

Convergence analysis of direct minimization and self-consistent iterations

Eric Cancès, Gaspard Kemlin, Antoine Levitt

▶ **To cite this version:**

Eric Cancès, Gaspard Kemlin, Antoine Levitt. Convergence analysis of direct minimization and self-consistent iterations. SIAM Journal on Matrix Analysis and Applications, 2021, 42 (1), 243-274 (32 p.). 10.1137/20M1332864 . hal-02546060v2

HAL Id: hal-02546060

https://inria.hal.science/hal-02546060v2

Submitted on 27 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CONVERGENCE ANALYSIS OF DIRECT MINIMIZATION AND SELF-CONSISTENT ITERATIONS

ERIC CANCÈS, GASPARD KEMLIN, ANTOINE LEVITT

Abstract. This article is concerned with the numerical solution of subspace optimization problems, consisting of minimizing a smooth functional over the set of orthogonal projectors of fixed rank. Such problems are encountered in particular in electronic structure calculation (Hartree-Fock and Kohn-Sham Density Functional Theory – DFT – models). We compare from a numerical analysis perspective two simple representatives, the damped self-consistent field (SCF) iterations and the gradient descent algorithm, of the two classes of methods competing in the field: SCF and direct minimization methods. We derive asymptotic rates of convergence for these algorithms and analyze their dependence on the spectral gap and other properties of the problem. Our theoretical results are complemented by numerical simulations on a variety of examples, from toy models with tunable parameters to realistic Kohn-Sham computations. We also provide an example of chaotic behavior of the simple SCF iterations for a nonquadratic functional.

1. Introduction

This paper is concerned with the convergence behavior of algorithms to solve the *subspace optimization problem*

$$\min\left\{E(P)\;\middle|\;P\in\mathbb{R}^{N_b\times N_b},\;P^2=P=P^*,\;\mathrm{Tr}(P)=N\right\} \tag{1.1}$$

consisting of optimizing a C^2 function $E:\mathbb{R}^{N_b\times N_b}\to\mathbb{R}$ over the set of rank-N orthogonal projectors P. Here P^* denotes the adjoint (transpose) of P. This problem can also be reformulated as

$$\min\left\{E\left(\sum_{i=1}^N\phi_i\phi_i^*\right)\;\middle|\;\phi_i\in\mathbb{R}^{N_b},\;\phi_i^*\phi_j=\delta_{ij}\quad\forall\,i,j\in\{1,\dots,N\}\right\}, \tag{1.2}$$

using an orthonormal basis $(\phi_i)_{i=1,\dots,N}$ for the subspace $\mathrm{Ran}(P)$. This problem is of interest in a number of contexts, such as matrix approximation, computer vision [1], and electronic structure theory [12, 28, 39, 40, 45, 56], the latter of which being the main motivation for this work.

Let $H(P)=\nabla E(P)$. The first order conditions for problem (1.1) is

$$PH(P)(1-P)=(1-P)H(P)P=0.$$

Up to an appropriate choice for the orthonormal basis $(\phi_i)_{i=1,\dots,N}$ of $\mathrm{Ran}(P)$, this yields

$$H(P)\phi_i=\varepsilon_i\phi_i, \tag{1.3}$$

which reveals an alternative interpretation of this problem as a *nonlinear eigenvector problem* (to be distinguished from *nonlinear eigenvalue problems* of the form $A(\varepsilon)\phi=0$, where $A:\mathbb{R}\to\mathbb{R}^{N_b\times N_b}$). In the case when $E(P)=\mathrm{Tr}(H_0P)$ for a fixed symmetric matrix H_0, one recovers the classical eigenvalue problem $H_0\phi_i=\varepsilon_i\phi_i$. At a minimizer of (1.1), the $(\varepsilon_i)_{i=1,\dots,N}$ are the lowest eigenvalues of H_0, counting multiplicities.

Problems of the form (1.1) are found in the Hartree-Fock and Kohn-Sham theories of electronic structure [28, 45], both approximations of the many-body Schrödinger equation. In this context, the ϕ_i are (discretized) *orbitals*, the projector P is the *density matrix*, and the energy $E(P)$ includes linear contributions from the kinetic and external potential energy of the electrons, as well as nonlinear terms arising from electron-electron interaction. Another notable problem of this form is the nonlinear Schrödinger or Gross-Pitaevskii equation for Bose-Einstein condensates [6], where $N=1$. In all these cases, the first-order condition (1.3) is interpreted as a *self-consistent* or *mean-field* equation: the particles behave as independent particles in an effective Hamiltonian $H(P)$ (also known as the Fock matrix) involving the

Date: October 27, 2020.

mean-field they create. In the rest of this paper, we will work on the formulation (1.1) without specifying E for generality.

The minimization problem (1.1) is compact but nonconvex: there exists at least one minimizer, but the minimizer might not be unique, and local minima might not be global ones. Solving this optimization problem is of considerable practical interest, and algorithms for doing so date back to the early days of quantum mechanics [26]. The first introduced and still most popular approach is the *self-consistent field* (SCF) method, which, in its original version [48, 54], works as follows: if P^k is the current iterate of the algorithm, P^{k+1} is found by solving (1.3) for the fixed matrix $H(P^k)$:

$$H(P^k)\phi_i^k = \varepsilon_i^k \phi_i^k, \quad (\phi_i^k)^* \phi_j^k = \delta_{ij}$$

with the ε_i^k sorted in non-decreasing order, and building P^{k+1} as

$$P^{k+1} = \sum_{i=1}^{N} \phi_i^k (\phi_i^k)^*.$$

This algorithm assumes the *Aufbau* property, which is that at a minimum P_* we have $P_* = \sum_{i=1}^N \phi_i \phi_i^*$ with ϕ_i a system of orthogonal eigenvectors associated with the lowest N eigenvalues of $H(P_*)$. This property holds for the (spin-unconstrained) Hartree-Fock model [4] and the Gross-Pitaevskii models without magnetic field [10], usually holds for molecular systems in the Kohn-Sham model, but does not hold in general for Gross-Pitaevskii models with strong magnetic fields.

This basic procedure converges for systems where the nonlinearity is weak, but fails to converge otherwise (see [14] for a comprehensive mathematical analysis of this behavior when the functional E is a sum of a linear and a quadratic term in P, which is the case for the Hartree-Fock model). A solution is to *damp* this procedure, and *mix* the iterates to accelerate convergence. This gives rise to a variety of SCF algorithms, among which Broyden-like and Anderson-like mixing algorithms [33, 44, 52, 57], the Direct Inversion in the Iterative Space (DIIS) algorithm [36, 50, 51], the Optimal Damping Algorithm [13] (ODA), and the Energy-DIIS (EDIIS) algorithm combining the latter two approaches [37].

A second class of algorithms solves the minimization problem (1.1) directly. The minimization set $\{P \in \mathbb{R}^{N_b \times N_b},\ P^2 = P^* = P,\ \operatorname{Tr} P = N\}$ is diffeomorphic to the Grassmann manifold of the N-dimensional vector subspaces of $\mathbb{R}^{N_b}$. This set is naturally equipped with the structure of a Riemannian manifold, and this allows the use of Riemann optimization algorithms [1, 21]. Direct minimization algorithms are preferred for the Gross-Pitaevskii model with magnetic fields [3, 18, 27, 29], for which the *Aufbau* principle is not satisfied in general. Gradient-type [2, 17, 49, 61, 66], Newton-type [5, 15, 68], and trust-region methods have also been designed to solve (1.1) for larger values of N. At the time of writing, direct minimization algorithms are less popular than SCF algorithms in electronic structure calculation, where N can be very large, but it is not clear whether this is for sound scientific reasons or because SCF algorithms have been implemented and optimized for decades in the main production codes, which has not been the case for direct minimization algorithms.

While the convergence of several SCF and direct minimization algorithms has been analyzed from a mathematical point of view (see e.g. [16, 38, 42, 53, 60, 64] and references therein), the two approaches have not been compared in a systematic way to our knowledge. The purpose of this paper is to contribute to fill this gap, by focusing on very simple representatives of each class, namely the damped SCF iteration and the gradient descent. We emphasize that neither of these two algorithms is a practical choice as is. The SCF iteration should be accelerated (for instance using the Anderson acceleration technique), and the gradient information in direct minimization methods should rather be used as part of a quasi-Newton method (such as the limited-memory BFGS algorithm [1]). Depending on the exact problem at hand, all these methods should be preconditioned to avoid issues related to small mesh sizes (which leads to a divergence of the kinetic energy term) and/or large computational domains (which can lead to a divergence of the Coulomb energy, or the confining potential). We refer to [63] for a recent review in the context of the Kohn-Sham equations for solids. Rather, in this paper, we aim to focus on the very simplest representative of each general strategy (SCF and direct minimization). The investigation of these two basic algorithms is informative on the strengths and weaknesses of the two classes, and is a first step in the analysis of more complex methods.

The paper is organized as follows. In Section 2, we recall some results about optimization on Grassmann manifolds, in particular the first and second order optimality conditions, and prove preparatory

lemmas. In Section 3, we present the two algorithms that are in the scope of this paper: a fixed-step gradient descent and a damped SCF algorithm. We prove their local convergence as long as the step is small enough and we derive convergence rates. We find that the convergence rates depend on the spectral radius of operators (acting on $\mathbb{R}^{N_b \times N_b}$) of the form $1-\beta J$, with β the fixed step and $J = \Omega_* + K_*$ for the gradient descent, $J = 1 + {\Omega_*}^{-1} K_*$ for the SCF algorithm, where the operators Ω_* and K_* are specified in the next section. Let us just mention at this stage that the lowest eigenvalue of Ω_* is equal to the spectral gap between the N^{th} and $(N+1)^{\text{st}}$ eigenvalues of $H(P_*)$, allowing us to analyze the convergence rates of the algorithms in terms of natural quantities of the problem. This also shows that the damped SCF algorithm can be seen as a matrix splitting of the fixed-step gradient descent algorithm.

In Section 4, we compare the two algorithms on several test problems. First, we focus on a toy model for which we can easily tune the gap and observe some fundamental differences between SCF and direct minimization algorithms, in agreement with the mathematical results established in Section 3. We also provide an example of chaos in SCF iterations, complementing the results of [14, 38] in the case of a non-quadratic objective functional E. Then, we analyse a 1D Gross-Pitaevskii model ($N = 1$) and its fermionic counterpart for $N = 2$, for which we investigate the behavior of the algorithms when the gap closes. We conclude with an example from electronic structure calculation: a Silicon crystal, in the framework of Kohn-Sham DFT, where we show in particular that accelerated SCF algorithms are less sensitive to small gaps than the simple damped SCF. Finally, in Section 5 we draw conclusions and outline perspectives for future work.

2. Optimization on Grassmann manifolds

We focus in this paper on the case of real symmetric matrices, but the study can be easily extended to complex hermitian matrices. Let $\mathcal{H} \coloneqq \mathbb{R}^{N_b \times N_b}_{\text{sym}}$ be the vector space of $N_b \times N_b$ real symmetric matrices endowed with the Frobenius inner product $\langle A, B \rangle_{\text{F}} \coloneqq \text{Tr}(AB)$. Let

$$\mathcal{M} \coloneqq \left\{ P \in \mathcal{H} \;\middle|\; P^2 = P \right\} \text{ and } \mathcal{M}_N \coloneqq \left\{ P \in \mathcal{H} \;\middle|\; P^2 = P,\ \text{Tr}(P) = N \right\}.$$

From a geometrical point of view, $\mathcal{M}$ is a compact subset of $\mathcal{H}$ with $N_b + 1$ connected components $\mathcal{M}_N$, $N = 0, \dots, N_b$, each of them being characterized by the value of $\text{Tr}(P)$, namely the rank of the orthogonal projector P, and being diffeomorphic to the Grassmann manifold $\text{Grass}(N, N_b)$ [1]. From now on, we fix the number of electrons N and we seek the local minimizers of the problem

$$\min_{P \in \mathcal{M}_N} E(P), \tag{2.1}$$

where $E : \mathcal{H} \to \mathbb{R}$ is a discretized energy functional, for which some examples are given below.

Example 2.1. As an example, we study a discrete Gross-Pitaevskii model in Section 4.4. Other models from electronic structure can be considered, such as the discretized Hartree-Fock or Kohn-Sham models, where the energy is of the form

$$E(P) \coloneqq \text{Tr}(H_0 P) + E_{\text{nl}}(P)$$

with H_0 being the core Hamiltonian (representing the kinetic energy and the external potential) and E_{nl} a nonlinear energy functional depending on the model (representing the interaction between electrons). For instance, for the Hartree-Fock model,

$$E_{\text{nl}}(P) \coloneqq \frac{1}{2} \text{Tr}(G(P)P) \qquad \text{where} \qquad (G(P))_{ij} \coloneqq \sum_{k,l=1}^{N_b} A_{ijkl} P_{kl} \quad \forall\, i, j = 1, \dots, N_b,$$

with A a symmetric tensor of order 4. For more details on these models or electronic structure in general, we refer to [12, 40, 56].

In plane-wave, finite differences, finite elements or wavelets electronic structure calculation codes, the size N_b of the discretized space is in practice much larger than the number N of electrons. Therefore, it is not practical to store and manipulate the (dense) matrix P. Instead, algorithms work on the orbitals $(\phi_i)_{i=1,\dots,N}$ introduced in (1.3). The density matrix P is then recovered as

$$P = \sum_{i=1}^{N} \phi_i \phi_i^*.$$

All the results in this article are presented in the density matrix framework. However, the algorithms we study can be expressed in a way that avoids ever forming the density matrix. We refer to [63] for details.

We will need two assumptions for our results.

Assumption 2.2. *The energy functional $E : \mathcal{H} \to \mathbb{R}$ is of class C^2 (twice continuously differentiable).*

Assumption 2.2 is true for Hartree-Fock models. For Kohn-Sham models, it is true when the density $\rho = \sum_{i=1}^{N} |\phi_i|^2$ is uniformly bounded avay from zero, which is the case for instance in condensed phase systems. Most of the results presented in this article are local in nature, and therefore this assumption can be relaxed to local regularity.

Assumption 2.3. *$P_* \in \mathcal{M}_N$ is a nondegenerate local minimizer of (2.1) in the sense that there exists some $\eta > 0$ such that, for $P \in \mathcal{M}_N$ in a neighborhood of P_*, we have*

$$E(P) \geqslant E(P_*) + \eta \left\|P - P_*\right\|_F^2 .$$

It is very hard in most practical situations to check this assumption, but it seems to be verified in practice. Notable exceptions are systems invariant with respect to continuous symmetry groups, in which case $E(P) = E(P_*)$ for all P in the orbit of P_* along the symmetry group. In this case, the assumption can not be true, and $\|P - P_*\|_{\mathrm{F}}$ must be replaced by the distance from P to the orbit of P_*. Our results can be extended to this case up to quotienting $\mathcal{H}$ by the symmetry group.

Throughout the paper, we will use the following notation:

- $H(P) \coloneqq \nabla E(P)$ is the gradient, and $H_* \coloneqq H(P_*)$;
- $K(P) \coloneqq \Pi_P \nabla^2 E(P) \Pi_P$ is the Hessian projected onto the tangent space at P, and $K_* \coloneqq K(P_*)$ (the projection Π_P is defined below in Proposition 2.4).

2.1. **First-order condition.** To study the first-order optimality conditions, we start by recalling some classical results about the geometry of the manifold $\mathcal{M}_N$.

Proposition 2.4. *$\mathcal{M}_N$ is a smooth real manifold and its tangent space $T_P\mathcal{M}_N$ at $P \in \mathcal{M}_N$ is given by*

$$T_P\mathcal{M}_N = \{X \in \mathcal{H} \mid PX + XP = X,\ \mathrm{Tr}(X) = 0\} = \{X \in \mathcal{H} \mid PXP = (1-P)X(1-P) = 0\} .$$

The orthogonal projection Π_P on $T_P\mathcal{M}_N$ for the Frobenius inner product is

$$\forall\ X \in \mathcal{H}, \quad \Pi_P(X) = PX(1-P) + (1-P)XP = [P, [P, X]], \tag{2.2}$$

where $[A, B] \coloneqq AB - BA$.

This classical result is proved in e.g. [1, Section 3.4]. Using the fact that $\mathcal{H} = \mathrm{Ran}(P) \oplus \mathrm{Ran}(1-P)$ and the induced decomposition of $P \in \mathcal{M}_N$ and $X \in \mathcal{H}$ as

$$P = \begin{bmatrix} I_N & 0 \\ 0 & 0 \end{bmatrix}, \quad X = \begin{bmatrix} (X)_{\mathrm{oo}} & (X)_{\mathrm{ov}} \\ (X)_{\mathrm{vo}} & (X)_{\mathrm{vv}} \end{bmatrix}, \tag{2.3}$$

the projection Π_P is given by

$$\Pi_P(X) = \begin{bmatrix} 0 & (X)_{\mathrm{ov}} \\ (X)_{\mathrm{vo}} & 0 \end{bmatrix} .$$

Here the subscript "o" (resp. "v") stand for *occupied* (resp. *virtual*). The first-order optimality condition at P_* is $\Pi_{P_*}(H_*) = 0$, which can be formulated as follows:

$$\boxed{\textbf{First-order optimality condition: } P_* H_* (1 - P_*) = (1 - P_*) H_* P_* = 0.} \tag{2.4}$$

Note that this condition can be rewritten as $[H_*, P_*] = 0$, showing that H_* and P_* can be codiagonalized. Let $(\phi_k)_{1 \leqslant k \leqslant N_b}$ be an orthonormal basis of eigenvectors of H_* associated with the eigenvalues $(\varepsilon_k)_{1 \leqslant k \leqslant N_b}$ sorted in ascending order. Then $P = \sum_{i \in \mathcal{I}} \phi_i \phi_i^*$, where $\mathcal{I} \subset \{1, \dots, N_b\}$, $|\mathcal{I}| = N$ is the set of occupied orbitals. The minimizer P_* is said to satisfy

- the *Aufbau* principle if $\mathcal{I} = \{1, \dots, N\}$;
- the strong *Aufbau* principle if $\mathcal{I} = \{1, \dots, N\}$ and if in addition $\varepsilon_N < \varepsilon_{N+1}$, in which case $P_* = \sum_{i=1}^{N} \phi_i \phi_i^*$.

2.2. **Second-order condition.** We derive here the second-order optimality condition from the nondegeneracy of the minimum (Assumption 2.3).

Let $X \in T_{P_*}\mathcal{M}_N$, I be a real interval containing 0 and $\gamma : I \to \mathcal{M}_N$ be a smooth path such that $\gamma(0) = P_*$ and $\dot\gamma(0) = X$. An example of a possible γ is given in Section 4.1. We expand

$$\begin{aligned} E\left(\gamma(t)\right) &= E(P_*) + t\left\langle H_*, X\right\rangle_{\mathrm{F}} + \frac{t^2}{2}\Big(\left\langle H_*, \ddot\gamma(0)\right\rangle_{\mathrm{F}} + \left\langle X, \nabla^2 E(P_*)X\right\rangle_{\mathrm{F}}\Big) + o(t^2) \\ &= E(P_*) + \frac{t^2}{2}\Big(\left\langle H_*, \ddot\gamma(0)\right\rangle_{\mathrm{F}} + \left\langle X, K_*X\right\rangle_{\mathrm{F}}\Big) + o(t^2) \end{aligned}$$

as H_* is orthogonal to $T_{P_*}\mathcal{M}_N$ at the minimum. Differentiating the relation $\gamma(t)^2 = \gamma(t)$ at $t = 0$, we get

$$P_*\ddot\gamma(0) + \ddot\gamma(0)P_* + 2X^2 = \ddot\gamma(0),$$

from which we obtain the following two relations on the diagonal blocks of $\ddot\gamma(0)$ in the decomposition (2.3):

$$\frac{1}{2}\left(\ddot\gamma(0)\right)_{\mathrm{oo}} = -\left(X^2\right)_{\mathrm{oo}} = -\left(X\right)_{\mathrm{ov}}\left(X\right)_{\mathrm{vo}}, \qquad \frac{1}{2}\left(\ddot\gamma(0)\right)_{\mathrm{vv}} = \left(X^2\right)_{\mathrm{vv}} = \left(X\right)_{\mathrm{vo}}\left(X\right)_{\mathrm{ov}}.$$

Thus, since $\left(H_*\right)_{\mathrm{vo}} = \left(H_*\right)_{\mathrm{ov}}^* = 0$ at the minimum, we have

$$\begin{aligned} \left\langle H_*, \ddot\gamma(0)\right\rangle_{\mathrm{F}} &= \mathrm{Tr}\left(\begin{bmatrix} \left(H_*\right)_{\mathrm{oo}} & 0 \\ 0 & \left(H_*\right)_{\mathrm{vv}} \end{bmatrix} \begin{bmatrix} \left(\ddot\gamma(0)\right)_{\mathrm{oo}} & \left(\ddot\gamma(0)\right)_{\mathrm{ov}} \\ \left(\ddot\gamma(0)\right)_{\mathrm{vo}} & \left(\ddot\gamma(0)\right)_{\mathrm{vv}} \end{bmatrix}\right) \\ &= 2\,\mathrm{Tr}\Big(-\left(H_*\right)_{\mathrm{oo}}\left(X\right)_{\mathrm{ov}}\left(X\right)_{\mathrm{vo}}\Big) + 2\,\mathrm{Tr}\Big(\left(H_*\right)_{\mathrm{vv}}\left(X\right)_{\mathrm{vo}}\left(X\right)_{\mathrm{ov}}\Big) \\ &= 2\,\mathrm{Tr}\Big(-\left(X\right)_{\mathrm{vo}}\left(H_*\right)_{\mathrm{oo}}\left(X\right)_{\mathrm{ov}}\Big) + 2\,\mathrm{Tr}\Big(\left(X\right)_{\mathrm{ov}}\left(H_*\right)_{\mathrm{vv}}\left(X\right)_{\mathrm{vo}}\Big) \\ &= \mathrm{Tr}\Big(X(\Omega_* X)\Big), \end{aligned}$$

where the operator $\Omega_* : T_{P_*}\mathcal{M}_N \to T_{P_*}\mathcal{M}_N$ is defined as

$$\begin{aligned} \Omega_* X &\coloneqq P_*X(1-P_*)H_* - H_*P_*X(1-P_*) + \mathrm{sym} \\ &= \begin{bmatrix} 0 & \left(X\right)_{\mathrm{ov}}\left(H_*\right)_{\mathrm{vv}} - \left(H_*\right)_{\mathrm{oo}}\left(X\right)_{\mathrm{ov}} \\ \left(H_*\right)_{\mathrm{vv}}\left(X\right)_{\mathrm{vo}} - \left(X\right)_{\mathrm{vo}}\left(H_*\right)_{\mathrm{oo}} & 0 \end{bmatrix}, \end{aligned} \tag{2.5}$$

where "sym" stands for the transpose of the previous expression. Introducing the operator

$$\Omega_* + K_* : T_{P_*}\mathcal{M}_N \to T_{P_*}\mathcal{M}_N, \tag{2.6}$$

one gets in the end

$$E\left(\gamma(t)\right) = E(P_*) + \frac{t^2}{2}\left\langle X, (\Omega_* + K_*)X\right\rangle_{\mathrm{F}} + o(t^2).$$

At the critical point P_*, the second order expansion of $E\left(\gamma(t)\right)$ only depends on $X = \dot\gamma(0)$, a common feature in constrained optimization. The operator $\Omega_* + K_*$ can be interpreted as the Hessian of the energy on the manifold, or alternatively as the partial Hessian of the Lagrangian on $\mathcal{H}$. The operator Ω_* represents the influence of the curvature of the manifold on the Hessian of E.

As P_* is a nondegenerate minimum in the sense of Assumption 2.3, we have the

$$\boxed{\textbf{Second-order optimality condition: } \forall\ X \in T_{P_*}\mathcal{M}_N, \quad \left\langle X, (\Omega_* + K_*)X\right\rangle_{\mathrm{F}} \geqslant \eta\left\|X\right\|_{\mathrm{F}}^2.} \tag{2.7}$$

Remark 2.5 (Structure of Ω_* and link with the *Aufbau* principle). Let P_* be a nondegenerate minimizer of (2.1) in the sense of Assumption 2.3. Denoting by A_{kl} the component along $\phi_k\phi_l^*$ of the matrix $A \in \mathcal{H}$, the operator Ω_* defined in (2.5) can alternatively be defined by

$$\forall\ X \in T_{P_*}\mathcal{M}_N, \quad (\Omega_* X)_{ia} = (\varepsilon_a - \varepsilon_i)X_{ia} \text{ and } (\Omega_* X)_{ai} = (\varepsilon_a - \varepsilon_i)X_{ai} \text{ for } i \in \mathcal{I},\ a \notin \mathcal{I},$$

where we have used the standard notation in chemistry of using the subscripts i for occupied and a for virtual orbitals ($\mathcal{I}$ is the set of occupied orbitals).

In the case when $E(D) = \mathrm{Tr}(HD)$ for some fixed symmetric matrix $H \in \mathcal{H}$ (linear eigenvalue problem), then $K_* = 0$ and so (2.7) is equivalent to the *Aufbau* principle. This equivalence does not hold true in general for nonlinear models: (2.7) is independent of the Aufbau principle, and η is unrelated to the gap $\nu = \min_{a\notin\mathcal{I}}\varepsilon_a - \max_{i\in\mathcal{I}}\varepsilon_i$ (equal to the lowest eigenvalue of the operator Ω_*). However in specific cases, such as the reduced Hartree-Fock or Gross-Pitaevskii model, where $K_* \geqslant 0$, we have $\eta \geqslant \nu$ and a positive gap is a sufficient (but not necessary) condition for optimality.

Remark 2.6 (Link with the Liouvillian). Another way to understand Ω_* is to use the Liouvillian $\mathcal{L}_{H_*}$ associated to H_*, which is defined by:

$$\forall\ A \in \mathcal{H}, \quad \mathcal{L}_{H_*} A \coloneqq [H_*, A].$$

The action of $\mathcal{L}_{H_*}$ has a simple expression in the eigenvector decomposition $(\varepsilon_k, \phi_k)_{1 \leqslant k \leqslant N_b}$ of H_*:

$$\forall\ 1 \leqslant k, l \leqslant N_b, \quad \mathcal{L}_{H_*}(\phi_k \phi_l^*) = (\varepsilon_k - \varepsilon_l)\phi_k \phi_l^*. \tag{2.8}$$

Thus, we have

$$\forall\ i \in \mathcal{I},\ a \notin \mathcal{I}, \quad \Omega_*(\phi_i \phi_a^* + \phi_a \phi_i^*) = (\varepsilon_a - \varepsilon_i)(\phi_i \phi_a^* + \phi_a \phi_i^*).$$

Hence, one can easily check that, using again the decomposition (2.3), we have

$$\forall\ X \in T_{P_*}\mathcal{M}_N, \quad \Omega_* X = -[P_*, \mathcal{L}_{H_*} X] = -[P_*, [H_*, X]]. \tag{2.9}$$

This definition also provides a canonical way to extend the operator Ω_*, originally defined on $T_{P_*}\mathcal{M}_N$, to the whole space $\mathcal{H}$.

2.3. **Fixed-point on a manifold.** Finally, we state a general abstract result that we will use to study the convergence of optimization algorithms on manifolds.

Lemma 2.7. *Let $\mathcal{M}$ be a smooth finite dimensional Riemannian manifold. Let $P_* \in \mathcal{M}$ and $f : U \to \mathcal{M}$ be a continuously differentiable mapping on a neighborhood U of P_* such that $f(P_*) = P_*$. Let $\mathrm{d}f(P_*) : T_{P_*}\mathcal{M} \to T_{P_*}\mathcal{M}$ be the derivative of f at P_*. If $\mathrm{d}f(P_*)$ verifies $r\left(\mathrm{d}f(P_*)\right) < 1$ where $r\left(\mathrm{d}f(P_*)\right)$ is the spectral radius of $\mathrm{d}f(P_*)$, 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 $r\left(\mathrm{d}f(P_*)\right)$, in the sense that for all $\theta > 0$ there exists $C_\theta > 0$ such that, for all P^0 close enough to P_*,*

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

Proof. We use the notation presented in [8, Chapter 1]. Up to a restriction of U to a smaller neighborhood of P_*, there exists a neighborhood V of 0 in $T_{P_*}\mathcal{M}$ and $g : V \to U$ a diffeomorphic parametrization of the manifold such that $g(0) = P_*$ and $\mathrm{d}g(0) = \mathrm{Id}$ (take for instance the restriction to V of the exponential map). Therefore, as f is continuously differentiable, there exists a neighborhood $\widetilde{V} \subset V$ of 0 in $T_{P_*}\mathcal{M}$ such that $F \coloneqq g^{-1} \circ f \circ g : \widetilde{V} \to V$ is a continuously differentiable map with fixed-point 0 and $\mathrm{d}F(0) = \mathrm{d}f(P_*)$. As $r\left(\mathrm{d}F(0)\right) = r\left(\mathrm{d}f(P_*)\right) < 1$, we can find a neighborhood $V' \subset V$ of 0 in $T_{P_*}\mathcal{M}$ such that F is a contraction in V' for some norm $\|\cdot\|_\theta$, with contraction factor $r(\mathrm{d}f(P_*)) + \theta$, $\theta > 0$ (see [32] for more details). Therefore, we can apply the Banach fixed-point theorem to F and we get that, for x^0 close enough to 0, $x^{k+1} = F(x^k)$ converges to 0. Finally, for $P^0 = g(x^0)$, $P^{k+1} = g(x^{k+1}) = g(F(x^k)) = f(g(x^k)) = f(P^k)$ converges to $P_* = g(0)$, with asymptotic rate $r(\mathrm{d}f(P_*))$. □

3. Algorithms and analysis of convergence

3.1. **Direct minimization.** The gradient descent algorithm consists in following the steepest descent direction with a fixed step β at each iteration point. As the iterations are constrained to stay on the manifold, we have to

(1) project the gradient on the tangent space with Π_{P^k} to bring the steepest descent line $P^k - \beta \Pi_{P^k}\left(\nabla E(P^k)\right)$ back to the manifold at first order;
(2) retract the steepest descent line defined in the tangent space onto the manifold $\mathcal{M}_N$ by a nonlinear retraction R mapping a neighborhood of $\mathcal{M}_N$ in $\mathcal{H}$ to $\mathcal{M}_N$.

An example of retraction is given in Section 4.1 and we will assume that the retraction R satisfies

Assumption 3.1. *$R : \mathcal{H} \to \mathcal{H}$ is of class C^2 and for all $P \in \mathcal{M}_N$ and $X \in \mathcal{H}$ small enough,*

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

These two successive operations are sketched in Figure 1 and the gradient descent algorithm is presented in Algorithm 1.

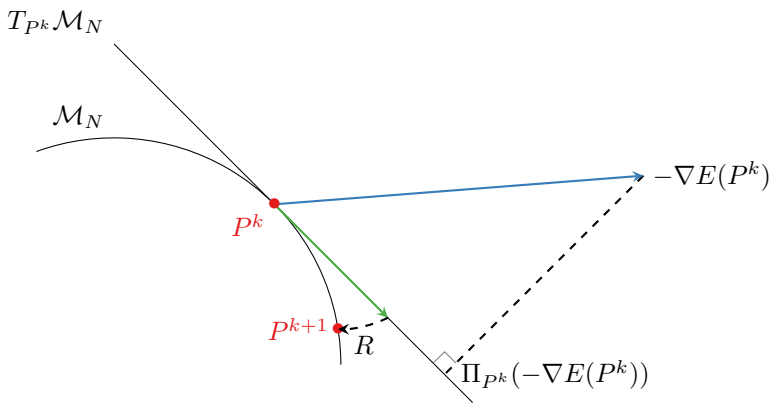


FIGURE 1 – Projection on the tangent space for the gradient descent, and retraction to the manifold.

ALGORITHM 1: Gradient descent

Data: $P^0 \in \mathcal{M}_N$
while *convergence not reached* **do**
  $P^{k+1} \coloneqq R\left(P^k - \beta \Pi_{P^k}\left(\nabla E(P^k)\right)\right)$;
end

At the continuous level, this algorithm can be seen as the discretization of the flow $\dot{P} = -\Pi_P \nabla E(P)$. Note that, by the use of the retraction R and Assumption 3.1, the projection step has no influence on the convergence of the algorithm for β small. Indeed, by Assumption 3.1,

$$\begin{aligned} \forall\ P \in \mathcal{M}_N, \quad R\left(P - \beta \Pi_P\left(\nabla E(P)\right)\right) &= P - \beta \Pi_P\left(\Pi_P\left(\nabla E(P)\right)\right) + O(\beta^2) \\ &= P - \beta \Pi_P\left(\nabla E(P)\right) + O(\beta^2) \end{aligned}$$

and thus the first order expansion is the same with or without the projection step. The reason we use this projection step is that it is convenient to interpret $\Pi_{P^k} \nabla E(P^k)$ as a residual.

The following theorems state that, for β small enough, Algorithm 1 globally converges in the sense that $\Pi_{P^k} \nabla E(P^k) \to 0$ and locally converges in the sense that $P^k \to P_*$ if P^0 is close enough to P_*.

Theorem 3.2. *Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 2.2 and $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 3.1. There exists $\beta_0 > 0$ such that for all $0 < \beta \leqslant \beta_0$ and all $P^0 \in \mathcal{M}_N$, the iterations*

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

satisfy the following properties:

(1) $(E(P^k))_{k\in\mathbb{N}}$ is a nonincreasing sequence converging to some critical value E_c of E on $\mathcal{M}_N$;
(2) when k goes to infinity, $\Pi_{P^k} \nabla E(P^k) \to 0$, $\|P^{k+1} - P^k\|_F \to 0$ and $d(P^k, A_c) \to 0$ where A_c is one of the connected components of $C(E_c) \coloneqq \{P \in \mathcal{M}_N \,|\, E(P) = E_c \text{ and } \Pi_P\left(\nabla E(P)\right) = 0\}$.

Proof. As $E : \mathcal{H} \to \mathbb{R}$ and $R : \mathcal{H} \to \mathcal{H}$ are C^2, and $\mathcal{M}_N$ is compact, we can use the expansion of Assumption 3.1 and obtain that there exists a constant $C \in \mathbb{R}_+$ such that for all $0 \leqslant \beta \leqslant 1$,

$$\forall\ k \in \mathbb{N}, \quad E(P^{k+1}) \leqslant E(P^k) - \beta \left\|\Pi_{P^k} \nabla E(P^k)\right\|_F^2 + C\beta^2 \left\|\Pi_{P^k} \nabla E(P^k)\right\|_F^2$$

Therefore, we have for $\beta > 0$ small enough,

$$\forall\ k \in \mathbb{N}, \quad E(P^{k+1}) \leqslant E(P^k) - \frac{\beta}{2} \left\|\Pi_{P^k} \nabla E(P^k)\right\|_F^2 .$$

This shows that the sequence $(E(P^k))_{k\in\mathbb{N}}$ is nonincreasing. As E is continuous on the compact set $\mathcal{M}_N$, $(E(P^k))_{k\in\mathbb{N}}$ is bounded and hence converges to some $E_c \in \mathbb{R}$. Moreover,

$$\sum_{k\in\mathbb{N}} \left\|\Pi_{P^k} \nabla E(P^k)\right\|_F^2 < \infty,$$

which implies that $\Pi_{P^k} \nabla E(P^k) \to 0$ when $k \to \infty$. It follows that $\|P^{k+1} - P^k\|_F \to 0$ when $k \to \infty$.

Let B be the non-empty compact set of accumulation points of $(P^k)_{k\in\mathbb{N}}$. By continuity of E and $P \mapsto \Pi_P \nabla E(P)$, it follows that $B \subset C(E_c)$. Assuming that $d(P^k, B)$ does not go to zero, we can extract a subsequence at finite distance of B which converges to a point in B, a contradiction. Assume that B is disconnected: it is then the union of two compact subsets B_1 and B_2 at positive distance from each other. Since $P^{k+1} - P^k \to 0$, there is an infinite number of points in $(P^k)_{k\in\mathbb{N}}$ at distance greater or equal to $\eta > 0$ from both B_1 and B_2, from which we can extract a point in B, a contradiction. It follows that B is connected, hence the result. □

This result implies in particular the convergence of the sequence $(P^k)_{k\in\mathbb{N}}$ in the generic case where critical points are isolated. If this is not the case but E and R are analytic, convergence can be shown following the approach in [38] based on Łojasiewicz inequality.

Theorem 3.3. *Let $E : \mathcal{H} \to \mathbb{R}$ satisfy Assumption 2.2 and Assumption 2.3 with P_* a local minimizer of (2.1). Let $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 3.1. Then, if $P^0 \in \mathcal{M}_N$ is close enough to P_*, the iterations*

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

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

Proof. In order to prove convergence, one can apply Lemma 2.7 to the function $f : \mathcal{M}_N \to \mathcal{M}_N$ defined by

$$f(P) \coloneqq R(P - \beta \Pi_P\left(\nabla E(P)\right)),$$

for which we know by the first order optimality condition that P_* is a fixed-point.

We compute explicitly $\mathrm{d}f(P_*)$ using the second-order optimality condition (2.7). To this end, take $X \in T_{P_*}\mathcal{M}_N$ and a smooth path $\gamma : I \to \mathcal{M}_N$ defined on a real interval I containing 0 such that $\gamma(0) = P_*$ and $\dot\gamma(0) = X$. We want to expand to the first order in t the following expression:

$$f(\gamma(t)) = R\left(\gamma(t) - \beta \Pi_{\gamma(t)}\left(\nabla E(\gamma(t))\right)\right).$$

First, we focus on the projection of $H(\gamma(t))$ on $T_{\gamma(t)}\mathcal{M}_N$:

$$\begin{aligned}
\Pi_{\gamma(t)} H(\gamma(t)) &= \gamma(t)H(\gamma(t))(1-\gamma(t)) + \mathrm{sym} \\
&= (P_* + tX)\left(H_* + t\left(\nabla^2 E(P_*)X\right)\right)(1 - P_* - tX) + \mathrm{sym} + O(t^2) \\
&= t\left[P_*\left(\nabla^2 E(P_*)X\right)(1-P_*) + \mathrm{sym}\right] + t\left[XH_*(1-P_*) - P_*H_*X + \mathrm{sym}\right] + O(t^2) \\
&= t(K_* + \Omega_*)X + O(t^2).
\end{aligned}$$

Inserting this into the expansion of $f(\gamma(t))$, using Assumption 3.1 and the fact that $\Pi_{\gamma(t)}X = \Pi_{P_*}X + O(t^2)$, gives

$$f(\gamma(t)) = R\big(\gamma(t) - \beta t(\Omega_* + K_*)X + O(t^2)\big) = P_* + t\big(X - \beta(\Omega_* + K_*)X\big) + O(t^2).$$

Therefore,

$$\mathrm{d}f(P_*)X = \big(1 - \beta(\Omega_* + K_*)\big)X.$$

As the second-order optimality condition (2.7) shows that $\Omega_* + K_*$ is positive definite on $T_{P_*}\mathcal{M}_N$, for β small enough, the spectral radius $r(\mathrm{d}f(P_*))$ of the derivative $\mathrm{d}f(P_*)$, is less than 1, which concludes the proof. □

3.2. **Damped self-consistent field.** The damped SCF algorithm is a damped version of the Roothaan algorithm [14, 38] and is presented in Algorithm 2, under the assumption that the strong *Aufbau* principle is satisfied, and represented in Figure 2. Note that it is well defined only if $\varepsilon_N^k < \varepsilon_{N+1}^k$ for all $k \in \mathbb{N}$. We introduce the nonlinear operators:

(1) $A(H) \coloneqq \mathbf{1}_{(-\infty,\varepsilon_N(H)]}(H)$, with $\varepsilon_N(H)$ the lowest N^{th} eigenvalue of H and where we recall that $\mathbf{1}_{(-\infty,\mu]}(H) \coloneqq \sum_{\varepsilon_i \le \mu} \phi_i \phi_i^*$, the ϕ_i's being orthonormal eigenvectors of H associated to the eigenvalues ε_i;

(2) $\Phi(P) = A(H(P))$ or, equivalently, $\Phi(P) \coloneqq \sum_{i=1}^{N} \phi_i \phi_i^*$ where the ϕ_i's are orthonormal eigenvectors associated to the lowest N eigenvalues of $H(P)$.

Algorithm 2: Damped SCF algorithm

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

while *convergence not reached* **do**

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

$$\Phi(P^k) \coloneqq \sum_{i=1}^{N} \phi_i^k \left(\phi_i^k\right)^*;$$

$$P^{k+1} \coloneqq R\left(P^k + \beta \Pi_{P^k}\left(\Phi(P^k) - P^k\right)\right);$$

end

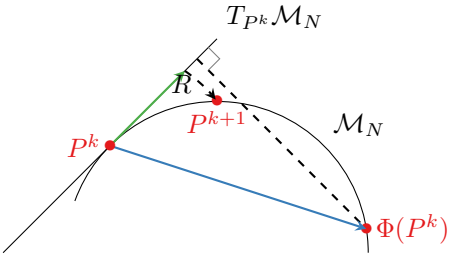


Figure 2 – Retraction for the damped SCF algorithm.

The following theorem states that, under the condition that there is a gap between the smallest N^{th} and $(N+1)^{\text{st}}$ eigenvalues of the Hamiltonian H_*, Algorithm 2 locally converges for β small enough.

Theorem 3.4. *Let $E : \mathcal{H} \to \mathbb{R}$ and $P_* \in \mathcal{M}_N$ satisfy Assumption 2.2 and Assumption 2.3 and $R : \mathcal{H} \to \mathcal{H}$ satisfy Assumption 3.1. Assume that P_* satisfies the strong* Aufbau *principle*

$$\Phi(P_*) = P_* \text{ and } \nu \coloneqq \varepsilon_{N+1} - \varepsilon_N > 0,$$

where $(\varepsilon_i)_{1 \leq i \leq N_b}$ are the eigenvalues of H_* ranked in nondecreasing order.

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

$$P^{k+1} \coloneqq R\left(P^k + \beta \Pi_{P^k}\left(\Phi(P^k) - P^k\right)\right)$$

are well-defined and P^k linearly converges to P_*, with asymptotic rate $r(1-\beta J_{\text{SCF}})$ where $J_{\text{SCF}} \coloneqq 1 + {\Omega_*}^{-1} K_*$.

Proof. In order to prove convergence, we apply Lemma 2.7 to the function $f : \mathcal{M}_N \to \mathcal{M}_N$ defined by

$$f(P) \coloneqq R(P + \beta \Pi_P\left(\Phi(P) - P\right)),$$

for which P_* is a fixed-point.

First, we compute the derivative of $\Phi = A \circ H$ at the minimizer P_* to get

$$\mathrm{d}\Phi(P_*) = \mathrm{d}A(H_*)\nabla^2 E(P_*).$$

Now, to compute $\mathrm{d}A(H_*)$, note that, as there is a gap $\varepsilon_{N+1} > \varepsilon_N$ at the minimum, we can find a contour $\mathcal{C}$ in the complex plane enclosing the lowest N eigenvalues of H_* (Figure 3) such that

$$A(H_*) = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H_*} \, \mathrm{d}z \tag{3.1}$$

(see [31, 34] for more details on spectral calculus, contour integrals and perturbation theory for functions of matrices). By continuity, we also have

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

for H in a neighborhood of H_*. Then, one can use the expression (3.1) of A and the expansion for H in a neighborhood of H_*

$$\forall \ z \in \mathcal{C}, \quad \frac{1}{z - H} = \frac{1}{z - H_*}(H - H^*)\frac{1}{z - H_*} + O(\|H - H_*\|_F^2)$$

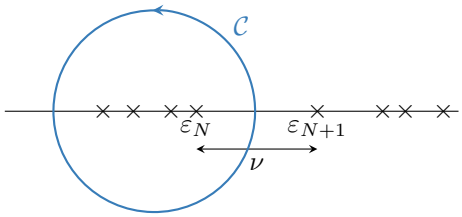


FIGURE 3 – Definition of A and graphical interpretation of the *Aufbau* principle and the existence of a gap.

to get

$$\begin{aligned}\forall\ h \in \mathcal{H}, \quad \mathrm{d}A(H_*)h &= \frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - H_*} h \frac{1}{z - H_*}\ \mathrm{d}z \\ &= \sum_{k=1}^{N_b} \sum_{l=1}^{N_b} \left(\frac{1}{2\pi\mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - \varepsilon_k} h_{kl} \frac{1}{z - \varepsilon_l}\ \mathrm{d}z \right) \phi_k \phi_l^*,\end{aligned}$$

where $h_{kl} = \phi_k^* h \phi_l$. Now, let us denote by $1 \leqslant i \leqslant N$ the occupied orbitals (ε_i is inside $\mathcal{C}$) and by $N+1 \leqslant a \leqslant N_b$ the virtual ones (ε_a is outside $\mathcal{C}$). Then,

$$\oint_{\mathcal{C}} \frac{1}{z - \varepsilon_i} \frac{1}{z - \varepsilon_a}\ \mathrm{d}z = \begin{cases} \dfrac{1}{\varepsilon_i - \varepsilon_a} & \text{if } 1 \leqslant i \leqslant N < a \leqslant N_b; \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the sum becomes

$$\mathrm{d}A(H_*)h = \sum_{i=1}^{N} \sum_{a=N+1}^{N_b} \frac{1}{\varepsilon_i - \varepsilon_a} \left(h_{ia} \phi_i \phi_a^* + h_{ai} \phi_a \phi_i^* \right) = -\Omega_*^{-1} \Pi_{P_*} h,$$

and we finally get

$$\forall\ X \in T_{P_*}\mathcal{M}_N, \quad \mathrm{d}\Phi(P_*)X = -{\Omega_*}^{-1} K_* X.$$

Now, we compute the derivative of f at point P_*. Let $X \in T_{P_*}\mathcal{M}_N$ and $\gamma : I \to \mathcal{M}_N$ a smooth path defined on a real interval I containing 0 such that $\gamma(0) = P_*$ and $\dot{\gamma}(0) = X$. First, we expand $\gamma(t)$ around 0 and Φ around $\gamma(0) = P_*$ to obtain

$$f(\gamma(t)) = P_* + t\Pi_{P_*} \left((1 - \beta) + \beta \mathrm{d}\Phi(P_*) \right) X + O\left(t^2\right).$$

Thus, for $X \in T_{P_*}\mathcal{M}_N$,

$$\mathrm{d}f(P_*)X = \left((1 - \beta) - \beta {\Omega_*}^{-1} K_* \right) X.$$

To conclude this proof, we compute the spectral radius of

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

First, notice that

$$1 + {\Omega_*}^{-1} K_* = 1 + {\Omega_*}^{-1/2} \left({\Omega_*}^{-1/2} K_* {\Omega_*}^{-1/2} \right) {\Omega_*}^{1/2}$$

and thus, $1 + {\Omega_*}^{-1} K_*$ and the symmetric operator $1 + {\Omega_*}^{-1/2} K_* {\Omega_*}^{-1/2}$ have the same eigenvalues. Moreover, using the second-order optimality condition (2.7), with $X = {\Omega_*}^{-1/2} Y$, we get

$$\begin{aligned}\forall\ Y \in T_{P_*}\mathcal{M}_N, \quad \left\langle Y, \left(1 + {\Omega_*}^{-1/2} K_* {\Omega_*}^{-1/2} \right) Y \right\rangle_{\mathrm{F}} &\geqslant \eta \left\langle Y, {\Omega_*}^{-1} Y \right\rangle_{\mathrm{F}} \\ &\geqslant \frac{\eta}{\|\Omega_*\|_{\mathrm{op}}} \|Y\|_{\mathrm{F}}^2,\end{aligned} \tag{3.2}$$

with $\|\Omega_*\|_{\mathrm{op}}$ the operator norm associated to $\|\cdot\|_{\mathrm{F}}$. Thus, all the eigenvalues of $1 + {\Omega_*}^{-1/2} K_* {\Omega_*}^{-1/2}$, hence of $1 + {\Omega_*}^{-1} K$, are real and positive. Consequently, for β small enough, the spectral radius $r\left(\mathrm{d}f(P_*)\right)$ is less than 1 and we conclude by applying Lemma 2.7. □

Remark 3.5 (Case when the *Aufbau* principle is not satisfied). In the case when the minimizer P_* does not verify the *Aufbau* principle, but does satisfy the condition that the eigenvalues of $1 + {\Omega_*}^{-1}K_*$ are positive (note that Ω_* is not positive when the *Aufbau* principle is not verified, but $1 + {\Omega_*}^{-1}K_*$ might still have only positive eigenvalues), the damped SCF still converges locally to P_* for $\beta > 0$ small enough if we change the way we select the occupied orbitals to build $\Phi(P)$ (in this case, we do not pick those associated to the smallest N eigenvalues of $H(P)$, but those corresponding to the occupied orbitals of P_*).

We conclude this section by proving the local convergence of the non-retracted variant of Algorithm 2.

ALGORITHM 3: Non-retracted damped SCF algorithm

Data: $P^0 \in \mathcal{M}_N$
while *convergence not reached* **do**
  solve $\begin{cases} H(P^k)\phi_i^k = \varepsilon_i^k \phi_i^k, \ \varepsilon_1^k \leqslant \cdots \leqslant \varepsilon_N^k < \varepsilon_{N+1}^k \leqslant \cdots \leqslant \varepsilon_{N_b}^k \\ \langle \phi_i^k, \phi_j^k \rangle = \delta_{ij}, \end{cases}$;
  $\Phi(P^k) \coloneqq \sum_{i=1}^{N} \phi_i^k \left(\phi_i^k\right)^*$;
  $P^{k+1} \coloneqq P^k + \beta \Pi_{P^k}\left(\Phi(P^k) - P^k\right)$;
end

Theorem 3.6. *Let $E : \mathcal{H} \to \mathbb{R}$ and P_* satisfy Assumption 2.2 and Assumption 2.3. Moreover, assume that*

$$\Phi(P_*) = P_* \text{ and } \nu \coloneqq \varepsilon_{N+1} - \varepsilon_N > 0 \text{ (strong Aufbau principle)},$$

where $(\varepsilon_i)_{1 \leqslant i \leqslant N_b}$ are the eigenvalues of H_* ranked in nondecreasing order.

Then, for $\beta > 0$ small enough and $P^0 \in \mathcal{H}$ close enough to P_* and with trace N, the iterations

$$P^{k+1} \coloneqq P^k + \beta\left(\Phi(P^k) - P^k\right) \tag{3.3}$$

are well-defined and P^k linearly converges to $P_* \in \mathcal{M}_N$, with asymptotic rate $\max(r(1 - \beta J_{\mathrm{SCF}}), 1 - \beta)$ where $J_{\mathrm{SCF}} \coloneqq 1 + {\Omega_*}^{-1}K_*$.

Note that the iterates P^k defined by (3.3) have trace N but do not lay on the manifold $\mathcal{M}_N$ in general.

Proof. The proof follows that of Theorem 3.4. This time, we need to compute the Jacobian matrix of $f : \mathcal{H} \ni P \mapsto P + \beta\left(\Phi(P) - P\right) \in \mathcal{H}$ at the minimizer $P_* \in \mathcal{M}_N$. As we work in the whole space $\mathcal{H}$, the Jacobian matrix has the form, in the decomposition $\mathcal{H} = T_{P_*}\mathcal{M}_N \oplus (T_{P_*}\mathcal{M}_N)^\perp$,

$$\mathrm{d}f(P_*) = \begin{bmatrix} 1 - \beta J_{\mathrm{SCF}} & \times \\ 0 & 1 - \beta \end{bmatrix},$$

where $J_{\mathrm{SCF}} = 1 + {\Omega_*}^{-1}K_*$ has been computed in the proof of Theorem 3.4. Hence, this time the algorithm converges to $P_* \in \mathcal{M}_N$ as long as β is such that $\max(r(1 - \beta J_{\mathrm{SCF}}), 1 - \beta) < 1$. □

In LDA and GGA Kohn-Sham models [45], the mean-field Hamiltonian $H(P)$ is actually a function $\widetilde{H}(\rho_P)$ of the density ρ_P associated with the density matrix P. Since the map $P \mapsto \rho_P$ is linear, (3.3) can be rewritten as

$$\rho^{k+1} = (1 - \beta)\rho^k + \beta\Psi(\rho^k),$$

where $\Psi(\rho) = \rho_{A(\widetilde{H}(\rho))}$. We can therefore interpret (3.3) as the equivalent density matrix formulation of this density mixing algorithm.

3.3. **Comparison.** In this section, we proved the local convergence of Algorithm 1 and Algorithm 2. and we obtained asymptotic convergence rates. On the tangent space, both Jacobian matrices are of the form $1 - \beta J$ where J has positive real spectrum and

- for the gradient descent: $J_{\mathrm{grad}} = K_* + \Omega_*$, which is self-adjoint for the Frobenius inner product;
- for the damped SCF algorithm if the strong *Aufbau* principle is satisfied at P_*: $J_{\mathrm{SCF}} = 1 + {\Omega_*}^{-1}K_*$, which is self-adjoint for the inner product $\langle \cdot, \cdot \rangle_{\Omega_*} \coloneqq \langle \Omega_* \cdot, \cdot \rangle_{\mathrm{F}}$.

One can notice that, *in the linear regime*, the SCF iterations correspond to a matrix splitting of the gradient iterations. Whether this results in a faster method or not depends not only on the relative conditioning of the iteration matrices but also on the relative cost of each step.

To have the fastest convergence, we want the eigenvalues of $1 - \beta J$ to be as close to 0 as possible. If we denote by λ_1 (resp. λ_N) the smallest (resp. largest) eigenvalue of J, the optimal step β_* is the minimizer of $\min_\beta \max\{|1-\beta\lambda_1|, |1-\beta\lambda_N|\}$, which is given by

$$\beta_* = \frac{2}{\lambda_1 + \lambda_N}.$$

Then, the rate of convergence is, with $\kappa \coloneqq \lambda_N/\lambda_1$ the spectral condition number of J,

$$r = \frac{\kappa - 1}{\kappa + 1}.$$

Now, we can evaluate the conditioning of J for the two algorithms :

- for the gradient descent, we have
$$\kappa(J_{\mathrm{grad}}) \leqslant \frac{\|\Omega_*\|_{\mathrm{op}} + \|K_*\|_{\mathrm{op}}}{\eta}, \tag{3.4}$$
where η is the coercivity constant in the nondegeneracy Assumption 2.3. First, the smaller η, the more difficult the convergence. Note however that there is no relationship in general between η and the gap ν. Second, the bigger $\|\Omega_*\|_{\mathrm{op}} = \varepsilon_{N_b} - \varepsilon_1$, the more difficult the convergence. In particular, for models arising from the discretization of partial differential equations, $\varepsilon_{N_b} - \varepsilon_1 \to \infty$ when the discretization is refined. In practice, this issue is solved by preconditioning (see Remark 3.7).
- for the damped SCF algorithm, a naive bound would be
$$\kappa(J_{\mathrm{SCF}}) \leqslant \|\Omega_*\|_{\mathrm{op}} \frac{1 + \nu^{-1}\|K_*\|_{\mathrm{op}}}{\eta}.$$
In this bound the right-hand side diverges when $\|\Omega_*\|_{\mathrm{op}} \to \infty$ as above, whereas the left-hand side may actually remain bounded. For instance, under the uniform coercivity assumption [11]
$$\forall\, X \in T_{P_*}\mathcal{M}_N, \quad \langle X, (\Omega_* + K_*)\, X\rangle_{\mathrm{F}} \geqslant \widetilde{\eta}\, \langle \Omega_* X, X\rangle_{\mathrm{F}}\,,$$
with $\widetilde{\eta}$ independent of N_b (which is often the case in practice), we have
$$\kappa(J_{\mathrm{SCF}}) \leqslant \frac{1 + \nu^{-1}\|K_*\|_{\mathrm{op}}}{\widetilde{\eta}}.$$
In contrast with the bound (3.4), we can see that the smaller the gap ν, the slower the convergence.

As a special case, if we consider the case where the Hessian $\nabla^2 E \equiv 0$, *i.e.* a linear eigenvalue problem. Then the SCF algorithm converges in one iteration, which is consistent with $J_{\mathrm{SCF}} = 1$. The gradient descent with optimal step locally converges with asymptotic rate $r = \frac{\kappa-1}{\kappa+1}$ where $\kappa = \frac{\varepsilon_{N_b} - \varepsilon_1}{\varepsilon_{N+1} - \varepsilon_N}$.

The convergence rates we derived in Theorem 3.3 and Theorem 3.4 are consistent with well-known convergence issues, for instance the failure of the simple damped SCF algorithm to converge for systems with small gaps [55] (although Section 4.5 shows this is not necessarily true for more sophisticated acceleration methods).

Remark 3.7 (Preconditioning). We discuss here the extension of Theorem 3.3 to the preconditioned gradient descent:

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

with $B : \mathcal{H} \to \mathcal{H}$ a symmetric positive definite preconditioner. If we denote by $\widetilde{B}_* \coloneqq \Pi_{P_*} B \Pi_{P_*}$ its restriction to the tangent plane, the Jacobian matrix of the gradient becomes $1 - \beta \widetilde{B}_*(\Omega_* + K_*)$ where $(\Omega_* + K_*)$ is positive definite (under Assumption 2.3) and the proof of local convergence for β small enough follows exactly in the same way, using the positive definiteness of $\widetilde{B}_*$ to show that $\widetilde{B}_*(\Omega_* + K_*)$ has real positive spectrum. The same analysis holds true for the preconditioned SCF algorithm. In practice, preconditioning is a crucial tool to accelerate iterations, in particular in order to achieve mesh- and domain-size independence of the number of iterations for discretized partial differential equations. However, we are interested here in the intrinsic aspects of each algorithm (direct minimization *vs* SCF)

and the influence of physical parameters (e.g. the gap ν), so that the study of preconditioned algorithms is not in the scope of this paper.

Remark 3.8 (Dielectric operator). In the context of Kohn-Sham density functional theory, the operator $\left(1 + K_* {\Omega_*}^{-1}\right)^{-1}$, the transpose of the inverse of the Jacobian of the simple SCF mapping, is known as the dielectric operator: it represents the infinitesimal change in the self-consistent Hamiltonian $H(P_*)$ in response to a change in the energy functional. Our results show that this operator is well-defined and has real positive spectrum, with no assumption on the sign of Hartree-exchange-correlation kernel K_*, recovering in an algebraic framework the results of [19, 25] obtained using a different variational principle.

4. Numerical tests

We present here some numerical experiments to illustrate our theoretical results, explore their limits and investigate the global behavior of the algorithms. First, we start by specifying the retraction R that we use in our numerical tests. In Section 4.2, we use a simple toy model for which we can control the gap and analytically compute the exact minimizer: this allows us to study the impact of the gap on the convergence of Algorithm 1 and Algorithm 2. In Section 4.3, we show that simple (nondamped) SCF iterations can exhibit chaotic behavior for some nonlinearities. Then, in Section 4.4, we report numerical tests for a 1D Gross-Pitaevskii model ($N = 1$) and its fermionic version for $N = 2$. Finally, in Section 4.5, we present results obtained with a more realistic case: Silicon in the framework of the Kohn-Sham DFT.

4.1. The retraction R.

We choose the following algorithm: for a given symmetric matrix $\widetilde{P}$ close to $\mathcal{M}_N$ with eigendecomposition $\widetilde{P} = V\widetilde{D}V^*$ with $\widetilde{D}$ diagonal and V orthogonal, we set the diagonal matrix D as

$$D_{ii} = \begin{cases} 1 & \text{if } \widetilde{D}_{ii} > 0.5 \\ 0 & \text{otherwise} \end{cases}.$$

and $R(\widetilde{P}) = VDV^*$. When $\widetilde{P}$ is close to $\mathcal{M}_N$, its eigenvalues are close to either 0 or 1. Given a contour $\mathcal{C}$ enclosing only the eigenvalues close to 1 , R has the following explicit expression

$$R(P) = \frac{1}{2\pi \mathrm{i}} \oint_{\mathcal{C}} \frac{1}{z - P} \, \mathrm{d}z. \tag{4.1}$$

Therefore, it follows from arguments similar to those used in the proof of Theorem 3.4 that R is analytic and satisfies Assumption 3.1.

4.2. A toy model with tunable spectral gap.

We work here in the very simple framework of real density matrices of order 2, *i.e.* the 2×2 real matrices P such that $P^* = P$, $P^2 = P$ and $\mathrm{Tr}(P) = 1$. Then, we consider the following minimization problem, with a parameter $\varepsilon \geqslant 0$: energy functional

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

for parameters $\varepsilon \geqslant 0$. The gradient and Hessian of E are

$$\begin{aligned} H_\varepsilon(P) &= 2\left(P - \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 0 \end{bmatrix}\right), \\ \nabla^2 E_\varepsilon(P) &= 2. \end{aligned}$$

Simple computations show that the set of rank-1 projectors on $\mathbb{R}^2$ can be parameterized as

$$\mathcal{M}_1 \coloneqq \left\{ P(a,b) = \begin{bmatrix} 1-a & b \\ b & a \end{bmatrix} \;\middle|\; a \in [0,1],\ b = \pm\sqrt{a(1-a)} \right\}.$$

The eigenvalues of H_ε at $P(a,b) \in \mathcal{M}_1$ are $\pm 2\sqrt{a^2 + (b-\varepsilon)^2}$. The gap is thus $\nu(a,b) \coloneqq 4\sqrt{a^2 + (b-\varepsilon)^2}$.

The case $\varepsilon = 0$. Here, the unique minimum is clearly

$$P(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{M}_1$$

and the gap is zero. Since $\nabla^2 E = 2$, this minimum satisfies Assumption 2.3 with $\eta = 2$.

The case $\varepsilon > 0$. We compute

$$E_\varepsilon(P(a,b)) = 2(a + \varepsilon^2 - 2\varepsilon b),$$

and therefore

$$E_\varepsilon\left(P(a, \sqrt{a(1-a)})\right) \leqslant E_\varepsilon\left(P(a, -\sqrt{a(1-a)})\right).$$

Hence, to compute the minimizer of the energy, we can restrict ourselves to the one-dimensional manifold

$$P(a) = \begin{bmatrix} 1-a & \sqrt{a-a^2} \\ \sqrt{a-a^2} & a \end{bmatrix}$$

with $a \in [0,1]$. Then, the energy is

$$E_\varepsilon(P(a)) = 2\left(a + \varepsilon^2 - 2\varepsilon\sqrt{a(a-1)}\right). \tag{4.2}$$

The first-order condition yields

$$a = \frac{1 \pm \sqrt{1 - \frac{4\varepsilon^2}{1+4\varepsilon^2}}}{2},$$

with the lowest energy achieved at

$$a(\varepsilon) \coloneqq \frac{1 - \sqrt{1 - \frac{4\varepsilon^2}{1+4\varepsilon^2}}}{2}.$$

The gap $\nu(\varepsilon) \coloneqq 4\sqrt{a(\varepsilon)^2 + \left(\sqrt{a(\varepsilon)\left(1-a(\varepsilon)\right)} - \varepsilon\right)^2}$ goes to 0 monotonically when $\varepsilon \to 0$. In particular, for $\varepsilon \approx 0$ we have $a(\varepsilon) \approx \varepsilon^2$ and $\nu(\varepsilon) \approx 4\varepsilon^2$. This model can thus be used to study the influence of the gap on the convergence of the two algorithms.

Influence of ε *on the convergence.* We run Algorithm 1 and Algorithm 2 with fixed β on this system. We start from a random point on the manifold $\mathcal{M}$. We take as convergence criterion $\left\|P^k - P(a(\varepsilon))\right\|_{\mathrm{F}} \leqslant 10^{-12}$, and consider the algorithm has failed if convergence was not achieved after $50,000$ iterations.

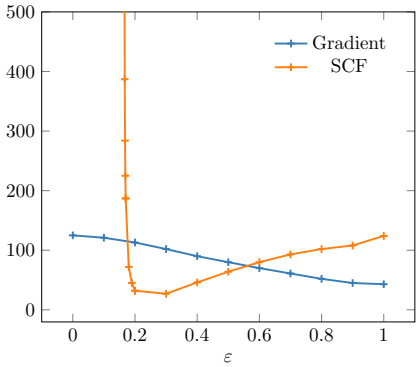
On Figure 4, we plotted the number of iterations to achieve convergence for each algorithm as a function of ε (without changing the starting point), for two different values of β: 10^{-1} and 10^{-3}. The results confirm the theory we developed in Section 3.3: the gap has a strong influence on the convergence behavior of the SCF algorithm. Indeed, as the gap decreases, smaller and smaller damping parameters must be used, and the number of iterations increases. In fact for this system, $1 + {\Omega_*}^{-1} K_*$ has a single eigenvalue equal to $1 + \frac{2}{\nu(\varepsilon)} \approx 1 + \frac{1}{2\varepsilon^2}$ for ε small. Thus we expect convergence for $\beta < 4\varepsilon^2$, and therefore a critical ε_{c} of ≈ 0.158 for $\beta = 10^{-1}$ and 0.0158 for $\beta = 10^{-3}$, with a number of iterations proportional to $\frac{1}{\varepsilon - \varepsilon_{\mathrm{c}}}$ when $\varepsilon > \varepsilon_{\mathrm{c}}$. The numerical results are in perfect agreement with this prediction. By contrast, the gradient algorithm is much less affected by the smallness of the gap, and converges in an essentially constant number of iterations: our prediction for the convergence rate of that method is $r = 1 - \beta(\nu(\varepsilon) + 2) \approx 1 - 2\beta$ for ε small, and therefore a number of iterations for convergence to 10^{-12} of 124 for $\beta = 10^{-1}$ and 1.3×10^4 for $\beta = 10^{-3}$, again in perfect agreement with the numerical results.

4.3. **Chaos in SCF iterations.** We consider in this section another toy model a model inspired from the one proposed in [43, Section 2.1]. Let $h \in \mathbb{R}^{3\times 3}_{\mathrm{sym}}$ and $J \in \mathbb{R}^{3\times 3}$ be the matrices defined by

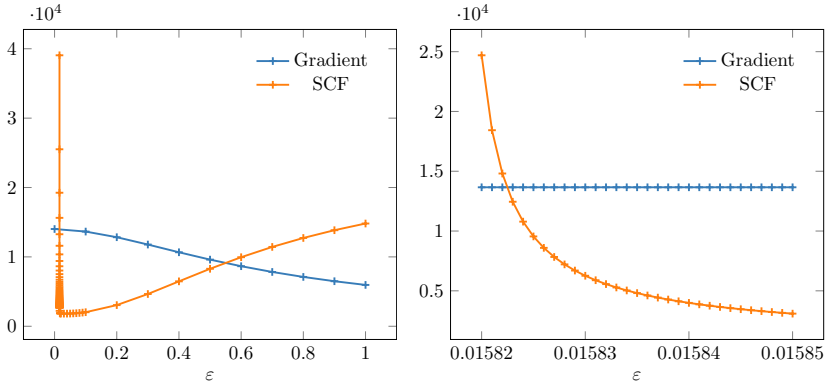
$$h \coloneqq \begin{bmatrix} 1.4299 & -0.2839 & -0.4056 \\ -0.2839 & 1.1874 & 0.2678 \\ -0.4056 & 0.2678 & 2.3826 \end{bmatrix} \quad \text{and} \quad J \coloneqq h^{-1}$$

and the energy functional $E_{c_1,c_2} : \mathbb{R}^{N_b \times N_b} \to \mathbb{R}$ defined by

$$E_{c_1,c_2}(P) = \mathrm{Tr}(hP) + \frac{c_1}{2}\rho_P^* J \rho_P - c_2 \sum_{j=1}^{3} \rho_{P,j}^{4/3}, \tag{4.3}$$



(a) Number of iterations to reach convergence for both algorithms as a function of ε for $\beta = 10^{-1}$.



(b) Number of iterations to reach convergence for both algorithms as a function of ε for $\beta = 10^{-3}$. On the left is a global view of the convergence for $\varepsilon \in [0, 1]$ and on the right, we zoom in the neighbourhood of ε_c.

FIGURE 4 – Comparison of the convergence of both algorithms depending on ε for two different values of the step β.

where $c_1, c_2 \in \mathbb{R}_+$ are nonnegative real parameter and where the density $\rho_P \in \mathbb{R}^3$ associated with the density matrix P is given by $\rho_{P,j} = P_{jj}$ for $j = 1, 2, 3$. This model is reminiscent of a Kohn-Sham LDA model, with h playing the role of the core Hamiltonian, J of the Hartree operator and the third term in E_{c_1,c_2} of the exchange-correlation energy. We seek the minimizers of E_{c_1,c_2} on $\mathcal{M}_1$.

We study the behavior of the simple SCF (Roothaan) iterations $P^k = \Phi(P^k)$ with initial guess $P^0 = \phi_0\phi_0^*$ with ϕ_0 a random vector of norm 1.

The case $c_2 = 0$. Here, the energy functional is the sum of a linear and a quadratic term. In this case, either $(P^k)_{k\in\mathbb{N}}$ converges to a critical point of the problem (in practice a local minimizer), or it has two different accumulation points P_{odd} and P_{even}, none of them being a critical point, and the iterates oscillates between the two, in the sense that $P^{2k+1} \to P_{\text{odd}}$ and $P^{2k} \to P_{\text{even}}$ when $k \to \infty$ [14, 38]. This is due to the fact that we have

$$P^{2k+1} = \operatorname{argmin}\left\{\widetilde{E}_{c_1,0}(P^{2k}, P),\ P \in \mathcal{M}_1\right\},$$
$$P^{2k+2} = \operatorname{argmin}\left\{\widetilde{E}_{c_1,0}(P, P^{2k+1}),\ P \in \mathcal{M}_1\right\},$$

with

$$\widetilde{E}_{c_1,0}(P, P') := \frac{1}{2}\operatorname{Tr}(hP) + \frac{1}{2}\operatorname{Tr}(hP') + \frac{c_1}{2}\rho_P^* J \rho'_P,$$

so that (P^{2k}, P^{2k+1}) converges to a minimizer of $\widetilde{E}_{c_1,0}$ on $\mathcal{M}_1 \times \mathcal{M}_1$. When c_1 is small, the simple SCF algorithm converges: for $c_1 = 0$, the matrix h has a nondegenerate lowest eigenvalue and the algorithm

converges in one iteration. When the value of c_1 increases, we observe numerically a bifurcation at a critical value $c_{1,*} \simeq 0.28$ after which the algorithm oscillates between two states.

The case $c_2 = 1$. Here the energy is not quadratic, and the previous theory does not apply. In Figure 5, we vary c_1 and plot the value of ρ_1 for the last 40 out of 1,500 SCF iterations. For this case, we still observe that the algorithm converges for c_1 small enough ($0 \leqslant c_1 < c_{1,*} \simeq 1.38$), and oscillates between two states for c_1 slightly larger than $c_{1,*}$. However, in contrast with the $c_2 = 0$ case, this is followed by a cascade of cycles of increasing periods, transitioning to a chaotic region, following the "period-doubling route to chaos" observed for other types of chaotic systems such as the logistic map [58].

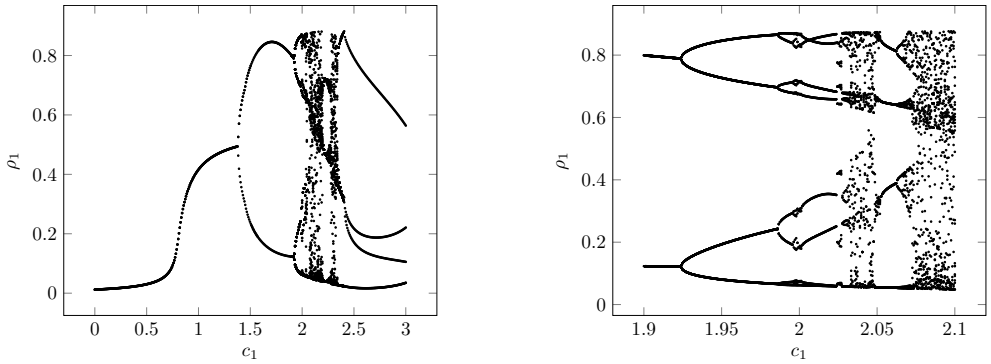


FIGURE 5 – Chaotic behavior of the simple SCF map for the energy functional $E_{c_1,1}$ defined by (4.3) and $N = 1$ as a function of c_1. On the left is a global view of the bifurcation and the right is a zoom on part of the interval on which we observe a chaotic behavior.

4.4. **Local convergence for a 1D nonlinear Schrödinger equation.** In this section, we present a simple 1D numerical experiment to validate on a more physically relevant system the sharpness of the convergence rates we derived in the previous section. As the goal here is to analyze the behavior of the simplest representative of each class of algorithms when physical parameters (such as the gap) vary, we chose unpreconditioned algorithms. We consider a discretized 1D Gross-Pitaevskii model ($N = 1$) on the torus, and its (non-physical) fermionic counterpart for $N = 2$. At the continuous level, the minimization set is

$$\left\{\gamma \in \mathcal{L}(L^2_{\mathrm{per}}),\ \gamma^2 = \gamma = \gamma^*,\ \mathrm{Tr}(\gamma) = N\right\},$$

where $\mathcal{L}(L^2_{\mathrm{per}})$ is the space of bounded operators on $L^2_{\mathrm{per}} := \{u \in L^2_{\mathrm{loc}}(\mathbb{R}) \mid u(\cdot - 1) = u(\cdot)\}$, and the energy functional is defined as

$$\mathcal{E}_\alpha(\gamma) = \mathrm{Tr}_{L^2_{\mathrm{per}}}\left(-\frac{1}{2}\Delta\gamma\right) + \int_0^1 V\rho_\gamma + \frac{\alpha}{2}\int_0^1 \rho_\gamma^2,$$

where ρ_γ is the density of the density matrix γ, $\alpha \in \mathbb{R}_+$ and V is an asymmetric double-well external potential chosen equal to

$$V(x) := -C\left(\exp\left(-c\cos\left(\pi(x - 0.20)\right)^2\right) + 2\exp\left(-c\cos\left(\pi(x + 0.25)\right)^2\right)\right), \tag{4.4}$$

with $c = 30$ and $C = 20$ (Figure 6).

The Euler-Lagrange equations of this minimization problem are

$$\gamma_* = \sum_{i=1}^N (\phi_i, \cdot)_{L^2_{\mathrm{per}}}\phi_i, \quad \rho_* = \sum_{i=1}^N |\phi_i|^2, \quad -\frac{1}{2}\Delta\phi_i + V\phi_i + \alpha\rho_*\phi_i = \varepsilon_i\phi_i, \quad \int_0^1 \phi_i\phi_j = \delta_{ij}. \tag{4.5}$$

We discretize this model using the finite difference method with a uniform grid of step size $\delta = 1/N_b$, which leads to the finite-dimensional model

$$\inf\left\{E_\alpha(P),\ P \in \mathcal{M}_N\right\}, \tag{4.6}$$

where

$$\forall\ P \in \mathcal{H}, \quad E_\alpha(P) := \mathrm{Tr}(hP) + \frac{\alpha}{2}\delta\sum_{i=1}^{N_b}\left(\frac{P_{ii}}{\delta}\right)^2, \tag{4.7}$$

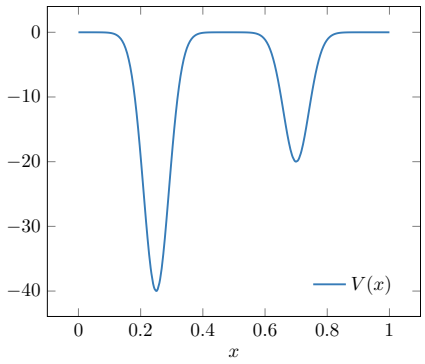


FIGURE 6 – V for $c = 30$ and $C = 20$.

the nonzero entries of the matrix $h \in \mathbb{R}^{N_b \times N_b}$ being given by

$$\forall\, 1 \leqslant i \leqslant N_b, \quad h_{ii} = \frac{1}{\delta^2} + V(i\delta), \quad h_{i,i+1} = h_{i,i-1} = -\frac{1}{2\delta^2}.$$

where we identify the sites 0 and N_b on the one hand and $N_b + 1$ and 1 on the other. With this discretization, the discrete density is then given by $\rho(i\delta) \approx \rho_i \coloneqq P_{ii}/\delta$. We compare the fixed-step gradient descent and damped SCF algorithms (Algorithm 1 and Algorithm 2) on this problem for various values of α, using as starting point the ground state for $\alpha = 0$. The functional E is smooth. To check Assumption 2.3 we notice that E is a convex functional of P, so that $\nabla^2 E(P) \geqslant 0$. Therefore, at any local minimizer satisfying the strong *Aufbau* principle, $\Omega_* + K_* \geqslant \Omega_* \geqslant \eta > 0$ and therefore Assumption 2.3 is satisfied. In the case where the *Aufbau* principle is not satisfied, Assumption 2.3 is not *a priori* always satisfied, so we check it *a posteriori* by computing the lowest eigenvalue of $\Omega_* + K_*$.

We prove in the Appendix the following lemma, which collects some of the mathematical properties of the discretized model under consideration. The proof of this lemma, given in the appendix, strongly relies on the properties of our particular model (one-dimensional difference equation with periodic boundary conditions and a specific nonlinearity).

Lemma 4.1 (Mathematical properties of (4.6))**.** *Let $\alpha \in \mathbb{R}_+$.*

(1) For $N = 1$, the optimization problem (4.6) has a unique minimizer P_*. In addition, P_* can be written as $P_* = \phi_* \phi_*^*$, with $\phi_* \in \mathbb{R}^{N_b}$, $\phi_*^* \phi_* = 1$, and ϕ_* positive componentwise, and P_* satisfies the strong Aufbau *principle.*

(2) For $2 \leqslant N \leqslant N_b$, with $N_b \neq 2N$ if $N_b \in 4\mathbb{N}^*$, the relaxed constrained optimization problem

$$\inf \left\{ E_\alpha(P),\ P \in \mathrm{CH}(\mathcal{M}_N) \right\}, \tag{4.8}$$

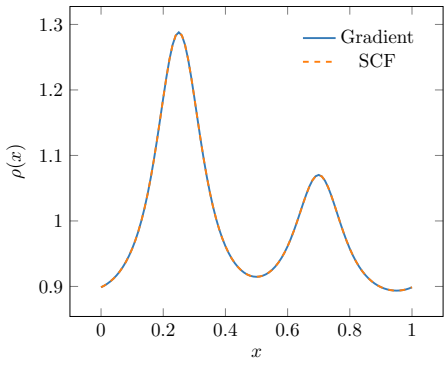
where $\mathrm{CH}(\mathcal{M}_N) = \{P \in \mathcal{H},\ P = P^*,\ 0 \leqslant P \leqslant 1,\ \mathrm{Tr}(P) = N\}$ is the convex hull of $\mathcal{M}_N$, has a unique minimizer P_*. Either $P_* \in \mathcal{M}_N$, in which case P_* is the unique minimizer of (4.6) and satisfies the Aufbau *principle, or $P_* \notin \mathcal{M}_N$, in which case the eigenvalues $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_b}$ of the mean field Hamiltonian matrix $H_* = \nabla E_\alpha(P_*)$ satisfy $\varepsilon_{N-1} < \varepsilon_N = \varepsilon_{N+1} < \varepsilon_{N+2}$ and none of the local minimizers to (4.6) satisfies the* Aufbau *principle.*

Note that the unique minimizer P_* to the relaxed constraint problem (4.8) can be computed using the optimal damping algorithm (ODA) introduced in [13]. As shown in the proof of Lemma 4.1, the only case when the minimizer P_* is not unique is the very particular case when $N_b \in 4\mathbb{N}^*$, $N_b = 2N$, and all the entries $[V_{\mathrm{eff}}]_i \coloneqq V(i\delta) + \alpha\delta^{-1}[P_*]_{ii}$ of the effective potential are equal. In the rest of this section, we consider the cases $N = 1$ and $N = 2$.

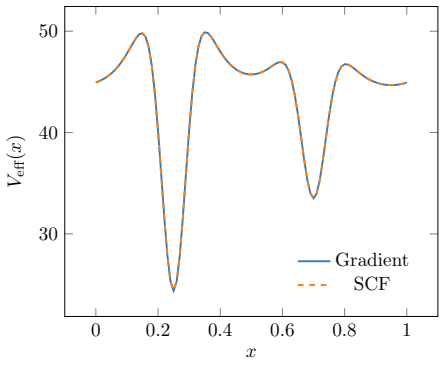
The case $N = 1$. It follows from Lemma 4.1, Theorem 3.3 and Theorem 3.4 that Algorithm 1 and Algorithm 2 locally converge to the unique minimizer P_* as long as β is chosen small enough. The resulting densities, effective potentials and convergence behavior of both algorithms are plotted in Figure 7 for $N_b = 100$. The SCF algorithm converges faster in terms of number of iterations, as a smaller β, hence more steps, are required for the gradient to converge. This is expected from the large spectral radius of the matrix h in the absence of preconditioning.

For the gradient algorithm, the convergence rate is consistent with the spectral radius of the Jacobian matrix $1 - \beta J_{\text{grad}}$. For the damped SCF algorithm with the ground state of the core Hamiltonian as starting point, surprisingly, we observe an asymptotic convergence rate slightly faster than that expected from the spectral radius of the Jacobian matrix $1 - \beta J_{\text{SCF}}$. Using a random perturbation of the ground state of the core Hamiltonian as starting point again gives a convergence rate consistent with the spectral radius.

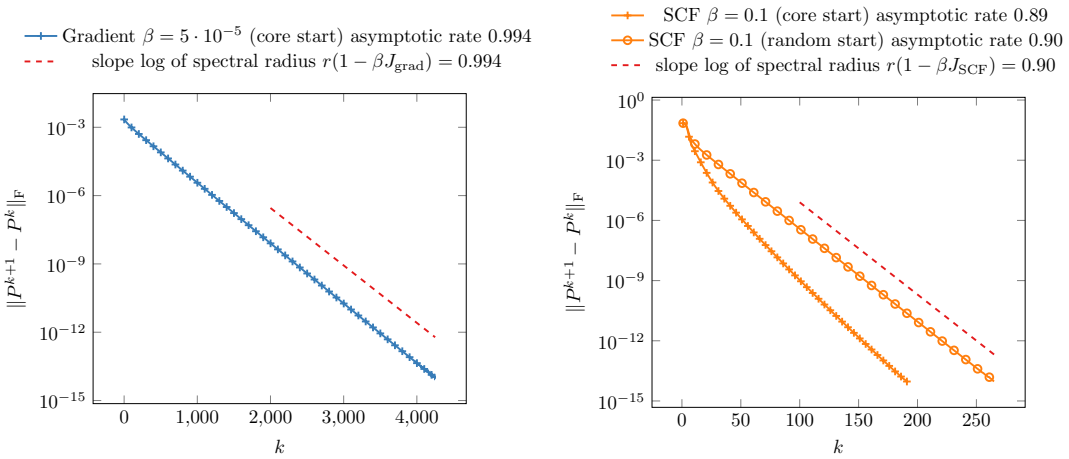
The explanation of this effect is to be found in the repartition of the error among the eigenvectors of J_{SCF}. Since both Ω_* and K_* are positive semidefinite operators, the eigenvalues of $J_{\text{SCF}} = 1 + {\Omega_*}^{-1} K_*$ are greater than 1, and the convergence for β small is limited by the modes associated with eigenvalues of J_{SCF} close to 1. These eigenvalues correspond to high eigenvalues of Ω_*, and therefore to highly oscillatory modes. When the initial guess is chosen as the ground state of the core Hamiltonian, these modes are only weakly excited and do not contribute to the observed convergence rate before convergence is achieved. When the initial guess is randomly perturbed, this effect is not present and the convergence rate is consistent with the spectral radius. For the gradient algorithm, the rate-limiting modes are associated with small eigenvalues of $\Omega_* + K_*$, which are not oscillatory, and this effect is not present either.



(a) Density ρ at convergence.



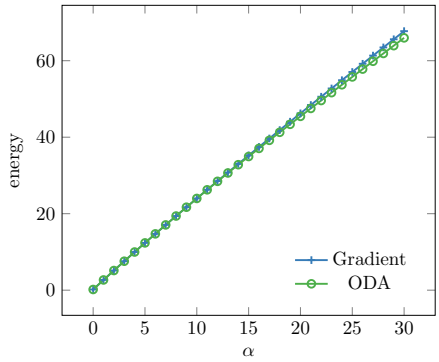
(b) Effective potential $V_{\text{eff}} = V + \alpha\rho$ at convergence.



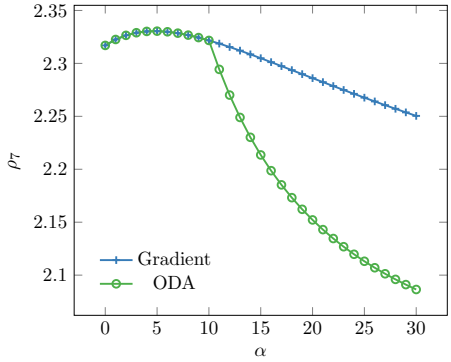
(c) Error decay for SCF and gradient descent algorithms (every few mark only is plotted for the sake of visibility). Marked lines are the evolution of the error $\left\|P^{k+1} - P^k\right\|_{\text{F}}$ and dashed lines represents the slope computed with the spectral radius of the Jacobian matrix, computed by finite differences.

FIGURE 7 – Convergence of Algorithm 1 and Algorithm 2 for $N = 1$, $\alpha = 50$ and $N_b = 100$.

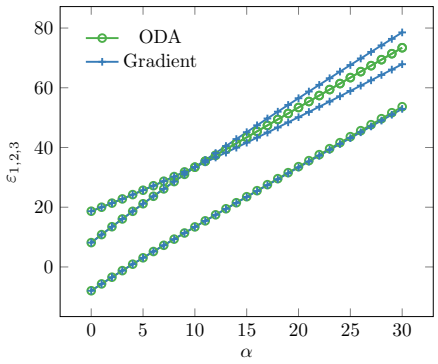
The case $N = 2$. Since the second smallest eigenvalue of the matrix h is strictly lower than the third one, for α small enough, the unique minimizer P_* of (4.8) is on $\mathcal{M}_2$ and satisfies the strong *Aufbau* principle, and both the gradient descent and SCF algorithm locally converge to P_*. For larger values of α, the two alternatives of Lemma 4.1 appear. We plot the energy, the density at an arbitrarily chosen point and the eigenvalues of the solutions P^{grad} and P^{ODA} obtained by the gradient and the ODA algorithm as a function of α in Figure 8, evidencing a bifurcation for a critical value of $\alpha_{\mathrm{c}} \simeq 10$, after which these two solutions are different.



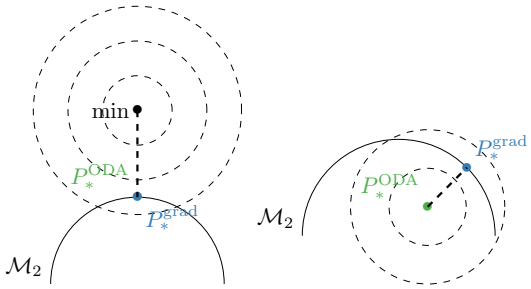
(a) Bifurcation on the energy.



(b) Bifurcation on the value of ρ_7 (7^{th} element of the discrete density for $N_b = 40$).



(c) Bifurcation on the first three eigenvalues of the Hamiltonians $H(P_*^{\mathrm{ODA}})$ and $H(P_*^{\mathrm{grad}})$. For ODA, after the collision, the second and third eigenvalues are degenerate.



(d) ODA minimum before (left) and after (right) the bifurcation. Dashed lines are the energy level sets around the minimum.

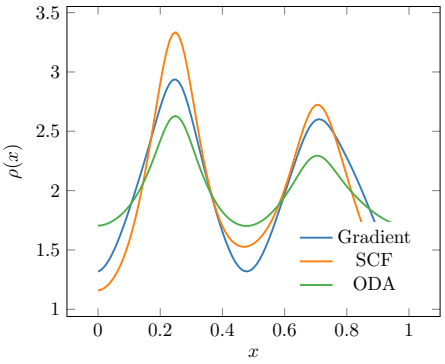
FIGURE 8 – Bifurcation on the energy, the density and the eigenvalues as a function of α for $N = 2$, $N_b = 40$.

For α lower than the critical value $\alpha_{\mathrm{c}} \simeq 10$, P_*^{ODA} is on the manifold $\mathcal{M}_2$ and satisfies the strong *Aufbau* principle. The algorithms all converge to this solution: $P_*^{\mathrm{grad}} = P_*^{\mathrm{SCF}} = P_*^{\mathrm{ODA}} = P_*$. However for $\alpha > \alpha_{\mathrm{c}}$ in the range tested, $P_*^{\mathrm{ODA}} \notin \mathcal{M}_2$. A geometrical interpretation of the bifurcation is sketched on Figure 8(d): the level sets of the function E_α are degenerate ellipsoids. Below α_{c}, the intersection of the nonempty closed convex set $\mathrm{CH}(\mathcal{M}_2)$ with the level set of E_α of lowest energy belongs to $\mathcal{M}_2$, while this is no longer the case beyond α_{c}.

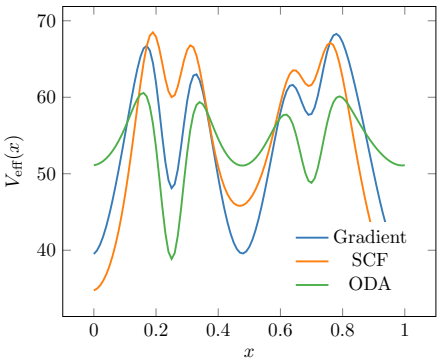
For $\alpha > \alpha_{\mathrm{c}}$, the solutions obtained by the ODA, gradient and SCF algorithm differ, as shown in Figure 9:

- the lowest second and third eigenvalues of $H(P_*^{\mathrm{ODA}})$ are degenerate ($\varepsilon_2 = \varepsilon_3$) and $P_*^{\mathrm{ODA}} \notin \mathcal{M}_2$. More precisely, we have
$$P_*^{\mathrm{ODA}} = \phi_1\phi_1^* + (1-f)\phi_2\phi_2^* + f\phi_3\phi_3^* \quad \text{with} \quad H(P_*^{\mathrm{ODA}})\phi_i = \varepsilon_i\phi_i, \quad \phi_i^*\phi_j = \delta_{ij}$$
with a fractional occupation $0 < f < 1$;
- Algorithm 1 and Algorithm 2 converge to two different limits P_*^{grad} and P_*^{SCF}, none of them satisfying the *Aufbau* principle:

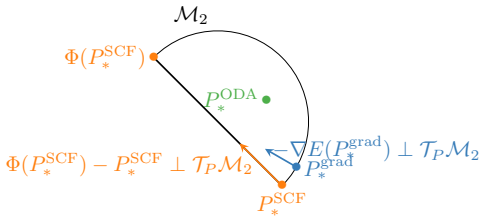
- P_*^{grad} is a local minimizer of E_α on $\mathcal{M}_2$, which does not satisfies the *Aufbau* principle. More precisely, P_*^{grad} is the orthogonal projector on the space generated by the eigenvectors associated with the lowest first and third eigenvalues of $H(P_*^{\text{grad}})$;
- P_*^{SCF} satisfies $\Pi_{P_*^{\text{SCF}}}\left(\Phi(P_*^{\text{SCF}}) - P_*^{\text{SCF}}\right) = 0$, but $\Phi(P_*^{\text{SCF}}) - P_*^{\text{SCF}} \neq 0$ and $[H(P_*^{\text{SCF}}), P_*^{\text{SCF}}] \neq 0$. The iterates are trapped as the search direction is orthogonal to the tangent space (Figure 9(c)). The limit point P_*^{SCF} is a spurious stationary state of the SCF iteration which is not physically relevant, not being a critical point of E_α.



(a) Densities of P_*^{grad}, P_*^{SCF} and P_*^{ODA}.



(b) Effective potential $V_{\text{eff}} = V + \alpha\rho_P$ for $P = P_*^{\text{grad}}$, P_*^{SCF} and P_*^{ODA}.



(c) Geometrical interpretation of the limiting points of the gradient descent, SCF et ODA algorithms.

	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

(d) Lowest three eigenvalues of $H(P)$ for $P = P_*^{\text{grad}}$, $P = P_*^{\text{SCF}}$ and $P = P_*^{\text{ODA}}$.

FIGURE 9 – Results obtained with the gradient descent, damped SCF and ODA algorithm for $N = 2$, $\alpha = 30$ and $N_b = 100$: the limiting points P_*^{grad} and P_*^{SCF} lay by construction on the manifold $\mathcal{M}_2$, while P_*^{ODA} does not (it only belongs to its convex hull $\text{CH}(\mathcal{M}_2)$).

4.5. **Kohn-Sham density functional theory.** We now investigate a more realistic computation: the electronic structure of a Silicon crystal using Kohn-Sham density functional theory (KS-DFT). We used the `DFTK.jl` code [30], which solves the equations of KS-DFT in a plane-wave basis under a pseudopotential approximation. All computations below use the local density approximation (LDA) of the exchange-correlation energy [35, 45], Goedecker-Teter-Hutter (GTH) pseudopotentials [24], a cutoff energy of 30 Hartrees, and Γ-only Brillouin zone sampling for simplicity, although the same behavior was observed with different exchange-correlation functionals and fine Brillouin zone discretizations. In all cases, the initial guess for the algorithms is a superposition of atom-centered densities. The DFTK code as well as the script used to produce these results are available at `https://dftk.org/`.

We consider the case of Silicon in its standard face-centered cubic phase. With the chosen pseudopotentials and without spin polarization, Silicon has four occupied orbitals: $N = 4$. We examine the convergence of algorithms as a function of the lattice constant a (the size of the computational domain). In the equilibrium state of Silicon ($a \approx 10.26$ Bohrs), there is a gap of about 0.08 Hartree between the occupied and virtual states. As the lattice constant a is increased, this gap decreases, until it closes at $a \approx 11.408$ Bohrs. We examine the convergence of self-consistent iterations, using both fixed-step damped density mixing with four values of the mixing parameter β ($\beta = 1$ – no damping –, $\beta = 0.5$, $\beta = 0.2$, $\beta = 0.1$), as well as the self-consistent iteration accelerated with Anderson acceleration (also known as

Pulay's DIIS method [16, 50, 51]). We plot the convergence of the density residual $\|\rho_{\Phi(P_n)} - \rho_{P_n}\|_2$ as a function of the iterations for three values of a, with decreasing gaps.

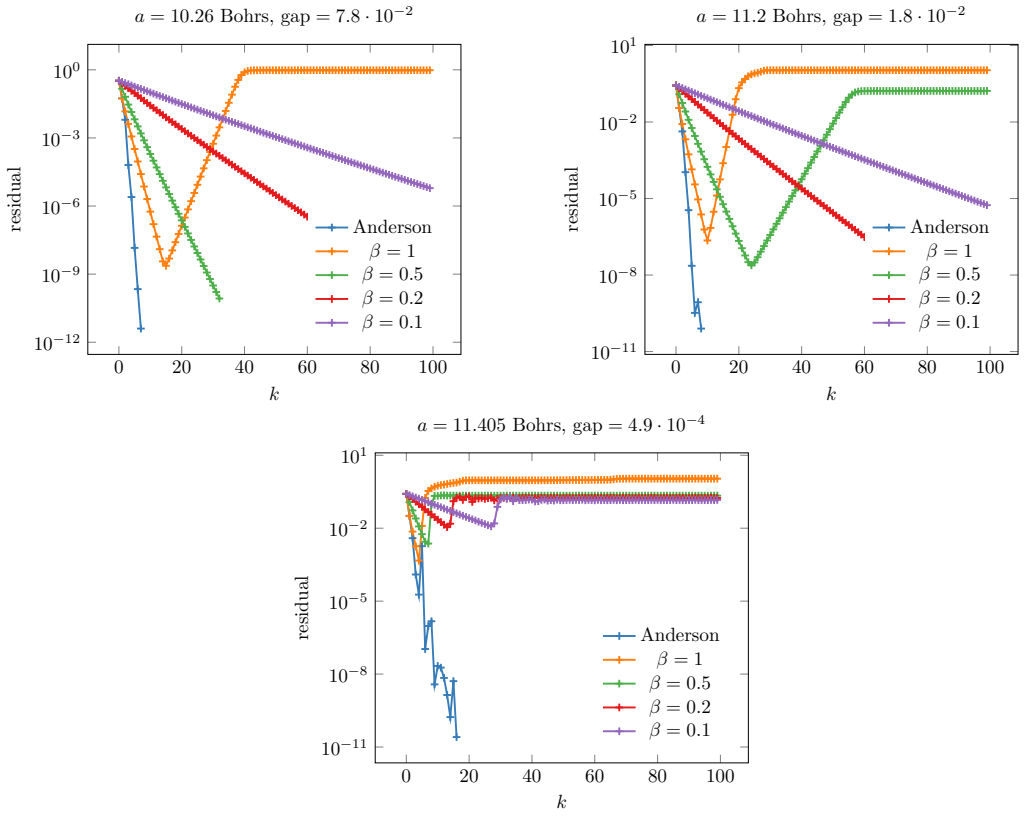


FIGURE 10 – Convergence curves of the density residual as a function of the number of iterations k for Silicon with different lattice constants a.

In the first case, with $a = 10.26$ Bohrs, the simple (undamped) SCF method appears to be converging for almost 20 iterations, but then diverges, until the density residual stabilizes at a positive value, as predicted in [14] for the Hartree-Fock model. The damped methods appear to converge. For $a = 11.2$ Bohrs, the damping method with $\beta = 0.5$ does not converge. When the lattice constant is further increased to 11.405 Bohrs, with a small gap of 4.9×10^{-4} Hartree, the fixed-step damped SCF iterations do not converge for the tested values of the damping parameter ($\beta = 1, 0.5, 0.2, 0.1$).

When it occurs, the transient behavior of apparent-convergence before eventual divergence is unusually long. For instance for $\beta = 1$ at $a = 10.26$ Bohrs, the method appears to be converging for almost 20 iterations, up to a reduction in residual of a factor 10^{-8}. In fact, it is consistent with an initial error of the order of machine precision (about 10^{-16} here) being amplified at a constant rate. The cause of this effect appears to be that the divergent modes break the natural inversion symmetry of the crystal in this particular case: we have checked that the divergence occurs much sooner if we break this symmetry by perturbing the positions of the atoms around their symmetric positions (at 9 iterations by perturbing the position of one atom by 10%). In practice, in the symmetric case, one way to overcome this issue is to ensure during the algorithm that, at each step, we have a symmetric solution. Note that this phenomenon is reminiscent of that observed in Figure 7, where all the modes were not fully excited, making the convergence faster than expected.

It is remarkable that the convergent methods (and even the divergent ones before their divergence) appear to have the exact same slope as with $a = 10.26$ Bohrs. This is consistent with our result: assuming the main effect of increasing the lattice constant is to decrease the gap, while keeping the lowest eigenvalues of $\Omega_*^{-1} K_*$ constant, then for β small enough the convergence is limited by the lowest, not the highest, eigenvalues of this operator.

In all these cases, Anderson acceleration was able to converge to a solution, even in presence of a very small gap, albeit in an irregular fashion. We attribute this to the well-known fact that, in the linear

regime, Anderson acceleration is equivalent to the GMRES algorithm. Since the GMRES algorithm is a Krylov method, it is robust to the presence of a large eigenvalue, and achieves convergence even though the underlying iteration is strongly divergent. This shows a limitation of our theoretical convergence rates, which do not capture the reduced sensitivity of accelerated methods to a small gap.

5. Conclusion

In this paper, we examined the convergence of two simple representatives in the class of direct minimization and SCF algorithms. We showed that both algorithms converge locally when the damping parameter is chosen small enough. We derived their convergence rates; we showed that the damped SCF algorithm is sensitive to the gap, while the gradient method is sensitive to the spectral radius of the Hamiltonian. We confirmed these results with numerical experiments. The goal here was not to propose efficient algorithms, but to analyze the behavior of the simplest representatives of each class. However, accelerated algorithms are generally found to follow the trend suggested by our theoretical results, although we showed that the Anderson-accelerated SCF algorithm was able to converge quickly even in the presence of a single very small gap.

In practice, should the SCF or direct minimization class of algorithms be preferred? The answer depends not only on the convergence rate studied in this paper, but also on the cost of each step, and the robustness of the algorithm. We examine two prototypical situations.

In quantum chemistry using Gaussian basis sets to solve the Hartree-Fock model or Kohn-Sham density functional theory using hybrid functionals, the rate-limiting step is often the computation of the Fock matrix $H(P)$. In this case, an iteration of a gradient descent and a damped SCF algorithm are of roughly equal cost. In most cases, solutions for isolated molecules satisfy the Aufbau principle, and the damped SCF algorithm, suitably robustified (for instance using the ODA algortithm) and accelerated (for instance with the DIIS algorithm), converges reliably and efficiently towards a solution. Direct minimization algorithms are then only useful in the cases where local or semilocal functionals are used [55] and the Aufbau principle is violated, or when SCF algorithms tend to converge to saddle points (for instance for computations involving spin).

In condensed-matter physics using plane-wave basis sets to solve the Kohn-Sham density functional theory with local or semilocal functionals, the matrices P and H are not stored explicitly. Solving the linear eigenproblem is then done using iterative block eigensolvers, which can be understood as specialized direct minimization algorithms in the case of a linear energy functional $E(P) = \mathrm{Tr}(H_0 P)$. In this case, direct minimization algorithms effectively merge the two loops of the SCF and linear eigensolver, and should therefore be more efficient. Another interest of direct minimization algorithms is their robustness, as the choice of a stepsize can be made in order to minimize the energy, unlike the damped SCF algorithm where choosing an appropriate damping parameter is often done empirically.

Despite this, direct minimization algorithms are rarely used in condensed-matter physics. The main reason seems to be that challenging problems are often metallic in character, and require the introduction of a finite temperature. Direct minimization algorithms then need to optimize over the occupations as well as the orbitals, a significantly more complex task [9, 23, 46]. A thorough comparison of the performance and robustness of direct minimization and self-consistent approach for these systems would be an interesting topic of inquiry. A number of implementation "tricks" commonly used to accelerate the convergence of iterative eigensolvers (for instance, using a block size larger than the number of eigenvectors sought) might also play a large role in performance comparison for the two classes of algorithms: understanding how to generalize these to direct minimization would be interesting.

We discussed in Remark 3.7 preconditioning for both direct minimization and SCF algorithms. The concept of preconditioning for Riemannian optimization problems seems not to have been explored much in the mathematical literature, except in some specific models and preconditioners (see for instance [7, 67] for the Gross-Pitaevskii model), and a deeper analysis of this would be interesting. In particular, this is necessary to extend the convergence theory presented in this paper to infinite-dimensional settings.

Appendix: proof of Lemma 4.1

For any $\alpha \in \mathbb{R}_+$ the energy functional E_α is smooth and convex on the nonempty compact convex set $\mathrm{CH}(\mathcal{M}_N)$. The set of minimizers to (4.8) is therefore nonempty, compact and convex. Let P_* and P'_* be two minimizers of E_α and let ρ_* and ρ'_* be their densities: $\rho_{*,i} = \delta^{-1}(P_*)_{ii}$ and $\rho'_{*,i} = \delta^{-1}(P'_*)_{ii}$. For all $\theta \in [0,1]$, we have

$$I_\alpha = E_\alpha(\theta P_* + (1-\theta)P'_*) = I_\alpha + \frac{\alpha}{2\delta}\sum_{i=1}^{N_b}\Big((\theta\rho_{*,i} + (1-\theta)\rho'_{*,i})^2 - (\theta\rho_{*,i}^2 + (1-\theta){\rho'_{*,i}}^2)\Big),$$

where $I_\alpha = E_\alpha(P_*) = E_\alpha(P'_*)$ is the minimum of (4.8). Since the function $\mathbb{R} \ni x \mapsto x^2 \in \mathbb{R}$ is strictly convex, we obtain that $\rho_* = \rho'_*$. Therefore, all the minimizers of (4.8) share the same density, hence the same mean-field Hamiltonian matrix H_*. If P_* is a minimizer of (4.8), it satisfies the first order optimality condition (Euler inequality)

$$\forall\ P \in \mathrm{CH}(\mathcal{M}_N), \quad \mathrm{Tr}(H_*(P - P_*)) \geqslant 0,$$

from which we infer by a classical argument that

$$P_* = \mathbb{1}_{(-\infty,\mu)}(H_*) + Q_* \quad \text{with} \quad 0 \leqslant Q_* \leqslant 1, \quad \mathrm{Ran}(Q_*) \subset \mathrm{Ker}(H_* - \mu), \quad \mathrm{Tr}(P_*) = N, \tag{5.1}$$

for some Fermi level $\mu \in \mathbb{R}$ (the Lagrange multiplier of the constraint $\mathrm{Tr}(P) = N$). Let $\varepsilon_1 \leqslant \cdots \leqslant \varepsilon_{N_b}$ be the eigenvalues of H_*, counting multiplicities. If $\varepsilon_N < \varepsilon_{N+1}$, then we necessarily have $P_* = \mathbb{1}_{(-\infty,\varepsilon_N]}(H_*)$, so that (4.8) has a unique minimizer, P_* is on $\mathcal{M}_N$ and therefore is also the unique minimizer of (4.6), and it satisfies the strong *Aufbau* principle.

Let us now consider the case when $\varepsilon_N = \varepsilon_{N+1} =: \mu$. Since the eigenvalue problem $H_*\psi = \mu\psi$ is a second-order difference equation

$$\frac{-\psi_{i+1} + 2\psi_i - \psi_{i-1}}{2\delta^2} + V_{\mathrm{eff},i}\psi_i = \mu\psi_i, \quad 1 \leqslant i \leqslant N_b, \tag{5.2}$$

(here and in the sequel we use the convention that $\psi_0 = \psi_{N_b}$ and $\psi_{N_b+1} = \psi_1$) with $V_{\mathrm{eff}} = V + \alpha\rho_*$, the eigenspace $\mathrm{Ker}(H_* - \mu)$ is at most of dimension 2. We therefore have $\varepsilon_{N-1} < \varepsilon_N = \varepsilon_{N+1} < \varepsilon_{N+2}$.

Using the variational characterization of the ground state eigenvalue, we have

$$\varepsilon_1 = \min_{\psi \in \mathbb{R}^{N_b},\, \psi^*\psi = 1} \psi^* H_* \psi \quad \text{with} \quad \psi^* H_* \psi = \sum_{i=1}^{N_b}\left|\frac{\psi_{i+1} - \psi_i}{\delta}\right|^2 + \sum_{i=1}^{N_b} V_{\mathrm{eff},i}|\psi_i|^2. \tag{5.3}$$

Since $||x| - |y|| \leqslant |x - y|$ for all $x, y \in \mathbb{R}$ with equality if and only if x and y have the same sign, we infer from (5.3), that all the entries of a ground state eigenvector of H_* have the same sign. In particular, two normalized ground state eigenvectors of H_* cannot be orthogonal. This implies that the ground state eigenvalue of H_* is simple, i.e. $\varepsilon_1 < \varepsilon_2$. The first statement of Lemma 4.1 staightforwardly follows from the results established so far.

To prove the second statement, assume that $N \geqslant 2$ and that (4.8) has two distinct minimizers P_* and P'_* sharing the same density. In view of (5.1), this can only occur if $\varepsilon_N = \varepsilon_{N+1} =: \mu$. Using an orthonormal basis (ϕ, ψ) of $\mathrm{Ker}(H_* - \mu)$ consisting of eigenvectors of P_*, we can assume without loss of generality that

$$\begin{aligned} P_* &= \mathbb{1}_{(-\infty,\mu)}(H_*) + (1-f)\phi\phi^* + f\psi\psi^*, \\ P'_* &= \mathbb{1}_{(-\infty,\mu)}(H_*) + (1-a)\phi\phi^* + a\psi\psi^* + b(\phi\psi^* + \psi\phi^*), \end{aligned}$$

with $0 \leqslant f \leqslant 1$, $0 \leqslant a \leqslant 1$ and $b^2 \leqslant a(1-a)$. Since P_* and P'_* have the same density, we have for all $1 \leqslant i \leqslant N_b$, $(1-f)\phi_i^2 + f\psi_i^2 = (1-a)\phi_i^2 + a\psi_i^2 + 2b\phi_i\psi_i$, that is $(a-f)\phi_i^2 - 2b\phi_i\psi_i - (a-f)\psi_i^2 = 0$.

If $a = f$, then $b \neq 0$ since $P_* \neq P'_*$ by assumption, so that $\phi_i\psi_i = 0$ for all $1 \leqslant i \leqslant N_b$. From (5.2), we see that it is not possible to have $\psi_i = \psi_{i+1} = 0$ (otherwise, ψ would be identically equal to zero), and the same holds true for ϕ. Therefore, N_b must be even, and either all the odd entries of ϕ and all the even entries ψ must vanish, or the other way round. We then infer from (5.2) that this implies that $\phi_{i+2} + \phi_i = \psi_{i+2} + \psi_i = 0$ for all $1 \leqslant i \leqslant N_b$, and that all the entries of V_{eff} are equal to $\mu - \delta^{-2}$. This implies that $N_b \in 4\mathbb{N}^*$ and that the states ϕ and ψ are given by $\phi_{2i} = c(-1)^i$, $\phi_{2i+1} = 0$, $\psi_{2i} = 0$, $\psi_{2i+1} = c'(-1)^i$ for all $1 \leqslant i \leqslant N_b$, where c and c' are normalization constants. By explicit diagonalization of the matrix $H_* = H(0) + (\mu - \delta^{-2})I_{N_b}$ one can check that the states ϕ and ψ are

therefore those spanning the two-dimensional space associated to the two-fold degenerate eigenvalues $\varepsilon_{N_b/2} = \varepsilon_{1+N_b/2}$ of H_*. This is only possible if $N = N_b/2$. The case $\alpha = f$ can thus be excluded for $2 \leqslant N \leqslant N_b$, with $N_b \neq 2N$ if $N_b \in 4\mathbb{N}^*$.

If $a \neq f$, we have for all $1 \leqslant i \leqslant N_b$, $\phi_i^2 - 2\gamma\phi_i\psi_i - \psi_i^2 = 0$, for $\gamma = \frac{b}{a-f}$, and up to replacing ψ with $-\psi$, we can assume without loss of generality that $\gamma \geqslant 0$. Denoting by $C_\pm := \gamma \pm \sqrt{1+\gamma^2}$ the roots of the polynomial $x^2 - 2\gamma x + 1$, with $C_+C_- = -1$, we obtain that for each $1 \leqslant i \leqslant N_b$, either $\phi_i = C_+\psi_i$ or $\phi_i = C_-\psi_i$. Using the discrete Schrödinger equation (5.2) satisfied by both ϕ and ψ, we see that if $\phi_i = C_+\psi_i$ and $\phi_{i+1} = C_+\psi_{i+1}$ for some $1 \leqslant i \leqslant N_b$, then $\phi = C_+\psi$, and likewise if C_+ is replaced by C_-. This is impossible since ϕ and ψ are orthonormal. Therefore, we must have $\phi_{2i} = C_+\psi_{2i}$ and $\phi_{2i+1} = C_-\psi_{2i+1}$ (or the other way around), and N_b must be even. Using again (5.2), this leads to $\phi_{i+2} + \phi_i = 0$ and $\psi_{i+2} + \psi_i = 0$ for all $1 \leqslant i \leqslant N_b$ and therefore, as in the previous case, that $N_b \in 4\mathbb{N}^*$, $N_b = 2N$, that all the entries of V_{eff} are equal and that ϕ and ψ span the two-dimensional space associated to the two-fold degenerate eigenvalues $\varepsilon_N = \varepsilon_{N+1}$ of H_*.

This proves that for $2 \leqslant N \leqslant N_b$, with $N_b \neq 2N$ if $N_b \in 4\mathbb{N}^*$, (4.8) has a unique minimizer P_*. If $P_* \in \mathcal{M}_N$, it is of course also the unique minimizer of (4.6), and P_* satisfies the *Aufbau* principle. Conversely, if $P'_* \in \mathcal{M}_N$ is a local minimizer of (4.6) satisfying the *Aufbau* principle, we have

$$\forall\ P \in \mathrm{CH}(\mathcal{M}_N), \quad \mathrm{Tr}(H(P'_*)(P - P'_*)) \geqslant 0,$$

which means that P'_* is a solution to the Euler inequality for (4.8), and therefore a global minimizer of this convex problem. Since the minimizer P_* of (4.8) is unique, we finally obtain that if $P_* \notin \mathcal{M}_N$, then none of the local minimizers of (4.6) satisfies the *Aufbau* principle.

6. Acknowledgement

The authors would like to thank Michael F. Herbst and Sami Siraj-Dine for fruitful discussions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 810367).

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[2] F. Alouges and C. Audouze. Preconditioned gradient flows for nonlinear eigenvalue problems and application to the Hartree-Fock functional. *Numerical Methods for Partial Differential Equations. An International Journal*, 25(2):380–400, 2009.

[3] X. Antoine, A. Levitt, and Q. Tang. Efficient spectral computation of the stationary states of rotating Bose-Einstein condensates by preconditioned nonlinear conjugate gradient methods. *Journal of Computational Physics*, 343:92–109, 2017.

[4] V. Bach, E. Lieb, L. M., and J. Solovej. There are no unfilled shells in unrestricted Hartree-Fock theory. *Physical Review Letters*, 72(19):2981–2983, 1994.

[5] G. Bacskay. A quadratically convergent Hartree-Fock (QC-SCF) method. Application to closed shell systems. *Chemical Physics*, 61(3):385–404, 1981.

[6] W. Bao and Y. Cai. Mathematical theory and numerical methods for Bose-Einstein condensation. *Kinetic and Related Models*, 6:1, 2013.

[7] W. Bao and Q. Du. Computing the Ground State Solution of Bose–Einstein Condensates by a Normalized Gradient Flow. *SIAM Journal on Scientific Computing*, 25(5):1674–1697, Jan. 2004.

[8] E. H. Brown and J. W. Milnor. Topology from the differentiable viewpoint. *American Mathematical Monthly*, 74(4):461, 1967.

[9] E. Cancès. Self-consistent field algorithms for Kohn-Sham models with fractional occupation numbers. *The Journal of Chemical Physics*, 114(24):10616–10622, 2001.

[10] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *Journal of Scientific Computing*, 45(1-3):90–117, 2010.

[11] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models. *ESAIM. Mathematical Modelling and Numerical Analysis*, 46(2):341–388, 2012.

[12] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: A primer. volume X of *Handbook of Numerical Analysis*, pages 3–270. North-Holland, Amsterdam, 2003.

[13] E. Cancès and C. Le Bris. Can we outperform the DIIS approach for electronic structure calculations? *Int. J. of Quantum Chem.*, 79(2):82–90, 2000.

[14] E. Cancès and C. Le Bris. On the convergence of SCF algorithms for the Hartree-Fock equations. *ESAIM. Mathematical Modelling and Numerical Analysis*, 34(4):749–774, 2000.

[15] G. Chaban, M. Schmidt, and M. Gordon. Approximate second order method for orbital optimization of SCF and MCSCF wavefunctions. *Theoretical Chemistry Accounts*, 97(1):88–95, 1997.
[16] M. Chupin, M.-S. Dupuy, G. Legendre, and E. Séré. Convergence analysis of adaptive DIIS algorithms with application to electronic ground state calculations. arXiv:2002.12850, 2020.
[17] X. Dai, Z. Liu, L. Zhang, and A. Zhou. A conjugate gradient method for electronic structure calculations. *SIAM Journal on Scientific Computing*, 39:A2702–A2740, 2017.
[18] I. Danaila and B. Protas. Computation of ground states of the Gross-Pitaevskii functional via Riemannian optimization. *SIAM Journal on Scientific Computing*, 39:B1102–B1129, 2017.
[19] P. Dederichs and R. Zeller. Self-consistency iterations in electronic-structure calculations. *Physical Review B*, 28(10):5462, 1983.
[20] W. E and J. Lu. *The Kohn-Sham Equation for Deformed Crystals*, volume 221 of *Mem. Amer. Math. Soc.* American Mathematical Society, 2013.
[21] A. Edelman, T. Arias, and S. Smith. The geometry of algorithms with orthonormality constraints. *J. Matrix Anal. Appl.*, 20:303–353, 1998.
[22] J. Francisco, J. Martinez, and L. Martinez. Globally convergent trust-region methods for self-consistent field electronic structure calculations. *Journal of Chemical Physics*, 121:10863–10878, 2004.
[23] C. Freysoldt, S. Boeck, and J. Neugebauer. Direct minimization technique for metals in density functional theory. *Physical Review B*, 79(24):241103, 2009.
[24] S. Goedecker, M. Teter, and J. Hutter. Separable dual-space Gaussian pseudopotentials. *Physical Review B*, 54(3):1703, 1996.
[25] X. Gonze. Towards a potential-based conjugate gradient algorithm for order-N self-consistent total energy calculations. *Physical Review B*, 54(7):4383, 1996.
[26] D. R. Hartree. The wave mechanics of an atom with a non-Coulomb central field. Part II. Some results and discussion. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(1):111–132, 1928.
[27] P. Heid, B. Stamm, and T. P. Wihler. Gradient flow finite element discretizations with energy-based adaptivity for the Gross-Pitaevskii equation. 2019.
[28] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. John Wiley & Sons, 2014.
[29] P. Henning and D. Peterseim. Sobolev gradient flow for the Gross-Pitaevskii eigenvalue problem: Global convergence and computational efficiency. arXiv:1812.00835, 2018.
[30] M. F. Herbst and A. Levitt. https://dftk.org. Mar. 2020.
[31] N. J. Higham. *Functions of Matrices*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, Jan. 2008.
[32] R. B. Holmes. A formula for the spectral radius of an operator. *American Mathematical Monthly*, 75(2):163, 1968.
[33] D. Johnson. Modified Broyden's method for accelerating convergence in self-consistent calculations. *Physical Review B*, 38(18):12807–12813, 1988.
[34] T. Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer-Verlag, Berlin Heidelberg, second edition, 1995.
[35] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.
[36] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996.
[37] K. Kudin, G. Scuseria, and E. Cancès. A black-box self-consistent field convergence algorithm: One step closer. *Journal of Chemical Physics*, 116:8255–8261, 2002.
[38] A. Levitt. Convergence of gradient-based algorithms for the Hartree-Fock equations. *ESAIM. Mathematical Modelling and Numerical Analysis*, 46(6):1321–1336, 2012.
[39] L. Lin and J. Lu. *A Mathematical Introduction to Electronic Structure Theory*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2019.
[40] L. Lin, J. Lu, and L. Ying. Numerical methods for Kohn–Sham density functional theory. *Acta Numerica*, 28:405–539, May 2019.
[41] X. Liu, X. Wang, Z. Wen, M. Ulbrich, and Y. Yuan. On the analysis of the discretized Kohn-Sham density functional theory. *SIAM Journal on Numerical Analysis*, 53:1758–1785, 2015.
[42] X. Liu, X. Wang, Z. Wen, and Y.-X. Yuan. On the convergence of the self-consistent field iteration in Kohn-Sham Density Functional Theory. *SIAM Journal on Matrix Analysis and Applications*, 35:546–558, 2013.
[43] X. Liu, Z. Wen, X. Wang, M. Ulbrich, and Y. Yuan. On the Analysis of the Discretized Kohn-Sham Density Functional Theory. *SIAM Journal on Numerical Analysis*, 53(4):1758–1785, Jan. 2015.
[44] L. Marks and D. Luke. Robust mixing for ab initio quantum mechanical calculations. *Physical Review B*, 78(7):075114, 2008.
[45] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2004.
[46] N. Marzari, D. Vanderbilt, and M. C. Payne. Ensemble density-functional theory for ab initio molecular dynamics of metals and finite-temperature insulators. *Physical review letters*, 79(7):1337, 1997.
[47] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261(5560):459–467, June 1976.
[48] R. McWeeny. The density matrix in self-consistent field theory. I. Iterative construction of the density matrix. *Proceedings of the Royal Society of London A*, 235:496–509, 1956.
[49] A. Mostofi, P. Haynes, C.-K. Skylaris, and M. Payne. Preconditioned iterative minimization for linear-scaling electronic structure calculations. *Journal of Chemical Physics*, 119(17):8842–8848, 2003.
[50] P. Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chemical Physics Letters*, 73(2):393–398, 1980.
[51] P. Pulay. Improved SCF convergence acceleration. *Journal of Computational Chemistry*, 3(4):556–560, 1982.

[52] D. Raczkowski, A. Canning, and L. W. Wang. Thomas-Fermi charge mixing for obtaining self-consistency in density functional calculations. *Physical Review B*, 64(12):121101, 2001.
[53] T. Rohwedder and R. Schneider. An analysis for the DIIS acceleration method used in quantum chemistry calculations. *Journal of Mathematical Chemistry*, 49(9):1889, 2011.
[54] C. Roothaan. New developments in molecular orbital theory. *Reviews of Modern Physics*, 23(2):69–89, 1951.
[55] E. Rudberg. Difficulties in applying pure Kohn-Sham density functional theory electronic structure methods to protein molecules. *Journal of Physics: Condensed Matter*, 24(7):072202, 2012.
[56] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical Methods for Electronic Structure Calculations of Materials. *SIAM Review*, 52(1):3–54, Jan. 2010.
[57] G. Srivastava. Broyden's method for self-consistent field convergence acceleration. *Journal of Physics A*, 17(6):L317–L321, 1984.
[58] S. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC press, 2001.
[59] L. Thorgersen, J. Olsen, D. Yeager, L. Jorgensen, P. Salek, and T. Helgaker. The trust-region self-consistent field method: Towards a black-box optimization in Hartree-Fock and Kohn-Sham theories. *Journal of Chemical Physics*, 121(1):16–27, 2004.
[60] P. Upadhyaya, E. Jarlebring, and E. Rubensson. A density matrix approach to the convergence of the self-consistent field iteration. *Numerical Algebra, Control and Optimization*, (2155-3289_2019_0_43), 2019.
[61] E. Vecharynski, C. Yang, and J. E. Pask. A projected preconditioned conjugate gradient algorithm for computing a large invariant subspace of a Hermitian matrix. *Journal of Computational Physics*, 290:73–89, 2015.
[62] N. Woods. *On the Nature of Self-Consistency in Density Functional Theory*. PhD thesis, University of Cambridge, 2018.
[63] N. Woods, M. Payne, and P. Hasnip. Computing the self-consistent field in Kohn–Sham density functional theory. *Journal of Physics: Condensed Matter*, 31(45):453001, Aug. 2019.
[64] C. Yang, W. Gao, and J. C. Meza. On the convergence of the self-consistent field iteration for a class of nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1773–1788, 2009.
[65] C. Yang, J. Meza, and L. Wang. A trust region direct constrained minimization algorithm for the Kohn-Sham equation. *SIAM Journal on Scientific Computing*, 29:1854–1875, 2007.
[66] X. Zhang, J. Zhu, Z. Wen, and A. Zhou. Gradient type optimization methods for electronic structure calculations. *SIAM Journal on Scientific Computing*, 36:265–289, 2014.
[67] Z. Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. Dec. 2019.
[68] Z. Zhao, Z. Bai, and X. Jin. A Riemannian Newton algorithm for nonlinear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 36:752–774, 2015.