Convergence of gradient-based algorithms for the Hartree-Fock equations

Antoine Levitt

CONVERGENCE OF GRADIENT-BASED ALGORITHMS FOR THE HARTREE-FOCK EQUATIONS

ANTOINE LEVITT

Abstract. The numerical solution of the Hartree-Fock equations is a central problem in quantum chemistry for which many algorithms exist. Attempts to justify these algorithms mathematically have been made, notably in [2], but no algorithm has yet been proved to convergence satisfactorily. In this paper, we prove the convergence of a natural gradient algorithm, using a gradient inequality for analytic functionals due to Łojasiewicz [7]. Then, expanding upon the analysis of [2], we prove convergence results for the Roothaan and Level-Shifting algorithms. In each case, our method of proof provides estimates on the convergence rate. We compare these with numerical results for the algorithms studied.

1. Introduction

In quantum chemistry, the Hartree-Fock method is one of the simplest approximations of the electronic structure of a molecule. By assuming minimal correlation between the N electrons, it reduces Schrödinger's equation, a linear partial differential equation on $\mathbb{R}^{3N}$, to the Hartree-Fock equations, a system of N coupled nonlinear equations on $\mathbb{R}^3$. This approximation makes it much more tractable numerically. It is used both as a standalone description of the molecule and as a starting point for more advanced methods, such as the Møller-Plesset perturbation theory, or multi-configuration methods. Mathematically, the Hartree-Fock method leads to a coupled system of nonlinear integro-differential equations, which are discretized by expanding the solution on a finite Galerkin basis. The resulting nonlinear algebraic equations are then solved iteratively, using a variety of algorithms, the convergence of which is the subject of this work.

The mathematical structure of the Hartree-Fock equations was investigated in the 70's, culminating in the proof of the existence of solutions by Lieb and Simon [5], later generalized by Lions [6]. On the other hand, despite their ubiquitous use in computational chemistry, the convergence of the various algorithms used to solve them is still poorly understood. A major step forward in this direction is the recent work of Cancès and LeBris [2]. Using the density matrix formulation, they provided a mathematical explanation for the oscillatory behavior observed in the simplest algorithm, the Roothaan method, and proposed the Optimal Damping Algorithm (ODA), a new algorithm inspired directly by the mathematical structure of the constraint set. This algorithm was designed to decrease the energy at each step, and linking the energy decrease to the difference of iterates allowed them to prove that this algorithm "numerically converges" in the weak sense that $\|D_k - D_{k-1}\| \to 0$.

However, this is still mathematically unsatisfactory, as it does not guarantee convergence, and merely prohibits fast divergence. The difficulty in proving convergence of the algorithms used to solve the Hartree-Fock equations lies in the lack of understanding of the second-order properties of the functional (for instance, there are no local uniqueness results available). In other domains, the convergence of gradient-based methods has been established using the Łojasiewicz inequality for analytic functionals [7] (see for instance [3, 8]). This method of proof has the advantage of not requiring any second-order information.

In this paper, we introduce a gradient-based algorithm to solve the Hartree-Fock equations. To our knowledge, this algorithm is new. Although it is by no means efficient and is unlikely to be of much practical use, it is the most natural generalisation of the classical gradient descent. It is the only algorithm for which we were able to prove unconditional convergence towards a solution of the Hartree-Fock equations. We do so, following the method of proof of [8], and obtain explicit estimates on its convergence rate. We also apply the method to the widely used Roothaan and Level-Shifting algorithms, although the conclusions are weaker.

Date: September 26, 2011.

2010 *Mathematics Subject Classification.* 35Q40,65K10.

Key words and phrases. Hartree-Fock equations, Łojasiewicz inequality, optimization on manifolds.

This work was partially supported by the NoNAP project (ANR-10-BLAN 0101) funded by the ANR.

The structure of this paper is as follows. We first introduce the Hartree-Fock problem in the mathematical setting of density matrices and prove a Łojasiewicz inequality on the constrained parameter space. We then introduce the gradient algorithm, and prove some estimates. We show the convergence and obtain convergence rates for this algorithm, then extend our method to the Roothaan and Level-Shifting algorithm, using an auxiliary energy functional following [2]. We finally test all these results numerically and compare the convergence of the algorithms.

2. Setting

We are concerned with the numerical solution of the Hartree-Fock equations. We will consider for simplicity of notation the spinless Hartree-Fock equations, where each orbital ϕ_i is a function in $L^2(\mathbb{R}^3, \mathbb{R})$, although our results are easily transposed to other variants such as General Hartree-Fock (GHF) and Restricted Hartree-Fock (RHF).

We will consider a Galerkin discretization space with finite orthonormal basis $(\chi_i)_{i=1\dots N_b}$. In this setting, the orbitals ϕ_i are expanded on the basis, and the operators we consider become $N_b \times N_b$ matrices.

The Hartree-Fock problem consists in minimizing the total energy of a N-body system. We describe the mathematical structure of the energy functional and the minimization set, and propose a natural gradient descent to solve this problem numerically.

2.1. The energy.

We consider the quantum N-body problem of N electrons in a potential field V (in most applications, V is the Coulombian potential created by a molecule or atom). In the spinless Hartree-Fock model, this problem is simplified by assuming that the N-body wavefunction ψ is a single Slater determinant of N orthogonal orbitals ϕ_i. A simple calculation then shows that the energy of the wavefunction ψ can be expressed as a function of the orbitals ϕ_i,

$$E(\psi) = \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{1}{2}(\nabla\phi_i)^2 + \int_{\mathbb{R}^3} V\rho + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y) - \tau(x,y)^2}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y,$$

where $\tau(x,y) = \sum_{i=1}^{N} \phi_i(x)\phi_i(y)$ and $\rho(x) = \tau(x,x)$.

The energy is then to be minimized over all sets of orthogonal orbitals ϕ_i. An alternative way of looking at this problem is to reformulate it using the density operator D. This operator, defined by its integral kernel $D(x,y) = \tau(x,y)$, can be seen to be the projection operator on the space spanned by the ϕ_i's. The energy can be written as a function of D only:

$$E(D) = \mathrm{Tr}((h + \frac{1}{2}G(D))D), \tag{2.1}$$

where

$$h = -\frac{1}{2}\Delta + V,$$
$$(G(D)\phi)(x) = \left(\rho \star \frac{1}{|\cdot|}\right)(x)\phi(x) - \int_y \frac{\phi(y)\tau(x,y)}{|x-y|}.$$

This time, the energy is to be minimized over all projection operators of rank N.

2.2. The manifold $\mathcal{P}$.

The Hartree-Fock energy is to be minimized over the set of density matrices

$$D \in \mathcal{P} = \{D \in M_{N_b}(\mathbb{R}), D^T = D, D^2 = D, \mathrm{Tr}\, D = N\}.$$

The manifold $\mathcal{P}$ is equipped with the canonical inner product in $M_{N_b}(\mathbb{R})$

$$\langle A, B\rangle = \mathrm{Tr}(A^T B).$$

We denote by $\|A\|_F = \sqrt{\langle A, A\rangle}$ the Frobenius (or Hilbert-Schmidt) norm of A and by $\|A\|_{\mathrm{op}} = \max_{\|x\|=1}\|Ax\|$ the operator norm of A. We write $\|A\|$ without subscript for the Frobenius norm, which is the most natural here.

The Riemannian structure of $\mathcal{P}$ allows us to compute the gradient of E. The tangent space $T_D\mathcal{P}$ at some point D is the set of Δ such that $D+\Delta$ verifies the constraints up to first order in Δ, that is, such that $\Delta^T = \Delta, D\Delta + \Delta D = \Delta, \mathrm{Tr}\,\Delta = 0$. Block-decomposing Δ on the two subspaces $\mathrm{im}D$ and $\ker D$, this implies that the tangent space $T_D\mathcal{P}$ is the set of matrices Δ of the form

$$\Delta = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix},$$

where A is an arbitrary $N_b \times N$ matrix.

Hence, the projection on the tangent space of an arbitrary symmetric matrix M is given by

$$\begin{aligned} P_D(M) &= DM(1-D) + (1-D)MD \\ &= [D,[D,M]]. \end{aligned}$$

Note that if M has decomposition $\begin{pmatrix} B & A^T \\ A & C \end{pmatrix}$, then $[D,[D,M]] = \begin{pmatrix} 0 & A^T \\ A & 0 \end{pmatrix}$ and $[D,M] = \begin{pmatrix} 0 & -A^T \\ A & 0 \end{pmatrix}$, which shows that $\left\|[D[D,M]]\right\| = \left\|[D,M]\right\|$, a property that will be useful in the sequel.

We can now compute the gradient of E. First, the unconstrained gradient in $M_{N_b}(\mathbb{R})$ is

$$\nabla E(D) = F_D = h + G(D),$$

the Fock operator describing the mean field generated by the electrons of D. We obtain the constrained gradient $\nabla_{\mathcal{P}} E$ by projecting onto the tangent space:

$$\begin{aligned} \nabla_{\mathcal{P}} E(D) &= P_D(\nabla E(D)) \\ &= [D,[D,F_D]]. \end{aligned}$$

2.3. **Łojasiewicz inequality.** The Łojasiewicz inequality for a functional f around a critical point x_0 is a local inequality that bounds the gradient of f from below by the value of the functional. This inequality quantifies the intuition that, around x_0, as long as $f(x)$ is larger than $f(x_0)$, there is a gradient we can use to get closer to x_0. Its only hypothesis is analyticity. In particular, no second order information is needed, and the inequality accommodates degenerate critical points.

2.3.1. *The classical Łojasiewicz inequality.*

Theorem 2.1 (Łojasiewicz inequality in $\mathbb{R}^n$)**.** *Let f be an analytic functional from $\mathbb{R}^n$ to $\mathbb{R}$. Then, for each $x_0 \in \mathbb{R}^n$, there is a neighborhood U of x_0 and two constants $\kappa > 0$, $\theta \in (0, 1/2]$ such that when $x \in U$,*

$$\left|f(x) - f(x_0)\right|^{1-\theta} \leq \kappa \left\|\nabla f(x)\right\|.$$

This inequality is trivial when x_0 is not a critical point. In the special case where the Hessian $\nabla^2 f(x_0)$ is invertible, a simple Taylor expansion shows that we can choose $\theta = \frac{1}{2}$ and $\kappa > \frac{1}{\sqrt{2\lambda_1}}$, where λ_1 is the lowest eigenvalue of $\nabla^2 f(x_0)$. When $\nabla^2 f(x_0)$ is singular (meaning that x_0 is a degenerate critical point), the analyticity hypothesis ensures that the derivatives cannot all vanish, and that a similar result holds with a smaller θ.

2.3.2. *Łojasiewicz inequality on $\mathcal{P}$.* We now extend this inequality to functionals defined on the manifold $\mathcal{P}$.

Theorem 2.2 (Łojasiewicz inequality on $\mathcal{P}$)**.** *Let f be an analytic functional from $\mathcal{P}$ to $\mathbb{R}$. Then, for each $D_0 \in \mathcal{P}$, there is a neighborhood U of D_0 and two constants $\kappa > 0$, $\theta \in (0, 1/2]$ such that when $D \in U$,*

$$\left|f(D) - f(D_0)\right|^{1-\theta} \leq \kappa \left\|\nabla_{\mathcal{P}} f(D)\right\|.$$

Proof. Let $D_0 \in \mathcal{P}$. Define the map R_{D_0} from $T_{D_0}\mathcal{P}$ to $\mathcal{P}$ by

$$\begin{aligned} R_{D_0}(\Delta) &= U D_0 U^T, \\ U &= \exp(-[D_0, \Delta]). \end{aligned}$$

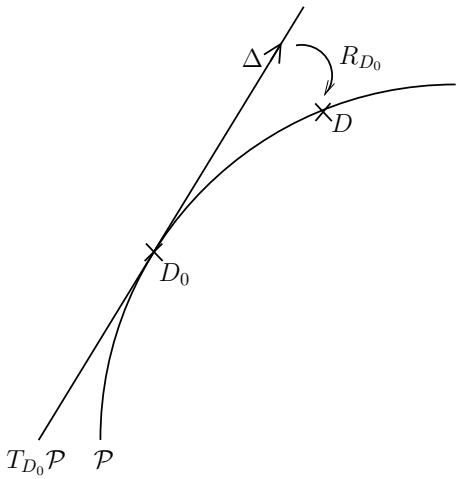

FIGURE 2.1. The map R_{D_0}

This map is analytic and verifies $R_{D_0}(0) = D_0$, $\mathrm{d}R_{D_0}(0) = \mathrm{id}_{T_{D_0}\mathcal{P}}$. Therefore, by the inverse function theorem, the image of a neighborhood of zero contains a neighborhood of D_0. We now compute the gradient of $\widetilde{f} = f \circ R_{D_0}$ at a point Δ, with $D = R_{D_0}(\Delta)$.

$$\begin{aligned}\widetilde{f}(\Delta+\delta) &= f(D) + \left\langle \nabla_{\mathcal{P}} f(D), \mathrm{d}R_{D_0}(\Delta)\delta \right\rangle + O(\delta^2)\\ &= f(D) + \left\langle \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D), \delta \right\rangle + O(\delta^2)\\ &= f(D) + \left\langle P_{D_0}\, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D), \delta \right\rangle + O(\delta^2)\end{aligned}$$

We deduce

$$\nabla_{T_{D_0}\mathcal{P}} \widetilde{f}(\Delta) = P_{D_0}\, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D).$$

We can now apply the Łojasiewicz inequality to $\widetilde{f}$, which is an analytic functional defined on the Euclidean space $T_{D_0}\mathcal{M}$. We obtain a neighborhood of zero in $T_{D_0}\mathcal{P}$, and therefore a neighborhood U of D_0 on which

$$\left|f(D) - f(D_0)\right|^{1-\theta} \leq \kappa \left\| P_{D_0}\, \mathrm{d}R_{D_0}(\Delta)^* \nabla_{\mathcal{P}} f(D) \right\|.$$

As $\mathrm{d}R^*_{D_0}$ is continuous in Δ, up to a change in the neighborhood U and the constant κ,

$$\left|f(D) - f(D_0)\right|^{1-\theta} \leq \kappa \left\| \nabla_{\mathcal{P}} f(D) \right\|.$$

□

3. The gradient method

3.1. **Description of the method.** The gradient flow

$$\begin{aligned}\frac{\mathrm{d}D}{\mathrm{d}t} &= -\nabla_{\mathcal{P}} E(D)\\ &= -[D,[D,F_D]]\end{aligned} \tag{3.1}$$

is a way of minimizing the energy over the manifold $\mathcal{P}$. The naive discretization

$$D_{k+1} = D_k - t[D_k,[D_k,F_k]]$$

is not suitable because it does not stay on $\mathcal{P}$. Instead, we must look for D_{k+1} on a curve on $\mathcal{P}$ that is tangent to $\nabla_{\mathcal{P}} E(D_n)$. A natural curve on $\mathcal{P}$ is the change of basis

$$D'(t) = U_t D U_t^T,$$

where U_t is a smooth family of orthogonal matrices. If we take

$$U_t = \exp(t[D,F_D]),$$

we get

$$\left.\frac{\mathrm{d}D'}{\mathrm{d}t}\right|_{t=0} = -[D,[D,F_D]],$$

so $D'(t)$ is a smooth curve on $\mathcal{P}$, tangent to the gradient flow at 0.

Our gradient method is then

$$D_{k+1} = U_t D_k U_{-t}, \tag{3.2}$$

with

$$U_t = \exp(t[D_k, F_k]). \tag{3.3}$$

We now prove a number of lemmas which are the main ingredients of the convergence proof. First, we bound the second derivative of the energy to obtain quantitative estimates on the energy decrease, then we link the difference of iterates $D_{k+1} - D_k$ to the gradient $\nabla_{\mathcal{P}} E(D_k)$, and finally we use the Łojasiewicz inequality to establish convergence.

3.2. **Derivatives.** We start from a point D_0 and compute the derivatives of E along the curve $D_t = U_t D_0 U_{-t}$. For ease of notation we will write $F_t = F(D_t)$ and $C_t = [D_t, F_t]$.

$$\begin{aligned}
\frac{\mathrm{d}D}{\mathrm{d}t} &= \frac{\mathrm{d}U_t}{\mathrm{d}t} D_0 U_{-t} + U_t D_0 \frac{\mathrm{d}U_{-t}}{\mathrm{d}t} \\
&= [C_0, D_t] \\
\frac{\mathrm{d}^n D}{\mathrm{d}t^n} &= \frac{\mathrm{d}^{n-1}}{\mathrm{d}t^{n-1}} [C_0, D_t] \\
&= \underbrace{[C_0, [C_0, \dots [C_0, D_t]]]}_{n \text{ times } C_0} \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \mathrm{Tr}(F_t [C_0, D_t]) \\
\left.\frac{\mathrm{d}E}{\mathrm{d}t}\right|_{t=0} &= -\|C_0\|^2 \\
\frac{\mathrm{d}^2 E}{\mathrm{d}t^2} &= \mathrm{Tr}(F_t [C_0, [C_0, D_t]]) + \mathrm{Tr}(G([C_0, D_t])[C_0, D_t])
\end{aligned}$$

3.3. **Control on the curvature.**

Lemma 3.1. *There exists $\alpha > 0$ such that for every D_0 and t,*

$$\left|\frac{\mathrm{d}^2 E}{\mathrm{d}t^2}\right|(t) \leq \alpha \|C_0\|^2 .$$

Proof.

$$\frac{\mathrm{d}^2 E}{\mathrm{d}t^2} = \mathrm{Tr}(F_t [C_0, [C_0, D_t]]) + \mathrm{Tr}(G([C_0, D_t])[C_0, D_t]). \tag{3.4}$$

First,

$$\begin{aligned}
\big\|F(D)\big\|_{\mathrm{op}} &\leq \frac{1}{2}\|-\Delta\|_{\mathrm{op}} + \|V\|_{\mathrm{op}} + \big\|G(D)\big\|_{\mathrm{op}} \\
&\leq \frac{1}{2}\|-\Delta\|_{\mathrm{op}} + 2(2N+Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}
\end{aligned}$$

by the Hardy inequality. Next, making use of the operator inequality $\mathrm{Tr}(AB) \leq \|A\|_{\mathrm{op}} \|B\|$, we show that

$$\mathrm{Tr}(F_t [C_0, [C_0, D_t]]) \leq 2\left(\frac{1}{2}\|-\Delta\|_{\mathrm{op}} + 2(2N+Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}\right)\|C_0\|^2 .$$

For the second term of (3.4),

$$\begin{aligned}
\mathrm{Tr}\left(G\left(\frac{\mathrm{d}D}{\mathrm{d}t}\right)\frac{\mathrm{d}D}{\mathrm{d}t}\right) &= \mathrm{Tr}(G([C_0, D_t])[C_0, D_t]) \\
&\leq \big\|G([C_0, D_t])\big\|_{\mathrm{op}} \mathrm{Tr}\left(\big|[C_0, D_t]\big|\right) \\
&\leq 4\sqrt{\|-\Delta\|_{\mathrm{op}}}\, \mathrm{Tr}\left(\big|[C_0, D_t]\big|\right)^2 \\
&\leq 16N\sqrt{\|-\Delta\|_{\mathrm{op}}}\|C_0\|^2 .
\end{aligned}$$

The result is now proved with

$$\alpha = \|-\Delta\|_{\mathrm{op}} + 4(6N+Z)\sqrt{\|-\Delta\|_{\mathrm{op}}}.$$

□

3.4. **Choice of the stepsize.** We can now expand the energy:

$$E(t) \leq E_0 - t\|C_0\|^2 + \frac{t^2}{2}\alpha\|C_0\|^2.$$

If we choose

$$t < \frac{2}{\alpha}, \tag{3.5}$$

we obtain a decrease of the energy

$$E(t) \leq E_0 - \beta\|C_0\|^2 \tag{3.6}$$

with $\beta = t - \frac{t^2}{2}\alpha > 0$.

The optimal choice of t with this bound on the curvature is $t = \frac{1}{\alpha}$, with $\beta = \frac{1}{2\alpha}$. Of course it could be that the actual optimal t is different, and could vary wildly, which is why we will not consider optimal stepsizes.

3.5. **Study of $D_{k+1} - D_k$.** We now prove that our iteration coincides with an unconstrained gradient method up to first order in t.

We say that $M \in o(\|N\|)$ when for all $\varepsilon > 0$, there is a neighborhood U of zero such that when $N \in U$, $\|M\| \leq \varepsilon\|N\|$. Note that this neighborhood U is not allowed to depend on N, meaning that the resulting bound is uniform, which will allow us to manipulate the remainders more easily.

Lemma 3.2. *For any D_0 and t,*

$$D_t = D_0 + t[C_0, D_0] + o(t\|C_0\|).$$

Proof.

$$D_t - D_0 - t[C_0, D_0] = \sum_{n=2}^{\infty} \frac{t^n}{n!} \underbrace{[C_0, [C_0, \dots [C_0, D_0]]]}_{n \text{ times } C_0}$$

$$\begin{aligned}\big\|D_t - D_0 - t[C_0, D_0]\big\| &\leq t\big\|[C_0, D_0]\big\| \sum_{n=2}^{\infty} t^{n-1}\|C_0\|^{n-1} \\ &\leq t\big\|[C_0, D_0]\big\| \frac{t\|C_0\|}{1 - t\|C_0\|}\end{aligned}$$

□

4. Convergence of the gradient algorithm

Theorem 4.1 (Convergence of the gradient algorithm)**.** *Let $D_0 \in \mathcal{P}$ be any density matrix and D_k be the sequences of iterates generated from D_0 by $D_{k+1} = U_t D_k U_{-t}$, with $t < \frac{2}{\alpha}$. Then D_k converges towards a solution of the Hartree-Fock equations.*

Proof. The energy $E(D)$ is bounded from below on $\mathcal{P}$, and therefore E_k converges to a limit E_∞. In the sequel we will work for convenience with $\widetilde{E}(D) = E(D) - E_\infty$ and drop the tildes. Immediately, summing 3.6 implies that C_k converges to 0, and therefore so does $D_k - D_{k-1}$. This is the result that Cancès and LeBris obtained for their ODA algorithm. Note that we only get that $\|D_k - D_{k-1}\|^2$ is summable, which alone is not enough to guarantee convergence (the harmonic series $x_k = \sum_{l=1}^{k} 1/l$ is a simple counter-example). To obtain convergence, we shall use the Łojasiewicz inequality.

Let us denote by Γ the level set $\Gamma = \{D \in \mathcal{P}, E(D) = 0\}$. It is non-empty and compact. We apply the Łojasiewicz inequality to every point of Γ to obtain a cover $(U_i)_{i\in\mathcal{I}}$ of Γ in which the Łojasiewicz inequality holds with constants κ_i, θ_i. By compactness, we extract a finite subcover from the U_i, from which we deduce $\delta > 0$, $\kappa > 0$ and $\theta \in (0, 1/2]$ such that whenever $d(D, \Gamma) < \delta$,

$$E(D)^{1-\theta} \leq \kappa\|C_D\|. \tag{4.1}$$

To apply the Łojasiewicz inequality to our iteration, it remains to show that $d(D_k, \Gamma)$ tends to zero. Suppose this is not the case. Then we can extract a subsequence, still denoted by D_k, such that

$d(D_k, \Gamma) \geq \varepsilon$ for some $\varepsilon > 0$. By compactness of $\mathcal{P}$ we extract a subsequence that converges to a D such that $d(D, \Gamma) \geq \varepsilon$ and (by continuity) $E(D) = 0$, a contradiction. Therefore $d(D_k, \Gamma) \to 0$, and for k larger than some k_0,

$$E(D_k)^{1-\theta} \leq \kappa \|C_k\| . \tag{4.2}$$

For $k \geq k_0$,

$$\begin{aligned} E(D_k)^\theta - E(D_{k+1})^\theta &\geq \frac{\theta}{E(D_k)^{1-\theta}} (E(D_k) - E(D_{k+1})) \\ &\geq \frac{\theta}{\kappa \|C_k\|} (E(D_k) - E(D_{k+1})) \\ &\geq \frac{\theta\beta}{\kappa} \|C_k\| \\ &\geq \frac{\theta\beta}{\kappa t} \|D_{k+1} - D_k\| + o(\|D_{k+1} - D_k\|) \end{aligned}$$

hence

$$\frac{\theta\beta}{\kappa t} \|D_{k+1} - D_k\| + o(\|D_{k+1} - D_k\|) \leq E(D_k)^\theta - E(D_{k+1})^\theta. \tag{4.3}$$

As the right-hand side is summable, so is the left-hand side, which implies that $\sum \|D_{k+1} - D_k\| < \infty$. D_k is therefore Cauchy and converges to some limit D_∞. $C_k \to 0$, so D_∞ is a critical point.

Note that now that we know that D_k converges to D_∞, we can replace the θ and κ in this inequality by the ones obtained from the Łojasiewicz inequality around D_∞ only. □

Let

$$e_k = \sum_{l=k}^{\infty} \|D_{l+1} - D_l\| .$$

This is a (crude) measure of the error at iteration number k. In particular, $\|D_k - D_\infty\| \leq e_k$.

Theorem 4.2 (Convergence rate of the gradient algorithm)**.**

(1) If $\theta = 1/2$ (non-degenerate case), then for any $\nu' < \frac{\beta}{2\kappa^2}$, there is a $c > 0$ such that

$$e_k \leq c(1 - \nu')^k. \tag{4.4}$$

(2) If $\theta \neq 1/2$ (degenerate case), then there exists $c > 0$ such that

$$e_k \leq c k^{-\frac{\theta}{1-2\theta}}. \tag{4.5}$$

Proof. Summing 4.3 from $l = k$ to ∞ with $k \geq k_0$, we obtain

$$\begin{aligned} e_k + o(\|e_k\|) &\leq \frac{\kappa t}{\theta\beta} E(D_k)^\theta \\ \left(\frac{\theta\beta}{\kappa t} e_k + o(\|e_k\|)\right)^{\frac{1-\theta}{\theta}} &\leq E(D_k)^{1-\theta} \\ &\leq \kappa \|C_k\| \\ &\leq \frac{\kappa}{t}(e_k - e_{k+1}) + o(\|e_k - e_{k+1}\|) \end{aligned}$$

Hence,

$$\begin{aligned} e_{k+1} &\leq e_k - \nu e_k^{\frac{1-\theta}{\theta}} + o(\|e_k\|^{\frac{1-\theta}{\theta}}), \text{ with} \\ \nu &= \frac{t}{\kappa} \left(\frac{\theta\beta}{\kappa t}\right)^{\frac{1-\theta}{\theta}} \end{aligned}$$

Two cases are to be distinguished. If $\theta = \frac{1}{2}$, the above inequality reduces to

$$e_{k+1} \leq (1 - \nu + o(1)) e_k$$

with $\nu = \frac{\beta}{2\kappa^2}$ and the result follows.

The case $\theta \neq 1/2$ is more involved. We define

$$y_k = c k^{-p},$$

which verifies

$$\begin{aligned} y_{k+1} &= c(k+1)^{-p} \\ &= ck^{-p}(1+1/k)^{-p} \\ &\geq ck^{-p}(1-\frac{p}{k}) \\ &\geq y_k(1-pc^{-\frac{1}{p}}y_k^{\frac{1}{p}}) \end{aligned}$$

We set $p = \frac{\theta}{1-2\theta}$ and c large enough so that $c > (\frac{\nu}{p})^{-p}$ and $y_{k_0} \geq e_{k_0}$. We then prove by induction $e_k \leq y_k$ for $k \geq k_0$. The result follows by increasing c to ensure that $e_k \leq y_k$, for $k < k_0$. □

In the non-degenerate case $\theta = 1/2$ (which was the case for all the numerical simulations we performed, see Section 7), the convergence is asymptotically geometric with rate $1-\nu$, where

$$\nu = \frac{\beta}{2\kappa^2}.$$

With the choice $t = \frac{1}{\alpha}$ suggested by our bounds, the convergence rate is

$$\nu = \frac{1}{4\kappa^2\alpha}.$$

5. Convergence of the Roothaan algorithm

The Roothaan algorithm (also called simple SCF in the literature) is based on the observation that a minimizer of the energy satisfies the *Aufbau* principle: D is the projector onto the first N eigenvectors of $F(D)$. A simple fixed-point algorithm is to take for D_{k+1} the projector onto the first N eigenvectors of $F(D_k)$. Unfortunately, this procedure does not always work: in some circumstances, oscillations between two states occur, and the algorithm never converges. This behavior was explained mathematically in [2], who noticed that the Roothaan algorithm minimizes the bilinear functional

$$E(D, D') = \mathrm{Tr}(h(D+D')) + \mathrm{Tr}(G(D)D')$$

with respect to its first and second argument alternatively. Thanks to the Łojasiewicz inequality, we can improve on their result and prove the convergence or oscillation of the method.

The bilinear functional verifies $E(D, D') = E(D', D)$, $E(D, D) = 2E(D)$. In fact, $\frac{1}{2}E(\cdot,\cdot)$ is the symmetric bilinear form associated with the quadratic form $E(\cdot)$. We use the uniform well-posedness hypothesis of [2], *i.e.* assume that there is a uniform gap of at least $\gamma > 0$ between the eigenvalues number n and $n+1$ of $F(D_k)$. Under this assumption, it can be proven [1] that there is a decrease of the bilinear functional at each iteration

$$\begin{aligned} E(D_{k+1}, D_{k+2}) &= E(D_{k+2}, D_{k+1}) \\ &= \min_{D\in\mathcal{P}} E(D, D_{k+1}) \\ &\leq E(D_k, D_{k+1}) - \gamma\|D_{k+2} - D_k\|^2 \end{aligned}$$

Since $E(\cdot,\cdot)$ is bounded from below on $\mathcal{P}\times\mathcal{P}$, this immediately shows that $D_k - D_{k+2} \to 0$, which shows that D_{2k} and D_{2k+1} converge up to extraction, which was noted in [2]. We now prove convergence of these two subsequences. To do that, we need to relate $D_k - D_{k+2}$ to the gradient of the bilinear functional. We first prove

Lemma 5.1. *For all F symmetric and $D \in \mathcal{P}$, if we let D' be the density matrix constructed by the Aufbau principle from F, then*

$$\big\|[D, F]\big\| \leq 2\|F\|_{op}\big\|D' - D\big\| + o\big(\big\|D' - D\big\|\big). \tag{5.1}$$

Proof. We choose a basis in which D is diagonal. We write $F = U\Lambda U^T$, with $U = I + \Delta$, and we expand the quantities of interest in powers of Δ. U is orthogonal, so $\Delta = -\Delta^T + o(\|\Delta\|)$, and $F = \Lambda + [\Delta, \Lambda] + o(\|\Delta\|)$. We now relate $D' - D$ and $[D, F]$ via Δ

$$\begin{aligned} D' - D &= \Delta D + D\Delta^T + o(\|\Delta\|) \\ &= [\Delta, D] + o(\|\Delta\|) \end{aligned}$$

$$\begin{aligned}[D,F] &= [D,\Lambda + [\Delta,\Lambda] + o(\|\Delta\|)] \\ &= [D,[\Delta,\Lambda]] + o(\|\Delta\|) \\ &= [[D,\Delta],\Lambda] + [\Delta,[D,\Lambda]] + o(\|\Delta\|) \\ &= [[D,\Delta],\Lambda] + o(\|\Delta\|)\end{aligned}$$

Therefore,

$$\big\|[D,F]\big\| \leq 2\|F\|_{\mathrm{op}}\big\|D'-D\big\| + o(\big\|D'-D\big\|).$$

□

Note that, to first order, if we write $\Delta = \begin{pmatrix} A & B \\ -B^T & C \end{pmatrix}$, then $\big\|[[D,\Delta],\Lambda]\big\|^2 = 2\sum_{i,j}(\lambda_{N+j}-\lambda_i)^2 B_{ij}^2$. Hence, for the inequality to be sharp, B must be concentrated on the lower i. This is not the case in numerical simulations, because the orbitals associated to the lowest energy levels converge faster (which we have not been able to explain). Therefore, in practice, the constant is much smaller.

Now, a reasoning similar to Theorem 2.2 shows that we can apply the Łojasiewicz inequality to $E(\cdot,\cdot)$ defined on $\mathcal{P}\times\mathcal{P}$ equipped with the natural Riemannian structure $\big\langle (D_1,D_1'),(D_2,D_2')\big\rangle = \langle D_1,D_2\rangle + \big\langle D_1',D_2'\big\rangle$. In this setting, the gradient is

$$\nabla_{\mathcal{P}\times\mathcal{P}}E(D,D') = \begin{pmatrix}[D,F(D')] \\ [D',F(D)]\end{pmatrix}$$

and therefore

$$\begin{aligned}\big\|\nabla_{\mathcal{P}\times\mathcal{P}}E(D_k,D_{k+1})\big\| &= \big\|[D_k,F(D_{k+1})]\big\| \\ &\leq 2\big\|F(D_{k+1})\big\|_{\mathrm{op}}\|D_{k+2}-D_k\| + o(\|D_{k+2}-D_k\|)\end{aligned}$$

Therefore, by the same compactness argument as before, the Łojasiewicz inequality

$$\begin{aligned}E(D_k,D_{k+1})^{1-\theta'} &\leq \kappa'\big\|\nabla_{\mathcal{P}\times\mathcal{P}}E(D_k,D_{k+1})\big\| \\ &\leq 2\kappa'\big\|F(D_{k+1})\big\|_{\mathrm{op}}\|D_{k+2}-D_k\| + o(\|D_{k+2}-D_k\|)\end{aligned}$$

holds for k large enough, with constants $\kappa'>0$ and $\theta'\in(0,\frac{1}{2}]$

The exact same reasoning as for the gradient algorithm proves the following theorems

Theorem 5.1 (Convergence/oscillation of the Roothaan algorithm)**.** *Let $D_0\in\mathcal{P}$ such that the sequence D_k of iterates generated by the Roothaan algorithms verifies the uniform well-posedness hypothesis with uniform gap $\gamma>0$. Then the two subsequences D_{2k} and D_{2k+1} are convergent. If both have the same limit, then this limit is a solution of the Hartree-Fock equations.*

Theorem 5.2 (Convergence rate of the Roothaan algorithm)**.** *Let D_k be the sequence of iterates generated by a uniformly well-posed Roothaan algorithm, and let*

$$e_k = \sum_{l=k}^{\infty}\|D_{l+2}-D_l\|\,.$$

Then,

(1) If $\theta'=1/2$ (non-degenerate case), then for any $\nu'<\frac{\gamma}{8\kappa'^2\|F\|_{op}^2}$, there is a $c>0$ such that

$$e_k \leq c(1-\nu')^k. \tag{5.2}$$

(2) If $\theta'\neq 1/2$ (degenerate case), then there exists $c>0$ such that

$$e_k \leq ck^{-\frac{\theta'}{1-2\theta'}}. \tag{5.3}$$

6. Level-shifting

The Level-Shifting algorithm was introduced in [9] as a way to avoid oscillation in the Roothaan algorithm. By shifting the energy levels (eigenvalues of F), one can force convergence, although denaturing the equations in the process. We now prove the convergence of this algorithm.

The same arguments as before apply to the functional

$$E^b(D, D') = \mathrm{Tr}(h(D + D')) + \mathrm{Tr}(G(D)D') + \frac{b}{2}\left\|D - D'\right\|^2$$
$$= \mathrm{Tr}(h(D + D')) + \mathrm{Tr}(G(D)D') - b\,\mathrm{Tr}(DD') + bN$$

with associated Fock matrix $F^b(D) = F(D) - bD$. The difference with the Roothaan algorithm is that for b large enough, there is a uniform gap $\gamma^b > 0$, and $D_k - D_{k+1}$ converges to 0 [2]. Therefore, we have the following theorems

Theorem 6.1 (Convergence of the Level-Shifting algorithm)**.** *Let $D_0 \in \mathcal{P}$. Then there exists a $b_0 > 0$ such that for every $b > b_0$, the sequence D_k of iterates generated by the Level-Shifting algorithm with shift parameter b verifies the uniform well-posedness hypothesis with uniform gap $\gamma > 0$ and converges.*

Theorem 6.2 (Convergence rate of the Level-Shifting algorithm)**.** *Let D_k be the sequence of iterates generated by the Level-Shifting with shift parameter $b > b_0$, and let*

$$e_k = \sum_{l=k}^{\infty} \|D_{l+2} - D_l\|\,.$$

Then,

(1) If $\theta' = 1/2$ (non-degenerate case), then for any $\nu' < \frac{\gamma^b}{8\kappa'^2\left\|F^b\right\|_{op}^2}$, there is a $c > 0$ such that

$$e_k \leq c(1 - \nu')^k. \tag{6.1}$$

(2) If $\theta' \neq 1/2$ (degenerate case), then there exists $c > 0$ such that

$$e_k \leq ck^{-\frac{\theta'}{1-2\theta'}}. \tag{6.2}$$

We can use this result to heuristically predict the behaviour of the algorithm when b is large. γ^b and $\left\|F^b\right\|_{\mathrm{op}}$ both scale as b for b large. We can take $\kappa' > \frac{1}{\sqrt{2\lambda_1}}$, where λ_1 is the smallest eigenvalue of $H_{\mathcal{P}\times\mathcal{P}}E(D^\infty, D^\infty)$. $H_{\mathcal{P}\times\mathcal{P}}\left\|D - D'\right\|^2$ has zero eigenvalues (for instance, $\left\|D - D'\right\|^2$ is constant along the curve $(D_t, D'_t) = (U_t D U_t^T, U_t D' U_t^T)$, where U_t is a family of orthogonal matrices), so λ_1 stays bounded as $b \to \infty$. Therefore, ν scales as $\frac{1}{b}$, which shows that b should not be too large for the algorithm to converge quickly.

7. Numerical results

We illustrate our results using a Galerkin discretization of the basis functions ϕ_i. The gradient method was implemented using the software Expokit [10] to compute matrix exponentials. The computational cost of a gradient step is not much higher than a step of the Roothaan algorithm, since the limiting step is computing the Fock matrix, not the exponential.

First, the Łojasiewicz inequality with exponent $\frac{1}{2}$ held in all the molecular systems and basis sets we encountered, suggesting that the minimizers are non-degenerate. Consequently, we never encountered sublinear convergence of any algorithm.

For a given molecular system and basis, we checked that the Level-Shifting algorithm converged as $(1 - \nu)^k$, where ν is proportional to $\frac{1}{b}$, which we predicted theoretically in Section 6. This means that the estimates we used have at least the correct scaling behavior.

Next, we compared the efficiency of the Roothaan algorithm and of the gradient algorithm, in the case where the Roothaan algorithm converges. Our analysis leads to the estimate $\nu = \frac{\gamma}{8\kappa'^2\|F\|_{\mathrm{op}}^2}$ for the Roothaan algorithm, and $\nu = \frac{1}{4\kappa^2\alpha}$ for the gradient algorithm with stepsize $t = \frac{1}{\alpha}$.

It is immediate to see that, up to a constant multiplicative factor, $\kappa' > \kappa$, $\gamma \leq \left\|F\right\|_{\mathrm{op}}$ and for the cases of interest $\alpha \approx \|F\|_{\mathrm{op}}$, so from our estimates we would expect the gradient algorithm to be faster than the Roothaan algorithm. However, in our tests the Roothaan algorithm was considerably faster than the gradient algorithm, even when the stepsize was adjusted at each iteration, for instance by a line search.

The reason for this is that, as we mentioned in the proof of Lemma 5.1, the inequality (5.1) is sharp only when $D_{k+1} - D_k$ has components on the lower eigenvectors. We conjecture that since the lower eigenvectors converge quickly, the bound of Lemma 5.1 is too large, and actual convergence is faster than what our estimates lead us to expect. A quantitative explanation of this effect is still an open question.

The outcome of these tests seems to be that the gradient algorithm is slower. It might prove to be faster in situations where the gap is small, or whenever κ' is much larger than κ. We have been unable to find concrete examples of such cases.

8. Conclusion, perspectives

In this paper, we introduced an algorithm based on the idea of gradient descent. By using the analyticity of the objective function and of the constraint manifold, we were able to derive a Łojasiewicz inequality, and use that to prove the convergence of the gradient method, under the assumption of a small enough stepsize. Next, expanding on the analysis of [2], we extended the Łojasiewicz inequality to a Lyapunov function for the Roothaan algorithm. By linking the gradient of this Lyapunov function to the difference in the iterates of the algorithm, we proved convergence (or oscillation), an improvement over previous results which only prove a weaker version of this. In this framework, the Level-Shifting algorithm can be seen as a simple modification of the Roothaan algorithm, and as such can be treated by the same methods. In each case, we were also able to derive explicit bounds on the convergence rates.

The strength of the Łojasiewicz inequality is that no higher-order hypothesis are needed for its use. As a consequence, the rates of convergence we obtain weaken considerably if the algorithm converges to a degenerate critical point. A more precise study of the local structure of critical points is necessary to understand why the algorithms usually exhibit geometric convergence. This is related to the problem of local uniqueness and is likely to be hard (and, indeed, to our knowledge has not been tackled yet).

Even though our results hide the complexity of the local structure in the constants of the Łojasiewicz inequality, they still provide insight as to the influence of the basis on the speed of convergence, and can be used to compare algorithms. All of our results use crucially the hypothesis of a finite-dimensional Galerkin space. For the gradient algorithm, we need it to ensure the existence of a stepsize that decreases the energy. This is analogous to a CFL condition for the discretization of the equation $\frac{\mathrm{d}D}{\mathrm{d}t} = -[D,[D,F_D]]$, and can only be lifted with a more implicit discretization of this equation. For the Roothaan and Level-Shifting algorithms, we use the finite dimension hypothesis in Lemma 5.1. As noted, this is not sharp, so it could be that the infinite-dimensional version of the Roothaan and Level-Shifting algorithms still converge. More research is needed to examine this.

Let us note that among the algorithms we considered, the gradient algorithm with sufficiently small step is the only one for which a completely satisfying local behavior can be expected. Indeed, the Roothaan algorithm can exhibit oscillations, and the Level-Shifting may converge towards a limit that is not an *Aufbau* solution of the Hartree-Fock equations. Although the gradient algorithm was found to be slower than other algorithms on the numerical tests we performed, it is more robust, and can probably outperform the other algorithms when the gap $\lambda_{N+1} - \lambda_N$ is small.

An algorithm that could achieve the speed of the fixed-point algorithms with the robustness granted by the energy monotonicity seems to be the ODA algorithm of Cancès and LeBris [2], along with variants such as EDIIS, or combinations of EDIIS and DIIS algorithms [4]. We were not able to examine these algorithms in this paper. At first glance, the ODA algorithm should fit into our framework (indeed, the ODA algorithm was built to satisfy an energy decrease inequality similar to 3.6). However, it works in a relaxed parameter space $\widetilde{P}$, and using the commutator to control the differences of iterates as we did only makes sense on $\mathcal{P}$, the border of $\widetilde{\mathcal{P}}$. Therefore, other arguments have to be used.

Also missing from this study is the study of other commonly used algorithms, such as DIIS or Bacskay's quadratically convergent algorithm. DIIS numerically exhibits a complicated behavior that is probably hard to explain analytically, and the QC algorithm requires a study of the second-order structure of the critical points.

9. Acknowledgments

The author would like to thank Eric Séré for his extensive help, Guillaume Legendre for the code used in the numerical simulations and Julien Salomon for introducing him to the Łojasiewicz inequality.

References

[1] E. Canc'es, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: a primer. *Handbook of numerical analysis*, 10:3–270, 2003.

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

[3] A. Haraux, M.A. Jendoubi, and O. Kavian. Rate of decay to equilibrium in some semilinear parabolic equations. *Journal of Evolution equations*, 3(3):463–484, 2003.

[4] K.N. Kudin, G.E. Scuseria, and E. Cancès. A black-box self-consistent field convergence algorithm: One step closer. *The Journal of chemical physics*, 116:8255, 2002.
[5] E.H. Lieb and B. Simon. The Hartree-Fock theory for Coulomb systems. *Communications in Mathematical Physics*, 53(3):185–194, 1977.
[6] P.L. Lions. Solutions of Hartree-Fock equations for Coulomb systems. *Communications in Mathematical Physics*, 109(1):33–97, 1987.
[7] S. Łojasiewicz. *Ensembles semi-analytiques*. Institut des Hautes Etudes Scientifiques, 1965.
[8] J. Salomon. Convergence of the time-discretized monotonic schemes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(01):77–93, 2007.
[9] V.R. Saunders and I.H. Hillier. A "Level–Shifting" method for converging closed shell Hartree–Fock wave functions. *International Journal of Quantum Chemistry*, 7(4):699–705, 1973.
[10] R.B. Sidje. Expokit: a software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, 24(1):130–156, 1998.

Université Paris-Dauphine, CEREMADE, Place du Maréchal Lattre de Tassigny, 75775 Paris Cedex 16, France.
E-mail address: `levitt@ceremade.dauphine.fr`