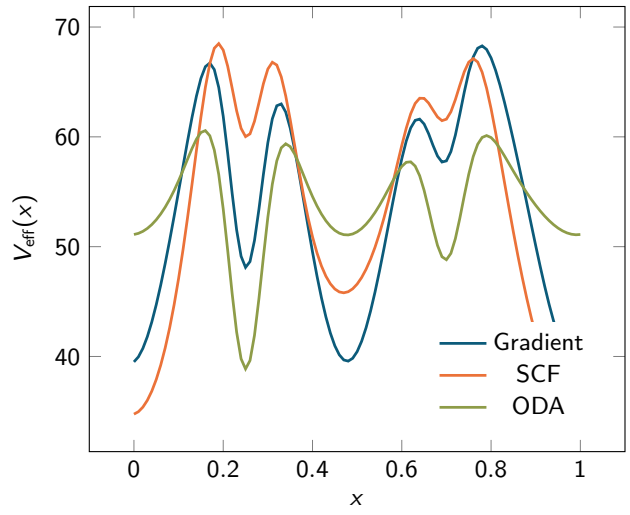
Finally,

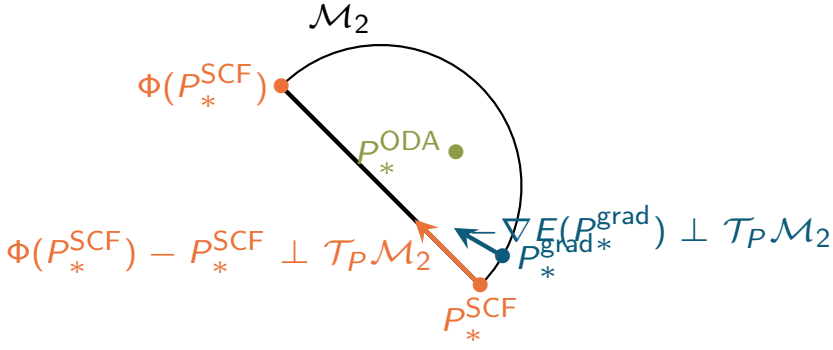
$$\forall\ X \in \mathcal{T}_{P_*}\mathcal{M}_N, \quad \mathrm{d}\mathcal{A}(P_*)X = \mathrm{d}A(H_*)\nabla^2 E(P_*)X = -\Omega_*^{-1}\Pi_{P_*}\nabla^2 E(P_*)\Pi_{P_*} X$$

$$\mathrm{d}\mathcal{A}(P_*)X = -\Omega_*^{-1} K_* X.$$

$N = 2,\ \alpha = 50$





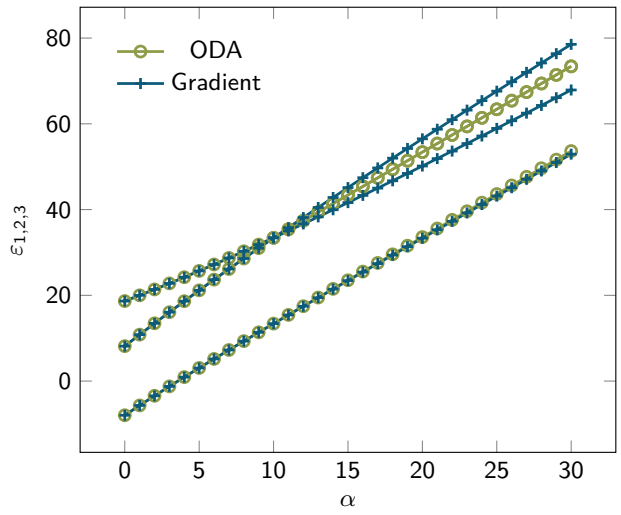


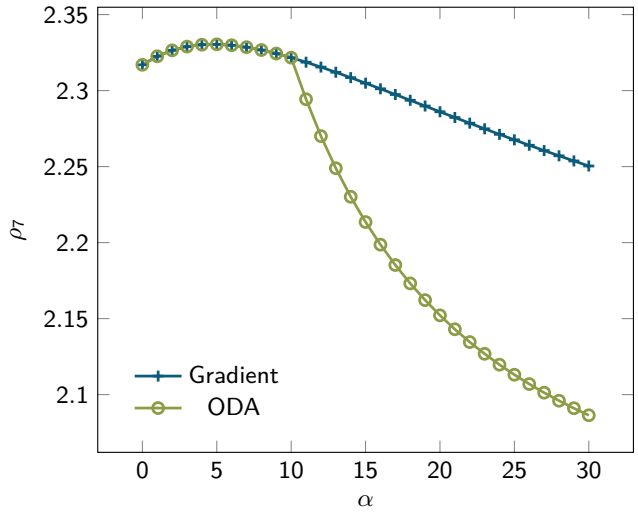
	Gradient	**SCF**	**ODA**
ε_1	52.9	51.3	53.6
ε_2	67.9	67.8	73.4
ε_3	78.5	79.6	73.4

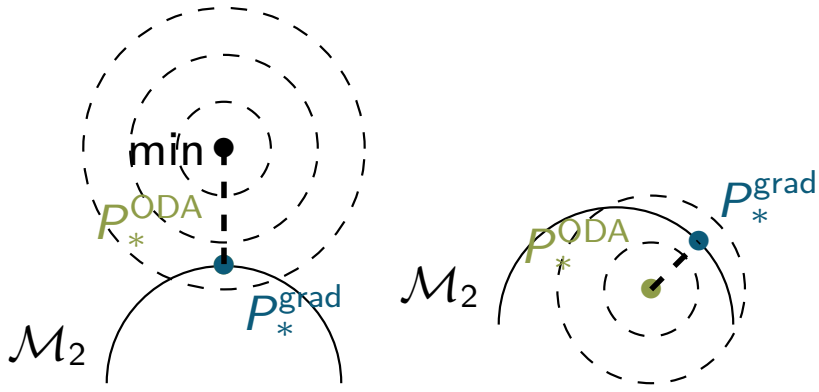
The solution reached by the gradient occupies ϕ_1 and ϕ_3, the one reached by ODA has fractional occupation numbers.

The SCF does not converge and we use ODA to explore the interior of the manifold.

$N = 2$, α varies







Before (left) and after (right) the bifurcation.