Preconditioned Gradient Flows for Nonlinear Eigenvalue Problems and Application to the Hartree-Fock Functional

François Alouges,1 Christophe Audouze2
1Laboratoire de Mathématiques, Université Paris–Sud XI, Bâtiment 425, Orsay 91405, France
2Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale de Paris Grande voie des vignes, Châtenay-Malabry 92295, France

Received 14 November 2007; accepted 23 February 2008
Published online 29 April 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI 10.1002/num.20347

In this article, we propose to solve numerically the problem of finding the smallest eigenvalues of a Hermitian operator (and the space spanned by the corresponding eigenvectors) by a gradient flow technique. This method is then applied to the Hartree-Fock problem. Improvements are also proposed in two directions: preconditioning of the dynamical system and development of a specific flow that enables to compute directly the eigenvectors. © 2008 Wiley Periodicals, Inc. Numer Methods Partial Differential Eq 25: 380–400, 2009

Keywords: quantum chemistry; diagonalization; gradient flows; Schrödinger equation; Hartree-Fock equations

I. INTRODUCTION

In quantum chemistry, the Hartree-Fock model comes from the Schrödinger equation when one looks for a solution that writes down as a unique Slater determinant (see [1] for instance). Considering the problem as a minimization on the $N$ orthonormal functions $(\phi_1, \ldots , \phi_N)$ composing the Slater determinant, one obtains a nonlinear eigenvalue problem where $(\phi_1, \ldots , \phi_N)$ span the space generated by the eigenfunctions corresponding to the $N$ smallest eigenvalues of a self-adjoint operator. In fact, the problem is really nonlinear since the operator depends itself on the solution $(\phi_1, \ldots , \phi_N)$. This is the so-called self-consistent principle (see Section B for a description of the Hartree-Fock model). After discretization, the design of iterative methods for solving the underlying finite dimensional self-consistent eigenvalue problem while keeping the orthonormality conditions on the vectors at all iterations is probably one of the major difficulties of the approach (see [2] for instance). Most of numerical methods commonly used by physicists and

Correspondence to: Christophe Audouze, Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale de Paris Grande voie des vignes, Châtenay-Malabry 92295, France (e-mail: christophe.audouze@ecp.fr)

© 2008 Wiley Periodicals, Inc.
chemists propose to solve this self-consistent principle by updating iteratively the eigenvectors and the self-adjoint operator. However, other more refined techniques exist such as the ODA algorithm [2] and the DIIS method [3, 4].

In this article based on the works developed in [5], we propose to tackle the problem in a different manner than usual since it does not require self-consistence updates. Namely, the viewpoint we adopt is the one of directly minimizing the energy by a gradient flow technique, suitably adapted in order to handle the orthonormality constraints on the eigenfunctions.

The article is divided as follows. Section II is devoted to the writing of standard vectorial gradient flows, first in the linear case (this is typically the problem which arises in a non self-consistent step of Hartree-Fock computations), and then in the nonlinear case when one considers the minimization of the Hartree-Fock functional. The question of preconditioning such gradient flows is then discussed in Section III. Section IV is devoted to the so-called triangular flows, which allow to compute not only the space spanned by the $k$ eigenvectors corresponding to the $k$ smallest eigenvalues, but also the eigenvectors themselves. The triangular flow is applied to the linear case in Section A, and then generalized to to the Hartree-Fock case in Section B. Eventually, numerical results are presented in Section V using all flow techniques presented before for a comparison, both in the linear case (Sections A and B) and nonlinear one (Section C). In this last Section, the full consistency Hartree-Fock problem is considered, in a three-dimensional periodic framework.

II. VECTORIAL GRADIENT FLOWS FOR NONLINEAR EIGENVALUE PROBLEMS

A. The Linear Eigenvalue Problem

Let $B$ be a real symmetric (or complex hermitian) matrix of size $n \times n$, for which we want to determine the $k$ ($\leq n$) smallest eigenvalues and associated eigenvectors. For the sake of simplicity, we first consider the case $k = 1$ and we assume that the first eigenvalue of $B$ is simple. Let us define the application

$$ F : \mathcal{M} \longrightarrow \mathbb{R} $$

$$ v \longmapsto \frac{1}{2} (Bv, v), $$

where $(\cdot, \cdot)$ stands for the usual euclidian inner product on $\mathbb{R}^n$, and $\mathcal{M}$ for the unit sphere in $\mathbb{R}^n$ which corresponds to the associated norm (that will be denoted by $\| \cdot \|$). By considering first-order variations, we obtain the tangent space to $\mathcal{M}$ at $v$,

$$ T_v \mathcal{M} = \{ w \in \mathbb{R}^n, (w, v) = 0 \} $$

and $DF(v) \cdot w = (Bv, w)$ $\forall w \in T_v \mathcal{M}$. Therefore, the equations defining the gradient

$$ \nabla F(v) \in T_v \mathcal{M}, $$

$$ (\nabla F(v), w) = DF(v) \cdot w, \quad \forall w \in T_v \mathcal{M} $$

lead to $\nabla F(v) = Bv + \lambda v$. From (1a), we have $\lambda = -(Bv, v)$ and the gradient flow obtained in this manner (usually called Oja flow) is written as

$$ \frac{dv}{dt} = -\nabla F(v) = -Bv + (Bv, v)v. $$

This gradient flow is studied in details in [6]. In particular, the norm of $v$ is preserved during time and there exists an exponential convergence to the first eigenvector of $B$.

**Theorem 2.1** (see [6]). If we denote by $\{u_i, \lambda_i\}$ the eigenpairs of $B$ such that $\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_n$, then for any initial condition $v(t_0) \in M$ such that $(v(t_0), u_1) \neq 0$, the flow (2) has a solution which converges towards $u_1$, with the estimate

$$\|v(t) - u_1\| \leq Ce^{-(\lambda_2 - \lambda_1)(t-t_0)}\|v(t_0) - u_1\|.$$  

**Remark 2.1.** One essentially gets the same flow using the map

$$G : v \in \mathbb{R}^n \setminus \{0\} \mapsto \frac{1}{2} \frac{(Bv, v)}{\|v\|^2},$$

with the following inner product $(w_1, w_2) = \frac{1}{\|w_1\|^2}(w_1, w_2)$ in $T_v(\mathbb{R}^n \setminus \{0\})$. In this case one gets

$$\frac{dv}{dt} = -Bv + \frac{(Bv, v)}{\|v\|^2}v$$  

which also preserves $\|v(t)\|^2$.

In the general case $1 \leq k \leq n$ where we want more than one eigenpair, we consider the mapping

$$F : St(k, n) \rightarrow \mathbb{R}$$

$$V \mapsto \frac{1}{2} \text{tr}(V^tBV),$$  

where $V^t$ stands for the transpose of $V$ (the adjoint in the hermitian case) and $St(k, n)$ is the Stiefel manifold (see [6]) defined by

$$St(k, n) = \{V \in \mathbb{R}^{n\times k}, V^tV = I_k\}.$$  

Again it is well known that minimizing $F$ over $St(k, n)$ leads to a matrix $V$ whose columns span the space generated by $k$ eigenvectors corresponding to the $k$ smallest eigenvalues of $B$. The associated gradient flow is given in [6] and writes

$$\frac{dV}{dt} = -BV + V^tBV,$$  

while the tangent space to $St(k, n)$ at $V$ is given by

$$T_V St(k, n) = \{\xi \in \mathbb{R}^{n\times k}, \xi^tV + V^t\xi = 0\},$$  

provided one chooses the euclidean inner product on $T_V St(k, n)$ given by

$$\langle \xi, \eta \rangle = \text{tr}(\xi^t\eta), \quad \forall \xi, \eta \in T_V St(k, n).$$
Equivalently, writing \( V \) in columns, \( V = (V_1, \ldots, V_k) \), leads to

\[
\frac{dV_i}{dt} = -BV_i(t) + \sum_{j=1}^{k} (BV_i(t), V_j(t))V_j(t), \quad \forall i = 1, \ldots k. \tag{9}
\]

For this vectorial gradient flow, it can be checked that the orthonormality constraints are kept during time. There also exists an exponential result of convergence of \( V(t) \in St(k, n) \) towards the subspace spanned by the eigenvectors associated to the \( k \) smallest eigenvalues of \( B \) (see [6]).

### B. The Hartree-Fock Equations

A nonlinear eigenvalue problem. In quantum chemistry, the Hartree-Fock model is used to describe the electronic structure of molecular systems and is an approximation of the \( N \) body problem associated to the Schrödinger equation. Here, the Hartree-Fock model that we consider is non-relativistic; we also do not take into account the spin for the sake of simplicity, which does not change the mathematical structure of the model (for physical aspects of the Hartree-Fock model, see [1, 7]), for mathematical studies see [8–10]). A molecule will be composed of \( N \) electrons and \( K \) nuclei, defined by their fixed positions \((\bar{x}_1, \ldots, \bar{x}_K)\) and their positive charges \((Z_1, \ldots, Z_K)\).

Writing down the wave function \( \psi \) of the \( N \) electrons as a unique Slater determinant

\[
\psi(x_1, \ldots, x_N) = \det(\phi_i(x_j)),
\]

and solving the Schrödinger equation for such functions leads to the minimization of the Hartree-Fock energy functional

\[
\mathcal{E}^{HF}(\varphi_1, \ldots, \varphi_N) = \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \varphi_i|^2 + V|\varphi_i|^2dx + \frac{1}{2} \sum_{i,j=1}^{N} \left[ D(\varphi_i^2, \varphi_j^2) - D(\varphi_i, \varphi_j) \right], \tag{10}
\]

where \( V(x) = -\sum_{k=1}^{K} \frac{Z_k}{|x - \bar{x}_k|} \) denotes the electrons-nuclei attractive interaction and \( D(u, v) \) is defined by

\[
D(u, v) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{u(x)v(y)}{|x - y|} dxdy.
\]

The \( N \) functions \( \varphi_i \), called the Hartree-Fock orbitals, belong to the space

\[
\mathcal{M} = \{(\varphi_1, \ldots, \varphi_N) \in (H^1(\mathbb{R}^3))^N, (\varphi_i, \varphi_j)_{L^2} = \delta_{ij}\}. \tag{11}
\]

From a physical point of view, the minimization of \( \mathcal{E}^{HF}(\varphi_1, \ldots, \varphi_N) \) over \( \mathcal{M} \) allows to get the most stable electronic configuration of the molecule.

Writing the Euler-Lagrange equations of (10) under the constraints (11) leads to a nonlinear system of \( N \) PDEs (the so-called Hartree-Fock equations)

\[
\begin{align*}
\bar{H}(\varphi)\varphi_i &= \varepsilon_i\varphi_i, \tag{12a} \\
(\varphi_i, \varphi_j)_{L^2} &= \delta_{ij}, \tag{12b}
\end{align*}
\]

where the operator \( \bar{H}(\varphi) \) is defined by

\[
\bar{H}(\varphi)\varphi_i = -\Delta \varphi_i + V\varphi_i + \left( \rho * \frac{1}{|x|} \right) \varphi_i - \int_{\mathbb{R}^3} \frac{\rho(x, y)}{|x - y|} \varphi_i(y)dy, \tag{13}
\]

with the electronic density $\rho(x) = \sum_{i=1}^{N} |\psi_i(x)|^2$ and $\rho(x, y) = \sum_{i=1}^{N} \psi_i(x)\psi_i^*(y)$. Concerning the Hartree-Fock equations, we refer to [8–10] for the existence of solutions. In particular, it is shown in [10] that for a positive ion, i.e. $Z = \sum_{k=1}^{K} Z_k > N - 1$, there exists a solution for (12a)–(12b) and that the corresponding eigenvalues $\varepsilon_1, \ldots, \varepsilon_N < 0$ are the $N$ smallest eigenvalues of the operator $\hat{H}(\psi)$.

In Section C, we will present numerical simulations for a simplified Hartree-Fock model which describes a periodic system of atoms. When the nuclei are arranged in a 1-periodic cubic lattice, the energy functional (10), defined now per unit cell, becomes

$$E_{\text{HF}}^C(\psi_1, \ldots, \psi_N) = \sum_{i=1}^{N} \int_C |\nabla \psi_i|^2 - N \int_C (\rho - N)G(\delta_{i1} - 1)$$

$$+ \frac{1}{2} \int_C (\rho - N)G(\rho - N) - \frac{1}{2} \sum_{i=1}^{N} \int_C (\psi_i\psi_j^* - \delta_{ij})G(\psi_i^*\psi_j - \delta_{ij}),$$

where the unit cell $C = [0, 1]^3$, which is electrically neutral, contains one nucleus of charge $Z_1 = N$ located at its center $\bar{x}_1 = (1/2, 1/2, 1/2)$, and $N$ electrons. To take into account the periodicity of the different physical quantities, we use the solution $\psi = G(u)$ of the Laplacian problem with periodic boundary conditions defined by

$$\begin{cases}
-\Delta \psi = 4\pi u & \text{on } C = [0, 1]^3, \\
\int_C \psi = 0, \\
\psi \text{ periodic},
\end{cases}$$

for which the Fourier series has the coefficients

$$\hat{\psi}(m, n, t) = \begin{cases}
\frac{4\pi \hat{u}(m, n, p)}{m^2 + n^2 + p^2} & \text{if } m^2 + n^2 + p^2 \neq 0, \\
0 & \text{otherwise}.
\end{cases}$$

In this context, we want to generalize the gradient flow technique explained before, as an alternative way to solve the Hartree-Fock problem. In that aim, we begin by discretizing the problem. We consider the following set

$$\mathcal{M}_h = \{(\psi_1, \ldots, \psi_N) \in H^N, (\psi_i, \psi_j)_{L^2} = \delta_{ij}\}$$

where $H$ is a finite dimension vectorial subspace of $H^1(\mathbb{R}^3)$. Then, we define the operator $\hat{H}(\psi) : H \rightarrow H$ as the Galerkin approximation of $\hat{H}$ on $H$ defined by

$$(\hat{H}(\psi)\theta, \psi)_{L^2} = (\hat{H}(\psi)\theta, \psi)_{L^2}, \quad \forall (\theta, \psi) \in H \times H.$$
Theorem 2.2. The $L^2$ gradient flow associated to the discretized Hartree-Fock problem writes as

$$\forall i \in \{1, \ldots, N\} \frac{d\varphi_i}{dt} = -\tilde{H}(\varphi)\varphi_i + \sum_{j=1}^{N} (\tilde{H}(\varphi)\varphi_i, \varphi_j)_{L^2} \varphi_j.$$  \hspace{1cm} (19)

Remark 2.2. The previous relation (19) coincides with the standard expression of a projected gradient, but as we are going to see later, it comes from the fact that $\tilde{H}(\varphi)$ is self-adjoint. Otherwise, the gradient flow would take a different form. It is also a slight generalization of the flow proposed in [6].

To prove the Theorem 2.2, we need the following lemma.

Lemma 2.1. The linear map $\Pi : H^N \mapsto T_{\varphi} \mathcal{M}_h$ defined by

$$(\Pi u)_i = u_i - \frac{1}{2} \sum_{j=1}^{N} ((u_i, \varphi_j)_{L^2} + (\varphi_i, u_j)_{L^2}) \varphi_j$$ \hspace{1cm} (20)

is a $L^2$-orthogonal (and thus self-adjoint) projector.

Proof. First, we can check that $\Pi$ takes its values in $T_{\varphi} \mathcal{M}_h$. We have

$$(\Pi u)_i, \varphi_j)_{L^2} = \frac{1}{2} ((u_i, \varphi_j)_{L^2} - (\varphi_i, u_j)_{L^2}) = - (\varphi_i, (\Pi u)_j)_{L^2},$$ \hspace{1cm} (21)

proving that $\Pi u \in T_{\varphi} \mathcal{M}_h$, from the relation (17). Moreover $\Pi|_{T_{\varphi} \mathcal{M}_h} = I$: as we have

$$(u_i, \varphi_j)_{L^2} + (\varphi_i, u_j)_{L^2} = 0, \quad \forall u \in T_{\varphi} \mathcal{M}_h,$$

we get $(\Pi u)_i = u_i$. Therefore $\Pi$ is a projector on $T_{\varphi} \mathcal{M}_h$. It is $L^2$-orthogonal since, $\forall u, v \in H^N$,

$$(u - \Pi u, \Pi v)_{(L^2)^N} = \sum_{i=1}^{N} ((u - \Pi u)_i, (\Pi v)_i)_{L^2}$$

$$= \frac{1}{2} \sum_{i,j=1}^{N} ((u_i, \varphi_j)_{L^2} + (\varphi_i, u_j)_{L^2}) (\varphi_j, (\Pi v)_i)_{L^2}$$

$$= 0$$

as it is a product of hermitian and skew-hermitian coefficients in $i, j$. \hfill \blacksquare

We are now able to prove the Theorem 2.2.

Proof (Theorem (2.2)). Writing

$$D\mathcal{E}^{\text{HF}}(\varphi) \cdot \delta \varphi = \sum_{i=1}^{N} (\tilde{H}(\varphi)\varphi_i, \delta \varphi_i)_{L^2},$$ \hspace{1cm} (22)
the system (18a)–(18b) can be set in the following form

\[ g = \Pi g = \Pi \tilde{H}(\varphi)\varphi. \]

This means

\[ g_i = \tilde{H}(\varphi)\varphi_i - \frac{1}{2} \sum_{j=1}^{N} ( (\tilde{H}(\varphi)\varphi_i, \varphi_j)_{L^2} + (\varphi_i, \tilde{H}(\varphi)\varphi_j)_{L^2} ) \varphi_j \]

\[ = \tilde{H}(\varphi)\varphi_i - \sum_{j=1}^{N} (\tilde{H}(\varphi)\varphi_i, \varphi_j)_{L^2} \varphi_j \]

as \( \tilde{H}(\varphi) \) is a self-adjoint operator. Writing

\[ \frac{d\varphi_i}{dt} = -g_i, \quad (23) \]

we finally get the \( L^2 \) Hartree-Fock gradient flow (19) proposed in the Theorem 2.2.

This gradient flow looks like the flow (9) obtained for the linear eigenvalue problem. We easily check that at convergence, the Hartree-Fock flow provides a critical point of the Hartree-Fock functional:

\[ \frac{d\varphi}{dt} = 0 \Rightarrow g = 0 \Rightarrow \tilde{H}(\varphi)\varphi = M\varphi, \]

which are the Hartree-Fock equations (projected on \( H^N \)), up to a unitary transformation on \( \varphi \). One can also see that the constraints (12b) on the Hartree-Fock orbitals are kept during time:

\[ \frac{d}{dt} (\varphi_i(t), \varphi_j(t))_{L^2} = \left( \frac{d\varphi_i}{dt}, \varphi_j \right)_{L^2} + \left( \varphi_i, \frac{d\varphi_j}{dt} \right)_{L^2} \]

\[ = -(g_i, \varphi_j)_{L^2} - (\varphi_i, g_j)_{L^2} \text{ from (23)} \]

\[ = 0 \text{ as } \Pi g = g. \]

Hence, we have built a method which provides a monotonic decrease of the Hartree-Fock energy towards a critical point (generically a local minimum), without using the self-consistence principle.

III. PRECONDITIONING

The choice of the metric in the tangent space is a very important point because it determines the form of the gradient flow. As a sake of example, let us consider the case where

\[ F : \mathbb{R}^n \to \mathbb{R} \]

\[ x \mapsto \frac{1}{2} (Bx, x) \]

where $B$ is a positive definite symmetric real matrix. Here $(\cdot, \cdot)$ denotes the classical euclidian inner product on $\mathbb{R}^n$. As no constraint is assumed ($\mathcal{M} = \mathbb{R}^n$), the gradient flow associated to the inner product $(\cdot, \cdot)$ is simply given by

$$\frac{dx}{dt} = -Bx(t). \quad (24)$$

Of course if we denote by $(\lambda_i, e_i)_i$ the eigenelements of $B$, writing $x(t)$ in the basis $(e_i)_i$ as $x(t) = \sum_i x_i(t)e_i$ leads to

$$x_i(t) = x_i(0)e^{-\lambda_it}, \quad (25)$$

reflecting that the components of $x$ decay to zero but not at the same speed. On the other hand, solving (24) with an explicit Euler scheme for instance, leads to

$$x^0 = x(0), \quad x^{m+1} = (I_n - \delta tB)x^m, \quad \forall m \geq 0,$$

where $\delta t$ is the time step and $x^m \sim x(m\delta t)$ for which the stability condition

$$\forall i, 0 < \delta t \lambda_i < 2 \quad (26)$$

must be satisfied. Therefore, the time step is constrained by the largest eigenvalue of $B$, and the number $N$ of iterations needed to get $\frac{\|x(t)\|}{\|x(0)\|} < \epsilon$ can be estimated as

$$N \sim \ln(\epsilon) \ln|1 - \delta t \lambda_1| \sim -\frac{\lambda_2}{2\lambda_1} \ln(\epsilon).$$

The worse $B$ is conditioned, the more iterations are needed. Now, if one takes another symmetric positive definite real matrix $A$ and uses the inner product $(x, y)_A = (Ax, y)$ to define the gradient, one gets the flow

$$\frac{dx}{dt} = -A^{-1}Bx(t). \quad (27)$$

In the ideal limit $A = B$, we get a flow which is independent of $B$. All the components of $x(t)$ tend to zero at the same speed and both the time step and the number of iterations needed to achieve convergence with the corresponding flow do not depend on the condition number of $B$ anymore. We see that changing the inner product to compute the gradient acts like a preconditioning of the system. We use this idea in our Hartree-Fock case; namely, we study the influence of a scalar product change on the tangent space onto the convergence speed.

From (13), we can see the Hartree-Fock operator as a perturbation of the Laplace operator. It is therefore natural to propose a $H^1$-flow defined by taking in (18b) the $H^1$ scalar product instead of the $L^2$ one. Namely we construct the gradient $(g_i)_i = 1, \ldots, N$ as the solution of

$$g \in T_\varphi\mathcal{M}_h, \quad (28a)$$

$$\int \sum_{i=1}^N g_i\delta\varphi_i + \int \sum_{i=1}^N \nabla g_i \cdot \nabla \delta\varphi_i = D\epsilon_{\text{HFF}}(\varphi) \cdot \delta\varphi, \quad \forall \delta\varphi = (\delta\varphi_1, \ldots, \delta\varphi_N) \in T_\varphi\mathcal{M}_h. \quad (28b)$$
which can also be written as

\[ g \in T_{\psi}M_h, \]  
\[ (\Pi (I - \Delta_h)g)_i = \tilde{H}(\psi_i) - \sum_{j=1}^{N} (\tilde{H}(\psi_i, \psi_j)_{L^2}) \psi_j, \]

where \( \Delta_h \) is the Galerkin approximation of \( \Delta \) on \( H \). This latter system can be numerically solved with a conjugate gradient technique since it is in the form

\[ \Pi (I - \Delta_h)g = f, \]

with \( f \in \text{Im}(\Pi) = T_{\psi}M_h \), and since \( \Pi (I - \Delta_h)\Pi \) is symmetric positive definite on \( \text{Im}(\Pi) \). A more detailed justification that one can use the conjugate gradient method for such system is available in [11], although the application is very different there. Let us note that the preconditioning (29a)–(29b) involves the eigenvalues of the Laplacian operator, but in a completely different way compared to the preconditioning technique proposed in [12].

IV. TRIANGULAR FLOWS

In the linear case, the \( L^2 \) and \( H^1 \) gradient flows presented before do not give the eigenvectors but only the space spanned by the eigenvectors corresponding to the \( k \) smallest eigenvalues of \( B \). Of course, thanks to a Lanczos procedure applied to one vector of \( V \in \mathbb{R}^{n \times k} \), one can retrieve the eigenelements of \( B \). Nevertheless, as we shall see in the numerical results, the convergence speed of all the eigenvalues is the same for both the \( L^2 \) and \( H^1 \) flows, even in the case of a diagonal matrix with a clustered spectrum. In a quantum chemistry framework, as the eigenvalues are typically focused around particular values, it would be useful to provide a numerical method for which the eigenspaces associated with these clusters of eigenvalues converge quickly. The aim of this Section is to solve directly this question with a particular flow, which does not only provide the eigenelements of \( B \) but whose convergence speed also depends of the presence of eigenvalues clusters. We derive the flow in the linear case while the extension to nonlinear problems is straightforward and will be done afterwards together with the preconditioning.

A. Linear Case

Again, we consider the case of finding the \( k \) eigenelements corresponding to the smallest eigenvalues of a \( n \times n \) hermitian matrix \( B \) with a flow on \( k \) vectors \( V_1, \ldots, V_k \) which belong at any time to the Stiefel manifold \( St(k, n) \). In that purpose, for each vector \( V_i \) we successively build an ODE which preserves the orthogonality constraints. Obviously, writing

\[ \frac{dV_1}{dt} = -BV_1 + (BV_1, V_1)V_1 \]

implies \( \|V_1(t)\| = 1 \) for \( t \geq 0 \), if \( \|V_1(0)\| = 1 \).

As a next step, for \( V_2 \) such that \( \|V_2(t)\| = 1 \) and \( (V_1, V_2) = 0 \), we take

\[ \frac{dV_2}{dt} = -BV_2 + \alpha V_1 + \beta V_2. \]
On the one hand, since \( B \) is hermitian, we have
\[
\text{Re} \left( V_2, \frac{dV_2}{dt} \right) = \frac{1}{2} \frac{d}{dt} \| V_2(t) \|^2 = 0 \Rightarrow \beta = (BV_2, V_2).
\]
On the other hand, \( \left( \frac{dV_2}{dt}, V_1 \right) = -(BV_2, V_1) + \alpha \) leads to
\[
\alpha = (BV_2, V_1) + \left( \frac{dV_2}{dt}, V_1 \right) = (BV_2, V_1) - \left( V_2, \frac{dV_1}{dt} \right) \quad \text{as} \quad (V_1, V_2) = 0
\]
\[
= 2(BV_2, V_1),
\]
from (30). In this way, we build recursively the following system of ODEs, for \( i = 1, \ldots, k \)
\[
\frac{dV_i}{dt} = -BV_i + 2 \sum_{j<i} (BV_i, V_j) V_j + (BV_i, V_i) V_i. \tag{32}
\]
The previous equations can be written in a more compact notation
\[
\frac{dV}{dt} = -BV + VT, \tag{33}
\]
where \( T \) is an upper triangular matrix defined by \( T_{ii} = (BV_i, V_i), \quad \forall i = 1, \ldots, k \) and
\[
T_{ij} = \begin{cases} 
2(BV_i, V_j) & i < j \\
0 & i > j.
\end{cases}
\]

For \( p \) such that \( 1 \leq p \leq k \), let define the sub-energy \( E_p(t) \) by
\[
E_p(t) = \frac{1}{2} \sum_{i=1}^{p} (BV_i(t), V_i(t)). \tag{34}
\]

**Proposition 4.1.** The \( k \) sub-energies \( (E_p)_{p=1,\ldots,k} \) are decreasing during time.

**Proof.** For \( p \) such that \( 1 \leq p \leq k \), we have
\[
\frac{d}{dt} E_p(t) = \text{Re} \left( \sum_{j=1}^{p} \left( BV_i, \frac{dV_i}{dt} \right) \right) \quad \text{as} \quad B' = B
\]
\[
= -\sum_{i=1}^{p} \| BV_i \|^2 + 2 \sum_{i=1}^{p} \sum_{j<i} |(BV_i, V_j)|^2 + \sum_{i=1}^{p} (BV_i, V_i)^2 \quad \text{from (32)}
\]
\[
= -\sum_{i=1}^{p} \| BV_i \|^2 + \sum_{i=1}^{p} \sum_{j=1}^{p} |(BV_i, V_j)|^2
\]
\[
= -\sum_{i=1}^{p} \left\| BV_i - \sum_{j=1}^{p} (BV_i, V_j) V_j \right\|^2 \leq 0. \tag{35}
\]
The previous relation (35) proves the decreasing of all the subenergies $E_p$ but also explains the different behavior of the triangular flow, compared to the $L^2$ or the preconditioned flows. When convergence is reached, i.e. when the ODEs systems are stationary, we have, for $i = 1, \ldots, k$, $B\mathbf{V}_i \in \text{Span}(\mathbf{V}_1, \ldots, \mathbf{V}_i)$ for the triangular flow (see (35)), while $B\mathbf{V}_i \in \text{Span}(\mathbf{V}_1, \ldots, \mathbf{V}_k)$ for the $L^2$ usual gradient flow (see (9)).

Eventually, we prove the following result about the convergence of the generated $p$-dimensional subspace towards the space spanned by the $p$ first eigenvectors. Namely, we show that the rate of convergence for $E_p$ only depends on the associated eigenvalue gap $(\lambda_{p+1} - \lambda_p)$ while for the $L^2$ gradient flow the convergence speed of all the $E_p$ ($1 \leq p \leq k$) depends on the last gap $(\lambda_{k+1} - \lambda_k)$.

**Theorem 4.1.** Let $B$ be a symmetric matrix of size $n$ for which we want to find the $k$ smallest eigenvalues and their associated eigenvectors. We denote by $\{\mathbf{u}_i, \lambda_i\}$ the eigenelements of $B$ and assume $\lambda_1 < \cdots < \lambda_n$. Let $\mathbf{V}_1, \ldots, \mathbf{V}_k$ be the vectors generated by the triangular flow (32). Then, any subspace $\text{Span}(\mathbf{V}_1(t), \ldots, \mathbf{V}_p(t))$ taken with $p \leq k$ converges towards $\text{Span}(\mathbf{u}_1, \ldots, \mathbf{u}_p)$ like $O(e^{-(\lambda_{p+1} - \lambda_p)t})$.

**Proof.** First, the existence of global solutions for the triangular flow directly follows from the Cauchy-Lipschitz Theorem. Moreover, since the functions $E_p$ decrease in time, and are non-negative, they converge to a limit

$$E_p^* = \lim_{t \to +\infty} E_p(t).$$

Next, the $\omega$-limit set of any trajectory $(\mathbf{V}_i(t))_{i=1,\ldots,k}$ defined by

$$\omega((\mathbf{V}_i(\cdot))_{i=1,\ldots,k}) = \bigcap_{T>0} \bigcup_{t>T} \{\mathbf{V}_1(t), \ldots, \mathbf{V}_k(t)\}$$

is such that $\forall (\mathbf{W}_1, \ldots, \mathbf{W}_k) \in \omega((\mathbf{V}_i(\cdot))_{i=1,\ldots,k})$, \n
$$\frac{1}{2} \sum_{i=1}^{p} (B\mathbf{W}_i, \mathbf{W}_i) = E_p^*.$$  

From the continuity of the flow with respect to its initial conditions and from (35), we also have that

$$\forall p \in \{1, \ldots, k\}, \left\|B\mathbf{W}_p - \sum_{j=1}^{p} (B\mathbf{W}_p, \mathbf{W}_j)\mathbf{W}_j\right\|^2 = 0.$$  

Taking $p = 1$ precisely says that $\mathbf{W}_1$ is a (unit) eigenvector of $B$, while, by induction, if $(\mathbf{W}_1, \ldots, \mathbf{W}_p)$ are eigenvectors of $B$, we have for $p + 1 \leq k$

$$B\mathbf{W}_{p+1} = \sum_{j=1}^{p+1} (B\mathbf{W}_{p+1}, \mathbf{W}_j)\mathbf{W}_j$$

$$= \sum_{j=1}^{p+1} (\mathbf{W}_{p+1}, B\mathbf{W}_j)\mathbf{W}_j$$

$$= (B\mathbf{W}_{p+1}, \mathbf{W}_{p+1})\mathbf{W}_{p+1},$$
since \((W_1, \ldots, W_k)\) are mutually orthogonal. Therefore, we deduce that \((W_1, \ldots, W_k)\) are all eigenvectors of \(B\). We call \(\mu_1, \ldots, \mu_k\) the corresponding eigenvalues and define \(\mu_{k+1} < \cdots < \mu_n\) the remaining eigenvalues so that \(\{\mu_1, \ldots, \mu_n\} = \{\lambda_1, \ldots, \lambda_n\}\). We have proved that possible \(\omega\)-limit sets consist of discrete points and therefore the dynamical system converges. Only stable equilibria can be generically reached and we look for such equilibria now. Thus, we linearize the system around \((W_1, \ldots, W_k)\). We find from (32)

\[
\frac{d\xi_i}{dt} = -B\xi_i + 2 \sum_{j<i} (B\xi_i, W_j)W_j + 2 \sum_{j<i} (BW_i, \xi_j)W_j + 2 \sum_{j<i} (BW_j, W_i)\xi_j
\]

while linearizing the orthogonality constraints leads to

\[
(\xi_i, W_j) + (W_j, \xi_i) = 0, \quad \forall 1 \leq i, j \leq k. \tag{40}
\]

This gives us

\[
\frac{d\xi_i}{dt} = -B\xi_i + 2 \sum_{j<i} (\mu_j - \mu_i)(\xi_i, W_j)W_j + \mu_i \xi_i
\]

\[
= \left( -B + 2 \sum_{j<i} (\mu_j - \mu_i)W_j \otimes W_j + \mu_i I \right) \xi_i. \tag{41}
\]

The eigenvalues of the involved matrix are (remember that \(\xi_i \perp W_i\))

\[
(\mu_1 - \mu_i), \ldots, (\mu_{i-1} - \mu_i), (-\mu_{i+1} + \mu_i), \ldots, (-\mu_n + \mu_i), \tag{42}
\]

which are all non positive if and only if

\[
\begin{cases}
\mu_j < \mu_i & \text{for } j < i, \\
\mu_j > \mu_i & \text{for } j > i.
\end{cases} \tag{43}
\]

This exactly means that \(\mu_1, \ldots, \mu_k\) are the \(k\) first eigenvalues of \(B\). It remains to prove the convergence estimate for the subspace spanned by \((V_i)_{i=1,\ldots,k}\). Taking, for \(l > i\), the scalar product of (41) with \(u_l\) leads to

\[
\frac{d}{dt} (\xi_i, u_l) = \left( -B + 2 \sum_{j<i} (\mu_j - \mu_i)W_j \otimes W_j + \mu_i I \right) (\xi_i, u_l)
\]

\[
= (-\lambda_l + \lambda_i)(\xi_i, u_l).
\]

This provides the desired estimate since \(\forall l > i, \lambda_l - \lambda_i \geq \lambda_{i+1} - \lambda_i\). \(\blacksquare\)
B. Nonlinear Case and Preconditioning

From the construction itself, the triangular flow gives an expression which is tangent to the Stiefel manifold. Therefore, by mimicking the strategy used before in the linear case, we may write a version for minimizing nonlinear functional or a $H^1$ version. Namely, the triangular flow easily extends to the Hartree-Fock functional by writing

$$\forall i \in \{1, \ldots, N\}, \quad \frac{d\phi_i}{dt} = -\tilde{H}(\varphi)\varphi_i + 2\sum_{j=1}^{i-1}(\tilde{H}(\varphi)\varphi_i, \varphi_j)_{L^2} \varphi_j + (\tilde{H}(\varphi)\varphi_i, \varphi_i)_{L^2} \varphi_i. \quad (44)$$

Similarly, the preconditioned version is given by

$$\frac{d\phi_i}{dt} = -g_i, \quad (45)$$

where $g = (g_1, \ldots, g_N)$ solves

$$g \in T_{\varphi}\mathcal{M}_b, \quad (46a)$$

$$(\Pi(I - \Delta_h)g)_i = \tilde{H}(\varphi)\varphi_i - 2\sum_{j=1}^{i-1}(\tilde{H}(\varphi)\varphi_i, \varphi_j)_{L^2} \varphi_j - (\tilde{H}(\varphi)\varphi_i, \varphi_i)_{L^2} \varphi_i. \quad (46b)$$

V. NUMERICAL RESULTS

To point out the difference of behavior for the convergence of the techniques exposed earlier, we have chosen three different test cases. The first one is the simplest possible case, where the problem is linear and we take for $B$ a diagonal matrix whose spectrum is given. In this case, it is possible to see the influence of clusters of eigenvalues on the convergence rate of all algorithms. Only the preconditioning is not tested since the problem cannot be viewed as a perturbation of the Laplace operator. Let us note that considering a diagonal matrix is not restrictive because the $L^2$ and the triangular flows are both invariant under any unitary transformation on the vectors of $V$. Therefore, the convergence speed of the numerical methods only depends of the spectrum of $B$. The second problem is still linear, and we have taken for $B$ a full matrix of size $n=283$, which comes from the discretization in a plane-wave basis of the Hamiltonian associated to a molecular system of two silicon atoms (this matrix has been generated with the help of the ABINIT code (Abinit is an open source ab-initio code downloadable from http://www.abinit.org, used by quantum chemists and physicists. The Abinit package, essentially developed by the University of Louvain-la-Neuve (PCPM) and the CEA-DAM, allows to compute electronic structures on molecular systems made of nuclei and electrons as on periodic systems of atoms)). In this situation, all the strategies have been tested including the preconditioning. Finally, the full Hartree-Fock model has been implemented. All the programs have been written in MATLAB, the ODE solver was a Runge-Kutta (RK) method of order 5, with an adaptive time-stepping given by a Richardson extrapolation [14, 15].

A. A Linear Case

In this case we consider a diagonal matrix of size $n = 200$ with the following clustered spectrum

$$\text{sp}(B) = \bigcup_{k=1}^{40} \{k - 0.01, k + 0.02, k + 0.03, k + 0.04, k + 0.05\}. \quad (47)$$

In Figs. 1 and 2 we have shown, in log$_{10}$ scale, the convergence rate of the energy

$$E_i = \frac{1}{2} \text{tr}(V^{(i)}' B V^{(i)})$$

with respect to the time $t_i$ of the RK algorithm. The exact energy, which is the half sum of the $k$ smallest eigenvalues of $B$ (see (4)), is denoted by $E$. In both $L^2$ and triangular flows, the convergence is very fast for $k = 5$ (see Table I). In that particular case, the number of eigenvalues which are expected corresponds exactly to the size of the eigenspace associated to the cluster of the $k$ smallest eigenvalues. The convergence speed is much slower when $k \leq 4$, meaning that the convergence depends on the gap between the $k$-th and the $(k + 1)$-th eigenvalue of $B$. We get the same behavior for $6 \leq k \leq 9$ and a faster convergence for $k = 10$, and so on.
To emphasize the difference of behavior between the $L^2$ and the triangular flows, we compare the convergence speed of different eigenvalues. In Fig. 3, we represent the convergence of the energy associated to the first cluster (corresponding to the five first smallest eigenvalues of $B$) and the convergence of the 6-th eigenvalue. As the $L^2$ gradient flow only computes the space spanned by the 6 first eigenvectors, the subenergies corresponding to spaces of lower dimension are approximated, as well as the corresponding subenergies. As pointed out in Section A, we can see in Fig. 3 that all the subenergies converge with the same speed. By contrast, the triangular flow provides the exact subenergies (here $E_{\text{clust}1}$) or eigenvalues (here $\lambda_6$). Moreover, Fig. 4 clearly shows the convergence rate of the energy associated to the first cluster which converges much more quickly than the 6-th eigenvalue, as proven previously in Theorem 4.1. As shown in Table II, the total number of RK iterations and CPU time are roughly equivalent for both methods.

<table>
<thead>
<tr>
<th>Number of eigenvalues $k$</th>
<th>$4$</th>
<th>$5$</th>
<th>$6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of RK iterations ($L^2$ flow)</td>
<td>$\sim 2300$</td>
<td>$\sim 100$</td>
<td>$\sim 1500$</td>
</tr>
<tr>
<td>Number of RK iterations (triangular flow)</td>
<td>$\sim 2100$</td>
<td>$\sim 100$</td>
<td>$\sim 1600$</td>
</tr>
</tbody>
</table>

FIG. 3. $L^2$ gradient flow: convergence of the first cluster $\log_{10} |E^{(i)}_{\text{clust1}} - E^{\text{lim}}_{\text{clust1}}|$ (in dashed lines) and $\log_{10} |\lambda_6^{(i)} - \lambda_6^{\text{lim}}|$ (in solid lines). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

FIG. 4. Triangular flow: convergence of the first cluster $\log_{10} |E^{(i)}_{\text{clust1}} - E_{\text{clust1}}|$ (in dashed lines) and $\log_{10} |\lambda_6^{(i)} - \lambda_6|$ (in solid lines). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

B. The Two Silicon Atoms

In this Section, we consider a full matrix of size $n = 283$ coming from the discretization in a plane-wave basis of the Hamiltonian associated to a system of two silicon atoms. Let us recall that this Hamiltonian has been generated thanks to the ab-initio code ABINIT [13]. We give numerical results for the different flows described earlier, namely the $L^2$ gradient flow (6) and the $H^1$ preconditioned one, as well as the triangular flow (33) and its preconditioned version. Since $B$ is a

<table>
<thead>
<tr>
<th></th>
<th>RK iterations</th>
<th>CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L^2$ gradient flow</td>
<td>$\sim 1400$</td>
<td>$\sim 14$ s</td>
</tr>
<tr>
<td>Triangular flow</td>
<td>$\sim 1200$</td>
<td>$\sim 13$ s</td>
</tr>
</tbody>
</table>

discretized Hamiltonian in a plane-wave basis, it can be viewed as a perturbation of the Laplace operator. It is then natural to consider the following $H^1$ gradient flow

$$G = \Pi G,$$  

$$(\Pi (I - \Delta_h) G)_i = -B V_i + \sum_{j=1}^{k} (B V_i, V_j) V_j$$

where $\Delta_h$ denotes the Galerkin approximation of the Laplace operator and $\Pi$ the projector on $T_{V}St(k,n)$. In Figs. 5–7, we represent the convergence history in $\log_{10}$ scale of the computed energy with the different flows, as well as the time step evolution. One can see the influence of the $H^1$ preconditioning, in the sense that less iterations are needed to reach the energy convergence (see Table III), since much larger time steps are allowed (see Fig. 6).
C. The Hartree-Fock Problem

In this Section, we present numerical simulations for the Hartree-Fock periodic problem, described in Section B. We have solved numerically the minimization problem of the periodic Hartree-Fock energy functional $E_{\text{HF}}^C(\psi_1, \ldots, \psi_N)$ over $\mathcal{M}_h$ [see Eqs. (14)–(16)], using the $L^2$, $H^1$ and triangular flows. In the periodic case, the $L^2$ gradient flow is given by

$$\frac{d\psi_i}{dt} = -\Pi \tilde{H}_C(\psi)\psi_i,$$

where $\Pi$ is defined by (20) and the gradient of the energy by

$$\tilde{H}_C(\psi)\psi_i = \frac{\partial E_{\text{HF}}^C}{\partial \psi_i} = (-\Delta - NG(\delta_{i1} - 1) + G(\rho - N))\psi_i - \sum_{j=1}^{N} \psi_j G(\psi_i\psi_j^* - \delta_{ij}).$$

The $H^1$ gradient flow writes $\frac{d\psi}{dt} = -g$, where $g$ solves (29) with $\tilde{H}_C$ instead of $\tilde{H}$. The triangular flow is given by (45)–(46) where $\tilde{H}$ is also replaced by $\tilde{H}_C$. As in the linear case, the numerical resolutions of these flows have been realized thanks to a Runge-Kutta method of order 5 with adaptive time stepping. The wave functions $\psi_i$ have been discretized on a finite differences regular grid; the spatial derivatives and the terms involving the Green kernel $G$ have been computed using a spectral method on a Fourier grid. Let us note that the resolution of the linear system (29) which provides the gradient $g$ is performed by a conjugate gradient algorithm, since $\Pi(I - \Delta_h)\Pi$ is symmetric positive definite on $T_\psi \mathcal{M}_h \cap H_0^1$.

| Number of RK iterations needed for $L^2$, $H^1$ and triangular flows, $k = 10$. |
|---------------------------------|----------------|----------------|
| RK iterations                  | CPU time       |
| $L^2$ gradient flow            | $\sim 200$     | $\sim 30$ s    |
| $H^1$ gradient flow            | $\sim 80$      | $\sim 170$ s   |
| Triangular flow                | $\sim 200$     | $\sim 30$ s    |
| Precond. triangular flow       | $\sim 160$     | $\sim 430$ s   |
In Fig. 8, we have shown the energy convergence of periodic systems composed of $N = 1$ electron per unit cell, for the different flows, with a regular discretization grid made of $32^3$ nodes. As in the linear case, we consider that the convergence is reached when the relative error on the energy is smaller than $10^{-9}$. Moreover, we give in Tables IV–VII the number of RK iterations and CPU time needed for the different flows, considered with various number of electrons per unit cell (1–5 electrons) and with several spatial discretization ($8^3$, $16^3$, and $32^3$ points). For numerical simulations which need more than 20 h, we consider that the computational cost is prohibitive and we denote this with a star symbol. We remark that preconditioning speeds up the convergence by a factor of about 15 in terms of RK iterations, for both gradient and triangular flows. For two electrons per unit cell for instance, the speed-up is even better. Indeed, in that case the speed of convergence is very slow for both the $L^2$ and triangular flows: after more than 20 hours of computations, only a relative error of $10^{-2}$ is reached on the energy convergence, highlighting the significant role played by preconditioning. As far as the CPU time is concerned, we get improvements by a factor up to 1.5. Indeed, the extra system (29) that needs now to be

**TABLE IV.** For different number of electrons per unit cell, number of RK iterations and CPU time needed for the $L^2$ gradient flow, on different regular spatial grids ($8^3$, $16^3$, and $32^3$ points).

<table>
<thead>
<tr>
<th>N</th>
<th>$e^{-}$</th>
<th>$2e^{-}$</th>
<th>$3e^{-}$</th>
<th>$4e^{-}$</th>
<th>$5e^{-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8^3$</td>
<td>180 it–38 s</td>
<td>1330 it–535 s</td>
<td>2510 it–610 s</td>
<td>2530 it–775 s</td>
<td>2175 it–867 s</td>
</tr>
<tr>
<td>$16^3$</td>
<td>550 it–290 s</td>
<td>4360 it–1 h 12</td>
<td>9080 it–3 h 45</td>
<td>6800 it–4 h</td>
<td>6525 it–4 h 50</td>
</tr>
<tr>
<td>$32^3$</td>
<td>1800 it–2 h 40</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

**TABLE V.** For different number of electrons per unit cell, number of RK iterations and CPU time needed for the $H^1$ gradient flow, on different regular spatial grids ($8^3$, $16^3$, and $32^3$ points).

<table>
<thead>
<tr>
<th>N</th>
<th>$e^{-}$</th>
<th>$2e^{-}$</th>
<th>$3e^{-}$</th>
<th>$4e^{-}$</th>
<th>$5e^{-}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8^3$</td>
<td>60 it–30 s</td>
<td>135 it–180 s</td>
<td>200 it–400 s</td>
<td>160 it–470 s</td>
<td>150 it–590 s</td>
</tr>
<tr>
<td>$16^3$</td>
<td>85 it–580 s</td>
<td>155 it–2920 s</td>
<td>240 it–2 h</td>
<td>200 it–2 h 40</td>
<td>205 it–3 h 50</td>
</tr>
<tr>
<td>$32^3$</td>
<td>110 it–1 h 40</td>
<td>230 it–10 h</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

solved to compute the new gradient, turns out to be rather costly. It is clear that one has to remedy to that by preconditionning this linear system, but obvious preconditionners that we have tried up to now are not yet satisfactory. This step is clearly the subject of ongoing research, if one wants to use these flows on realistic problems (for instance with more electrons per unit cell and with refined spatial grids).

VI. CONCLUSION

In this article, we have proposed to solve nonlinear eigenvalue problems by gradient flow techniques and have applied this methodology to Hartree-Fock equations. The interesting feature of this approach is that the algorithm is very reliable since the constraints are directly taken into account and nothing particular needs to be done to ensure the self-consistence of the Hartree-Fock equations. Moreover, the equations obtained with these flows may be numerically solved by standard ODE techniques such as Runge-Kutta methods. We also provide improvements in two directions (which can also be used together) : preconditioning and triangular flows. Both strategies seem interesting since the first one significantly reduces the number of Runge-Kutta iterations while the second one also provides the eigenvectors. For the time being a clear improvement is observed concerning the number of Runge-Kutta steps needed to obtain the convergence of the method, but the CPU time only still remains slightly improved. Improving the linear system at the heart of the $H^1$ flow is still to be done, e.g. by preconditionning the conjugate gradient routine.

References